Fast Methods for Mixed-Integer PDE-Constrained Optimization

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium
(Dr. rer. nat.)

von Mirko Hahn, M. Sc.

geb. am 30.10.1990 in Köln

genehmigt durch die Fakultät für Mathematik
der Otto-von-Guericke-Universität Magdeburg

Gutachter: Dr. rer. nat. Sebastian Sager

Dr. rer. nat. Michael Ulbrich

eingereicht am: 17. Februar 2025

Verteidigung am: 24. September 2025

Abstract

Mixed-integer optimization is often counted among the more difficult subfields of mathematical optimization. This is because integers cannot be changed in arbitrarily small steps, which makes most kinds of derivative-based iterative optimization methods ineffective. Exact solution methods for problems with integers are often *enumerative* in nature. In the worst case, they enumerate all feasible values of the integer variables.

This generally requires an exponential amount of resources. In many cases, this makes the problem intractable. Intuitively, problems with an infinite number of integer-valued variables, e.g., an integer-valued function, should also be intractable. Such problems occur, for instance, in the fields of mixed-integer optimal control (MIOC) and mixed-integer PDE-constrained optimization (MIPDECO). Both of these fields deal with solutions of differential equations, which reside in infinite-dimensional vector spaces and can depend on infinite-dimensional integer-valued functions as parameters.

We argue that, in the case of binary-valued functions, these problems are not intractable. They are more tractable than even regular mixed-integer problems. We develop a theoretical framework within which the space of binary-valued measurable functions can be treated almost as if it was a vector space. This framework then enables us to transfer well-known nonlinear optimization (NLP) methods to this space.

We use this framework to transfer a subset of basic NLP theory to our setting. Based on this, we transfer two well-known NLP algorithms, the steepest descent method and the quadratic penalty method, to our setting.

We lay some groundwork for future expansion on these and other existing solution methods for MIOC and MIPDECO problems. We propose an adaptive variant of the well-known sum-up rounding (SUR) method and a step finding method for problems with special ordered set constraints of type 1 (SOS1).

To demonstrate the viability of our transferred NLP methods, we solve two basic test problems and compare the result to existing solutions from literature. Finally, we propose some avenues of future expansion upon our theoretical foundation.

Zusammenfassung

Die gemischt-ganzzahlige Optimierung wird oft den schwierigeren Teilgebieten der mathematischen Optimierung zugeordnet. Grund dafür ist, dass Ganzzahlen nicht in beliebig kleinen Schritten angepasst werden können, was die meisten ableitungsbasierten schrittweisen Optimierungsverfahren ineffektiv macht. Exakte Lösungsverfahren für ganzzahlige Probleme sind oft *enumerativ*. Im schlimmsten Fall zählen sie alle zulässigen Belegungen der ganzzahligen Variablen auf.

Allgemein erfordert dies eine exponentielle Menge an Resourcen. Das macht Probleme in vielen Fällen praktisch unlösbar. Intuitiv sollten Probleme mit einer unendlichen Anzahl von ganzzahligen Variablen, beispielsweise in Form einer ganzzahligen Funktion, ebenfalls unlösbar sein. Solche Probleme treten beispielsweise in den Feldern der gemischt-ganzzahligen Optimalsteuerung (MIOC) und der gemischt-ganzzahligen PDE-beschränkten Optimierung (MIPDECO) auf. Diese beiden Felder befassen sich mit Lösungen von Differenzialgleichungen, die Elemente von unendlichdimensionalen Vektorräumen sind und von unendlichdimensionalen ganzzahligen Parameterfunktionen abhängen können.

Wir argumentieren, dass diese Probleme im Fall binärwertiger Funktionen nicht unlösbar sind. Sie sind sogar besser lösbar, als gewöhnliche gemischt-ganzzahlige Optimierungsprobleme. Wir entwickeln einen theoretischen Rahmen, in dem der Raum der binärwertigen Funktionen fast so behandelt werden kann, als wäre er ein Vektorraum. Dieser Rahmen erlaubt es uns, bekannte Methoden aus der nichtlinearen Optimierung (NLP) in diesen Raum zu übertragen.

Wir benutzen diesen Rahmen, um eine Untermenge der grundlegenden NLP-Theorie in unser Setting zu übertragen. Darauf aufbauend übertragen wir zwei bekannte NLP-Algorithmen, die Methode des steilsten Abstiegs und die quadratische Penalty-Methode, in unser Setting.

Wir leisten einige Vorarbeiten für zukünftige Erweiterungen dieser und anderer Lösungsmethoden für MIOC- und MIPDECO-Probleme. Wir schlagen eine adaptive Variante der bekannten Sum-Up Rounding (SUR) Methode und eine Schrittbestimmungsmethode für Optimierungsprobleme mit Special Ordered Set Nebenbedingungen von Typ 1 (SOS1) vor.

Um die Realisierbarkeit unserer übertragenen NLP-Methoden zu zeigen, lösen wir zwei einfache Testprobleme an und vergleichen die Ergebnisse mit existierenden Lösungen aus der Literatur. Zuletzt schlagen wir einige Möglichkeiten vor, wie das theoretische Fundament, das wir gelegt haben, in Zukunft erweitert werden könnte.

Contents

Chapter 1

Introduction and Background

Although many of its algorithmic approaches trace back to older solution methods for systems of equations and inequalities, algorithmic optimization as its own subdiscipline of mathematics is relatively young, having only truly come into its own around the middle of the twentieth century with the development of the first algorithms capable of solving larger linear optimization problems for logistical planning.

Since then, the field has advanced by leaps and bounds to the point where a full survey would take a prohibitively long time. Today, algorithms capable of solving quadratic problems, general nonlinear problems, problems with uncertainties, and problems with non-convex objectives and constraints are widely available. In terms of scale, nonlinear optimization problems with billions of variables are regularly solved for the purposes of machine learning, and entire computer networks are dedicated to the joint solution of optimization problems with special distributed optimization algorithms (see, e.g., [Gor+22; Yan+19]).

Under the large umbrella of solvable problem classes, we find two that are of particular interest to us: infinite-dimensional problems and integer problems. The difficulty in integer optimization is that, because integer values are discrete from one another, gradual improvement is not generally a suitable method for integer optimization. Instead, integer problems often have to be solved with enumerative methods. In the worst case, the amount of computational resources required to solve integer and mixed-integer optimization problems is exponential in the size of the problem. Therefore, even small problems can be very difficult to solve. This was theoretically substantiated by the discovery that linear optimization with binary variables is NP-hard [Coo71; Kar72].

The difficulty of infinite-dimensional optimization arises from the fact that, there is no such thing as a "small" infinite-dimensional problem. Every infinite-dimensional optimization problem is nominally infinitely large. In practice, all infinite-dimensional problems are, at some point, approximated with finite-dimensional approximations. However, these approximations tend to be of relatively high dimension. Therefore, if a subset of the optimization variables has to assume integer values, the approximation problem becomes intractable very quickly.

Infinite-dimensional problems with binary- or integer-valued variables are, however, of great practical relevance. They occur frequently when discrete de-

cisions are optimized that differ based on spatial or temporal location. For instance, they can be found in optimal control of switched systems under ordinary differential equations (ODEs) or differential algebraic equations (DAEs), and the design of physical structures well-suited for certain applications under partial differential equations (PDEs) [BS04; NS13]. Again, applications in which such problems occur are too numerous to meaningfully survey here. Infinite-dimensional integer-valued optimization problems can, for instance, be found in the optimization gear and propulsion choices in motor vehicles [Kir+10; Rob+21], traffic light patterns [Sor16; Bet+21; Le+22], optimized medical treatment (as indicated, e.g., in [Jos+20; Geb+23]), as well as the optimal design of various kinds of structures, such as truss structures [BS04], fluid vessels and aerodynamic structures [MP04], and electromagnetic scatterers [LMW21], among many other applications.

To avoid the computational workloads required to solve integer optimization problems and mitigate them where they are unavoidable, a variety of mitigation techniques have been created. In order to understand these methods, we first have to recognize that most integer optimization problems can fundamentally be thought of as being binary-valued problems. This is because any finite set of discrete choices can be encoded by the same number of binary indicator variables, linked by a constraint of mutual exclusion (sometimes known as a "special ordered set of type 1" or "SOS1" constraint). In optimal control, this transformation is commonplace and is sometimes referred to as "partial outer convexification." We will therefore treat optimization with binary variables as equivalent to optimization with arbitrary integer variables. The advantage of spatially or temporally distributed binary variables as opposed to integer variables is that they can be interpreted as indicators of sets. Partial outer convexification can be problematic if the number of choices per point is very high. However, we will not consider problems with many pointwise binary choices here.

From this analogy between distributed binary variables and sets arises the first category of solution approaches that we want to mention here: shape and switching time optimization (see, e.g., [LT03; ZA15] for introductions and surveys). Both approaches can be seen as analogous to one another except that one stems from the application domain of optimal design while the other stems from optimal control. Shape optimization replaces the binary indicator variable of a set with continuous parameters describing the shape of its boundary. Because these parameters are continuous, the combinatorial explosion associated with integer optimization can be avoided completely. However, the parameterization of the set boundary is generally only possible under additional assumptions on the shape of that boundary.

On its own, the shape optimization approach makes it relatively difficult to change the topology of an object. It can be difficult to split sets into multiple components, merge separate components, or introduce holes into the interior of a set. In the field of optimal design, *topology optimization* (see, e.g., [BS04; NS13]) has been developed as a way to address this flaw. Topology optimization works with a so-called "topological derivative" that predicts the change of an objective with respect to the introduction of infinitesimal holes in the interior of a set. The topological derivative can be used in a variety of ways. For instance, it can be used to heuristically introduce holes in a structure [EKS94], or it can be used as part of a *level set method* [Dij+13; AA06]. There have been attempts to unify shape and topological derivatives into a single construct (see, e.g., [Ams11;

Céa+00]). We do not know of a direct analogue to topology optimization in the field of optimal control.

Level set methods in the sense of [Dij+13] deserve mention in and of themselves. They are not necessarily an optimization method of their own, but rather a way to parameterize sets. At the foundation of level set methods lies the recognition that sets can be derived from functions by taking the sublevel set with respect to a fixed level. The optimization variables are then the parameters describing that function. Level set methods usually prescribe a function space with desirable properties as part of the modelling process. The derivatives in level set optimization are closely related to topological derivatives.

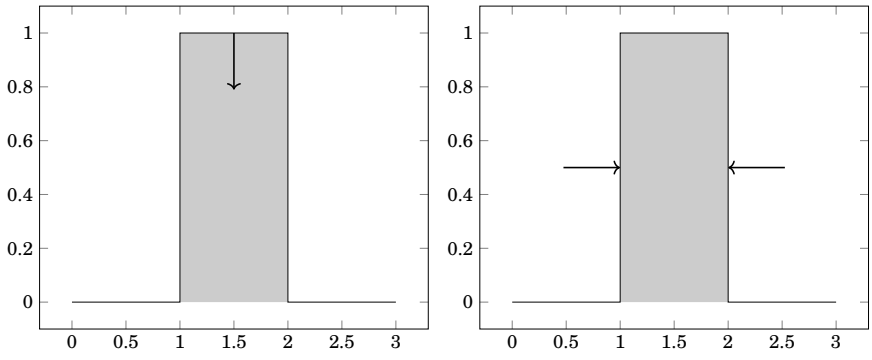
All of the methods described thus far can be thought of as so-called "optimize-then-discretize" methods. They are originally conceived as methods to solve infinite-dimensional optimization problems whose behavior is then approximated with finite-dimensional operations to obtain a practically implementable algorithm. An alternative approach is to first discretize the infinite-dimensional optimization problem and then solve the resulting large-scale integer optimization problem with more conventional finite-dimensional optimization methods. This is a relatively popular approach because it does not require the development of alternative optimization methods. Instead, well-established methods from finite-dimensional optimization can be brought to bear. The issue with this approach is that it produces discretized problems with a very large number of integer variables. Therefore, it can easily suffer from combinatorial explosion.

One approach to mitigating the computational workload of the discretize-then-optimize approach are so-called relaxation methods (see, e.g., [Sag06; Sag+06; SJK11; HS13]). In the context of mixed-integer optimal control, a "relaxation method," is an approximate solution method in which the discretized problem is first solved with relaxed integrality constraints. In a second step, the behavior observed in the relaxed solution is then approximated with an integer solution. In contrast to the original problem, the approximation problem generally has a much simpler, more easily standardized structure that allows for optimized solution algorithms. The methods to accomplish the second step range from specialized "rounding" algorithms such as sum-up rounding (SUR) [Sag06, Sec. 5.1] and next-forced rounding (NFR) [Jun14, Sec. 4.4.2] to full Branch and Bound solvers [Bür+20] that are highly optimized for the specific problem type.

1.1 MOTIVATION

At its outset, the research effort underlying this thesis was intended to find rounding schemes suitable for PDE-constrained optimization problems. Originally, rounding schemes such as SUR and NFR were developed for optimal control problems and presume the existence of a single time axis along which the effects of control perturbations travel in only one direction. We note that the transfer of rounding schemes to fairly general PDE-constrained settings has since been independently accomplished by the work of Paul Manns and Christian Kirches [MK20].

The main subject of this thesis derives instead from a different line of thinking. Quality guarantees for relaxation methods are often derived from a priori estimates of the approximation error. These estimates often substantially overestimate the actual error and therefore demand an impractically high degree of uniform mesh refinement. If we solve the approximation problem on a coarser



(a) Scaling in value (becomes non-binary) (b) Scaling in measure (stays binary)

Figure 1.1: Different ways of "scaling" a binary-valued function.

mesh, then we risk obtaining a solution of unacceptably low quality. This is particularly relevant in PDE-constrained settings because higher spatial dimensions lead to a much quicker escalation in the problem dimension with increased spatial resolution.

In existing relaxation methods, it is generally implied that a sub-par solution would be discarded and a better solution would be found by re-solving the approximation problem on a finer control grid. However, one might ask whether we can improve on a sub-par solution rather than discarding it entirely. Naturally, once we have designed a procedure capable of improving an existing binary-valued solution, it is a small additional step to apply this procedure iteratively to create an entire optimization algorithm.

To improve an existing solution, we would have to be able to predict how certain adjustments to an existing solution would change the objective function value. In derivative-based optimization, such predictions are generally made using truncations of the Taylor sequence and are only valid over small distances. As we have indicated earlier, this is problematic in the context of integer optimization because switching between integer values requires jumps of a certain minimal length.

The foundation of our approach is remarkably simple. In vector space optimization, steps are shortened by scaling them. If we think of a binary variable, this means that we would scale the value of the variable. For binary-valued variables, this would almost inevitably violate the integrality constraint. However, because we work with binary variables that exist on arbitrarily refinable grids, we have the option of shortening steps by restricting them to smaller subsets of the overall step. We can think of this as "scaling in measure" as opposed to "scaling in value" because we are reducing the measure of the indicated set rather than reducing the pointwise value of the function. The difference between the two modes of scaling is illustrated in Figure 1.1.

Our objective is to describe a space in which iterative improvement based on derivatives is possible without ever violating integrality constraints. We then want to transport, as proof of concept, some gradient-based optimization methods to this space. We ultimately find that metric spaces of measurable sets are suitable spaces for this purpose. The type of optimization problem under

discussion in this thesis is therefore roughly

$$\begin{aligned} &\inf_{U} F(U) \\ &\text{s.t. } U \in \mathscr{U}, \end{aligned} \tag{1.1}$$

where $U \in \Sigma$ is a measurable set from a suitable measure space (X, Σ, μ), $F : \Sigma \to \mathbb{R}$ is a meaningfully differentiable functional, and $\mathscr{U} \subseteq \Sigma$ is a family of measurable sets described by a suitable set of constraints.

In iterative optimization, step size and step size control are important to ensure global convergence. We measure step size by the measure of the symmetric difference between two sets. Therefore, we cannot distinguish between sets that only differ by a nullset. To account for this, we operate in the space of residual classes with respect to "being equal up to a nullset." We refer to this as a "similarity space" and to the individual residual classes as "similarity classes."

Working in a metric space is difficult because metric spaces lack much of the structure that is usually used to perform optimization. However, we will demonstrate that suitable similarity spaces retain much of that structure and can therefore still be treated with the same methods as conventional optimization.

1.2 RELATION TO OTHER WORK

This work enters into a rich tapestry of pre-existing work originating from various fields. In this section, we explain how our work relates to other pre-existing work.

Relaxation Methods

As we have stated, this work starts with the question whether we could iteratively improve upon the output of a relaxation method. However, our approach does not have much in common with relaxation methods. Most relaxation methods follow a discretize-then-optimize approach, whereas we follow an optimize-then-discretize approach. In principle, it is possible to design infinite-dimensional relaxation methods. In Section 3.2.2.2, we briefly discuss how a mesh-free rounding problem could be solved. However, this is not central to this thesis and we do not implement any of the methods described in that section.

There is extensive work proving convergence of relaxation methods for various types of ODEs and PDEs. For instance, [Sag06; SJK11] provide convergence proofs for ODEs, [HS13] proves convergence for semilinear parabolic PDEs, [MK20] first transfers the concept to elliptic PDEs with spatially distributed controls. Studies of convergence behavior and approximation rates are made, for instance, in [KMU21]. Some of the theory underlying these results has successfully been applied to create convergence proofs for algorithms similar to those presented here [Man+23].

The algorithms that we put forward in Section 3.2.2.2 suggest that our theoretical framework could be applied to the field of relaxation methods. However, it is difficult to tell whether this would substantially advance the field.

Shape Optimization

Shape optimization and switching time optimization are more similar to our approach because they do not presume a specific mesh. However, they sometimes

assume fixed dimensionality of their boundary parameterization. They also often make additional assumptions about the smoothness of the boundary.

We diverge from shape and switching time optimization in that we do not assume fixed topology. However, as we will see, our derivatives switch signs at the boundary of control sets, which means that there is almost always a way to improve the objective by locally shifting the set boundary. In this way, shape optimization could be used as an optimization method within our framework. In this thesis, however, we do not constrain our improvement steps in this way.

There is an alternative approach in switching time optimization that is sometimes known as "method of competing Hamiltonians" [BL85; Jun14; Boc+17]. In Section 2.4.3, we discuss how to calculate derivatives in ODE-constrained optimization problems. There, we see that our derivatives are the same as those used by the method of competing Hamiltonians. This means that our methods could be seen as a generalization of that method in the context of ODE-constrained optimal control.

Topology Optimization

Out of the established optimization methods mentioned thus far, topology optimization is by far the closest to our methodology. Topological derivatives are derivatives with respect to perturbations that introduce small "holes" in a domain (see, e.g., [NS13, Sec. 1.1]). While the definition of the topological derivative is almost always stated in a manner that is dependent on the shape of those holes, the Lebesgue differentiation theorem makes it quite plausible that the derivative that we use is a topological derivative. In our case, the derivative must always be independent of the shape of the hole.

There have been a variety of efforts to unify the concepts of shape and topological derivative (see, e.g., [NS13]), which may be related to the observation that our derivative can be used to discern favorable deformations in a set's boundary. The optimization methods used in the field of topology optimization appear to be less direct transfers of conventional optimization methods, and more focused on heuristics, fixed point iterations, and level set methods [EKS94; Céa+00; AA06].

The most notable parallel between our work and the field of topology optimization is found in its earliest stages. In [CGM74], the authors attempt to derive an optimization method for binary-valued functions from the gradient descent method. This is almost exactly what our unconstrained optimization algorithm achieves. However, this method does not appear to ever have found widespread adoption in the field. We suspect that this is because the authors do not use a globalization scheme, but rather explicitly calculate step sizes, which presupposes an unrealistic amount of prior knowledge about the objective function. Our methods are stated as trust region methods and can therefore discover appropriate step sizes dynamically. This greatly simplifies their implementation. It is possible that, because [CGM74] was written in 1974 and therefore predates the first use of the term "trust region", which is sometimes ascribed to [Sor82], the authors were not familiar with the trust region method at the time.

One distinction that is certain is that [CGM74] does not have the same theoretical foundation that we have in the theory of measure space geodesics. This foundation makes it possible to transfer almost arbitrary optimization

methods to the measure space setting with very little method-specific theoretical work.

1.3 CONTRIBUTIONS TO PRIOR WORK

Some early results of the research underlying this thesis were published in [HLS22]. This notably includes some of the conceptual framework of differentiable set functionals (Section 2.4), methods to calculate derivatives for ODE- and PDE-constrained problems (Sections 2.4.2 to 2.4.4), an early variant of the unconstrained descent framework (Section 3.1.3) that does not incorporate any form of error control, and steepest descent step finding (Section 3.1.4).

The paper is joint work with Sven Leyffer and Sebastian Sager. The initial suggestion of researching trust region methods for binary optimization was made by Sven Leyffer, while the development of a theoretical foundation for them, the construction of concrete algorithms, their theoretical analysis, and their implementation was performed by the author of this thesis. Sven Leyffer also strongly advocated the use of a simplified geodesic-free suboptimality estimator. Aside from this, the paper was written by this thesis' author with only occasional input from Sven Leyffer. Sebastian Sager contributed via advice, funding, supervision, and proof-reading of the final manuscript.

This thesis substantially expands on the work presented in [HLS22]. The paper does not include theoretical arguments that require measure space geodesics. The theory of measure space geodesics that we discuss in Section 2.3 is a later development. All algorithms presented in this thesis account for errors in functional and gradient evaluation. We also allow for multivariate problems through "layering" (see Section 2.1.3), while the paper is limited to univariate problems.

For specific problem types, a convergence proof for the binary trust region steepest descent method was performed in [Man+23]. This argument is primarily attributable to Paul Manns. Therefore, we do not discuss it further in this thesis. The thesis' author's contribution to this paper is limited to the theoretical foundation and implementation of the binary trust region steepest descent method itself.

A related strand of research has spun off from the starting point of binary trust region steepest descent method in [Sha+20; LM22]. This is mostly focused on the use of trust region methods without mesh refinement as a heuristic and the imposition of complex constraints on the finite-dimensional trust region subproblem. Although the author has had some involvement in [Sha+20], this is entirely outside of the scope of this thesis because it violates the thesis' fundamental maxim of avoiding fixed discretization wherever possible.

1.4 OUTLINE AND CONTRIBUTION

The goal of this thesis is to transfer well-known iterative optimization schemes to the setting of binary-valued optimal control. We limit ourselves to binary-valued variables because we want to exploit the correspondence between binary-valued functions and sets. Partial outer convexification guarantees that this does not substantially restrict the range of applications for our algorithms. Finally, we do not investigate real "mixed-integer" problems in the sense that we do not include any continuous variables in our theory. However, because our methods are so similar to conventional NLP methods, it is not difficult to conceive of "hybrid"

methods that combine one of our set-valued optimization algorithms with its corresponding conventional NLP solver to solve true mixed-integer problems. In this sense, we consider the theory in this thesis an essential prerequisite for fast mixed-integer solvers for ODE- and PDE-constrained problems.

We develop all of our theory in what we refer to as "similarity spaces." Similarity spaces are residual spaces obtained by equating measurable sets when they are equal "up to a nullset." These spaces are quite unfamiliar as a setting for optimization because they are not vector spaces. Rather, we will see that, under the right circumstances, they are geodesic metric spaces with a commutative group structure whose metric is invariant under translation. This combination of properties means that we can work with these spaces almost as if they were actual normed vector spaces. We can transfer much of vector space optimization theory into these spaces in some form.

Of course, working on the basis of conceptual analogy can be quite treacherous and we must take great care to ensure that all of our theoretical arguments rest on a robust foundation. Accordingly, Chapter 2, which is dedicated to investigating the theoretical properties of similarity spaces and developing an extensive toolkit for theoretical arguments about similarity spaces, is quite long. We start with simple foundations in Section 2.2 and quickly expand into the extensive theory of measure space geodesics, which we develop in Section 2.3. In Sections 2.4 and 2.5, we briefly deal with the concepts of differentiability and convexity. The entirety of Chapter 2 occupies a large fraction of this thesis and can be seen as one of its primary contributions. Although the fact that geodesics exist in measure spaces is sometimes alluded to in literature, the author is not aware of any discussion of their properties that would remotely approach the scope of Section 2.3.

In Chapter 3, we turn our attention to the primary goal of the thesis: transferring NLP solution algorithms to similarity spaces. Section 3.1 is largely an extension of the work presented in [HLS22], though Sections 3.1.1 and 3.1.2 substantially expand upon prior work with their discussion of error control and working with error-prone evaluation methods. In Section 3.2, we discuss constrained optimization methods. This is an entirely new contribution of this thesis. In Section 3.2.1, we use the theory of measure space geodesics to develop an analogue of the Karush-Kuhn-Tucker theorem for optimization in similarity spaces. We then proceed to develop a quadratic penalty method based on the unconstrained trust region loop from Section 3.1. In Section 3.2.2, we propose a different kind of constraint that is based on boolean set operations. We develop a step-finding method for a hypothetical optimization algorithm that works on set variables whose values form a partition of the ground set. In Sections 3.2.2.2 and 3.2.2.3, we propose adaptive variants of the sum-up rounding method that could be used in conjunction with measure space geodesics to create relaxation methods that do not require a fixed control mesh.

The optimization algorithms that we develop in Chapter 3 are the second major contribution of this thesis. We provide a reusable implementation of them in the PYCOIMSET Python package, whose public release accompanies the submission of this thesis. PYCOIMSET is a central part of the work underlying this thesis. The package's state at the time of submission has been archived for purposes of scientific record-keeping [Hah25a; Hah25b].

In Chapter 4, we apply PYCOIMSET to two test problems to demonstrate the practical applicability of the algorithms proposed in the previous chapter.

Section 4.1 presents an unconstrained instance of the Lotka-Volterra fishing problem, which was already used as an example in [HLS22]. We dedicate some time to developing an error control method capable of estimating the errors relevant to the trust region steepest descent method and briefly discuss the results. In Section 4.2, we solve a PDE-constrained topology optimization problem with a scalar inequality constraint. We apply both unconstrained and constrained optimization to it and briefly discuss the results.

In Chapter 5, we discuss the strengths and weaknesses of our approach, enumerate some starting points for future research, and give some closing remarks.

This thesis includes several appendices. Sections A.1, A.2 and B.1 all lay out partial theoretical frameworks and arguments, or subroutines of hypothetical algorithms. These fragments are left from unsuccessful attempts to extend the theory developed in the main body of the thesis. At some point, these attempts were abandoned, but the given fragments are sufficiently well-developed and interesting that they should not be discarded thoughtlessly. These appendices are intended to assist future researchers who might attempt to complete these extensions. Chapter C briefly describes the interface and a few of the core architectural choices underlying PyCoimset.

1.5 NOTATION AND TERMINOLOGY

We adopt the standard symbols $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$ for natural numbers, integers, rational numbers, and real numbers, respectively. We do not consider 0 to be a natural number. We use $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$ to signify the set of all natural numbers and zero. We use the following shorthands for index sets:

$$[n] := \{i \in \mathbb{N} \,|\, i \leq n\} \quad \forall n \in \mathbb{Z},$$
$$[n]_0 := \{i \in \mathbb{N}_0 \,|\, i \leq n\} \quad \forall n \in \mathbb{Z}.$$

For subsets of these number classes that satisfy a certain predicate, we use a shorthand notation in which the restricting predicate is given as a subscript. For instance, we write

$$\mathbb{R}_{\geq 0} := \{x \in \mathbb{R} \,|\, x \geq 0\}.$$

Starting in Chapter 2, we rely heavily on residual classes and quotient spaces. Let X be a set and let $\sim$ be an equivalence relation on X. Then we define

$$[x]_\sim := \left\{y \in X \,\middle|\, x \sim y\right\} \quad \forall x \in X,$$
$$X/_\sim := \left\{[x]_\sim \,\middle|\, x \in X\right\}.$$

We refer to $[x]_\sim$ as the *residual class* or *equivalence class* of x with respect to $\sim$ and to $X/_\sim$ as the *quotient space* of X with respect to $\sim$. We generally use the symbol $\sim$ for equivalence relations and $\prec$, $\preceq$, $\succ$, or $\succeq$ for order relations. We attach indices to these symbols to distinguish between different relations of the same kind.

In Chapter 2, we take care to distinguish between measurable sets and their corresponding equivalence classes, to which we refer as "similarity classes." This is to ensure that our theoretical foundation does not rest on a confusion of terms. However, we show that, both can mostly be used interchangeably. Therefore, in later chapters, we are less stringent with this distinction. As a general rule, if an argument involves any "set" that is only well-defined up to a nullset, then

all involved objects should be assumed to also be similarity classes rather than concrete sets.

We sometimes speak of the value $f(x)$ of a measurable function f in a point x. It is somewhat philosophically disputable whether this is appropriate. In most cases, the functions of which we speak in this thesis are only well-defined up to differences on nullsets. Therefore, they technically do not have well-defined pointwise values. Instead, we would have to make all arguments based solely on the similarity classes of the function's sublevel sets, which are always well-defined objects for measurable functions. We do not do so as a concession to comprehensibility and reader comfort. The inclined reader may verify that all arguments that we make using pointwise function values could be made equally well using only the similarity classes of sublevel sets.

The letter λ is conventionally assigned to the Lebesgue measure, but may clash with conventional notations for Lagrange parameters. We use $\mathscr{B}(X)$ to signify the Borel-σ-algebra on X, which is the σ algebra generated by all relatively open subsets of X. $\mathscr{L}(X)$ signifies the Lebesgue-σ-algebra on X. We note that in some contexts, we also use $\mathscr{L}(V, W)$ to signify the set of all bounded linear mappings from a vector space V to a vector space W.

We regularly work with L^p spaces, which are normed vector spaces of functions whose p-th power is a Lebesgue-integrable function. Because we work with Lebesgue-integrable functions on fairly arbitrary measure spaces, we annotate these spaces with the applicable σ-algebra and measure: $L^p(\Sigma, \mu)$. Here, we omit the universal set X because it is implicitly given as the union over all sets in Σ. To conserve space, we sometimes also use the shorthand $L^p_\mu(\Sigma) \coloneqq L^p(\Sigma, \mu)$. In more classical ODE or PDE application contexts, we sometimes omit the σ-algebra in favor of simply stating the functions domain. In these cases, use of a Lebesgue-σ-algebra is implied. For integrable functions that do not map to $\mathbb{R}$, we add an additional argument to the end of the argument list to specify the function's codomain.

In Chapter 2, we introduce the "locally inverted difference variation" (LIDV, see Definition 2.4.3), which is a modified distance measurement between signed measures. For signed measures ϕ, ψ and a measurable set A, the A-inverted difference variation between ϕ and ψ is written as

$$(\phi \underset{A}{\ominus} \psi).$$

Although it is written as a difference, the LIDV is best thought of as the absolute value of a difference. In contrast to most differences, it is commutative. We note that, despite our use of the term "distance measurement", the LIDV only acts as a metric under very specific circumstances. These circumstances will be present in most contexts in which we use it.

Starting in Chapter 3, we define and describe algorithms. Some of these algorithms have subroutines. We distinguish between two types of subroutines: *functions* and *procedures*. The main distinction between the two is that functions always return the same output for a given input, whereas procedures may depend on cached states from prior invocations or random variables and may therefore return different outputs on different invocations with the same inputs. We make this distinction primarily to avoid introducing an excessive amount of opaque state variables, while still allowing for some of the peculiarities of highly optimized implementations.

Chapter 2

Theoretical Foundation

Before we formulate concrete optimization algorithms, we first have to understand how to work within our search space. Our goal is to transfer iterative optimization schemes such as gradient descent to spaces of measurable sets. As we have already indicated, spaces of measurable sets are not vector spaces. Therefore, it is not immediately obvious how to move from one solution to the next.

In this chapter, our goal is to understand the structure of that space and how to move from one iterate to the next within it. This presupposes some prior knowledge about measure theory.

2.1 MEASURE SPACES AND INTEGRALS

A general introduction to measure theory is far beyond the scope of this thesis. We refer to comprehensive textbooks such as [Bog07; Coh13; Kub15; Shi18] for an introduction. We primarily draw upon [Bog07] because it is exceptionally exhaustive.

At the foundation of most measure theory is the σ-algebra. A σ-algebra is a family of sets that is closed under complementation and countable union, and contains the empty set $\emptyset$. Because σ-algebras are closed under complementation, they also always contain the universal set, which is the complement of $\emptyset$. For us, closedness under countable union is the most significant property of a σ-algebra because it allows for a set to be composed of infinitesimally fine parts. In our setting, where optimization steps are essentially sets, this allows for steps to be divisible into smaller sub-steps, which allows for steps to be shortened.

Definition 2.1.1 (Measurable Spaces).
Let X be a set, and let $\Sigma \subseteq 2^X$ be a subset of the power set of X such that

(1) $\emptyset \in \Sigma$;

(2) $A^{\complement} \in \Sigma \ \forall A \in \Sigma$ (closed under complementation);

(3) $\bigcup_{i=1}^{\infty} A_i \in \Sigma \ \forall (A_i)_{i \in \mathbb{N}} \in \Sigma^{\mathbb{N}}$ (closed under countable union).

Then we refer to Σ as a *σ-algebra* and to (X, Σ) as a *measurable space*. ◁

We note that closedness under complementation and countable union also implies closedness under countable intersection.

The intersection of σ-algebras is always a σ-algebra. This holds for arbitrary infinite intersections. For a given universal set X and a family of subsets $\mathscr{F} \subseteq 2^X$, the intersection of all σ-algebras containing $\mathscr{F}$ is the smallest σ-algebra containing $\mathscr{F}$. We refer to this as the σ-algebra *generated by* the *generator* $\mathscr{F}$ and write:

$$\sigma(\mathscr{F}) \coloneqq \bigcap_{\substack{\Sigma\ \sigma\text{-algebra} \\ \mathscr{F}\subseteq\Sigma}} \Sigma.$$

In topological spaces, the σ-algebra generated by all open sets has a special role and we refer to it as the *Borel-σ-algebra* $\mathscr{B}(X)$.

Given a subset $Y \subseteq X$, and a σ-algebra Σ over X, we can derive a σ-algebra over Y by taking the intersections of all sets in X with Y. If Y is in Σ, then the result is a sub-σ-algebra of Σ. If X is a topological space, then $\mathscr{B}(Y)$ is the σ-algebra formed by intersecting all sets in $\mathscr{B}(X)$ with Y. This is an important convention on our part. In this case, $\mathscr{B}(Y)$ is the σ-algebra generated by the *relatively open* sets as opposed to the *open* sets. This means that we implicitly switch to the relative topology. Whenever we speak of a Borel-σ-algebra on a subset of our current universal set X, we assume this switch. Otherwise, we will make an explicit note of a change in topology.

We can extend the Borel-σ-algebra on $\mathbb{R}$ by allowing $\pm\infty$ as set elements. Effectively, this yields a σ-algebra containing the same sets as $\mathscr{B}(\mathbb{R})$ as well as all of their unions with subsets of $\{\pm\infty\}$. We refer to this σ-algebra as $\overline{\mathscr{B}}(\mathbb{R})$.

We refer to a σ-algebra as *countably generated* if it has a countable generator. The Borel-σ-algebra on $\mathbb{R}^n$ is countably generated.

2.1.1 Measures

In measurable spaces, there are special maps known as *measures*. There are several varieties of measures. However, they all have in common that they are additive over disjoint unions, including over countable unions. There are some sources (see, e.g., [Bog07]) that additionally distinguish between finitely additive measures and countably additive measures, but for our purposes, all measures are countably additive.

Definition 2.1.2 (Measure).
Let $(X,\ \Sigma)$ be a measurable space. A mapping $\mu\colon \Sigma \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ is called a *(positive) measure* if

(1) $\mu(\emptyset) = 0$;

(2) $\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$ for all sequences $(A_i)_{i\in\mathbb{N}}$ in Σ such that $A_i \cap A_j = \emptyset$ for all $i, j \in \mathbb{N}$ with $i \neq j$.

In this case $(X,\ \Sigma,\ \mu)$ is referred to as a *measure space*. If μ does not assume the value ∞, then μ and $(X,\ \Sigma,\ \mu)$ are both referred to as *finite*. If there exists a sequence $(A_i)_{i\in\mathbb{N}}$ in Σ such that

(1) $\mu(A_i) < \infty$, and

(2) $X = \bigcup_{i=1}^{\infty} A_i$,

then μ and (X, Σ, μ) are referred to as *σ-finite*. ◁

We note that countable additivity is also sometimes referred to as *σ-additivity*. Because they are additive on arbitrarily small countable partitions of sets, measures behave a bit like linear maps in vector spaces. If we add two disjoint sets together, then their measures add. If we "scale" a set by choosing a smaller subset, then the measure of the original set is split into the measure of the "scaled" set and the measure of its complement. Measures are useful as substitutes for norms. However, we can make them more similar to linear maps by allowing negative values.

Definition 2.1.3 (Signed Measure).
Let (X, Σ) be a measurable space. A mapping $\varphi\colon \Sigma \to \mathbb{R} \cup \{\pm\infty\}$ is called a *signed measure* if

(1) $\varphi(\emptyset) = 0$;

(2) $\varphi\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \varphi(A_i)$ for all sequences $(A_i)_{i\in\mathbb{N}}$ in Σ such that $A_i \cap A_j = \emptyset$ for all $i, j \in \mathbb{N}$ with $i \neq j$.

If φ does not assume the value $\pm\infty$, then ϕ is referred to as *finite*. ◁

It is evident due to additivity that a well-defined signed measure can never assume both ∞ and $-\infty$ as a value at the same time. Signed measures are a generalization of positive measures. Similarly, finite signed measures can be generalized to measures that assume vector values. Such measures are not commonly discussed in measure theoretical literature. Therefore, we refer to the following definition from [Hof71].

Definition 2.1.4 (Vector Measures).
Let (X, Σ) be a measurable space, and let E be a locally convex Hausdorff space. A mapping $\nu\colon \Sigma \to E$ is called a *vector measure* or *E-measure* if

(1) $\varphi(\emptyset) = 0_E$;

(2) $\varphi\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \varphi(A_i)$ for all sequences $(A_i)_{i\in\mathbb{N}}$ in Σ such that $A_i \cap A_j = \emptyset$ for all $i, j \in \mathbb{N}$ with $i \neq j$. ◁

While Definition 2.1.4 is stated in terms of locally convex topological vector spaces, most sources assume a Banach space (see, e.g., [BS20; Coh13]) when introducing vector-valued integrals.

Sets of measure zero are known as *nullsets* with an annotation indicating with respect to which measure they are nullsets if necessary. If we treat measures like norms, then nullsets have norm zero. Therefore, we have to treat all nullsets as if they were the empty set. When we deal with multiple measures at the same time, they have to roughly agree on which sets are nullsets. The relevant property here is "absolute continuity." In order to introduce this property, we first have to learn about the Hahn-Jordan decomposition and total variation. The following steps are taken from [BS20, Sec. 2.6].

Theorem 2.1.5 (Hahn Decomposition [BS20]).
Let φ be a signed measure on a measurable space (X, Σ). Then there exists $X^- \in \Sigma$ such that with $X^+ := X \setminus X^-$, we have

$$\varphi(A \cap X^-) \leq 0 \text{ and } \varphi(A \cap X^+) \geq 0 \quad \forall A \in \Sigma. \qquad \triangleleft$$

The Hahn decomposition is a partition of the universal set into subsets X^+ and X^- such that φ is non-negative on X^+ and non-positive on X^-. The Jordan decomposition is a corresponding decomposition of the signed measure itself.

Corollary 2.1.6 (Jordan Decomposition [BS20]).
Let φ, X^-, and X^+ satisfy Theorem 2.1.5. Let

$$\varphi^+(A) \coloneqq \varphi(A \cap X^+) \text{ and } \varphi^-(A) \coloneqq -\varphi(A \cap X^-) \quad \forall A \in \Sigma.$$

Then φ^+ and φ^- are positive measures and $\varphi = \varphi^+ - \varphi^-$. ◁

The Hahn decomposition is a decomposition of sets, while the Jordan decomposition is a decomposition of measures. They are strictly speaking different decompositions. However, because they are so closely related, they are sometimes collectively referred to as the *Hahn-Jordan decomposition* or the *Jordan-Hahn decomposition*.

Definition 2.1.7 (Total Variation [BS20]).
The measures φ^+ and φ^- constructed in Corollary 2.1.6 are called the *positive* and *negative parts of φ*, respectively. The measure

$$|\varphi| \coloneqq \varphi^+ + \varphi^-$$

is called the *total variation of φ*. The quantity $\|\varphi\| \coloneqq |\varphi|(X)$ is called the *variation* or *variation norm* of the signed measure φ. If $\|\varphi\| < \infty$, then we refer to φ as a signed measure of *bounded variation*. ◁

The term "variation norm" is not misleading in this case. The set of all signed measures of bounded variation over a given measurable space (X, Σ) is indeed a normed real vector space. It is, in fact, a Banach space.

It can be shown that

$$\begin{aligned}
\mu^+(A) &= \sup\{\mu(B) \,|\, B \subseteq A,\ B \in \Sigma\},\\
\mu^-(A) &= \sup\{-\mu(B) \,|\, B \subseteq A,\ B \in \Sigma\},\\
|\mu|(A) &= \sup\Big\{\sum_{i=1}^{\infty} \big|\mu(A_i)\big| \,\Big|\, (A_i)_{i\in\mathbb{N}} \in \Sigma^{\mathbb{N}} \text{ pairwise disjoint partition of } A\Big\}
\end{aligned}$$

for all $A \in \Sigma$. This definition can be expanded to vector measures.

Definition 2.1.8 (Total Variation of Vector Measures [Hof71]).
Let (X, Σ) be a measurable space, let E be a locally convex topological vector space, let $q\colon E \to \mathbb{R}$ be a seminorm on E, and let $\nu\colon \Sigma \to E$ be a vector measure on Σ. Then the measure $|\nu|_q$ with

$$|\nu|_q(A) \coloneqq \sup\Big\{\sum_{i=1}^{\infty} q\big(\nu(A_i)\big) \,\Big|\, (A_i)_{i\in\mathbb{N}} \in \Sigma^{\mathbb{N}} \text{ pairwise disjoint partition of } A\Big\}$$

for all $A \in \Sigma$ is called the *q-variation* of ν. If E is a normed vector space and q is the primary norm in E, then we refer to $|\nu|_q$ as the *total variation* of ν and write $|\nu| \coloneqq |\nu|_q$. In this case, we refer to $\|\nu\| \coloneqq |\nu|(X)$ as the *variation* or *variation norm* of ν. ◁

For signed measures, we can think of nullsets as falling into two categories: some nullsets consist of parts with strictly positive and strictly negative measure such that the measures of both parts cancel each other out, while other nullsets only have subsets of measure zero. If two signed measures agree on which sets fall into the latter category, then we call them "equivalent" or "absolutely continuous" with respect to one another.

Definition 2.1.9 (Absolute Continuity [BS20]).
Let μ and ν be two signed measures on a shared measurable space (X, Σ). We call ν *absolutely continuous* with respect to μ if

$$|\nu|(A) = 0 \quad \forall A \in \Sigma \colon |\mu|(A) = 0.$$

In this case, we write $\nu \ll \mu$. If $\nu \ll \mu$ and $\mu \ll \nu$, then we call μ and ν *equivalent*.◁

We note that this definition also translates to vector measures. Including this in Definition 2.1.9 complicates the definition because formulating the same definition for vector measures is not strictly a generalization. After all, non-finite signed measures are nominally not vector measures.

One important attribute of measure space is "atomicity." A measure is said to be "atomic" if there are atoms. An atom is a set of strictly positive measure which cannot be split into disjoint pieces of strictly positive measure, thus making its measure indivisible.

Definition 2.1.10 (Atoms and Atomlessness).
Let (X, Σ, μ) be a measure space. A set $A \in \Sigma$ is called a μ-*atom* if $\mu(A) > 0$ and for all $B \in \Sigma$ with $B \subseteq A$, we either have $\mu(B) = 0$ or $\mu(B) = \mu(A)$.

The measure μ is called *atomless* if there are no μ-atoms. In this case, we also refer to the measure space as atomless. ◁

Atomlessness is central to our entire endeavor, because it allows us to cut sets into chunks of arbitrary size, which is the process with which we replace scaling. We particularly emphasize the role of [Bog07, Thm. 1.12.9] and [Bog07, Cor. 1.12.10] for our work.

2.1.2 Lebesgue Integrals and Density Functions

Measures are fundamental to the definition of the Lebesgue integral. A full introduction to the Lebesgue integral is significantly beyond the scope of this thesis, so we will only cite a few important definitions and theorems here. For complete introductions, we refer to [Bog07; BS20; Coh13]. The first important definition is that of a "measurable function."

Definition 2.1.11 (Measurable Functions).
Let (X, Σ_X) and (Y, Σ_Y) be measurable spaces. A function $f \colon X \to Y$ is called Σ_X-Σ_Y-*measurable* if

$$f^{-1}(A) \in \Sigma_X \quad \forall A \in \Sigma_Y.$$

If Σ_Y is a Borel-σ-algebra on Y, then we refer to f as being Σ_X-*measurable*, because this is the convention in most sources on Lebesgue integrals. In cases where it is clear what Σ_X is, we simply call f *measurable*.

If both Σ_X and Σ_Y are either Borel- or Lebesgue-σ-algebras, then we use simplified expression such as *Lebesgue-Borel-measurable*, *Borel-Borel-measurable*, *Lebesgue-measurable*, or *Borel-measurable*. ◁

For Σ_X-Σ_Y-measurable functions f, the *σ-algebra generated by f* is the σ-algebra generated by the preimages of members of Σ_Y, which is equal to the family of preimages:

$$\sigma(f) \coloneqq \sigma\Big(\big\{f^{-1}(A)\,\big|\,A \in \Sigma_Y\big\}\Big) = \big\{f^{-1}(A)\,\big|\,A \in \Sigma_Y\big\}.$$

It can be shown that the Borel-σ-algebra on $\mathbb{R}^n$ is the σ-algebra generated by all Cartesian products of n intervals. The Lebesgue-σ-algebra $\mathscr{L}(\mathbb{R}^n)$ is an extension of $\mathscr{B}(\mathbb{R}^n)$. It can be shown that there is exactly one positive measure that assigns each Cartesian product of intervals the product of the length of those intervals as a measure. This measure is referred to as the *Lebesgue measure λ*. The Lebesgue-σ-algebra $\mathscr{L}(\mathbb{R}^n)$ is obtained by adding all sets to $\mathscr{B}(\mathbb{R}^n)$ that lie strictly between two Borel sets whose difference has Lebesgue measure zero.

For our purposes, Borel-measurable functions are of particular interest because they can be chained without losing measurability. Furthermore, the distinction between Lebesgue- and Borel-measurable sets becomes irrelevant once one ignores nullsets, because every Lebesgue-measurable set is infinitesimally close to a Borel set.

As opposed to dividing the domain into infinitesimal pieces, as it is done to approximate the Riemann integral, the Lebesgue integral divides the codomain. The codomain is generally a subset of $\mathbb{R}$ and is divided into intervals. We then add up the product between the upper and lower bounds of the interval and the measure of the interval's preimage to obtain upper and lower approximations for the integral. The advantage of this approach is that it is agnostic with respect to the domain and only operates on the codomain, which means that many properties of the Lebesgue integral and measurable functions can be proven irrespective of what domain a function is defined on.

Lebesgue-integrable functions are grouped into L^p spaces. We write

$$L^p(\Sigma,\,\mu) \coloneqq \left\{f\colon X \to \mathbb{R}\,\middle|\, f\ \Sigma\text{-Borel-measurable},\ \|f\|_{L^p} \coloneqq \left(\int_X |f|^p\,\mathrm{d}\mu\right)^{\frac{1}{p}} < \infty\right\}$$

for a measure space $(X,\,\Sigma,\,\mu)$, where X is implicitly given by the biggest set contained in Σ, and $1 < p < \infty$. For $p = \infty$, the space L^p is given by

$$L^\infty(\Sigma,\,\mu) \coloneqq \big\{f\colon X \to \mathbb{R}\,\big|\, f\ \Sigma\text{-Borel-measurable},\ \|f\|_{L^\infty} \coloneqq \operatorname*{ess\,sup}_{x\in X}|f| < \infty\big\}.$$

Generally speaking, the L^p spaces are not contained within each other unless $(X,\,\Sigma,\,\mu)$ is a finite measure space, in which case $L^p(\Sigma,\,\mu) \subseteq L^q(\Sigma,\,\mu)$ for $p \geq q$. All L^p spaces are Banach spaces, but L^2 is especially important, because it is also a Hilbert space with the inner product

$$\langle f,\,g\rangle_{L^2} \coloneqq \int_X fg\,\mathrm{d}\mu.$$

Note that we only consider real vector spaces here.

There exists a variety of useful theoretical results about L^p spaces, such as the Hölder and Minkowski inequalities, the monotone convergence theorem, Fatou's lemma, the dominated convergence theorem, Vitali's theorem, and many more. We will not discuss these here, though we will cite them if necessary. We refer to [Bog07] as a source for these theorems unless otherwise noted.

We will not discuss weakly differentiable function here. Their use is important in the theory of partial differential equations. However, they are otherwise irrelevant to our work. Above all, they have no particular bearing on the theory of optimization in measure spaces.

Of particular interest within the context of this discussion are integrable functions whose integral corresponds to certain signed measures. Let (X, Σ, μ) be a measure space and let $f \in L^1(\Sigma, \mu)$. Then it can be shown that $\varphi\colon \Sigma \to \mathbb{R}$ with

$$\varphi(U) \coloneqq \int_U f \,\mathrm{d}\mu \quad \forall U \in \Sigma$$

is a finite signed measure over (X, Σ). The characterization of this situation is generally referred to as the *Radon-Nikodym theorem*.

Theorem 2.1.12 (Radon-Nikodym [Bog07, Thm. 3.2.2]).
Let (X, Σ) be a measurable space, and let φ and ν be finite signed measures over (X, Σ). Then we have $\nu \ll \varphi$ if and only if there exists $f \in L^1(\Sigma, \varphi)$ such that

$$\nu(U) = \int_U f \,\mathrm{d}\varphi.$$

◁

We note that this variant of the Radon-Nikodym theorem also applies when both measures are signed. However, we are mostly interested in the case in which φ is a non-negative measure. The function f whose existence is established by the Radon-Nikodym theorem is often referred to as the "density function" or "Radon-Nikodym derivative" of ν with respect to φ. Indeed, Radon-Nikodym derivatives will play a large role in bringing our derivatives into a form that we can use in an algorithmic context.

As a final introductory note on measurable functions, we cite a property that is not widely discussed in fields adjacent to PDE-constrained optimization, but has an interesting application with respect to measure space geodesics (see Section 2.3.4): the *Doob-Dynkin property* [Bog07, vol. 2, p. 51]. The Doob-Dynkin property is an interesting example of a property that only depends on the codomain of the function. The nature of the domain is entirely irrelevant.

Definition 2.1.13 (Doob-Dynkin Property [Bog07]).
A measurable space (X, Σ_X) is said to have the *Doob-Dynkin property* if for every pair of measurable spaces (E, Σ_E) and (F, Σ_F), every Σ_E-Σ_F-measurable function $f \coloneqq E \to F$, and every Σ_E-Σ_X-measurable g with $\sigma(g) \subseteq \sigma(f)$, there exists a Σ_F-Σ_X-measurable function $h\colon F \to X$ such that $g = h \circ f$. ◁

What the Doob-Dynkin property essentially means is that if we have two measurable functions f and g such that g generates a smaller σ-algebra than f and the codomain of g has the Doob-Dynkin property, then we can construct a measurable function h such that $g = h \circ f$. As is noted in [Bog07], $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ has the Doob-Dynkin property.

2.1.3 "Layering" and Multivariate Problems

At first glance, Problem 1.1 on page 5 does not appear to account for problems with multiple set-valued variables. However, we can show that multivariate problems are a special case of univariate problems. It is well-known that measure spaces

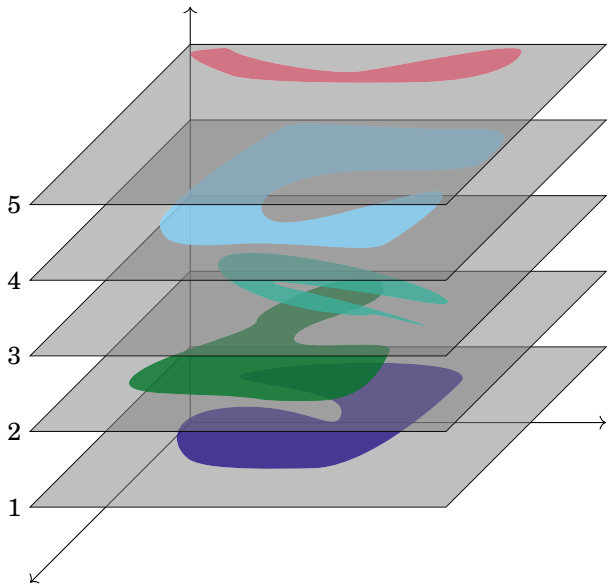


Figure 2.1: Illustration of "layering" with the vertical axis as "index dimension."

can be combined into *product spaces* by forming Cartesian products of measurable sets and extending the *product measure* from a measure that assigns products of measures to rectangles.

The process required to generalize from univariate to multivariate problems is slightly different. We will refer to this process as "layering." Let subsequently $\big((X_i,\, \Sigma_i,\, \mu_i)\big)_{i\in[n]}$ be a tuple of measure spaces with $n \in \mathbb{N}$. The idea of layering is to introduce an *index dimension* in which the n universal sets X_i are layered, one over the other, as depicted in Figure 2.1. In this section, we will show that the result can very easily be made into a measure space of its own.

Remark 2.1.14.
Layering is proposed in [Fre09, 214L–214N, 214Xh–214Xk] under the name "direct sum." The author of this thesis is not aware of any other sources that use this term. Layered measure spaces substantially simplify this thesis by obviating the need for a separate discussion of multivariate problems. In conjunction with the fact that [Fre09] does not provide proofs for many of the claims made, this justifies a more in-depth discussion of the concept. We prefer the term "layering" over "direct sum" to stress that the process of forming a layered measure space is distinct from forming a direct sum in the algebraic sense, though [Fre09] is correct in pointing out that the intent behind both is analogous. ◁

First, we define the layered universal set

$$X \coloneqq \bigcup_{i=1}^{n} \big(\{i\} \times X_i\big) \subseteq [n] \times \Big(\bigcup_{i=1}^{n} X_i\Big).$$

and the layered σ-algebra

$$\Sigma \coloneqq \Big\{ U \subseteq X \;\Big|\; \underbrace{\big\{x \in X_i \,\big|\, (i,\, x) \in U\big\}}_{=:U_i} \in \Sigma_i \;\; \forall i \in [n] \Big\}.$$

In this definition, we refer to U_i as the *i-th layer* of U. Finally, with the definition of the i-th layer, it is relatively simple to define a measure on (X, Σ). We define $\mu\colon \Sigma \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ as

$$\mu(U) \coloneqq \sum_{i=1}^{n} \mu_i(U_i) \quad \forall U \in \Sigma.$$

Of course, we still have to prove that Σ is a σ-algebra and that μ is a measure.

Definition 2.1.15 (Layered Measure Spaces).
Let $n \in \mathbb{N}$, and let (X_i, Σ_i, μ_i) be measure spaces for $i \in [n]$. Then we refer to the tuple (X, Σ, μ) where μ is a mapping $\mu\colon \Sigma \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ such that

$$\begin{aligned} X &\coloneqq \bigcup_{i=1}^{n} \big(\{i\} \times X_i\big), \\ \Sigma &\coloneqq \Big\{ U \subseteq X \,\Big|\, U_i \in \Sigma_i \ \forall i \in [n] \Big\}, \\ \mu(U) &\coloneqq \sum_{i=1}^{n} \mu_i(U_i) \qquad\qquad \forall U \in \Sigma, \end{aligned}$$

and

$$U_i \coloneqq \big\{ x \in X_i \,\big|\, (i, x) \in U \big\} \qquad \forall U \in \Sigma,\ i \in [n]$$

as the *layered measure space* or simply the *layering* of $\big((X_i, \Sigma_i, \mu_i)\big)_{i \in [n]}$. For each $U \in \Sigma$ and $i \in [n]$, we refer to U_i as the *i-th layer* of U. Similarly, we refer to

$$\Sigma_i = \{U_i \,|\, U \in \Sigma\} \quad \forall i \in [n]$$

as the *i-th layer* of Σ. ◁

We stress that layered measure spaces are not a special case of product measure spaces. In a product space, each layer of X would have to be the same. However, the layers of a layered space can be different spaces that draw from different universal set and use different measures.

What we are doing here is easier to understand if we think of subsets of the layered set X as *relations* between indices in $[n]$ and points in their respective X_i. The i-th layer of a set is the set of all points that "relate" to the index i. The layered σ-algebra Σ consists of relations U where the sets relating to the indices i are in their respective σ-algebras Σ_i. In many ways, this behaves like a tuple of measurable sets. However, this construct allows us to treat the entire tuple as one measurable set, which means that we do not need to develop any separate theory for tuples of measurable sets in order to solve multivariate problems.

In addition to the fact that the layered space is actually a measure space, we also have to show that the process of layering does not break desirable properties of the layer algebras Σ_i. As a prerequisite, we note that it is evident that for any layered set

$$U = \bigcup_{i=1}^{n} \big(\{i\} \times U_i\big)$$

with $U_i \in \Sigma_i$ for all $i \in [n]$, the i-th layer of U is exactly U_i. This is very straightforward and simplifies a lot of the remaining proofs.

Theorem 2.1.16 (Layered Measure Spaces).
Let $n \in \mathbb{N}$, let $\big((X_i,\ \Sigma_i,\ \mu_i)\big)_{i\in[n]}$ be a tuple of measure spaces, and let $(X,\ \Sigma,\ \mu)$ refer to the layering of $\big((X_i,\ \Sigma_i,\ \mu_i)\big)_{i\in[n]}$. Then $(X,\ \Sigma,\ \mu)$ is a measure space. ◁

PROOF. PART 1 (Σ IS A σ-ALGEBRA). We have

$$\emptyset = \bigcup_{i=1}^{n} \big(\{i\} \times \emptyset\big) \in \Sigma.$$

For every $U \in \Sigma$, there exist $U_i \in \Sigma_i$ for $i \in [n]$ such that

$$U = \bigcup_{i=1}^{n} \big(\{i\} \times U_i\big).$$

We find that

$$\begin{aligned}
U^{\complement} &= \Big(\bigcup_{i=1}^{n} \big(\{i\} \times U_i\big)\Big)^{\complement} \\
&= \bigcap_{i=1}^{n} \big(\{i\} \times U_i\big)^{\complement} \\
&= \bigcap_{i=1}^{n} \Big(\big(\{i\} \times U_i^{\complement}\big) \cup \bigcup_{\substack{j=1 \\ j\neq i}}^{n} \big(\{j\} \times X_j\big)\Big) \\
&= \bigcup_{i=1}^{n} \Big(\big(\{i\} \times U_i^{\complement}\big) \cap \bigcap_{\substack{j=1 \\ j\neq i}}^{n} \big(\{i\} \times X_i\big)\Big) \\
&= \bigcup_{i=1}^{n} \big(\{i\} \times U_i^{\complement}\big) \\
&\in \Sigma.
\end{aligned}$$

Let $(U_i)_{i\in\mathbb{N}}$ be a tuple in Σ. For each $i \in \mathbb{N}$ and $j \in [n]$, let $U_{i,j} \in \Sigma_j$ be such that

$$U_i = \bigcup_{j=1}^{n} \big(\{j\} \times U_{i,j}\big) \quad \forall i \in \mathbb{N}.$$

Then we have

$$\begin{aligned}
\bigcup_{i=1}^{\infty} U_i &= \bigcup_{i=1}^{\infty} \bigcup_{j=1}^{n} \big(\{j\} \times U_{i,j}\big) \\
&= \bigcup_{j=1}^{n} \bigcup_{i=1}^{\infty} \big(\{j\} \times U_{i,j}\big) \\
&= \bigcup_{j=1}^{n} \Big(\{j\} \times \underbrace{\bigcup_{i=1}^{\infty} U_{i,j}}_{\in\Sigma_j}\Big) \\
&\in \Sigma.
\end{aligned}$$

Thus, Σ contains $\emptyset$ and is closed under complementation and countable union. Σ is therefore a σ-algebra.

PART 2 (μ IS A MEASURE). It is evident that μ is non-negative. We have

$$\mu(\emptyset) = \sum_{i=1}^{n} \mu_i(\underbrace{\emptyset_i}_{i\text{-th layer}}) = \sum_{i=1}^{n} \mu_i(\emptyset) = 0.$$

Let $(U_i)_{i\in\mathbb{N}}$ be a sequence in Σ such that $U_i \cap U_j = \emptyset$ for $i \neq j$. For $i \in \mathbb{N}$ and $j \in [n]$, let $U_{i,j} \in \Sigma_j$ be such that

$$U_i = \bigcup_{j=1}^{n} \big(\{j\} \times U_{i,j}\big) \quad \forall i \in \mathbb{N}.$$

Then we have

$$\begin{aligned}
\mu\Big(\bigcup_{i=1}^{\infty} U_i\Big) &= \sum_{j=1}^{n} \mu_j\Big(\bigcup_{i=1}^{\infty} U_{i,j}\Big) \\
&= \sum_{j=1}^{n} \sum_{i=1}^{\infty} \mu_j(U_{i,j}) \\
&= \sum_{i=1}^{\infty} \sum_{j=1}^{n} \mu_j(U_{i,j}) \\
&= \sum_{i=1}^{\infty} \mu(U_i).
\end{aligned}$$

Here, exchanging the finite sum and the series is valid because all measures involved are non-negative and therefore, convergence (which is implied by boundedness) of the series in one arrangement implies convergence of the series in the other arrangement. □

Having proven that layered measure spaces are measure spaces, we turn our attention to the preservation of properties of the layer σ-algebras. The properties of measure spaces that we will primarily be interested in are atomlessness, finiteness, σ-finiteness, and being countably generated. For each of these properties, we can show that it can be transferred from the individual layers to the layered space in the sense that the layered space has the property if and only if all layers individually have it.

Theorem 2.1.17 (Property Inheritance for Layered Measure Spaces). *Let $n \in \mathbb{N}$, let $\big((X_i, \Sigma_i, \mu_i)\big)_{i\in[n]}$ be a tuple of measure spaces, and let (X, Σ, μ) refer to the layering of $\big((X_i, \Sigma_i, \mu_i)\big)_{i\in[n]}$. Then the following statements hold:*

1. *(X, Σ, μ) is atomless if and only if (X_i, Σ_i, μ_i) is atomless for all $i \in [n]$;*
2. *(X, Σ, μ) is finite if and only if (X_i, Σ_i, μ_i) is finite for all $i \in [n]$;*
3. *(X, Σ, μ) is σ-finite if and only if (X_i, Σ_i, μ_i) is σ-finite for all $i \in [n]$;*
4. *(X, Σ, μ) is countably generated if and only if (X_i, Σ_i, μ_i) is countably generated for all $i \in [n]$.* ◁

PROOF. PART 1 (ATOMLESSNESS). First, we consider a case where the layered space $(X,\ \Sigma,\ \mu)$ is not atomless. Let $U \in \Sigma$ be a μ-atom. By definition, we have $\mu(U) > 0$, which implies that U has at least one layer of strictly positive measure. Let $i \in [n]$ be such that the i-th layer U_i has strictly positive measure $\mu_i(U_i) > 0$. Our claim is that U_i is a μ_i-atom.

Let $V_i \in \Sigma_i$ such that $V_i \subset U_i$. If $0 < \mu(V_i) < \mu(U_i)$, then we could define a layered set

$$V \coloneqq \big(\{i\} \times V_i\big) \cup \bigcup_{\substack{j=1 \\ j \neq i}}^{n} \big(\{j\} \times U_j\big) \in \Sigma.$$

V would evidently satisfy $V \subseteq U$ because $V_i \subseteq U_i$. Furthermore, we would have $\mu(V) \geq \mu_i(V_i) > 0$ and

$$\mu(V) = \mu_i(V_i) + \sum_{\substack{j=1 \\ j \neq i}} \mu_j(U_j) < \sum_{j=1}^{n} \mu_j(U_j).$$

However, this is impossible because U is a μ-atom. This demonstrates by contradiction that such a subset V_i cannot exist and that, U_i must therefore be a μ_i-atom. Thus, if $(X,\ \Sigma,\ \mu)$ is not atomless, then there exists at least one layer i such that $(X_i,\ \Sigma_i,\ \mu_i)$ is also not atomless.

Next, we discuss the case in which at least one layer is not atomless. Let $i \in [n]$ be such that $(X_i,\ \Sigma_i,\ \mu_i)$ is not atomless. Let $U_i \in \Sigma_i$ be a μ_i-atom. Our claim is that $U \coloneqq \{i\} \times U_i$ is a μ-atom.

Let $V \in \Sigma$ be such that $V \subseteq U$. Because each layer of V is a subset of the corresponding layer of U, we have $V_j = \emptyset$ for $j \neq i$ and $V_i \subseteq U_i$. Because U_i is a μ_i-atom, we have

$$\mu(V) = \underbrace{\mu_i(V_i)}_{\in \{0,\ \mu_i(U_i)\}} \in \big\{0,\ \mu_i(U_i)\big\} = \big\{0,\ \mu(U)\big\}.$$

This proves that U is a μ-atom and demonstrates that $(X,\ \Sigma,\ \mu)$ is also not atomless.

PART 2 (FINITENESS). First, we consider the case where $\mu(X) < \infty$. For every $i \in [n]$, we have

$$\mu_i(X_i) = \mu\big(\{i\} \times X_i\big) \leq \mu(X) < \infty,$$

which shows that $(X_i,\ \Sigma_i,\ \mu_i)$ is finite. Conversely, if $\mu_i(X_i) < \infty$ for all $i \in [n]$, then

$$\mu(X) = \sum_{i=1}^{n} \mu_i(X_i) < \infty,$$

which demonstrates that $(X,\ \Sigma,\ \mu)$ is finite.

PART 3 (σ-FINITENESS). First, we consider the case where $(X,\ \Sigma,\ \mu)$ is σ-finite. Let $(U_i)_{i \in \mathbb{N}}$ be a sequence such that $\mu(U_i) < \infty$ and

$$X = \bigcup_{i=1}^{\infty} U_i.$$

Let $j \in [n]$. We know that X_j is the j-th layer of X. We have

$$X_j = \Big(\bigcup_{i=1}^{\infty} U_i\Big)_j = \bigcup_{i=1}^{\infty} U_{i,j}$$

where $U_{i,j}$ is the j-th layer of U_i, and

$$\mu_j(U_{i,j}) = \mu\big(\{j\} \times U_{i,j}\big) \leq \mu(U_i) < \infty \quad \forall i \in \mathbb{N}.$$

This demonstrates that $(X_j,\ \Sigma_j,\ \mu_j)$ is σ-finite. Conversely, let $(X_j,\ \Sigma_j,\ \mu_j)$ be σ-finite for all $j \in [n]$. For every $j \in [n]$, let $(U_{j,i})_{i\in\mathbb{N}}$ be a sequence in Σ_j such that $\mu_j(U_{j,i}) < \infty$ for all $i \in \mathbb{N}$ and

$$X_j = \bigcup_{i=1}^{\infty} U_{j,i}.$$

For every $i \in \mathbb{N}$, we define

$$V_i \coloneqq \bigcup_{j=1}^{n} \big(\{j\} \times U_{j,i}\big).$$

We have

$$\mu(V_i) = \sum_{j=1}^{n} \underbrace{\mu_j(U_{j,i})}_{<\infty} < \infty \quad \forall i \in \mathbb{N}$$

and

$$\bigcup_{i=1}^{\infty} V_i = \bigcup_{j=1}^{n} \Big(\{j\} \times \bigcup_{i=1}^{n} U_{j,i}\Big) = \bigcup_{j=1}^{n} \big(\{j\} \times X_j\big) = X.$$

Therefore, $(X,\ \Sigma,\ \mu)$ is also σ-finite.

Part 4 (Countable Generation "$\Longleftarrow$"). For $i \in [n]$, let $(U_{i,j})_{j\in\mathbb{N}} \in \Sigma_i^{\mathbb{N}}$ be a countable generator of Σ_i. We show that

$$\mathscr{U} \coloneqq \big\{\{i\} \times U_{i,j} \,\big|\, i \in [n],\ j \in \mathbb{N}\big\} \cup \big\{\{i\} \times X_i \,\big|\, i \in [n]\big\}$$

is a generator of Σ. It is evident that

$$\mathbb{N} \ni k \mapsto \left\{k - n \cdot \left\lfloor \frac{k-1}{n} \right\rfloor\right\} \times \begin{cases} X_{k-n\cdot\left\lfloor\frac{k-1}{n}\right\rfloor} & \text{if } k \leq n, \\ U_{k-n\cdot\left\lfloor\frac{k-1}{n}\right\rfloor,\ \left\lceil\frac{k}{n}\right\rceil-1} & \text{if } k > n \end{cases}$$

is an enumeration of $\mathscr{U}$. This sequence first iterates through the sets $\{i\} \times X_i$ for $i \in [n]$ in ascending order. It then iterates through all $\{i\} \times U_{i,j}$ for $i \in [n]$ and $j \in \mathbb{N}$ in order of ascending j and, in an inner loop, ascending i.

For all $i \in [n]$ and $j \in \mathbb{N}$, we have $U_{i,j} \in \Sigma_i$ and therefore $\{i\} \times U_{i,j} \in \Sigma$. We also have $X_i \in \Sigma_i$ and therefore $\{i\} \times X_i \in \Sigma$. This means that $\mathscr{U} \subseteq \Sigma$ and therefore $\sigma(\mathscr{U}) \subseteq \Sigma$.

To prove the converse inclusion, it is sufficient to show that for each $i \in [n]$, the family

$$\widetilde{\Sigma}_i \coloneqq \big\{U \in \Sigma_i \,\big|\, \{i\} \times U \in \sigma(\mathscr{U})\big\}$$

is equal to Σ_i. The inclusion $\Sigma \subseteq \sigma(\mathscr{U})$ then follows because $\sigma(\mathscr{U})$ is closed under finite union.

We have $\widetilde{\Sigma}_i \subseteq \Sigma_i$ by definition. We also have $X_i \in \widetilde{\Sigma}_i$ and $U_{i,j} \in \widetilde{\Sigma}_i$ by definition of $\mathscr{U}$. Because $(U_{i,j})_{j\in\mathbb{N}}$ is a generator of Σ_i, this means that if $\widetilde{\Sigma}_i$ is a σ-algebra, then $\widetilde{\Sigma}_i = \Sigma_i$.

We trivially have $\{i\} \times \emptyset = \emptyset \in \sigma(\mathscr{U})$ and therefore $\emptyset \in \widetilde{\Sigma}_i$. Let $U \in \widetilde{\Sigma}_i$. Because $\sigma(\mathscr{U})$ is a σ-algebra and therefore closed under set differences, we have

$$\{i\} \times U^{\complement} = \underbrace{\big(\{i\} \times X_i\big)}_{\in\sigma(\mathscr{U})} \setminus \underbrace{\big(\{i\} \times U\big)}_{\in\sigma(\mathscr{U})} \in \sigma(\mathscr{U})$$

which implies that $U^{\complement} \in \widetilde{\Sigma}_i$. Finally, given a sequence $(V_j)_{j\in\mathbb{N}} \in \widetilde{\Sigma}_i^{\mathbb{N}}$, we have

$$\{i\} \times \Big(\bigcup_{j=1}^{\infty} V_j\Big) = \bigcup_{j=1}^{\infty} \underbrace{\big(\{i\} \times V_j\big)}_{\in\sigma(\mathscr{U})} \in \sigma(\mathscr{U})$$

and therefore $\bigcup_{j=1}^{\infty} V_j \in \widetilde{\Sigma}_i$. We have thus demonstrated that $\widetilde{\Sigma}_i$ is a σ-algebra with respect to the universal set X_i. Because $\widetilde{\Sigma}_i$ contains $U_{i,j}$ for $j \in \mathbb{N}$, $\widetilde{\Sigma}_i$ is at least as large as Σ_i. However, we have $\widetilde{\Sigma}_i \subseteq \Sigma_i$ by definition. Therefore, we have $\widetilde{\Sigma}_i = \Sigma_i$ for all $i \in [n]$. According to the definition of $\widetilde{\Sigma}_i$, this means that

$$\{i\} \times U \in \sigma(\mathscr{U}) \quad \forall i \in [n],\, U \in \Sigma_i.$$

Because $\sigma(\mathscr{U})$ is closed under finite union, we have

$$\bigcup_{i=1}^{n} \big(\{i\} \times V_i\big) \in \sigma(\mathscr{U}) \quad \forall (V_i)_{i\in[n]} \colon V_i \in \Sigma_i \ \forall i \in [n]$$

and therefore $\Sigma \subseteq \sigma(\mathscr{U})$. In conjunction with the prior inclusion, this yields the overall equality

$$\sigma(\mathscr{U}) = \Sigma$$

and therefore demonstrates that $\mathscr{U}$ is a generator of Σ and that Σ is countably generated.

PART 5 (COUNTABLE GENERATION "$\Longrightarrow$"). The converse implication is easier to show. Let $(U_j)_{j\in\mathbb{N}} \in \Sigma^{\mathbb{N}}$ be a generator of Σ. For given $i \in [n]$, we define

$$\mathscr{U}_i \coloneqq \{U_{j,i} \mid j \in \mathbb{N}\}$$

where $U_{j,i}$ is the i-th layer of U_j for all $j \in \mathbb{N}$. The family $\mathscr{U}_i$ is evidently countable. Our claim is that $\sigma(\mathscr{U}_i) = \Sigma_i$.

We first note that by definition of the layered σ-algebra, we have $U_{j,i} \in \Sigma_i$ for all $j \in \mathbb{N}$, which implies $\sigma(\mathscr{U}_i) \subseteq \Sigma_i$.

We prove the converse inclusion by contradiction. Let us assume that $\sigma(\mathscr{U}_i) \neq \Sigma_i$. We define

$$\widetilde{\Sigma}_i \coloneqq \sigma(\mathscr{U}_i).$$

Under this assumption, there would exist $V \in \Sigma_i \setminus \widetilde{\Sigma}_i$. We could then define a smaller layered σ-algebra

$$\widetilde{\Sigma} \coloneqq \Big\{ \bigcup_{k=1}^{n} \big(\{k\} \times U_k\big) \,\Big|\, U_i \in \widetilde{\Sigma}_i,\, U_k \in \Sigma_k \text{ for } k \neq i \Big\}.$$

The family $\widetilde{\Sigma}$ would also be a σ-algebra according to Theorem 2.1.16. However, it would not contain the set

$$\big(\{i\} \times V\big) \cup \bigcup_{\substack{k=1 \\ k\neq i}}^{n} \big(\{k\} \times X_k\big),$$

which is contained in Σ. Therefore, $\widetilde{\Sigma}$ would be a strictly smaller σ-algebra than Σ. Because $\widetilde{\Sigma}_i = \sigma(\mathscr{U}_i)$, we would have $U_{j,i} \in \widetilde{\Sigma}_i$ for all $j \in \mathbb{N}$. For the remaining

layers $k \neq i$, we have $U_{j,k} \in \Sigma_k$ for all $j \in \mathbb{N}$ because $U_j \in \Sigma$. By definition of $\widetilde{\Sigma}$, this would mean that

$$U_j \in \widetilde{\Sigma} \quad \forall j \in \mathbb{N}.$$

However, because $\widetilde{\Sigma}$ is a σ-algebra, if it contained all U_j for $j \in \mathbb{N}$, that would mean that the σ-algebra generated by $(U_j)_{j\in\mathbb{N}}$ would be a subset of $\widetilde{\Sigma}$. It would therefore be a strict subset of Σ, which would contradict our assumption that $(U_j)_{j\in\mathbb{N}}$ is a generator of Σ.

The contradiction shows that the assumption that $\widetilde{\Sigma}_i \neq \Sigma_i$ must be wrong. Therefore, we have $\widetilde{\Sigma}_i = \Sigma_i$. By definition of $\widetilde{\Sigma}_i$, this means that $\mathscr{U}_i$ is a generator of Σ_i and that Σ_i is therefore countably generated. □

This demonstrates that the application of layering does not interfere with the relevant properties of the σ-algebras under discussion in this thesis. Next, we turn our attention to measurable functions on layered spaces. As we would expect, these functions can also be considered as if they were tuples of measurable functions.

Definition 2.1.18 (Layers of Functions).
Let $n \in \mathbb{N}$, let X_i be a set for each $i \in [n]$, let Y be a set, let

$$X := \bigcup_{i=1}^{n} \big(\{i\} \times X_i\big)$$

and let $f : X \to Y$. For each $i \in [n]$, we refer to $f_i : X_i \to Y$ with

$$f_i(x) := f(i,x) \quad \forall x \in X_i$$

as the *i-th layer of f*. ◁

Theorem 2.1.19 (Measurability of Layered Functions).
Let $n \in \mathbb{N}$, let (X_i, Σ_i) be a measurable space for each $i \in [n]$, let X and Σ refer to the layerings of $(X_i)_{i\in[n]}$ and $(\Sigma_i)_{i\in[n]}$, respectively, let $(Y, \widetilde{\Sigma})$ be a measurable space, and let $f : X \to Y$. Then f is Σ-$\widetilde{\Sigma}$-measurable if and only if its i-th layer f_i is Σ_i-$\widetilde{\Sigma}$-measurable for every $i \in [n]$. ◁

Proof. Let first f be Σ-$\widetilde{\Sigma}$-measurable. By definition, that means that

$$f^{-1}(B) \in \Sigma \quad \forall B \in \widetilde{\Sigma}.$$

Let $i \in [n]$ be fixed. For every $B \in \widetilde{\Sigma}$, we have

$$f_i^{-1}(B) = \big\{x \in X_i \,\big|\, f(i,x) \in B\big\} = \big\{x \in X_i \,\big|\, (i,x) \in f^{-1}(B)\big\} = \big(f^{-1}(B)\big)_i \in \Sigma_i.$$

Now, we consider the case where f_i is Σ_i-$\widetilde{\Sigma}$-measurable for every $i \in [n]$. For every $B \in \widetilde{\Sigma}$, we have

$$\begin{aligned} f^{-1}(B) &= \big\{(i,\ x) \in X \,\big|\, f(i,x) \in B\big\} \\ &= \bigcup_{i=1}^{n} \Big(\{i\} \times \big\{x \in X_i \,\big|\, f(i,x) \in B\big\}\Big) \\ &= \bigcup_{i=1}^{n} \Big(\{i\} \times \big\{x \in X_i \,\big|\, f_i(x) \in B\big\}\Big) \\ &= \bigcup_{i=1}^{n} \big(\{i\} \times \underbrace{f_i^{-1}(B)}_{\in\Sigma_i}\big) \\ &\in \Sigma. \end{aligned}$$

□

Due to the additivity of the Lebesgue integral, integrals of layered functions are simply sums of component integrals.

Theorem 2.1.20 (Integrals of Layered Functions).
Let $n \in \mathbb{N}$, let $\big((X_i, \Sigma_i, \mu_i)\big)_{i\in[n]}$ be a tuple of measure spaces, let (X, Σ, μ) refer to the layering of $\big((X_i, \Sigma_i, \mu_i)\big)_{i\in[n]}$, and let $f : X \to \mathbb{R}$ be measurable. Then f is integrable if and only if all of its layers are integrable. In this case, the integral satisfies

$$\int_X f \,\mathrm{d}\mu = \sum_{i=1}^{n} \int_{X_i} f_i \,\mathrm{d}\mu_i. \qquad \triangleleft$$

PROOF. This mostly follows from the additivity of the Lebesgue integral. As we had previously shown, f is measurable if and only if all f_i are measurable. We have

$$\begin{aligned}\int_X |f| \,\mathrm{d}\mu &= \sum_{i=1}^{n} \int_{\{i\}\times X_i} |f| \,\mathrm{d}\mu \\ &= \sum_{i=1}^{n} \int_{X_i} |f_i| \,\mathrm{d}\mu_i\end{aligned}$$

which proves that f is integrable if and only if all f_i are integrable. We can do this because the Lebesgue integral is always well defined for measurable non-negative functions. Once integrability is established, we can make the same reformulation for the signed functions:

$$\begin{aligned}\int_X f \,\mathrm{d}\mu &= \sum_{i=1}^{n} \int_{\{i\}\times X_i} f \,\mathrm{d}\mu \\ &= \sum_{i=1}^{n} \int_{X_i} f_i \,\mathrm{d}\mu_i.\end{aligned} \qquad \square$$

Due to the existence of layering and the ease with which the relevant properties and integrals can be transferred from the layers to the layered space, we will ignore multivariate problems for most of the remainder of this thesis and consider them a special case of univariate problems unless explicitly otherwise noted.

2.2 METRIC SPACES OF MEASURABLE SETS

In this section, we explore the metric structure of measure spaces. Metric spaces are spaces in which there is a meaningful concept of "distance." Much of general nonlinear optimization relies on the use of model functions which locally approximate the behavior of nonlinear functions. The metric structure is necessary to quantify the "locality" of such models.

We allow for metric spaces in which points can be infinitely far apart. This is somewhat unusual. Some theoretical works about metric spaces only deal with real-valued metrics. There do, however, exist some works, such as [BBI01], that do allow for infinite distances.

We will see that in many cases, finiteness is not necessary for our theory. We therefore do not assume it unless doing so is theoretically necessary. This could aid in future applications of measure space geodesics outside of this thesis.

Definition 2.2.1 (Metric Space).
A *metric space* is a tuple $(X,\ d)$ of a set X and a map $d\colon X \times X \to \mathbb{R} \cup \{\infty\}$ such that

(1) $d(x,y) \geq 0\ \forall x,y \in X$ with $d(x,y) = 0$ if and only if $x = y$,

(2) $d(x,y) = d(y,x)\ \forall x,y \in X$,

(3) $d(x,z) \leq d(x,y) + d(y,z)\ \forall x,y,z \in X$.

If d satisfies Properties 2.2.1 (1) to 2.2.1 (3), then d is called a *metric* on X. A set X such that there exists a metric $d\colon X \times X \to \mathbb{R}$, then X is called *metrizable*. ◁

The concept of a local model may also be valid in a more general topological space. However, metrics quantify "distance" and allow model error to be bounded by a function of distance. Having metric structure is therefore likely close to being the minimal assumption necessary for the search space of an optimization method that uses local approximations.

It is sometimes valid to relax the conditions of Definition 2.2.1 by allowing $d(x,y) = 0$ for $x \neq y$ in Property 2.2.1 (1). In this case, $(X,\ d)$ is a *pseudometric space*. Without modification, measure spaces only have a pseudometric structure.

Definition 2.2.2 (Pseudometric Space).
A *pseudometric space* is a tuple $(X,\ d)$ of a set X and a map $d\colon X \times X \to \mathbb{R} \cup \{\infty\}$ such that

(1) $d(x,x) = 0\ \forall x \in X$,

(2) $d(x,y) = d(y,x)\ \forall x,y \in X$,

(3) $d(x,z) \leq d(x,y) + d(y,z)\ \forall x,y,z \in X$.

If d satisfies Properties 2.2.2 (1) to 2.2.2 (3), then d is called a *pseudometric*. ◁

The non-negativity of d is not directly stated in Definition 2.2.2. It is, however, implied by Properties 2.2.2 (1) to 2.2.2 (3), because we have

$$d(x,y) \overset{\text{2.2.2 (2)}}{=} \frac{1}{2}\big(d(x,y) + d(y,x)\big) \overset{\text{2.2.2 (3)}}{\geq} \frac{1}{2} d(x,x) \overset{\text{2.2.2 (1)}}{=} 0 \qquad \forall x,y \in X.$$

2.2.1 Symmetric Difference

Intuitively, a meaningful concept of the *distance* between two sets A,B should measure the "difference" between both sets. However, the regular set difference is not sufficient for this purpose. The set difference $A \setminus B$ only includes those points in A that are not contained within B. Therefore, $\mu(A \setminus B)$ measures the points that have to be added to B to obtain A. It completely ignores points that have to be subtracted from B.

The difference between two sets is better captured by the "symmetric difference" which is defined as the disjoint union between the points that have to be added and the points that have to be subtracted.

Definition 2.2.3 (Symmetric Difference).
Let A,B be sets. The *symmetric difference* of A and B is given by

$$A \triangle B \coloneqq (A \setminus B) \cup (B \setminus A).$$ ◁

The union in Definition 2.2.3 is disjoint, which means that the measure of $A \triangle B$ is the sum of the measures of $A \setminus B$ and $B \setminus A$. Working with the symmetric difference is relatively simple because measure spaces have the underlying algebraic structure of a commutative ring in which the symmetric difference acts as an addition and the intersection acts as a multiplication. We briefly recapitulate the most significant properties of the symmetric difference because they are significant for the remainder of our discussion.

Lemma 2.2.4 (Properties of the Symmetric Difference).
Let A, B, C be sets. We have

$$A \triangle B = (A \cup B) \setminus (A \cap B), \tag{2.1}$$
$$A \triangle \emptyset = A, \tag{2.2}$$
$$A \triangle A = \emptyset, \tag{2.3}$$
$$A \triangle B = B \triangle A, \tag{2.4}$$
$$A \cap (B \triangle C) = (A \cap B) \triangle (A \cap C), \tag{2.5}$$
$$(A \triangle B) \triangle C = A \triangle (B \triangle C). \tag{2.6}$$

◁

PROOF. Equation (2.1) follows from

$$\begin{aligned} A \triangle B &= (A \setminus B) \cup (B \setminus A) \\ &= \big(A \setminus (A \cap B)\big) \cup \big(B \setminus (A \cap B)\big) \\ &= (A \cup B) \setminus (A \cap B). \end{aligned}$$

This makes the remaining claims much easier to prove. We have

$$A \triangle \emptyset = (A \cup \emptyset) \setminus (A \cap \emptyset) = A \setminus \emptyset = A,$$

which proves Equation (2.2). Similarly, we have

$$A \triangle A = (A \cup A) \setminus (A \cap A) = A \setminus A = \emptyset,$$

which proves Equation (2.3). Equation (2.4) follows from the commutativity of set union and intersection. The distributivity stated in Equation (2.5) follows from the distributivity of intersections over unions, intersections, and differences because

$$\begin{aligned} A \cap (B \triangle C) &= A \cap \big((B \cup C) \setminus (B \cap C)\big) \\ &= \big(A \cap (B \cup C)\big) \setminus \big(A \cap (B \cap C)\big) \\ &= \big((A \cap B) \cup (A \cap C)\big) \setminus \big((A \cap B) \cap (A \cap C)\big) \\ &= (A \cap B) \triangle (A \cap C). \end{aligned}$$

In order to prove Equation (2.6), we transform both of its sides to a common expression. For the left hand side, we have

$$\begin{aligned} (A \triangle B) \triangle C &= \big((A \triangle B) \cup C\big) \setminus \big((A \triangle B) \cap C\big) \\ &= (A \cup B \cup C) \setminus \big((A \cap B) \setminus C\big) \setminus \big((A \triangle B) \cap C\big) \\ &= (A \cup B \cup C) \setminus \big((A \cap B) \setminus C\big) \setminus \big((A \triangle B) \cap C\big) \\ &= (A \cup B \cup C) \setminus (A \cap B \cap C^{\complement}) \setminus \big((A \cap B^{\complement} \cap C) \cup (A^{\complement} \cap B \cap C)\big) \\ &= \big((A \cup B \cup C) \cap (A^{\complement} \cup B^{\complement} \cup C)\big) \setminus (A \cap B^{\complement} \cap C) \setminus (A^{\complement} \cap B \cap C) \\ &= (A \cup B \cup C) \cap (A^{\complement} \cup B^{\complement} \cup C) \cap (A^{\complement} \cup B \cup C^{\complement}) \cap (A \cup B^{\complement} \cup C^{\complement}). \end{aligned}$$

For the right hand side, we find

$$\begin{aligned} A \triangle (B \triangle C) &= (B \triangle C) \triangle A \\ &= (B \cup C \cup A) \cap (B^\complement \cup C^\complement \cup A) \cap (B^\complement \cup C \cup A^\complement) \cap (B \cup C^\complement \cup A^\complement) \\ &= (A \cup B \cup C) \cap (A^\complement \cup B^\complement \cup C) \cap (A^\complement \cup B \cup C^\complement) \cap (A \cup B^\complement \cup C^\complement) \end{aligned}$$

which proves Equation (2.6). □

By combining these properties with commonly known properties of the intersection operator, the ring structure of measurable spaces becomes apparent.

Lemma 2.2.5 (Ring Structure of Measurable Spaces).
Let (X, Σ) be a measurable space. Then the tuple $(\Sigma, \triangle, \cap)$ is a commutative ring. Specifically, $(\Sigma, \triangle)$ is an Abelian group with the neutral element $\emptyset$ in which the inverse element for $A \in \Sigma$ is A itself, and $(\Sigma, \cap)$ is a commutative monoid with neutral element X. ◁

PROOF. Lemma 2.2.4 shows that $(\Sigma, \triangle)$ is an Abelian group with neutral element $\emptyset$. Since $A \subseteq X$ for all $A \in \Sigma$, $(\Sigma, \cap)$ is a commutative monoid with neutral element X. Distributivity is also shown in Lemma 2.2.4. □

Using the Abelian group structure of $(\Sigma, \triangle)$, we can define the distance between two sets and show that it is a pseudometric.

Lemma 2.2.6 (Pseudometric on Measure Spaces).
Let (X, Σ, μ) be a measure space. The map $d_\mu \colon \Sigma \times \Sigma \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ with

$$d_\mu(A,B) := \mu(A \triangle B) \qquad \forall A, B \in \Sigma \tag{2.7}$$

is a pseudometric. ◁

PROOF. For all $A, B, C \in \Sigma$, we have

$$\begin{aligned} d_\mu(A,A) &= \mu(A \triangle A) \\ &= \mu(\emptyset) \\ &= 0, \end{aligned}$$

$$\begin{aligned} d_\mu(A,B) &= \mu(A \triangle B) \\ &= \mu(B \triangle A) \\ &= d_\mu(B,A), \end{aligned}$$

$$\begin{aligned} d_\mu(A,C) &= \mu(A \triangle C) \\ &= \mu(A \triangle B \triangle B \triangle C) \\ &= \mu\big((A \triangle B) \triangle (B \triangle C)\big) \\ &\leq \mu\big((A \triangle B) \cup (B \triangle C)\big) \\ &\leq \mu(A \triangle B) + \mu(B \triangle C) \\ &= d_\mu(A,B) + d_\mu(B,C). \end{aligned}$$

□

2.2.2 Similarity Spaces

A pseudometric space can be turned into a full metric space by treating elements that have zero distance from one another as the same elements. On a technical level, this means that we consider the set of residual classes with respect to an equivalence relation that relates elements that have zero distance with each other. We refer to this as "being similar to one another." Due to the way in which we define distance, this means that we treat sets that differ by a nullset as if they were the same.

In the theory of Lebesgue integration and measure theory more generally, this kind of "equality up to a nullset" is used commonly. It makes a lot of sense in the context of derivative-based optimization because continuous mappings cannot assign different values to points that have zero distance. Tying distance to measure, however, does limits the scope of our theory. If we treat sets as equal if their difference does not have strictly positive measure, then we cannot model any problem in which nullsets have an impact on the objective function or constraints. For instance, this becomes an issue in the case of total variation bounds, because they limit the *surface area* or *circumference* of a set rather than its volume.

Definition 2.2.7 (Set Similarity).
Let $(X,\ \Sigma,\ \mu)$ be a measure space. We refer to the relation $\sim_\mu \subseteq \Sigma \times \Sigma$ with

$$A \sim_\mu B \iff d_\mu(A,B) = 0 \qquad \forall A,B \in \Sigma$$

as *μ-similarity*. If $A \sim_\mu B$, then A and B are called *μ-similar*. ◁

Next, we show that μ-similarity is an equivalence relation. Following, e.g., [Gor16, Sec. 5.2], this implies that the equivalence class $[\emptyset]_{\sim_\mu}$ is an ideal in Σ. This then implies that the operators $\triangle$ and $\cap$ are well-defined operators on the quotient space $\Sigma/_{\sim_\mu}$ and give it an analogous ring structure.

Lemma 2.2.8 (Set Similarity as Equivalence Relation).
Let $(X,\ \Sigma,\ \mu)$ be a measure space. Then the μ-similarity relation $\sim_\mu \subseteq \Sigma \times \Sigma$ is an equivalence relation. ◁

PROOF. This follows directly from the fact that d_μ is a pseudometric, which we had proven in Lemma 2.2.6. Reflexivity follows from Property 2.2.2 (1), symmetry follows from Property 2.2.2 (2), and transitivity follows from Property 2.2.2 (3) and the fact that d_μ is non-negative. □

Theorem 2.2.9 (Properties of the Quotient Space).
Let $(X,\ \Sigma,\ \mu)$ be a measure space, and let $A,B \in \Sigma$. We have

$$\mu(A') = \mu(A) \qquad \forall A' \in [A]_{\sim_\mu}, \tag{2.8}$$

$$A' \triangle B' \sim_\mu A \triangle B \qquad \forall A' \in [A]_{\sim_\mu}, B' \in [B]_{\sim_\mu}, \tag{2.9}$$

$$A' \cap B' \sim_\mu A \cap B \qquad \forall A' \in [A]_{\sim_\mu}, B' \in [B]_{\sim_\mu}, \tag{2.10}$$

$$A' \cup B' \sim_\mu A \cup B \qquad \forall A' \in [A]_{\sim_\mu}, B' \in [B]_{\sim_\mu}, \tag{2.11}$$

$$A' \setminus B' \sim_\mu A \setminus B \qquad \forall A' \in [A]_{\sim_\mu}, B' \in [B]_{\sim_\mu}, \tag{2.12}$$

$$(A')^\complement \sim_\mu A^\complement \qquad \forall A' \in [A]_{\sim_\mu}. \tag{2.13}$$

Therefore, $\triangle$, $\cap$, $\cup$, $\setminus$, $\cdot^{\complement}$, and μ are well-defined as maps on $\Sigma/_{\sim_\mu}$ and $(\Sigma/_{\sim_\mu}, \triangle, \cap)$ is a commutative ring with additive neutral $[\emptyset]_{\sim_\mu}$ and multiplicative neutral $[X]_{\sim_\mu}$, wherein every $[A]_{\sim_\mu} \in \Sigma/_{\sim_\mu}$ is its own additive inverse. $\lhd$

Proof. Part 1 (Equations (2.8) to (2.10)). Let $A, A' \in \Sigma$ and $A \sim_\mu A'$. We have

$$\mu(A') = \mu\big(\underbrace{A \triangle (A \triangle A')}_{\subseteq A \cup (A \triangle A')}\big) \leq \mu(A) + \underbrace{\mu(A \triangle A')}_{=0} = \mu(A).$$

By swapping the roles of A and A', we have $\mu(A) \leq \mu(A')$ and thus $\mu(A') = \mu(A)$, which proves Equation (2.8). Let further $B, B' \in \Sigma$ with $B \sim_\mu B'$. We have

$$\mu\big((A' \triangle B') \triangle (A \triangle B)\big) = \mu\big(\underbrace{(A \triangle A') \triangle (B \triangle B')}_{\subseteq (A \triangle A') \cup (B \triangle B')}\big) \leq \mu(A \triangle A') + \mu(B \triangle B') = 0.$$

Because the measure of $(A' \triangle B') \triangle (A \triangle B)$ is always non-negative, we have $A' \triangle B' \sim_\mu A \triangle B$, which proves Equation (2.9). Similarly, we have

$$\begin{aligned}
\mu\big((A' \cap B') \triangle (A \cap B)\big) &= \mu\big(\big((A' \cap B') \setminus (A \cap B)\big) \cup \big((A \cap B) \setminus (A' \cap B')\big)\big) \\
&= \mu\big(\big((A' \cap B') \cap (A \cap B)^{\complement}\big) \cup \big((A \cap B) \cap (A' \cap B')^{\complement}\big)\big) \\
&= \mu\big(\big((A' \cap B') \cap (A^{\complement} \cup B^{\complement})\big) \cup \big((A \cap B) \cap ((A')^{\complement} \cup (B')^{\complement})\big)\big) \\
&= \mu\big(\underbrace{(A' \cap B' \setminus A)}_{\subseteq A' \setminus A} \cup \underbrace{(A' \cap B' \setminus B)}_{\subseteq B' \setminus B} \cup \underbrace{(A \cap B \setminus A')}_{\subseteq A \setminus A'} \cup \underbrace{(A \cap B \setminus B')}_{\subseteq B \setminus B'}\big) \\
&\leq \mu\big((A' \setminus A) \cup (A \setminus A')\big) + \mu\big((B' \setminus B) \cup (B \setminus B')\big) \\
&= \mu(A' \triangle A) + \mu(B' \triangle B) \\
&= 0 + 0,
\end{aligned}$$

which proves Equation (2.10).

Part 2 (Equations (2.11) to (2.13)). We first note that for all $A \in \Sigma$, we have $A^{\complement} = X \setminus A$. Since $A \subseteq X$ for all $A \in \Sigma$, this implies $A^{\complement} = X \triangle A$. Therefore, if $A, A' \in \Sigma$ with $A \sim_\mu A'$, we have

$$(A')^{\complement} = X \triangle A' \overset{(2.9)}{\sim_\mu} X \triangle A = A^{\complement},$$

which proves Equation (2.13). Set subtraction can be expressed as an intersection with a complement. Let $A, A', B, B' \in \Sigma$ be such that $A \sim_\mu A'$ and $B \sim_\mu B'$. Then we have

$$A' \setminus B' = A' \cap \underbrace{(B')^{\complement}}_{\sim_\mu B^{\complement}} \sim_\mu A \cap B^{\complement} = A \setminus B,$$

which proves Equation (2.12). Finally, let $A, A', B, B' \in \Sigma$ with $A \sim_\mu A'$ and $B \sim_\mu B'$. We have

$$A' \cup B' = \big((A')^{\complement} \cap (B')^{\complement}\big)^{\complement} \overset{(2.13)}{\sim_\mu} (A^{\complement} \cap B^{\complement})^{\complement} = A \cup B,$$

which proves Equation (2.11).

Part 3 (Ring Structure). Because the operations $\triangle$ and $\cap$ function in exactly the same manner in the space of residual classes, we can prove the ring structure by the same arguments as we used in the proof of Lemma 2.2.5. $\square$

Equation (2.8) shows that μ is well-defined when applied to residual classes $[A]_{\sim_\mu}$. Similarly, the remaining equations in Theorem 2.2.9 allow us to use the common set operations on such residual classes as if they were sets. By combining Equations (2.8) and (2.9), we see that $d_\mu \colon \Sigma/_{\sim_\mu} \times \Sigma/_{\sim_\mu} \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ with

$$d_\mu\big([A]_{\sim_\mu}, [B]_{\sim_\mu}\big) \coloneqq d_\mu(A,B) = \mu(A \mathbin{\triangle} B) \quad \forall A, B \in \Sigma$$

is well-defined. By definition of $\sim_\mu$, we have $d_\mu([A]_{\sim_\mu}, [B]_{\sim_\mu}) = 0$ if and only if $[A]_{\sim_\mu} = [B]_{\sim_\mu}$ for all $A, B \in \Sigma$. This ensures that $(\Sigma/_{\sim_\mu},\ d_\mu)$ is a metric space.

Theorem 2.2.10 (Metric on Measure Spaces).
Let $(X,\ \Sigma,\ \mu)$ be a measure space. The map $d_\mu \colon \Sigma/_{\sim_\mu} \times \Sigma/_{\sim_\mu} \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ with

$$d_\mu\big([A]_{\sim_\mu}, [B]_{\sim_\mu}\big) \coloneqq \mu(A \mathbin{\triangle} B) \qquad \forall A, B \in \Sigma$$

is a metric. ◁

PROOF. According to Equations (2.8) and (2.9), d_μ is well-defined. For the remaining properties, we refer to the same theorem. Let subsequently $A, B, C \in \Sigma/_{\sim_\mu}$ and let $\widetilde{A} \in A$, $\widetilde{B} \in B$, and $\widetilde{C} \in C$ be representatives of A, B, and C, respectively. The estimate $d_\mu(A,B) \geq 0$ follows because μ is a measure. If $A = B$, then we have $\widetilde{A} \in B$ and therefore

$$d_\mu(A,B) = \mu(\widetilde{A} \mathbin{\triangle} \widetilde{B}) \overset{(2.9)}{=} \mu(\widetilde{A} \mathbin{\triangle} \widetilde{A}) = 0.$$

Conversely, if $d_\mu(A,B) = 0$, then we have

$$\mu(\widetilde{A} \mathbin{\triangle} \widetilde{B}) = d_\mu(A,B) = 0$$

and therefore $\widetilde{A} \sim_\mu \widetilde{B}$. This implies

$$A = [\widetilde{A}]_{\sim_\mu} = [\widetilde{B}]_{\sim_\mu} = B.$$

Symmetry and triangle inequality follow from Lemma 2.2.6. We have

$$d_\mu(A,B) = \mu(\widetilde{A} \mathbin{\triangle} \widetilde{B}) = \mu(\widetilde{B} \mathbin{\triangle} \widetilde{A}) = d_\mu(B,A)$$

and

$$d_\mu(A,C) = \mu(\widetilde{A} \mathbin{\triangle} \widetilde{C}) \leq \mu(\widetilde{A} \mathbin{\triangle} \widetilde{B}) + \mu(\widetilde{B} \mathbin{\triangle} \widetilde{C}) = d_\mu(A,B) + d_\mu(B,C). \qquad \square$$

The space $(\Sigma/_{\sim_\mu},\ d_\mu)$ is therefore a metric space. We will subsequently refer to this space as the *similarity space* of $(X,\ \Sigma,\ \mu)$.

Definition 2.2.11 (Similarity Spaces).
Let $(X,\ \Sigma,\ \mu)$ be a measure space. We refer to the metric space $(\Sigma/_{\sim_\mu},\ d_\mu)$ as the *similarity space of* $(X,\ \Sigma,\ \mu)$ or the *μ-similarity space of* Σ. For each $A \in \Sigma$, we refer to the residual class $[A]_{\sim_\mu}$ as the *μ-similarity class of A*. We omit the measure or measure space from these designations when it is clear from the context which measure space is used. ◁

Because the difference between representatives of a similarity class is always a nullset, the integral of a function over different representatives of the same residual class is always the same.

Proposition 2.2.12 (Integrals over Similarity Classes).
Let (X, Σ, μ) be a measure space, let $A, B \in \Sigma$ with $A \sim_\mu B$, and let the function $f : X \to \mathbb{R} \cup \{\pm\infty\}$ be measurable. Then we have

$$\int_A |f| \, \mathrm{d}\mu = \int_B |f| \, \mathrm{d}\mu.$$

If $\int_A |f| \, \mathrm{d}\mu < \infty$, then we also have

$$\int_A f \, \mathrm{d}\mu = \int_B f \, \mathrm{d}\mu. \qquad \triangleleft$$

PROOF. If $A \sim_\mu B$, then we have $\mu(A \setminus B) = \mu(B \setminus A) = 0$ because $A \setminus B \subseteq A \triangle B$ and $B \setminus A \subseteq A \triangle B$. It therefore follows that

$$\int_A |f| \, \mathrm{d}\mu - \int_B |f| \, \mathrm{d}\mu = \int_{A \setminus B} |f| \, \mathrm{d}\mu - \int_{B \setminus A} |f| \, \mathrm{d}\mu = 0 - 0 = 0.$$

Similarly, we have

$$\int_A f \, \mathrm{d}\mu - \int_B f \, \mathrm{d}\mu = \int_{A \setminus B} f \, \mathrm{d}\mu - \int_{B \setminus A} f \, \mathrm{d}\mu = 0 - 0 = 0$$

if the integrals exist. □

Proposition 2.2.12 shows that we can use similarity classes as domains of integration without fear of ambiguity. To evaluate an integral over a similarity class, we can simply calculate the integral over any representative of that class. The last well-definedness result that we will discuss in this section concerns the well-definedness of measures other than μ, signed measures, and vector measures on similarity spaces.

Proposition 2.2.13 (Measures on Similarity Spaces).
Let (X, Σ, μ) be a measure space; let $\nu : \Sigma \to X$ be either a measure ($X = \mathbb{R}_{\geq 0} \cup \{\infty\}$), a signed measure ($X = \mathbb{R} \cup \{\pm\infty\}$), or a vector measure ($X$ is a Banach space); and let $\nu \ll \mu$. Then we have

$$\nu(A) = \nu(B) \quad \forall A, B \in \Sigma : A \sim_\mu B. \qquad \triangleleft$$

PROOF. Let $A, B \in \Sigma$ be such that $A \sim_\mu B$. By definition, we have $\mu(A \triangle B) = 0$. If ν is a positive measure, then $\nu \ll \mu$ implies $\nu(A \triangle B) = 0$, i.e., $A \sim_\nu B$. According to Theorem 2.2.9, this implies $\nu(A) = \nu(B)$.

If ν is a signed or vector measure, then $\nu \ll \mu$ means that $|\nu| \ll \mu$ where $|\nu|$ is the variation of ν. Thus, we have $|\nu|(A \triangle B) = 0$. We can then partition $A \triangle B$ into $A \setminus B$ and $B \setminus A$ and obtain

$$\big\|\nu(A \setminus B)\big\| + \big\|\nu(B \setminus A)\big\| \leq |\nu|(A \triangle B) = 0$$

which implies $\nu(A \setminus B) = \nu(B \setminus A) = 0$. By using the additivity of ν, we obtain

$$\nu(A) = \nu(A \cap B) + \nu(A \setminus B) = \nu(A \cap B) = \nu(A \cap B) + \nu(B \setminus A) = \nu(B). \qquad \square$$

This means that we can apply positive, signed, and vector measures to μ-equivalence classes as if they were sets as long as those measures are absolutely continuous with respect to μ. For finite measures, this is already implied by the Radon-Nikodym theorem and Proposition 2.2.12. However, it is important to note that these maps are well-defined as maps on $\Sigma/_{\sim_\mu}$ even if the density function does not exist because they are not finite.

Up to this point, there is still one gap in our effort to show that set operations are well-defined on similarity classes. Theorem 2.2.9 only proves the well-definedness of finite unions and intersections. However, we are interested in countable unions and intersections as well because they are useful to find limits in measure spaces.

Lemma 2.2.14 (Countable Set Operations in Similarity Spaces).
Let (X, Σ, μ) be a measure space, and let $A_i, B_i \in \Sigma$ with $A_i \sim_\mu B_i$ for all $i \in \mathbb{N}$. Then we have

$$\bigcap_{i=1}^{\infty} A_i \sim_\mu \bigcap_{i=1}^{\infty} B_i, \tag{2.14}$$

$$\bigcup_{i=1}^{\infty} A_i \sim_\mu \bigcup_{i=1}^{\infty} B_i. \tag{2.15}$$

◁

PROOF. We have

$$\begin{aligned}\Big(\bigcap_{i=1}^{\infty} A_i\Big) \setminus \Big(\bigcap_{i=1}^{\infty} B_i\Big) &= \bigcap_{i=1}^{\infty} \Big(A_i \setminus \bigcap_{j=1}^{\infty} B_i\Big) \\ &\subseteq \bigcap_{i=1}^{\infty} (A_i \setminus B_i) \\ &\subseteq \bigcap_{i=1}^{\infty} (A_i \triangle B_i),\end{aligned}$$

which implies that

$$\mu\Big(\Big(\bigcap_{i=1}^{\infty} A_i\Big) \setminus \Big(\bigcap_{i=1}^{\infty} B_i\Big)\Big) = 0.$$

By exchanging the roles of $(A_i)_i$ and $(B_i)_i$, we obtain

$$\mu\Big(\Big(\bigcap_{i=1}^{\infty} B_i\Big) \setminus \Big(\bigcap_{i=1}^{\infty} A_i\Big)\Big) = 0.$$

By combining these results, we find that

$$\mu\Big(\Big(\bigcap_{i=1}^{\infty} A_i\Big) \triangle \Big(\bigcap_{i=1}^{\infty} B_i\Big)\Big) \leq \mu\Big(\Big(\bigcap_{i=1}^{\infty} A_i\Big) \setminus \Big(\bigcap_{i=1}^{\infty} B_i\Big)\Big) + \mu\Big(\Big(\bigcap_{i=1}^{\infty} B_i\Big) \setminus \Big(\bigcap_{i=1}^{\infty} A_i\Big)\Big) = 0,$$

which proves Equation (2.14). Equation (2.15) follows from Equations (2.13) and (2.14) because

$$\bigcup_{i=1}^{\infty} A_i = \Big(\bigcap_{i=1}^{\infty} A_i^{\complement}\Big)^{\complement} \sim_\mu \Big(\bigcap_{i=1}^{\infty} B_i^{\complement}\Big)^{\complement} = \bigcup_{i=1}^{\infty} B_i.$$

□

2.2.3 Essential Subsets and Disjointness

Next, we broaden the definition of "subset" to include *essential subsets* or sets that are subsets except for a nullset.

Definition 2.2.15 (Essential Subset).
Let $(X,\ \Sigma,\ \mu)$ be a measure space. The *μ-essential subset relation* $\subseteq_\mu \subseteq \Sigma/_{\sim_\mu} \times \Sigma/_{\sim_\mu}$ is given by

$$A \subseteq_\mu B :\Longleftrightarrow \mu(A \setminus B) = 0 \quad \forall A, B \in \Sigma/_{\sim_\mu}.$$

If $A \subseteq_\mu B$, then we refer to A as a *μ-essential subset* or simply an *essential subset* of B. ◁

We could also define the μ-essential subset relation on sets rather than on similarity classes. However, we will only apply it in the context of similarity classes. Regardless, the μ-essential subset relation is very strongly linked to the conventional subset relation in the sense that an essential subset relationship can also be thought of as a conventional subset relationship between suitably chosen representatives of the similarity classes.

Lemma 2.2.16 (Existence Of Proper Subset Representatives).
Let $(X,\ \Sigma,\ \mu)$ be a measure space, and let $A, B \in \Sigma/_{\sim_\mu}$. Then we have

$$A \subseteq_\mu B \iff \left(\exists \widetilde{A} \in A,\ \widetilde{B} \in B \colon \widetilde{A} \subseteq \widetilde{B}\right).$$

More specifically, if $A \subseteq_\mu B$, then we have

$$\widetilde{A} \cap \widetilde{B} \in A \quad \forall \widetilde{A} \in A,\ \widetilde{B} \in B, \tag{2.16}$$

$$\widetilde{A} \cup \widetilde{B} \in B \quad \forall \widetilde{A} \in A,\ \widetilde{B} \in B. \tag{2.17}$$

◁

Proof. First, we consider the case in which there exist $\widetilde{A} \in A$ and $\widetilde{B} \in B$ such that $\widetilde{A} \subseteq \widetilde{B}$. In this case, we have

$$\mu(A \setminus B) = \mu(\widetilde{A} \setminus \widetilde{B}) = \mu(\emptyset) = 0$$

and therefore $A \subseteq_\mu B$. Next, we address the converse case. Let $A, B \in \Sigma/_{\sim_\mu}$ be such that $A \subseteq_\mu \Sigma$. Due to the reflexivity of equivalence relations, all elements of a quotient space are non-empty. We therefore have $A \neq \emptyset$ and $B \neq \emptyset$. If we take any two representatives $\overline{A} \in A$ and $\overline{B} \in B$, then we could choose suitable representatives by setting either

$$\widetilde{A} := \overline{A} \in A, \qquad \widetilde{B} := \overline{A} \cup \overline{B} \overset{(2.17)}{\in} B,$$

or

$$\widetilde{A} := \overline{A} \cap \overline{B} \overset{(2.16)}{\in} A, \qquad \widetilde{B} := \overline{B} \in B.$$

Either choice would guarantee that $\widetilde{A} \subseteq \widetilde{B}$. It is therefore sufficient to prove Equations (2.16) and (2.17). To this end, let $\widetilde{A} \in A$ and $\widetilde{B} \in B$ be representives of

their respective similarity classes. We have

$$\begin{aligned}\mu\big((\widetilde{A}\cap\widetilde{B})\triangle\widetilde{A}\big) &= \overbrace{\mu\big((\widetilde{A}\cap\widetilde{B})\setminus\widetilde{A}\big)}^{=0}+\mu\big(\widetilde{A}\setminus(\widetilde{A}\cap\widetilde{B})\big)\\ &= \mu\big(\widetilde{A}\setminus(\widetilde{A}\cap\widetilde{B})\big)\\ &= \mu(\widetilde{A}\setminus\widetilde{B})\\ &= \mu(A\setminus B)\\ &= 0,\end{aligned}$$

which proves Equation (2.16). We also have

$$\begin{aligned}\mu\big((\widetilde{A}\cup\widetilde{B})\triangle\widetilde{B}\big) &= \mu\big((\widetilde{A}\cup\widetilde{B})\setminus\widetilde{B}\big)+\overbrace{\mu\big(\widetilde{B}\setminus(\widetilde{A}\cup\widetilde{B})\big)}^{=0}\\ &= \mu(\widetilde{A}\setminus\widetilde{B})\\ &= 0,\end{aligned}$$

which proves Equation (2.17). This proves both Equations (2.16) and (2.17). This then establishes the converse implication as we had discussed before. □

The representatives satisfying the true subset relation are not unique. Because they are constructed from arbitrary representatives of the similarity classes, we are free to add or subtract nullsets that do not interfere with the subset relation. The important aspect of Lemma 2.2.16 is that whenever we encounter μ-essential subsets, we can choose representatives that are true subsets or supersets of one another. This complements Theorem 2.2.9 and Lemma 2.2.14. Taken together, these results allow us to almost freely switch between similarity classes and specifically chosen representatives in our proofs, depending on whichever is more suitable for our purposes at the time.

We primarily want to use Equations (2.16) and (2.17) inductively on sequences of *essentially increasing* or *essentially decreasing* similarity classes. Because σ-algebras are closed with respect to countable union and intersection, we can find *limit sets* of increasing or decreasing set sequences. The two equations allow us to transfer this type of limit process to similarity spaces.

Normally, even if a given sequence of similarity classes is, for instance, essentially increasing, not every sequence of representatives of those classes is truly increasing. Therefore, not every sequence of representatives has a limit set. By using Lemma 2.2.16, we can turn an arbitrary sequence of representatives into an increasing one and therefore find a limit set whose similarity class then becomes the limit of the original sequence of similarity classes.

Lemma 2.2.17 (Existence of Monotonic Representative Sequences).
Let $(X,\ \Sigma,\ \mu)$ be a measure space. Let $(A_i)_{i\in\mathbb{N}}\subseteq \Sigma/_{\sim_\mu}$ be a sequence of similarity classes, and let $\overline{A}_i\in A_i$ for all $i\in\mathbb{N}$.

If $A_i\subseteq_\mu A_{i+1}$ for all $i\in\mathbb{N}$, then there exists a sequence $(\widetilde{A}_i)_{i\in\mathbb{N}}$ such that $\widetilde{A}_i\in A_i$, $\widetilde{A}_i\supseteq\overline{A}_i$, and $\widetilde{A}_i\subseteq\widetilde{A}_{i+1}$ for all $i\in\mathbb{N}$.

If $A_i\supseteq_\mu A_{i+1}$ for all $i\in\mathbb{N}$, then there exists a sequence $(\widetilde{A}_i)_{i\in\mathbb{N}}$ such that $\widetilde{A}_i\in A_i$, $\widetilde{A}_i\subseteq\overline{A}_i$, and $\widetilde{A}_i\supseteq\widetilde{A}_{i+1}$ for all $i\in\mathbb{N}$. ◁

PROOF. This follows by complete induction from Equations (2.16) and (2.17). The proofs of both statements are very similar. They only differ in the equation used. We therefore limit ourselves to proving the case where $A_i \subseteq_\mu A_{i+1}$ for all $i \in \mathbb{N}$. In this case, we use Equation (2.17).

We define $\widetilde{A}_1 \coloneqq \overline{A}_1$. This satisfies $\overline{A}_1 \subseteq \widetilde{A}_1$ and $\widetilde{A}_1 \in [\overline{A}_i]_{\sim_\mu} = A_i$. Let $i \in \mathbb{N}$ such that $\widetilde{A}_i \in A_i$ is defined. According to Equation (2.17), Because $A_{i+1} \subseteq_\mu A_i$ and $\overline{A}_{i+1} \in A_{i+1}$, we can define

$$\widetilde{A}_{i+1} \coloneqq \overline{A}_{i+1} \cup \widetilde{A}_i \overset{(2.17)}{\in} [\overline{A}_{i+1}]_{\sim_\mu} = A_{i+1}.$$

This choice guarantees that $\widetilde{A}_{i+1} \in A_{i+1}$, $\widetilde{A}_i \subseteq \widetilde{A}_{i+1}$, and $\overline{A}_{i+1} \subseteq \widetilde{A}_{i+1}$. By complete induction, these conditions hold for all $i \in \mathbb{N}$.

The case where $A_i \supseteq_\mu A_{i+1}$ for all $i \in \mathbb{N}$ follows by an analogous argument using Equation (2.16). □

By combining Lemma 2.2.17 with the well-definedness of countable union and intersection in similarity spaces we find that every sequence of *essentially increasing* or *essentially decreasing* similarity classes approaches a unique *limit class* which contains the limit sets of all sequences of representatives. This result is of major significance for the coming steps. In Section 2.3, we use the convergence of monotonic sequences in conjunction with continuity for a variety of well-definedness arguments.

Finally, we introduce the concept of *essential disjointness*, which generalizes the concept of disjointness. This is fairly straightforward.

Definition 2.2.18 (Essential Disjointness).
Let (X, Σ, μ) be a measure space. Two similarity classes $A, B \in \Sigma/_{\sim_\mu}$ are called *μ-essentially disjoint* if and only if

$$\mu(A \cap B) = 0. \qquad \triangleleft$$

Any sequence of pairwise essentially disjoint similarity classes has a pairwise disjoint sequence of representatives. This allows us to transfer the σ-additivity of measures from sets to similarity classes.

Lemma 2.2.19 (Existence of Disjoint Representative Sequences).
Let (X, Σ, μ) be a measure space. A sequence $(A_i)_{i\in\mathbb{N}} \subseteq \Sigma/_{\sim_\mu}$ is pairwise essentially disjoint if and only if there exists a sequence $(\widetilde{A}_i)_{i\in\mathbb{N}} \subseteq \Sigma$ of pairwise disjoint measurable sets such that $\widetilde{A}_i \in A_i$ for all $i \in \mathbb{N}$.

Specifically, let $(A_i)_{i\in\mathbb{N}} \subseteq \Sigma/_{\sim_\mu}$ be a sequence of pairwise μ-essentially disjoint similarity classes, and let $(\overline{A}_i)_{i\in\mathbb{N}} \subseteq \Sigma$ with $\overline{A}_i \in A_i$ for all $i \in \mathbb{N}$. Then the sequence $(\widetilde{A}_i)_{i\in\mathbb{N}}$ with

$$\widetilde{A}_i \coloneqq \overline{A}_i \setminus \bigcup_{j=1}^{i-1} \overline{A}_j \quad \forall i \in \mathbb{N}$$

satisfies

$$\widetilde{A}_i \subseteq \overline{A}_i \quad \forall i \in \mathbb{N}, \tag{2.18}$$

$$\widetilde{A}_i \in A_i \quad \forall i \in \mathbb{N}, \tag{2.19}$$

$$\widetilde{A}_i \cap \widetilde{A}_j = \emptyset \quad \forall i, j \in \mathbb{N} \colon i \neq j. \tag{2.20}$$

$\triangleleft$

PROOF. First, we address the case in which there exists a pairwise disjoint sequence $(\widetilde{A}_i)_{i\in\mathbb{N}} \subseteq \Sigma$ with $\widetilde{A}_i \in A_i$ for all $i \in \mathbb{N}$. For all $i, j \in \mathbb{N}$ with $i \neq j$, we have

$$\mu(A_i \cap A_j) = \mu(\widetilde{A}_i \cap \widetilde{A}_j) = \mu(\emptyset) = 0$$

which means that A_i and A_j are essentially disjoint.

To prove the converse implication, we note again that every similarity class contains at least one set. Therefore, every sequence $(A_i)_{i\in\mathbb{N}}$ of similarity classes has at least one sequence $(\overline{A}_i)_{i\in\mathbb{N}}$ of representatives. It is therefore sufficient to prove the more specific statement about the construction of $(\widetilde{A}_i)_{i\in\mathbb{N}}$ from $(\overline{A}_i)_{i\in\mathbb{N}}$. Equation (2.18) follows directly from the definition of $\widetilde{A}_i$. Let $i \in \mathbb{N}$. Because $\widetilde{A}_i \subseteq \overline{A}_i$, we have

$$\begin{aligned}
\mu(\widetilde{A}_i \triangle \overline{A}_i) &= \mu(\overline{A}_i \setminus \widetilde{A}_i) \\
&\leq \mu\Big(\overline{A}_i \cap \bigcup_{j=1}^{i-1} \overline{A}_j\Big) \\
&= \mu\Big(\bigcup_{j=1}^{i-1} (\overline{A}_i \cap \overline{A}_j)\Big) \\
&\leq \sum_{j=1}^{i-1} \mu(\overline{A}_i \cap \overline{A}_j) \\
&= \sum_{j=1}^{i-1} \underbrace{\mu(A_i \cap A_j)}_{=0} \\
&= 0
\end{aligned}$$

and therefore $[\widetilde{A}_i]_{\sim_\mu} = [\overline{A}_i]_{\sim_\mu} = A_i$. By applying this result to all $i \in \mathbb{N}$, we obtain Equation (2.19). Let $i, j \in \mathbb{N}$ with $i \neq j$. Without loss of generality, let $i < j$. We have

$$\widetilde{A}_i \cap \widetilde{A}_j = \underbrace{\Big(\overline{A}_i \setminus \bigcup_{k=1}^{i-1} \overline{A}_k\Big)}_{\subseteq \overline{A}_i} \cap \underbrace{\Big(\overline{A}_j \setminus \bigcup_{k=1}^{j-1} \overline{A}_k\Big)}_{\subseteq \overline{A}_j \setminus \overline{A}_i} = \emptyset.$$

This proves Equation (2.20). □

2.2.4 Measurable Functions

Most of the measurable functions that we work with are only well-defined up to differences on nullsets. This is usually acceptable because it does not affect integrals and ensures similarity between level sets. We state this explicitly here so that we can reference this elementary result later.

Theorem 2.2.20 (Essential Equality Of Measurable Functions).
Let (X, Σ, μ) be a measure space, and let $f, g\colon X \to \mathbb{R} \cup \{\pm\infty\}$ be measurable functions. The following statements are equivalent:

(1) $\{f \leq t\} \sim_\mu \{g \leq t\}$ for all $t \in \mathbb{R} \cup \{\pm\infty\}$;

(2) $f^{-1}(B) \sim_\mu g^{-1}(B)$ for all $B \in \overline{\mathscr{B}}(\mathbb{R})$;

(3) there exists a μ-nullset $N \in \Sigma$ such that $f|_{X\setminus N} = g|_{X\setminus N}$. ◁

Proof. Part 1 ((1) $\implies$ (2)). This part of the proof is based on the good-set principle. Let

$$\mathscr{F} \coloneqq \left\{B \in \overline{\mathscr{B}}(\mathbb{R}) \,\middle|\, f^{-1}(B) \sim_\mu g^{-1}(B)\right\}.$$

The premise for this part of the proof is that $\mathscr{F}$ contains all extended intervals $[-\infty,\ t]$ for $t \in \mathbb{R} \cup \{\pm\infty\}$. By showing that $\mathscr{F}$ is a σ-algebra we can show that $\mathscr{F}$ contains the σ-algebra generated by those extended intervals, which is $\overline{\mathscr{B}}(\mathbb{R})$.

In particular, we know that $\mathscr{F}$ includes $[-\infty,\ \infty] = \mathbb{R} \cup \{\pm\infty\}$, which is the complement of $\emptyset$. We therefore do not need to prove that $\mathscr{F}$ contains $\emptyset$. It is sufficient to show that $\mathscr{F}$ is closed under complementation and countable union, which is straightforward because both operations commute with the preimage.

For all $B \in \mathscr{F}$, we have

$$\begin{aligned} f^{-1}(B^\complement) &= f^{-1}\Big(\big(\mathbb{R} \cup \{\pm\infty\}\big) \setminus B\Big) \\ &= f^{-1}\big(\mathbb{R} \cup \{\pm\infty\}\big) \setminus f^{-1}(B) \\ &\sim_\mu g^{-1}(\mathbb{R} \cup \{\pm\infty\}) \setminus g^{-1}(B) \\ &= g^{-1}(B^\complement). \end{aligned}$$

For $(B_i)_{i\in\mathbb{N}} \in \mathscr{F}^{\mathbb{N}}$, we have

$$f^{-1}\Big(\bigcup_{i=1}^{\infty} B_i\Big) = \bigcup_{i=1}^{\infty} f^{-1}(B_i) \sim_\mu \bigcup_{i=1}^{\infty} g^{-1}(B_i) = g^{-1}\Big(\bigcup_{i=1}^{\infty} B_i\Big).$$

Therefore, we have both $B^\complement \in \mathscr{F}$ and $\bigcup_{i=1}^{\infty} B_i \in \mathscr{F}$. Thus, $\mathscr{F}$ is a σ-algebra.

Because $\{-\infty\} = [-\infty,\ -\infty] \in \mathscr{F}$, $(-\infty,\ t] = [-\infty,\ t] \setminus [-\infty,\ -\infty] \in \mathscr{F}$ for all $t \in \mathbb{R}$, and

$$\{\infty\} = [-\infty,\ \infty] \setminus \bigcup_{i=1}^{\infty} [-\infty,\ i] \in \mathscr{F},$$

the σ-algebra generated by the intervals $(-\infty,\ t]$ for $t \in \mathbb{R}$, $\{-\infty\}$, and $\{\infty\}$ is a subset of $\mathscr{F}$. That σ-algebra is $\overline{\mathscr{B}}(\mathbb{R})$.

Part 2 ((2) $\implies$ (3)). Let $f^{-1}(B) \sim_\mu g^{-1}(B)$ for all $B \in \overline{\mathscr{B}}(\mathbb{R})$. Let

$$\begin{aligned} N_1 &\coloneqq \big(f^{-1}(\{\infty\}) \mathbin{\triangle} g^{-1}(\{\infty\})\big) \cup \big(f^{-1}(\{-\infty\}) \mathbin{\triangle} g^{-1}(\{-\infty\})\big), \\ N_2 &\coloneqq \left\{x \in X \,\middle|\, |f(x)| < \infty,\ |g(x)| < \infty,\ f(x) \neq g(x)\right\}. \end{aligned}$$

For any $x \in X$ with $f(x) \neq g(x)$ and $|f(x)| = \infty$ or $|g(x)| = \infty$, we have either $f(x) = \infty$ and $g(x) \neq \infty$, $g(x) = \infty$ and $f(x) \neq \infty$, $f(x) = -\infty$ and $g(x) \neq -\infty$, or $g(x) = -\infty$ and $f(x) \neq -\infty$. In every one of these cases, we have $x \in N_1$. Since $\{\infty\} \in \overline{\mathscr{B}}(\mathbb{R})$ and $\{-\infty\} \in \overline{\mathscr{B}}(\mathbb{R})$, N_1 is a μ-nullset. For N_2, we can now restrict ourselves to

$$Y \coloneqq \left\{x \in X \,\middle|\, |f(x)| < \infty,\ |g(x)| < \infty\right\} \in \Sigma.$$

On Y, we can safely examine the difference between f and g, which is a measurable function $f - g \colon Y \to \mathbb{R}$. This was previously problematic because differences between infinities are not always well-define. Evidently, we have

$$N_2 = \left\{x \in Y \,\middle|\, f(x) - g(x) \neq 0\right\}$$

and

$$N_2 = \bigcup_{i=1}^{\infty} \left\{\left|f|_Y - g|_Y\right| > \frac{1}{2^i}\right\}.$$

We now perform a proof by contradiction. If $\mu(N_2) > 0$ held, then there would exist $i_0 \in \mathbb{N}$ such that

$$\mu\Bigg(\underbrace{\left\{\left|f|_Y - g|_Y\right| > \frac{1}{2^{i_0}}\right\}}_{=:Z}\Bigg) > 0.$$

The newly defined set $Z \subseteq Y$ is measurable and is chosen such that f and g differ by at least $\frac{1}{2^{i_0}}$ everywhere on Z. Due to our initial assumption, we would have $\mu(Z) > 0$.

We now partition the set Y into "level slices" with respect to the value of f:

$$L_j \coloneqq (f|_Y)^{-1}\left(\left[\frac{j}{2^{i_0+1}}, \frac{j+1}{2^{i_0+1}}\right)\right) \quad \forall j \in \mathbb{Z}.$$

Here the thickness of the slice is very significant. By making each slice half as thick as the minimal difference between f and g, we guarantee that the values of f and g must lie in different slices in every point in Z. Furthermore, because the L_j form a partition of Y and $Z \subseteq Y$, $\mu(Z) > 0$ would imply that there exists $j_0 \in \mathbb{Z}$ such that $\mu(L_{j_0} \cap Z) > 0$.

Next, we partition the set $L_{j_0} \cap Z$ into two measurable subsets according to the sign of $f - g$:

$$\begin{aligned} L_+ &\coloneqq (L_{j_0} \cap Z) \cap \{f - g > 0\} \\ &= (L_{j_0} \cap Z) \cap \{f > g\}, \end{aligned}$$

$$\begin{aligned} L_- &\coloneqq (L_{j_0} \cap Z) \cap \{f - g < 0\} \\ &= (L_{j_0} \cap Z) \cap \{f < g\}. \end{aligned}$$

Here, we can ignore $\{f = g\}$ because Z is chosen such that $f \neq g$ everywhere on Z. Because of this, L_+ and L_- partition $L_{j_0} \cap Z$, and $\mu(L_{j_0} \cap Z) > 0$ would imply that $\mu(L_+) > 0$ or $\mu(L_-) > 0$. These two cases are analogous to one another and we discuss them separately.

Case 1 ($\mu(L_+) > 0$). For all $x \in L_+$, we have

$$g(x) < f(x) - \frac{1}{2^{i_0}} < \frac{j_0 + 1}{2^{i_0+1}} - \frac{2}{2^{i_0+1}} = \frac{j_0 - 1}{2^{i_0+1}}.$$

This means that

$$L_+ \subseteq \left\{g < \frac{j_0}{2^{i_0+1}}\right\} \setminus \left\{f < \frac{j_0}{2^{i_0+1}}\right\} \subseteq \left\{g < \frac{j_0}{2^{i_0+1}}\right\} \triangle \left\{f < \frac{j_0}{2^{i_0+1}}\right\}. =: M_+.$$

We would therefore have $\mu(M_+) > 0$, which would contradict our initial assumption because M_+ is the symmetric difference between the preimages of the extended interval $\left[-\infty, \frac{j_0}{2^{i_0+1}}\right)$ under f and g, respectively. However, because that extended interval is in $\overline{\mathscr{B}}(\mathbb{R})$, the premise of this part of the proof implies that those preimages are similar. ◁

Case 2 ($\mu(L_+) = 0$). In this case, we would have $\mu(L_-) > 0$. For $x \in L_-$, we have

$$g(x) > f(x) + \frac{1}{2^{i_0}} \geq \frac{j_0 + 2}{2^{i_0+1}},$$

which means that

$$L_- \subseteq \left\{g \geq \frac{j_0+1}{2^{i_0+1}}\right\} \setminus \left\{f \geq \frac{j_0+1}{2^{i_0+1}}\right\} \subseteq \left\{g \geq \frac{j_0+1}{2^{i_0+1}}\right\} \triangle \left\{f \geq \frac{j_0+1}{2^{i_0+1}}\right\} =: M_-,$$

from which we could then infer that $\mu(M_-) > 0$. This would contradict the premise of this part of the proof because M_- is the symmetric difference between preimages of an extended interval under f and g. ◁

In either case, we obtain a contradiction, which proves that the initial assumption of our contradiction argument, namely that $\mu(N_2) > 0$, must be false. We therefore have $\mu(N_1) = \mu(N_2) = 0$, which implies

$$\mu(\{f \neq g\}) = \mu(N_1 \cup N_2) \leq \mu(N_1) + \mu(N_2) = 0,$$

and therefore $f = g$ almost everywhere. The nullset on which they differ is $N \coloneqq N_1 \cup N_2$.

Part 3 ((3) $\implies$ (1)). Let $N \in \Sigma$ be a nullset such that $f(x) = g(x)$ for all $x \in X \setminus N$. Let further $t \in \mathbb{R} \cup \{\pm\infty\}$. For all $x \in \{f \leq t\} \triangle \{g \leq t\}$ we have either $f(x) \leq t$ and $g(x) > t$, or $g(x) \leq t$ and $f(x) > t$. In either case, we have $f(x) \neq g(x)$ and therefore $x \in N$. It follows that

$$\mu\Big(\underbrace{\{f \leq t\} \triangle \{g \leq t\}}_{\subseteq N}\Big) \leq \mu(N) = 0$$

and therefore $\{f \leq t\} \sim_\mu \{g \leq t\}$. □

Equality $f = g$ pointwise almost everywhere implies equality of integrals, i.e.,

$$\int_B |f| \,\mathrm{d}\mu = \int_B |g| \,\mathrm{d}\mu \quad \forall B \in \Sigma, \tag{2.21}$$

$$\int_B f \,\mathrm{d}\mu = \int_B g \,\mathrm{d}\mu \quad \forall B \in \Sigma \colon \int_B |f| \,\mathrm{d}\mu < \infty. \tag{2.22}$$

Equation (2.21) implies that f and g are integrable over the same sets. This will be relevant when we discuss pushforward and pullback functions in Section 2.3.3.

2.3 MEASURE SPACE GEODESICS

In contrast to vector spaces, similarity spaces and other metric spaces do not have a concept of "direction." If we have a step that seems suitable to improve the objective, we cannot increase or decrease it in order to obtain more or less of the desirable effect. This is because metric spaces generally do not allow for *scaling*. This is a problem for optimization methods that work with local model functions such as linearizations.

The algorithms that we develop later on were originally conceived as trust region methods (see [HLS22]). Much of the following section is inspired by frequent questions as to whether or not these algorithms can be written as line search algorithms. The answer to those questions is decidedly "yes," but has to be preceded by an answer to the more fundamental question of what a "line" is in a space that does not allow for scaling.

Although there is no general equivalent to scaling a step up, i.e., making it longer, we can restore the ability to scale a step down by using a construct known as a "geodesic." A geodesic is, by definition, a path that realizes the distance between any two points in its parameter interval.

Definition 2.3.1 (Geodesic).
Let $(X,\ d)$ be a metric space, let $I \subseteq \mathbb{R}$ be an interval. A *geodesic* is a map $\gamma\colon I \to X$ such that there exists a constant $C \geq 0$ with

$$d\big(\gamma(s), \gamma(t)\big) = C \cdot |s - t| \qquad \forall s, t \in I. \tag{2.23}$$

A geodesic is called *minimizing* if and only if Equation (2.23) is satisfied with $C = 1$. We refer to suitable constants C as *geodesic constants* of γ, and to I as the *parameter interval* of γ. ◁

The existence of geodesics with $C = 0$, $I = \emptyset$, or $I = \{t\}$ for a single $t \in \mathbb{R}$ is a set of pernicious edge cases that can make discussing measure space geodesics very tedious. To avoid this pitfall, we state most of the results in this section for *minimizing geodesics* $\gamma\colon I \to X$ *for non-empty* $I \subseteq \mathbb{R}$, which excludes the first two edge cases. Because geodesics for which $C \neq 0$ can always be affinely reparameterized such that $C = 1$, this is not a significant restriction. We refer to Theorem 2.3.48 on page 121 for a detailed proof.

It is evident that, due to the triangle inequality, geodesics are always shortest continuous paths between any two points along their length. The shortest path between two points is not unique. As we will show in Theorem 2.3.47 on page 117, there are usually infinitely many ways to rearrange a geodesic. Accordingly, there are usually infinitely many geodesics connecting two points.

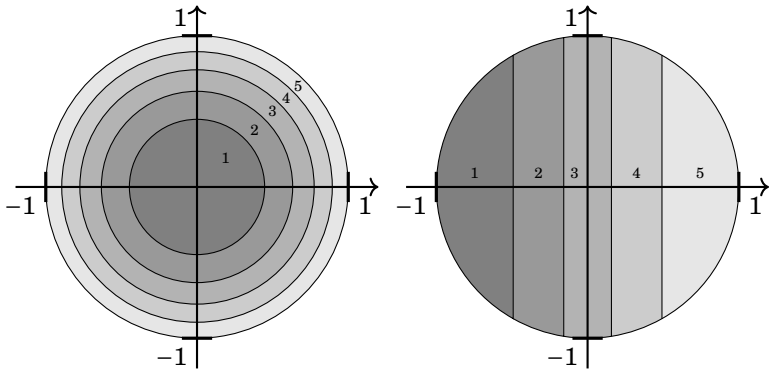
The primary goal of this section is Theorem 2.3.71 on page 150, which provides a strong criterion for measure spaces which are geodesic and therefore suitable for our algorithms. Along the way, we develop a theoretical toolkit that simplifies working with measure space geodesics. Much of this theory was originally developed in an unsuccessful attempt to prove Hypothesis 2.5.7 on page 208. Some of it also finds application in Chapter 3. We present all of it in hopes that interested readers may find future use in it.

This section begins with Section 2.3.1, in which we establish basic properties of geodesics in similarity spaces. We discuss these properties prior to proving the existence of these geodesics because their discussion simplifies later discussions of geodesic construction and handling.

In Section 2.3.2, we describe two types of geodesic interpolation: dense and sparse. Both allow us to define a geodesic by specifying its value at a countable set of support points rather than for every real parameter value. This can be useful if the values need to be modified using an iterative procedure prior to being used as support points. In *dense* interpolation, we demand that the set of all support points is dense in the parameter interval. *Sparse* interpolation adds the ability to "fill in" gaps between the support points.

In Section 2.3.3, we show that every geodesic in a similarity space corresponds to a measurable function whose sublevel sets encode the geodesic. We refer to these functions as *geodesic level set functions* or "GLSFs." GLSFs allow us to examine the variation of a geodesic over an arbitrary Borel set of parameters rather than just an interval. This lays the theoretical groundwork for "rearranging" geodesics. GLSFs also allow us to introduce pullback and pushforward functions which allow us to examine functions "along" geodesics.

In Section 2.3.4, we discuss several ways to modify and combine existing geodesics to obtain new ones. This is useful in some construction methods and is used later to prove the existence of geodesics.



(a) Expanding concentric circle (b) Ascending X coordinate

Figure 2.2: Illustration of two geodesics connecting the empty set with the unit circle. Depicted are states at five equidistant parameter points. Due to the geodesic property the change in volume is the same in each segment.

In Section 2.3.5, we describe the construction of special types of geodesics in similarity spaces. These construction methods are mostly simple extensions of our prior work.

With construction and modification methods, we prove necessary and sufficient conditions for a geodesic structure in measure spaces in Section 2.3.6.

2.3.1 Properties

The step between two similarity classes is best expressed by their symmetric difference. To get from similarity class $A \in \Sigma/_{\sim_\mu}$ to $B \in \Sigma/_{\sim_\mu}$, we must remove $A \setminus B$ from A and then add $B \setminus A$ to the result. This is the same as forming the symmetric difference between A and the step $A \triangle B$.

There is no unique way to perform "half of" this step. In order to do so, we must first fix an order in which the changes are to be performed. Once such an order is fixed, we could perform the first half of the ordered changes. A measure space geodesic encodes such an order. Figure 2.2 demonstrates how even a simple step like that from $\emptyset$ to the unit ball in $\mathbb{R}^2$ can be broken down into multiple different orders.

The first of these geodesics, depicted in Figure 2.2a, builds the unit ball up as a sequence of expanding concentric balls around the origin:

$$\gamma_1(t) \coloneqq \left[B_{\sqrt{t/\pi}}(0)\right]_{\sim_\lambda} \quad \forall t \in I$$

where $I \coloneqq [0, \pi]$ is the parameter interval and λ is the Lebesgue measure. The radius of the ball is precisely such that

$$\lambda\big(\gamma_1(t_2) \triangle \gamma_1(t_1)\big) = \lambda\big(\gamma_1(t_2)\big) - \lambda\big(\gamma_1(t_1)\big) = \pi \cdot \left(\sqrt{\frac{t_2}{\pi}}\right)^2 - \pi \cdot \left(\sqrt{\frac{t_1}{\pi}}\right)^2 = t_2 - t_1$$

for $0 \le t_1 \le t_2 \le \pi$. Because the rate of volume is constant, the increase in radius naturally becomes slower over time.

The second geodesic, depicted in Figure 2.2b, includes points in order of ascending X coordinate. The measure of the intermediate sets is that of a circular

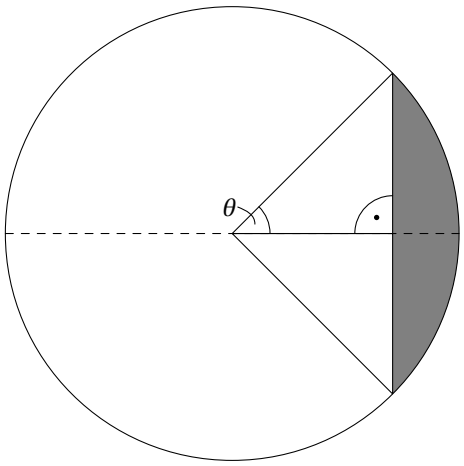


Figure 2.3: Illustration of the parameterization of the geodesic in Figure 2.2b on the previous page. The area of the shaded circular section is a bijective function of $\theta \in [0,\ \pi]$. By forming the inverse, we express θ as a function of the area. As $\theta \to \pi$, the shaded section covers the entire circle.

segment, which is most easily calculated using angles. For a given angle $\theta \in [0,\ \pi]$, the symmetric circular section depicted in Figure 2.3 covers the unit circle for $\theta \to \infty$. We have to parameterize this symmetric circular section by area. The area of one half is calculated by subtracting a right triangle from a circular sector:

$$A(\theta) \coloneqq 2 \cdot \left(\frac{1}{2} \cdot \theta - \sin\theta \cdot \cos\theta \right) = \theta - \sin(2\theta) = \frac{1}{2} \cdot \big(2\theta - \sin(2\theta)\big)$$

The function $A : [0,\ \pi] \to [0,\ \pi]$ is strictly monotonic and therefore bijective. Using the inverse of A, we can write

$$\gamma_2(t) \coloneqq \Big[B_1(0) \cap \big\{ (x,\ y) \,\big|\, x \leq -\cos(A^{-1}(t)) \big\} \Big]_{\sim_\mu} \qquad \forall t \in I.$$

We then have

$$\lambda\big(\gamma_2(t_2) \triangle \gamma_2(t_1)\big) = \lambda\big(\gamma_2(t_2)\big) - \lambda\big(\gamma_2(t_1)\big) = A(A^{-1}(t_2)) - A(A^{-1}(t_1)) = t_2 - t_1.$$

We can see that both γ_1 and γ_2 are geodesics connecting the same two sets. We stress that neither of these geodesics is "canonically correct" or "better" than the other. There is no inherent hierarchy of preference in geodesics. Therein lies the primary challenge of working with geodesics: finding the best geodesic for your purposes is not always straightforward.

2.3.1.1 MONOTONICITY AND TOTAL VARIATION

While geodesics can arrange changes in an almost arbitrary order, one thing is strictly forbidden: once a change is made, it cannot be undone. This is implied by the geodesic condition. For instance, if a geodesic were to make a change and then immediately undo it, the combined parameter interval would be twice as long while there would be no net change associated with it. We can also express this as a monotonicity condition of the following form: *The set of changes associated with a growing sub-interval of a geodesic's parameter interval is always increasing*.

Lemma 2.3.2 (Monotonicity of Measure Space Geodesics).
Let $(X,\ \Sigma,\ \mu)$ be a measure space, let $I \subseteq \mathbb{R}$ be an interval, and let $\gamma\colon I \to \Sigma/_{\sim_\mu}$ be a geodesic. For $a,b,c,d \in I$ with $a \le b \le c \le d$, we have

$$\gamma(b) \mathbin{\triangle} \gamma(c) \subseteq_\mu \gamma(a) \mathbin{\triangle} \gamma(d).$$ ◁

Proof. Let $a,b,c,d \in I$ with $a \le b \le c \le d$, and let $C \ge 0$ be a geodesic constant associated with γ. Let further

$$D \coloneqq \big(\gamma(b) \mathbin{\triangle} \gamma(c)\big) \mathbin{\triangle} \big(\gamma(a) \mathbin{\triangle} \gamma(d)\big).$$

The measure μ is monotonic and subadditive in the sense that

$$\begin{aligned} \mu(A) &\le \mu(B) && \forall A,B \in \Sigma/_{\sim_\mu}\colon A \subseteq_\mu B, \\ \mu(A \cup B) &\le \mu(A) + \mu(B) && \forall A,B \in \Sigma/_{\sim_\mu}. \end{aligned}$$

Because the symmetric difference is a subset of the union of two sets, this allows us to make the following estimate:

$$\begin{aligned} \mu(D) &\le \mu\Big(\big(\gamma(a) \mathbin{\triangle} \gamma(d)\big) \cup \big(\gamma(b) \mathbin{\triangle} \gamma(c)\big)\Big) \\ &\le \mu\big(\gamma(a) \mathbin{\triangle} \gamma(d)\big) + \mu\big(\gamma(b) \mathbin{\triangle} \gamma(c)\big) \\ &= C \cdot (d - a + c - b). \end{aligned}$$

The symmetric difference can be written as a disjoint union of set differences. Due to the additivity of μ, this yields

$$\begin{aligned} \mu(D) &= \mu\Big(\big(\big(\gamma(a) \mathbin{\triangle} \gamma(d)\big) \setminus \big(\gamma(b) \mathbin{\triangle} \gamma(c)\big)\big) \cup \big(\big(\gamma(b) \mathbin{\triangle} \gamma(c)\big) \setminus \big(\gamma(a) \mathbin{\triangle} \gamma(d)\big)\big)\Big) \\ &= \mu\Big(\big(\gamma(a) \mathbin{\triangle} \gamma(d)\big) \setminus \big(\gamma(b) \mathbin{\triangle} \gamma(c)\big)\Big) + \mu\Big(\big(\gamma(b) \mathbin{\triangle} \gamma(c)\big) \setminus \big(\gamma(a) \mathbin{\triangle} \gamma(d)\big)\Big). \end{aligned}$$

Therefore we have

$$\begin{aligned} \mu\big(\big(\gamma(b) \mathbin{\triangle} \gamma(c)\big) \setminus \big(\gamma(a) \mathbin{\triangle} \gamma(d)\big)\big) &= \mu(D) - \mu\big(\big(\gamma(a) \mathbin{\triangle} \gamma(d)\big) \setminus \big(\gamma(b) \mathbin{\triangle} \gamma(c)\big)\big) \\ &\le \mu(D) - \mu\big(\gamma(a) \mathbin{\triangle} \gamma(d)\big) + \mu\big(\gamma(b) \mathbin{\triangle} \gamma(c)\big) \\ &= \mu(D) - C \cdot (d - a) + C \cdot (c - b) \\ &\le C \cdot (d - a + c - b) - C \cdot (d - a + c - b) \\ &= 0 \end{aligned}$$

which proves that $\gamma(b) \mathbin{\triangle} \gamma(c) \subseteq_\mu \gamma(a) \mathbin{\triangle} \gamma(d)$. □

This monotonicity has far-reaching impact on the theory of measure space geodesics. First of all, we can use it to show that changes corresponding to disjoint parameter intervals of the same geodesic must be essentially disjoint. This is the "changes cannot be undone" result that we had mentioned earlier. Because every set is its own inverse with respect to the symmetric difference, making the same change twice is the same as undoing the change.

Lemma 2.3.3 (Essentially Disjointness of Geodesic Variation).
Let $(X,\ \Sigma,\ \mu)$ be a measure space, let $I \subseteq \mathbb{R}$ be an interval, and let $\gamma\colon I \to \Sigma/_{\sim_\mu}$ be a geodesic. Let further $a,b,c,d \in I$ with $a \le b \le c \le d$. Then we have

$$\mu\Big(\big(\gamma(a) \mathbin{\triangle} \gamma(b)\big) \cap \big(\gamma(c) \mathbin{\triangle} \gamma(d)\big)\Big) = 0.$$

In other words, the changes corresponding to the parameter intervals $[a,\ b]$ *and* $[c,\ d]$ *are essentially disjoint.* ◁

PROOF. Let $C \geq 0$ be a geodesic constant of γ. Let $a, b, c, d \in I$ with $a \leq b \leq c \leq d$. We first address the case where $b = c$. We have

$$\begin{aligned}
C \cdot (d-a) &= \mu\big(\gamma(a) \triangle \gamma(d)\big) \\
&= \mu\Big(\big(\gamma(a) \triangle \gamma(b)\big) \triangle \big(\underbrace{\gamma(b)}_{=\gamma(c)} \triangle \gamma(d)\big)\Big) \\
&\leq \mu\Big(\big(\gamma(a) \triangle \gamma(b)\big) \cup \big(\gamma(c) \triangle \gamma(d)\big)\Big) \\
&= \underbrace{\mu\big(\gamma(a) \triangle \gamma(b)\big)}_{=C\cdot(b-a)} + \underbrace{\mu\big(\gamma(c) \triangle \gamma(d)\big)}_{=C\cdot(d-c)=C\cdot(d-b)} \\
&\qquad - \mu\Big(\big(\gamma(a) \triangle \gamma(b)\big) \cap \big(\gamma(c) \triangle \gamma(d)\big)\Big) \\
&= C \cdot (d-a) - \mu\Big(\big(\gamma(a) \triangle \gamma(b)\big) \cap \big(\gamma(c) \triangle \gamma(d)\big)\Big).
\end{aligned}$$

Because $d - a \geq 0$ and because μ is a non-negative measure, this implies

$$\mu\Big(\big(\gamma(a) \triangle \gamma(b)\big) \cap \big(\gamma(c) \triangle \gamma(d)\big)\Big) = 0.$$

Next, we address the case in which $b < c$. According to Lemma 2.3.2, we have $\gamma(c) \triangle \gamma(d) \subseteq_\mu \gamma(b) \triangle \gamma(d)$. We can invoke the first case to show that

$$\mu\Big(\big(\gamma(a) \triangle \gamma(b)\big) \cap \big(\gamma(c) \triangle \gamma(d)\big)\Big) \leq \mu\Big(\big(\gamma(a) \triangle \gamma(b)\big) \cap \big(\gamma(b) \triangle \gamma(d)\big)\Big) = 0.$$

This implies that

$$\mu\big(\big(\gamma(a) \triangle \gamma(b)\big) \cap \big(\gamma(c) \triangle \gamma(d)\big)\big) = 0.$$ □

So far, we have not specified what type of interval the parameter interval I should be. From a theoretical standpoint, it would be desirable to have a closed parameter interval. It is evident that the geodesic property implies that geodesics are continuous. We could therefore consider continuations of geodesics to the boundaries of their parameter intervals without running into any issues. However, this would preclude the consideration of geodesics with unbounded parameter intervals.

It turns out that measure space geodesics have an odd property: they can have continuous continuations to infinity. This oddity arises from the properties of σ-algebras and is therefore not transferrable to other kinds of metric spaces. To investigate this property, we first have to investigate "variation."

For closed intervals $[a,\ b] \subseteq I$, we have a clear concept of the *variation* of γ, i.e., the totality of changes made by γ, over the interval $[a,\ b]$. For open and unbounded intervals, we can find the variation by exploiting the fact that the variation over an increasing sequence of closed intervals is also increasing according to Lemma 2.3.2. Increasing sequences of similarity classes have a well-defined limit in their countable union.

We can use this to derive a similarity class that encompasses the totality of changes made by a geodesic over its entire parameter interval. In essence, this is the total "step" made by the geodesic. We refer to this similarity class as the geodesic's *total variation*.

Definition 2.3.4 (Total Variation Of Geodesics).
Let $(X,\ \Sigma,\ \mu)$ be a measure space, let $I \subseteq \mathbb{R}$ be an interval, and let $\gamma\colon I \to \Sigma/_{\sim_\mu}$ be a geodesic. The *total variation of* γ is a similarity class $\mathrm{TV}(\gamma) \in \Sigma/_{\sim_\mu}$ such that

$$\gamma(s) \triangle \gamma(t) \subseteq_\mu \mathrm{TV}(\gamma) \quad \forall s,t \in I, \tag{2.24}$$

$$\mathrm{TV}(\gamma) \subseteq_\mu V \qquad \forall V \in \Sigma/_{\sim_\mu}\colon \gamma(s) \triangle \gamma(t) \subseteq_\mu V\ \forall s,t \in I. \tag{2.25}$$

In other words, $\mathrm{TV}(\gamma)$ is the smallest similarity class that contains all symmetric differences between similarity classes that occur along the geodesic γ. ◁

For simplicity, we will refer to a similarity class that satisfies Equation (2.24) as "encompassing" of the total variation because it encompasses all variations of a geodesic over sub-intervals of its parameter interval. We also refer to classes that satisfy Equation (2.25) as "bounding" for the total variation because they have to be contained within all encompassing classes.

Proposition 2.3.5 (Existence and Uniqueness of the Total Variation).
Let $(X,\ \Sigma,\ \mu)$ be a measure space, let $I \subseteq \mathbb{R}$ be an interval, and let $\gamma\colon I \Sigma/_{\sim_\mu}$ be a geodesic. Then there exists a unique similarity class $\mathrm{TV}(\gamma) \in \Sigma/_{\sim_\mu}$ that satisfies Definition 2.3.4.

Let $(s_i)_{i\in\mathbb{N}} \in I^{\mathbb{N}}$ and $(t_i)_{i\in\mathbb{N}} \in I^{\mathbb{N}}$ such that $s_i \to \inf I$ and $t_i \to \sup I$ for $i \to \infty$. Then we have

$$\mathrm{TV}(\gamma) = \bigcup_{i=1}^{\infty} \big(\gamma(s_i) \triangle \gamma(t_i)\big). \tag{2.26}$$

◁

Proof. Part 1 (Uniqueness). Let $U, V \in \Sigma/_{\sim_\mu}$ satisfy be similarity classes that qualify as total variations of γ according to Definition 2.3.4. This means that, let both U and V are both encompassing and bounding. Because encompassing classes have to encompass all bounding classes, we have both $U \subseteq_\mu V$ and $V \subseteq_\mu U$. It follows that $U = V$ because

$$\mu(U \triangle V) = \underbrace{\mu(U \setminus V)}_{=0} + \underbrace{\mu(V \setminus U)}_{=0} = 0.$$

Part 2 (Existence and Equation (2.26)). We first address the edge case in which $I = \emptyset$. In this case, there exist no $s,t \in I$ and every similarity class is encompassing. This includes $[\emptyset]_{\sim_\mu}$, which does not have any essential subsets except for itself. Therefore, $[\emptyset]_{\sim_\mu}$ is minimally encompassing and we have

$$\mathrm{TV}(\gamma) = [\emptyset]_{\sim_\mu}.$$

For non-trivial geodesics, we calculate the total variation by using Equation (2.26). Let $(s_i)_{i\in\mathbb{N}}$ and $(t_i)_{i\in\mathbb{N}}$ be sequences in I such that $s_i \to \inf I$ and $t_i \to \sup I$ for $i \to \infty$. We write

$$V \coloneqq \bigcup_{i=1}^{\infty} \big(\gamma(s_i) \triangle \gamma(t_i)\big) \in \Sigma/_{\sim_\mu}.$$

Let $U \in \Sigma/_{\sim_\mu}$ be any encompassing similarity class. Because V is a union of variations over closed intervals, we have

$$V = \bigcup_{i=1}^{\infty} \underbrace{\big(\gamma(s_i) \triangle \gamma(t_i)\big)}_{\subseteq_\mu U} \subseteq_\mu U.$$

Because this holds for all encompassing classes U, it follows that V is bounding for the total variation.

To prove that V is also encompassing, let $q, r \in I$ act as the bounds of an interval $[q,\ r]$. Without loss of generality, let $q \leq r$. Let $\varepsilon > 0$ be a constant that can be arbitrarily close to 0, and let $C \geq 0$ denote a geodesic constant of γ. Because $q \geq \inf I$ and $s_i \to \inf I$ for $i \to \infty$, there exists $i_1 \in \mathbb{N}$ such that

$$s_i \leq q + \frac{\varepsilon}{2(C+1)} \quad \forall i \geq i_1.$$

Similarly, because $r \leq \sup I$ and $t_i \to \sup I$ for $i \to \infty$, there exists $i_2 \in \mathbb{N}$ such that

$$t_i \geq r - \frac{\varepsilon}{2(C+1)} \quad \forall i \geq i_2.$$

We have to allow for an additional margin to account for the possibility that $q = \inf I$ or $r = \sup I$. In that case, the sequences can approach q or r, respectively, without ever passing it. Let $i_3 := \max\{i_1,\ i_2\}$. We find that

$$\begin{aligned}\big(\gamma(q) \triangle \gamma(r)\big) \setminus V &= \big(\gamma(q) \triangle \gamma(r)\big) \setminus \bigcup_{i=1}^{\infty} \big(\gamma(s_i) \triangle \gamma(t_i)\big) \\ &\subseteq_\mu \big(\gamma(q) \triangle \gamma(r)\big) \setminus \big(\gamma(s_{i_3}) \triangle \gamma(t_{i_3})\big).\end{aligned}$$

By using the monotonicity of γ, we can infer that

$$\begin{aligned}\mu\Big(\big(\gamma(q) \triangle \gamma(r)\big) \setminus V\Big) \leq\ & \begin{cases} \mu\big(\gamma(q) \triangle \gamma(s_{i_3})\big) & \text{if } q \leq s_{i_3}, \\ 0 & \text{if } q > s_{i_3} \end{cases} \\ & + \begin{cases} \mu\big(\gamma(t_{i_3}) \triangle \gamma(r)\big) & \text{if } t_{i_3} \leq r, \\ 0 & \text{if } t_{i_3} > r \end{cases} \\ &= C \cdot \max\{0,\ s_{i_3} - q\} + C \cdot \max\{0,\ r - t_{i_3}\} \\ &\leq 2C \cdot \frac{\varepsilon}{2(C+1)} \\ &< \varepsilon.\end{aligned}$$

Because this holds for all $\varepsilon > 0$, we obtain

$$\mu\Big(\big(\gamma(q) \triangle \gamma(r)\big) \setminus V\Big) = 0$$

and therefore $\gamma(q) \triangle \gamma(r) \subseteq_\mu V$. The fact that this holds for all $q, r \in I$ means that V is encompassing. □

Proposition 2.3.5 shows an inherent limitation of geodesics. Because the parameter space of a geodesic is a real interval, it can always be expressed as a countable union of closed intervals. The total variation of a geodesic must therefore be a countable union of variations over closed intervals. All variations of γ over bounded closed sub-intervals of I have finite measure, which means that the total variation of a geodesic is always σ-finite. This is an early indicator of Section 2.3.6, in which we will show that σ-finiteness is necessary for a measure space to be geodesic.

2.3.1.2 TRANSLATION AND LIMIT POINTS

In school settings, vectors are often described as "arrows," that have a specific length and direction, but no specific origin point. Instead, they can freely be shifted around and affixed to an arbitrary origin point. In this section, we show that measure space geodesics behave similarly with the symmetric difference as an analogous operation to vector addition.

This analogy justifies thinking of measure space geodesics as directions rather than just paths. This also makes them an excellent theoretical substitute for vectors.

Lemma 2.3.6 (Translation of Geodesics).
Let (X, Σ, μ) be measure spaces for $i \in [n]$, let $I \subseteq \mathbb{R}$ be an interval, let $\gamma \colon I \to \Sigma/_{\sim_\mu}$ be a geodesic, and let $A \in \Sigma/_{\sim_\mu}$. Then the map $\gamma \triangle A \colon I \to \Sigma/_{\sim_\mu}$ with

$$(\gamma \triangle A)(t) \coloneqq \gamma(t) \triangle A \quad \forall t \in I$$

is a geodesic. If $C \geq 0$ is a geodesic constant of γ, then C is also a geodesic constant of $\gamma \triangle A$. ◁

PROOF. Let $s, t \in I$. We have

$$(\gamma \triangle A)(s) \triangle (\gamma \triangle A)(t) = \gamma(s) \triangle A \triangle \gamma(t) \triangle A = \gamma(s) \triangle \gamma(t).$$

Therefore, if $C \geq 0$ is a geodesic constant of γ, then we have

$$\mu\big((\gamma \triangle A)(s) \triangle (\gamma \triangle A)(t)\big) = \mu\big(\gamma(s) \triangle \gamma(t)\big) = C \cdot |s - t|.$$

This shows that $\gamma \triangle A$ is a geodesic with geodesic constant C. □

The fact that we can translate a geodesic may not appear profound at first, but it is not self-evident. Metric spaces do not generally have an addition operation. Therefore, there are no a priori prescriptions dictating that the distance between two points has to remain the same if both are shifted in the same way. The reason why a translated measure space geodesic is still a geodesic is because the underlying metric d_μ is translation invariant. In a metric space with a group structure in which the metric is not invariant under translation, it is generally not possible to consider geodesics as *directions* independent of their concrete location in space.

Because variation is defined by the differences between points on a geodesic, translation has no impact on variation. If we think of the total variation as a set of changes that is applied to the origin point to arrive at the destination point, then geodesics impose an order in which these changes are applied. The following lemma shows that the total variation does not change. However, the proof, much like that of Lemma 2.3.6 is based on the fact that it is evident that variations on closed sub-intervals are unaffected by translation.

Lemma 2.3.7 (Translation Invariance of the Total Variation).
Let (X, Σ, μ) be a measure space, let $I \subseteq \mathbb{R}$ be an interval, let $\gamma \colon I \to \Sigma/_{\sim_\mu}$, and let $A \in \Sigma/_{\sim_\mu}$. Then we have

$$\mathrm{TV}(\gamma \triangle A) = \mathrm{TV}(\gamma).$$ ◁

PROOF. Once more, we have to treat the case in which $I = \emptyset$ differently. If $I = \emptyset$, we have $\mathrm{TV}(\gamma \triangle A) = [\emptyset]_{\sim_\mu} = \mathrm{TV}(\gamma)$. If $I \neq \emptyset$, then there exist sequences $(s_i)_{i\in\mathbb{N}} \in I^{\mathbb{N}}$ and $(t_i)_{i\in\mathbb{N}} \in I^{\mathbb{N}}$ such that $s_i \to \inf I$ and $t_i \to \sup I$ for $i \to \infty$. We then have

$$\begin{aligned}\mathrm{TV}(\gamma \triangle A) &= \bigcup_{i=1}^{\infty} \big((\gamma \triangle A)(s_i) \triangle (\gamma \triangle A)(t_i)\big) \\ &= \bigcup_{i=1}^{\infty} \big(\gamma(s_i) \triangle A \triangle A \triangle \gamma(t_i)\big) \\ &= \bigcup_{i=1}^{\infty} \big(\gamma(s_i) \triangle \gamma(t_i)\big) \\ &= \mathrm{TV}(\gamma).\end{aligned}$$

□

By translating a geodesic, we do not change variation. However, we do change the "origin point" to which the changes are applied. The concept of an origin point seems intuitive for a geodesic whose parameter interval includes its infimum. For bounded intervals that do not include their infimum, we could conceive of a continuity-based argument to extend the geodesic to the infimum. However, the intuitive concept of an origin point breaks down when we remember that in infinite measure spaces, geodesics can have unbounded parameter intervals.

Here, we come to one of the most interesting theoretical peculiarities about measure space geodesics: *every measure space geodesic connects two well-defined end points, regardless of whether it is infinitely long or not*. The term "end point" is somewhat inappropriate here because an infinitely long geodesic does, by definition, not have "ends." We address this linguistic pitfall by referring to these points as "limit points" instead. Specifically, we call the left limit point "origin point" and the right one "destination point."

Definition 2.3.8 (Limit Points Of Geodesics).
Let (X, Σ, μ) be a measure space, let $I \subseteq \mathbb{R}$ be an interval, and let $\gamma\colon I \to \Sigma/_{\sim_\mu}$ be a geodesic. We refer to a similarity class $A \in \Sigma/_{\sim_\mu}$ as the *origin point* or *origin* of γ if and only if

$$\gamma(s) \triangle A \subseteq_\mu \gamma(t) \triangle A \quad \forall s,t \in I\colon s \leq t, \tag{2.27}$$

$$\bigcap_{i=1}^{\infty} \big(\gamma(s_i) \triangle A\big) = [\emptyset]_{\sim_\mu} \qquad \forall (s_i)_{i\in\mathbb{N}} \subseteq I^{\mathbb{N}}\colon s_i \xrightarrow{i\to\infty} \inf I. \tag{2.28}$$

Similarly, we refer to a similarity class $B \in \Sigma/_{\sim_\mu}$ as the *destination point* or *destination* of γ if and only if

$$\gamma(s) \triangle B \supseteq_\mu \gamma(t) \triangle B \qquad \forall s,t \in I\colon s \leq t, \tag{2.29}$$

$$\bigcap_{i=1}^{\infty} \big(\gamma(t_i) \triangle B\big) = [\emptyset]_{\sim_\mu} \qquad \forall (t_i)_{i\in\mathbb{N}} \subseteq I\colon t_i \xrightarrow{i\to\infty} \sup I. \tag{2.30}$$

Collectively, we refer to the origin and destination points of a geodesic as its *limit points*. ◁

The choice of the term "limit point" is not arbitrary. Limit points are indeed limits of sequences of similarity classes. The geodesic property is of significant importance to the construction of limit points. We have already vaguely hinted

at "continuity arguments" to continue geodesics to the extremes of their parameter intervals. For points on the boundary of the parameter interval that are not included within the interval, such arguments require completeness. For continuations to infinity, however, this is not sufficient and we have to use the monotonicity of geodesics to argue further.

If we take any fixed point $t_0 \in I$ and we select an increasing subsequence of $(t_i)_{i\in\mathbb{N}}$ such that $t_0 \leq t_i$ for all $i \in \mathbb{N}$, then the sequence $\big(\gamma(t_0) \triangle \gamma(t_i)\big)_{i\in\mathbb{N}}$ is also μ-essentially increasing. Being an essentially increasing sequence of similarity classes, Lemma 2.2.17 shows that this sequence has a well-defined limit in the countable union

$$\bigcup_{i=1}^{\infty} \big(\gamma(t_0) \triangle \gamma(t_i)\big)$$

which is in $\Sigma/_{\sim_\mu}$ according to Lemma 2.2.14. If $t_i \to \sup I$ for $i \to \infty$, then we can then interpret

$$B_\gamma \coloneqq \gamma(t_0) \triangle \bigcup_{i=1}^{\infty} \big(\gamma(t_0) \triangle \gamma(t_i)\big)$$

as the destination point of γ. An origin point may similarly be derived by choosing a decreasing sequence $(s_i)_{i\in\mathbb{N}} \in I^{\mathbb{N}}$ that approaches to the infimum of I.

This is, intuitively, why origin and destination points exist and we will argue this more formally in Propositions 2.3.11 and 2.3.12. However, we must also bear in mind that our construction method, as stated, depends on the starting point t_0 as well as the approximating sequences $(s_i)_{i\in\mathbb{N}}$ and $(t_i)_{i\in\mathbb{N}}$, respectively. We will have to show that the choice of these parameters does not affect the constructed result.

Before we do this, we briefly highlight a different aspect of this construction method. When we say "$(s_i)_{i\in\mathbb{N}}$ approaches $\inf I$," then this intuitively evokes the image of a sequence that converges to a finite number. There is, however, no aspect of the construction described above that requires the infimum to be finite. The construction method works equally well if the infimum is $-\infty$ and $s_i \to -\infty$ for $i \to \infty$. We will also see that this does not affect well-definedness. This enables us to state most of our theory of measure space geodesics without having to presume finiteness of the underlying measure space.

We now rigorously prove the well-definedness of geodesic limit points. For our later work, it is convenient to begin with the origin point and derive the destination point from the origin point and the total variation. We begin by defining the *canonical form* of a geodesic, which is a translated form of a geodesic where the origin point is normalized to $[\emptyset]_{\sim_\mu}$.

Definition 2.3.9 (Canonical Form of a Geodesic).
Let $(X,\ \Sigma,\ \mu)$ be a measure space, and let $I \subseteq \mathbb{R}$ be an interval. We say that a geodesic $\gamma\colon I \to \Sigma/_{\sim_\mu}$ is in *canonical form* if and only if we have

$$\gamma(s) \subseteq_\mu \gamma(t) \qquad \forall s,\ t \in I : s \leq t, \tag{2.31}$$

$$\gamma(t) \subseteq_\mu \mathrm{TV}(\gamma) \quad \forall t \in I. \tag{2.32}$$

If γ is in canonical form, we refer to γ as a *canonical geodesic*. ◁

Careful comparison between Definitions 2.3.8 and 2.3.9 quickly indicates that a canonical geodesic is a geodesic whose origin point is $[\emptyset]_{\sim_\mu}$. A semi-explicit formula for the canonical form of a geodesic can be given by using the total

variation of a restriction of the geodesic γ. Essentially, the canonical form of a geodesic is the geodesic formed by the totality of all changes performed up to a certain parameter point.

Theorem 2.3.10 (Canonical Form of a Geodesic).
Let (X, Σ, μ) be a measure space, let $I \subseteq \mathbb{R}$ be an interval, and let $\gamma\colon I \to \Sigma/_{\sim_\mu}$ be a geodesic. Then the geodesic $\check{\gamma}\colon I \to \Sigma/_{\sim_\mu}$ with

$$\check{\gamma}(t) \coloneqq \mathrm{TV}(\gamma|_{I\cap(-\infty,\, t]})$$

is in canonical form and satisfies

$$\check{\gamma}(s) \triangle \check{\gamma}(t) = \gamma(s) \triangle \gamma(t) \quad \forall s,t \in I. \tag{2.33}$$

We refer to $\check{\gamma}$ as the canonical form of γ. ◁

PROOF. PART 1 (EQUATION (2.33); $\check{\gamma}$ IS A GEODESIC). If $I = \emptyset$, then every map $\check{\gamma}\colon I \to \Sigma/_{\sim_\mu}$ is a geodesic and Equation (2.33) is trivially always satisfied. If $I \neq \emptyset$, then let $s,t \in I$. Without loss of generality, we assume that $s \leq t$. Let further $(s_i)_{i\in\mathbb{N}} \in I^{\mathbb{N}}$ with $s_i \leq s$ for all $i \in \mathbb{N}$ and $s_i \to \inf I$ for $i \to \infty$. According to Proposition 2.3.5 and Lemma 2.3.3, we have

$$\begin{aligned}
\check{\gamma}(t) &= \mathrm{TV}(\gamma|_{I\cap(-\infty,\, t]}) \\
&= \bigcup_{i=1}^{\infty} \underbrace{\big(\gamma(s_i) \triangle \gamma(t)\big)}_{\supseteq_\mu \gamma(s_i)\triangle\gamma(s)} \\
&= \bigcup_{i=1}^{\infty} \Big(\big(\gamma(s) \triangle \gamma(t)\big) \triangle \big(\gamma(s_i) \triangle \gamma(s)\big)\Big) \\
&= \bigcup_{i=1}^{\infty} \Big(\big(\gamma(s) \triangle \gamma(t)\big) \cup \big(\gamma(s_i) \triangle \gamma(s)\big)\Big) \\
&= \big(\gamma(s) \triangle \gamma(t)\big) \cup \underbrace{\bigcup_{i=1}^{\infty} \big(\gamma(s_i) \triangle \gamma(s)\big)}_{=\mathrm{TV}(\gamma|_{I\cap(-\infty,\, s]})=\check{\gamma}(s)} \\
&= \big(\gamma(s) \triangle \gamma(t)\big) \triangle \check{\gamma}(s).
\end{aligned}$$

By forming the symmetric difference with $\check{\gamma}(s)$, we obtain Equation (2.33). For the distance between two points along the canonical form, this implies that

$$\mu\big(\check{\gamma}(s) \triangle \check{\gamma}(t)\big) = \mu\big(\gamma(s) \triangle \gamma(t)\big) = C \cdot |s-t|$$

for every geodesic constant $C \geq 0$ of γ. This demonstrates that $\check{\gamma}$ is also a geodesic.

PART 2 ($\check{\gamma}$ IS IN CANONICAL FORM). Let $s,t \in I$. Without loss of generality let $s \leq t$. Let further $(s_i)_{i\in\mathbb{N}} \subseteq I^{\mathbb{N}}$ with $s_i \leq s$ for all $i \in \mathbb{N}$ and $s_i \to \inf I$ for $i \to \infty$. Then we have

$$\check{\gamma}(t) = \mathrm{TV}(\gamma|_{I\cap(-\infty,\, t]}) = \bigcup_{i=1}^{\infty} \big(\gamma(s_i) \triangle \gamma(t)\big) = \bigcup_{i=1}^{\infty} \big(\check{\gamma}(s_i) \triangle \check{\gamma}(t)\big) \subseteq_\mu \mathrm{TV}(\check{\gamma}).$$

The monotonicity of measure space geodesics implies that

$$\check{\gamma}(s) = \bigcup_{i=1}^{\infty} \underbrace{\big(\gamma(s_i) \triangle \gamma(s)\big)}_{\subseteq_\mu \gamma(s_i)\triangle\gamma(t)} \subseteq_\mu \bigcup_{i=1}^{\infty} \big(\gamma(s_i) \triangle \gamma(t)\big) = \check{\gamma}(t).$$

Together, both of these equations show that $\check{\gamma}$ is in canonical form. □

The well-definedness of the canonical form of γ is a direct consequence of the well-definedness of the total variation. From now on, we adopt the universal notation $\check{\gamma}$ for the canonical form of a geodesic γ.

The fact that $\check{\gamma}(s) \triangle \check{\gamma}(t) = \gamma(s) \triangle \gamma(t)$ holds for all $s, t \in I$ suggests that $\check{\gamma}$ is a translation of γ. This is easy to show and implies the uniqueness of the canonical form. For non-empty parameter intervals $I \subseteq \mathbb{R}$, the similarity class by which γ must be translated to obtain $\check{\gamma}$ is uniquely defined and is the sole origin point of γ.

Proposition 2.3.11 (Existence and Uniqueness of Origin Points).
Let (X, Σ, μ) be a measure space, let $I \subseteq \mathbb{R}$, and let $\gamma \colon I \to \Sigma/_{\sim_\mu}$ be a geodesic. Let further $A \in \Sigma/_{\sim_\mu}$. The following three statements are equivalent:

(1) A is an origin point of γ;

(2) $\gamma \triangle A = \check{\gamma}$;

(3) $\gamma \triangle A$ is canonical.

There always exists at least one origin point of γ. If $I \neq \emptyset$, then γ has exactly one origin point $A_\gamma \in \Sigma/_{\sim_\mu}$. This origin point satisfies

$$A_\gamma = \gamma(t_0) \triangle \mathrm{TV}(\gamma|_{I \cap (-\infty,\, t_0]}) \quad \forall t_0 \in I.$$

The canonical form $\check{\gamma}$ is the only canonical geodesic that is a translation of γ. $\lhd$

Proof. Part 1 ((1) $\Longrightarrow$ (2)). Let $A \in \Sigma/_{\sim_\mu}$ be an origin point of γ. we have to show that $\gamma(t) \triangle A = \check{\gamma}(t) = \mathrm{TV}(\gamma|_{I \cap (-\infty,\, t]})$ for all $t \in I$. For $I = \emptyset$, this is trivially true. For $I \neq \emptyset$, let $t \in I$. Because $t \in I$, there exists $(s_i)_{i \in \mathbb{N}} \in I^{\mathbb{N}}$ with $s_i \leq t$ for all $i \in \mathbb{N}$ and $s_i \to \inf I$ for $i \to \infty$. Because A is an origin point, Definition 2.3.8 dictates that

$$\begin{aligned}
\check{\gamma}(t) &= \mathrm{TV}(\gamma|_{I \cap (-\infty,\, t]}) \\
&= \bigcup_{i=1}^{\infty} \big(\gamma(s_i) \triangle \gamma(t)\big) \\
&= \bigcup_{i=1}^{\infty} \Big(\big(\underbrace{\gamma(s_i) \triangle A}_{\subseteq_\mu \gamma(t) \triangle A}\big) \triangle \big(\gamma(t) \triangle A\big)\Big) \\
&= \bigcup_{i=1}^{\infty} \Big(\big(\gamma(t) \triangle A\big) \setminus \big(\gamma(s_i) \triangle A\big)\Big) \\
&= \big(\gamma(t) \triangle A\big) \setminus \underbrace{\bigcap_{i=1}^{\infty} \big(\gamma(s_i) \triangle A\big)}_{\overset{2.3.8}{=} [\emptyset]_{\sim_\mu}} \\
&= \gamma(t) \triangle A.
\end{aligned}$$

This holds for all $t \in I$. Therefore, we have $\gamma \triangle A = \check{\gamma}$.

Part 2 ((2) $\Longrightarrow$ (3)). Let A be such that $\gamma \triangle A = \check{\gamma}$. Theorem 2.3.10 shows that $\check{\gamma}$ is canonical. Therefore, $\gamma \triangle A$ is canonical.

Part 3 ((3) $\Longrightarrow$ (1)). Let $A \in \Sigma/_{\sim_\mu}$ be such that $\gamma \triangle A$ is canonical. According to Definition 2.3.9, we have

$$\gamma(s) \triangle A \subseteq_\mu \gamma(t) \triangle A \quad \forall s,\, t \in I : s \leq t,$$

which means that A satisfies Equation (2.27). Let $(s_i)_{i\in\mathbb{N}} \in I^{\mathbb{N}}$ be a sequence with $s_i \to \inf I$ for $i \to \infty$. If such a sequence exists, then we may assume its existence to imply $I \neq \emptyset$. There then similarly exists a sequence $(t_i)_{i\in\mathbb{N}} \in I^{\mathbb{N}}$ with $t_i \geq s_i$ for all $i \in \mathbb{N}$ and $t_i \to \sup I$ for $i \to \infty$. Because $t_i \geq s_i$ for all $i \in \mathbb{N}$, Equation (2.27) implies

$$\big(\gamma(t_i) \triangle A\big) \triangle \underbrace{\big(\gamma(s_i) \triangle A\big)}_{\subseteq_\mu \gamma(t_i)\triangle A} = \big(\gamma(t_i) \triangle A\big) \setminus \big(\gamma(s_i) \triangle A\big) \quad \forall i \in \mathbb{N}.$$

Because $\gamma \triangle A$ is canonical, Definition 2.3.9 implies

$$\begin{aligned}
\bigcap_{i=1}^{\infty} \underbrace{\big(\gamma(s_i) \triangle A\big)}_{\subseteq_\mu \mathrm{TV}(\gamma\triangle A)} &\subseteq_\mu \mathrm{TV}(\gamma \triangle A) \\
&= \bigcup_{i=1}^{\infty} \big(\gamma(t_i) \triangle \gamma(s_i)\big) \\
&= \bigcup_{i=1}^{\infty} \Big(\big(\gamma(t_i) \triangle A\big) \triangle \underbrace{\big(\gamma(s_i) \triangle A\big)}_{\subseteq_\mu \gamma(t_i)\triangle A}\Big) \\
&= \bigcup_{i=1}^{\infty} \Big(\big(\gamma(t_i) \triangle A\big) \setminus \big(\gamma(s_i) \triangle A\big)\Big) \\
&\subseteq_\mu \bigcup_{i=1}^{\infty} \Big(\big(\gamma(t_i) \triangle A\big) \setminus \bigcap_{j=1}^{\infty} \big(\gamma(s_j) \triangle A\big)\Big) \\
&= \Big(\bigcup_{i=1}^{\infty} \big(\gamma(t_i) \triangle A\big)\Big) \setminus \Big(\bigcap_{i=1}^{\infty} \big(\gamma(s_i) \triangle A\big)\Big).
\end{aligned}$$

We see that the intersection on the left hand side is an essential subset of a similarity class from which it is simultaneously essentially disjoint. This is only possible if

$$\bigcap_{i=1}^{\infty} \underbrace{\big(\gamma(s_i) \triangle A\big)}_{\subseteq_\mu \mathrm{TV}(\gamma\triangle A)} = [\emptyset]_{\sim_\mu},$$

which is exactly Equation (2.28).

PART 4 (UNIQUENESS OF $\check{\gamma}$). If $\gamma \triangle A$ is canonical, then A is an origin point of γ and therefore $\gamma \triangle A = \check{\gamma}$. Therefore, $\check{\gamma}$ is the only translation of γ that is canonical.

PART 5 (EXISTENCE AND UNIQUENESS OF A). We first consider the edge case where $I = \emptyset$. Due to the fact that both Equations (2.27) and (2.28) are trivially true for all $A \in \Sigma/_{\sim_\mu}$, every such A is an origin point. Specifically, $[\emptyset]_{\sim_\mu}$ is an appropriate choice of origin point. Uniqueness is not guaranteed in this case.

Next, we consider the case in which $I \neq \emptyset$. Because $\check{\gamma}$ is the only translation of γ that is canonical, an origin point $A \in \Sigma/_{\sim_\mu}$ must satisfy

$$A = \gamma(t) \triangle \check{\gamma}(t) = \gamma(t) \triangle \mathrm{TV}(\gamma|_{I\cap(-\infty,\, t]}) \quad \forall t \in I.$$

As there exists at least one $t_0 \in I$, we define

$$A_\gamma \coloneqq \gamma(t_0) \triangle \check{\gamma}(t_0)$$

for an arbitrary $t_0 \in I$. Let $t \in I$ be given and let $(s_i)_{i\in\mathbb{N}} \subseteq I^{\mathbb{N}}$ be a sequence with $s_i \leq \min\{t,\ t_0\}$ for all $i \in \mathbb{N}$ and $s_i \to \inf I$ for $i \to \infty$. Due to the monotonicity of measure space geodesics, we have

$$\big(\gamma(t) \triangle \gamma(t_0)\big) \cap \big(\gamma(t_0) \triangle \gamma(s_i)\big) = \begin{cases} [\emptyset]_{\sim_\mu} & \text{if } t \geq t_0, \\ \gamma(t) \triangle \gamma(t_0) & \text{if } t < t_0 \end{cases}$$

for all $i \in \mathbb{N}$. We note that this case distinction is independent of i. The symmetric difference between these similarity classes is therefore either always a disjoint union or always a set difference, both of which distribute over unions. We can then make the following reformulation:

$$\begin{aligned} \gamma(t) \triangle A_\gamma &= \big(\gamma(t) \triangle \gamma(t_0)\big) \triangle \check{\gamma}(t_0) \\ &= \big(\gamma(t) \triangle \gamma(t_0)\big) \triangle \bigcup_{i=1}^{\infty} \big(\gamma(t_0) \triangle \gamma(s_i)\big) \\ &= \bigcup_{i=1}^{\infty} \Big(\big(\gamma(t) \triangle \gamma(t_0)\big) \triangle \big(\gamma(t_0) \triangle \gamma(s_i)\big)\Big) \\ &= \bigcup_{i=1}^{\infty} \big(\gamma(t) \triangle \gamma(s_i)\big) \\ &= \check{\gamma}(t). \end{aligned}$$

Having proven that $\gamma \triangle A_\gamma = \check{\gamma}$ and that A_γ is an origin point, we still need to show that A_γ is unique. If $B \in \Sigma/_{\sim_\mu}$ is another origin point of γ, then $\gamma \triangle B = \check{\gamma} = \gamma \triangle A_\gamma$. With the same $t_0 \in I$ as before, we have

$$\gamma(t_0) \triangle B = \check{\gamma}(t_0) = \gamma(t_0) \triangle A_\gamma.$$

By subtracting $\gamma(t_0)$ from both sides, we obtain

$$B = \gamma(t_0) \triangle \gamma(t_0) \triangle A_\gamma = A_\gamma. \qquad \square$$

We do not need to replicate the effort of Proposition 2.3.11 for destination points. As one would expect, if the origin point A_γ is where the geodesic γ "starts" and $\mathrm{TV}(\gamma)$ comprises exactly all changes made by γ over its parameter interval, then the destination point of γ should be $B_\gamma \coloneqq A_\gamma \triangle \mathrm{TV}(\gamma)$. This simplifies existence and uniqueness proofs for the destination point significantly.

Proposition 2.3.12 (Existence and Uniqueness of Destination Points). *Let $(X,\ \Sigma,\ \mu)$ be a measure space, let $I \subseteq \mathbb{R}$ be an interval, let $\gamma\colon I \to \Sigma/_{\sim_\mu}$ be a geodesic, and let $B \in \Sigma/_{\sim_\mu}$. We have*

$$B \text{ is a destination point of } \gamma \iff B \triangle \mathrm{TV}(\gamma) \text{ is an origin point of } \gamma$$

Every geodesic has a destination point. If $I \neq \emptyset$, then the destination point B_γ of γ is unique. $\triangleleft$

Proof. Part 1 (Reformulation of $\mathrm{TV}(\gamma)$). Let $(s_i)_{i\in\mathbb{N}} \in I^{\mathbb{N}}$ be a decreasing sequence such that $s_i \to \inf I$ for $i \to \infty$, let $(t_i)_{i\in\mathbb{N}} \in I^{\mathbb{N}}$ be an increasing

sequence such that $s_i \leq t_i$ for all $i \in \mathbb{N}$ and $t_i \to \sup I$ for $i \to \infty$, and let $A \in \Sigma/_{\sim_\mu}$ be an origin point of γ. Then we can rewrite $\mathrm{TV}(\gamma)$ as follows

$$\begin{aligned}
\mathrm{TV}(\gamma) &= \bigcup_{i=1}^{\infty} \big(\gamma(s_i) \triangle \gamma(t_i)\big) \\
&= \bigcup_{i=1}^{\infty} \Big(\big(\underbrace{\gamma(s_i) \triangle A}_{\subseteq_\mu \gamma(t_i) \triangle A} \big) \triangle \big(\gamma(t_i) \triangle A\big) \Big) \\
&= \bigcup_{i=1}^{\infty} \Big(\big(\gamma(s_i) \triangle A\big) \triangle \big(\gamma(t_i) \triangle A\big) \Big) \\
&= \bigcup_{i=1}^{\infty} \Big(\big(\gamma(t_i) \triangle A\big) \setminus \big(\gamma(s_i) \triangle A\big) \Big) \\
&= \Big(\bigcup_{i=1}^{\infty} \big(\gamma(t_i) \triangle A\big) \Big) \setminus \Big(\bigcap_{i=1}^{\infty} \big(\gamma(s_i) \triangle A\big) \Big).
\end{aligned}$$

The validity of the last reformulation in this chain is best shown by mutual essential inclusion, which can be shown by showing true inclusion with an appropriately chosen set of class representatives. Let $S_i \in \gamma(s_i) \triangle A$ and $T_i \in \gamma(t_i) \triangle A$ for $i \in \mathbb{N}$ be chosen such that $S_{i+1} \subseteq S_i \subseteq T_i \subseteq T_{i+1}$ holds for all $i \in \mathbb{N}$. We have

$$\begin{aligned}
x \in \bigcup_{i=1}^{\infty} (T_i \setminus S_i) &\Longrightarrow (\exists i_0 \in \mathbb{N} \colon x \in T_{i_0} \wedge x \notin S_{i_0}) \\
&\Longrightarrow x \in \bigcup_{i=1}^{\infty} T_i \wedge x \notin \bigcap_{i=1}^{\infty} S_{i_0} \\
&\Longrightarrow x \in \Big(\bigcup_{i=1}^{\infty} T_i \Big) \setminus \Big(\bigcap_{i=1}^{\infty} S_{i_0} \Big).
\end{aligned}$$

This establishes a subset relationship

$$\bigcup_{i=1}^{\infty} (T_i \setminus S_i) \subseteq \Big(\bigcup_{i=1}^{\infty} T_i \Big) \setminus \Big(\bigcap_{i=1}^{\infty} S_{i_0} \Big)$$

that implies an essential subset relationship between the respective similarity classes in accordance with Lemma 2.2.16:

$$\bigcup_{i=1}^{\infty} \Big(\big(\gamma(t_i) \triangle A\big) \setminus \big(\gamma(s_i) \triangle A\big) \Big) \subseteq_\mu \Big(\bigcup_{i=1}^{\infty} \big(\gamma(t_i) \triangle A\big) \Big) \setminus \Big(\bigcap_{i=1}^{\infty} \big(\gamma(s_i) \triangle A\big) \Big).$$

Conversely, for all $x \in \big(\bigcup_{i=1}^{\infty} T_i\big) \setminus \big(\bigcap_{i=1}^{\infty} S_i\big)$, there exists $i_1 \in \mathbb{N}$ with $x \in T_{i_1}$ and $i_2 \in \mathbb{N}$ with $x \notin S_{i_2}$. Let $i_0 \coloneqq \max\{i_1,\ i_2\}$. Because of the monotonicity of the sequences $(s_i)_{i \in \mathbb{N}}$ and $(t_i)_{i \in \mathbb{N}}$, and because of the monotonicity of geodesics, we have

$$x \in \underbrace{T_{i_1}}_{\subseteq T_{i_0}} \setminus \underbrace{S_{i_2}}_{\supseteq S_{i_0}} \subseteq T_{i_0} \setminus S_{i_0} \subseteq \bigcup_{i=1}^{\infty} (T_i \setminus S_i).$$

We therefore have $\big(\bigcup_{i=1}^{\infty} T_i\big) \setminus \big(\bigcap_{i=1}^{\infty} S_i\big) \subseteq \bigcup_{i=1}^{\infty} (T_i \setminus S_i)$, which implies

$$\Big(\bigcup_{i=1}^{\infty} \big(\gamma(t_i) \triangle A\big) \Big) \setminus \Big(\bigcap_{i=1}^{\infty} \big(\gamma(s_i) \triangle A\big) \Big) \subseteq_\mu \bigcup_{i=1}^{\infty} \Big(\big(\gamma(t_i) \triangle A\big) \setminus \big(\gamma(s_i) \triangle A\big) \Big).$$

By combining both essential inclusions, we have equality of the similarity classes:

$$\bigcup_{i=1}^{\infty}\Big(\big(\gamma(t_i)\triangle A\big)\setminus\big(\gamma(s_i)\triangle A\big)\Big)=\Big(\bigcup_{i=1}^{\infty}\big(\gamma(t_i)\triangle A\big)\Big)\setminus\Big(\bigcap_{i=1}^{\infty}\big(\gamma(s_i)\triangle A\big)\Big).$$

We have thus shown that

$$\mathrm{TV}(\gamma)=\Big(\bigcup_{i=1}^{\infty}\big(\gamma(t_i)\triangle A\big)\Big)\setminus\underbrace{\Big(\bigcap_{i=1}^{\infty}\big(\gamma(s_i\triangle A\big)\Big)}_{=[\emptyset]_{\sim_\mu}}=\bigcup_{i=1}^{\infty}\big(\gamma(t_i)\triangle A\big).$$

We can make a similar reformulation of $\mathrm{TV}(\gamma)$ with respect to a destination point. Let $B\in\Sigma/_{\sim_\mu}$ be a destination point of γ. Then we have

$$\begin{aligned}
\mathrm{TV}(\gamma)&=\bigcup_{i=1}^{\infty}\Big(\big(\gamma(s_i)\triangle B\big)\triangle\underbrace{\big(\gamma(t_i)\triangle B\big)}_{\subseteq_\mu\gamma(s_i)\triangle B}\Big)\\
&=\bigcup_{i=1}^{\infty}\Big(\big(\gamma(s_i)\triangle B\big)\setminus\big(\gamma(t_i)\triangle B\big)\Big)\\
&=\Big(\bigcup_{i=1}^{\infty}\big(\gamma(s_i)\triangle B\big)\Big)\setminus\underbrace{\Big(\bigcap_{i=1}^{\infty}\big(\gamma(t_i)\triangle B\big)\Big)}_{=[\emptyset]_{\sim_\mu}}\\
&=\bigcup_{i=1}^{\infty}\big(\gamma(s_i)\triangle B\big).
\end{aligned}$$

Part 2 ($\Longleftarrow$). Let $A:=B\triangle\mathrm{TV}(\gamma)\in\Sigma/_{\sim_\mu}$ be an origin point of γ. Let $s,t\in I$ with $s\le t$. As always, if such s,t exist at all, then $I\neq\emptyset$. There then exists an increasing sequence $(t_i)_{i\in\mathbb{N}}\in I^{\mathbb{N}}$ with $t_i\ge t$ for all $i\in\mathbb{N}$ and $t_i\to\sup I$ for $i\to\infty$. According to Part 1, we have

$$\begin{aligned}
\gamma(t)\triangle B&=\gamma(t)\triangle A\triangle\mathrm{TV}(\gamma)\\
&=\underbrace{\gamma(t)\triangle A}_{\subseteq_\mu\gamma(t_i)\triangle A\ \forall i}\triangle\bigcup_{i=1}^{\infty}\big(\gamma(t_i)\triangle A\big)\\
&=\bigcup_{i=1}^{\infty}\Big(\big(\gamma(t_i)\triangle A\big)\setminus\big(\gamma(t)\triangle A\big)\Big)\\
&=\bigcup_{i=1}^{\infty}\Big(\big(\gamma(t_i)\triangle A\big)\triangle\big(\gamma(t)\triangle A\big)\Big)\\
&=\bigcup_{i=1}^{\infty}\underbrace{\big(\gamma(t_i)\triangle\gamma(t)\big)}_{\subseteq_\mu\gamma(t_i)\triangle\gamma(s)}\\
&\subseteq_\mu\bigcup_{i=1}^{\infty}\big(\gamma(t_i)\triangle\gamma(s)\big)\\
&=\bigcup_{i=1}^{\infty}\Big(\big(\gamma(t_i)\triangle A\big)\triangle\underbrace{\big(\gamma(s)\triangle A\big)}_{\subseteq_\mu\gamma(t_i)\triangle A\ \forall i}\Big)\\
&=\gamma(s)\triangle A\triangle\bigcup_{i=1}^{\infty}\big(\gamma(t_i)\triangle A\big)
\end{aligned}$$

$$= \gamma(s) \triangle B$$

which proves Equation (2.29). To prove Equation (2.30), let $(t_i)_{i\in\mathbb{N}} \in I^{\mathbb{N}}$ be a sequence with $t_i \to \sup I$ for $i \to \infty$. Because we consider an intersection over all $i \in \mathbb{N}$, we may assume without loss of generality that $(t_i)_{i\in\mathbb{N}}$ is increasing. According to Part 1, we have

$$\mathrm{TV}(\gamma) = \bigcup_{i=1}^{\infty} \big(\gamma(t_i) \triangle A\big).$$

We have

$$\begin{aligned}
\bigcap_{i=1}^{\infty} \big(\gamma(t_i) \triangle B\big) &= \bigcap_{i=1}^{\infty} \big(\gamma(t_i) \triangle A \triangle \mathrm{TV}(\gamma)\big) \\
&= \bigcap_{i=1}^{\infty} \Big(\underbrace{\gamma(t_i) \triangle A}_{\subseteq_\mu \bigcup_{j=1}^{\infty} (\gamma(t_j) \triangle A)} \triangle \bigcup_{j=1}^{\infty} \big(\gamma(t_j) \triangle A\big)\Big) \\
&= \bigcap_{i=1}^{\infty} \Big(\Big(\bigcup_{j=1}^{\infty} \big(\gamma(t_j) \triangle A\big)\Big) \setminus \big(\gamma(t_i) \triangle A\big)\Big) \\
&= \bigcap_{i=1}^{\infty} \Big(\Big(\bigcup_{j=1}^{\infty} \big(\gamma(t_j) \triangle A\big)\Big) \cap \big(\gamma(t_i) \triangle A\big)^{\complement}\Big) \\
&= \bigcup_{j=1}^{\infty} \big(\gamma(t_j) \triangle A\big) \cap \bigcap_{i=1}^{\infty} \big(\gamma(t_i) \triangle A\big)^{\complement} \\
&= \bigcup_{j=1}^{\infty} \big(\gamma(t_j) \triangle A\big) \cap \Big(\bigcup_{i=1}^{\infty} \big(\gamma(t_i) \triangle A\big)\Big)^{\complement} \\
&= [\emptyset]_{\sim_\mu}
\end{aligned}$$

which proves Equation (2.30).

PART 3 ($\Longrightarrow$). Once more, let $A := B \triangle \mathrm{TV}(\gamma)$. The arguments for this are very similar to those given in Part 2 and we abbreviate them here. For $s, t \in I$ with $s \le t$, we can select $(s_i)_{i\in\mathbb{N}}$ decreasing with $s_i \to \inf I$ and $s_i \le s$ for $i \in \mathbb{N}$. The premise is that B is a destination point. Therefore, $s_i \le s$ implies $\gamma(s) \triangle B \subseteq_\mu \gamma(s_i) \triangle B$ and we have

$$\begin{aligned}
\gamma(s) \triangle A &= \gamma(s) \triangle B \triangle \bigcup_{i=1}^{\infty} \big(\gamma(s_i) \triangle B\big) \\
&\overset{(2.29)}{=} \bigcup_{i=1}^{\infty} \Big(\big(\gamma(s_i) \triangle B\big) \setminus \big(\gamma(s) \triangle B\big)\Big) \\
&= \bigcup_{i=1}^{\infty} \Big(\big(\gamma(s_i) \triangle B\big) \triangle \big(\gamma(s) \triangle B\big)\Big) \\
&= \bigcup_{i=1}^{\infty} \big(\gamma(s_i) \triangle \gamma(s)\big) \\
&\subseteq_\mu \bigcup_{i=1}^{\infty} \big(\gamma(s_i) \triangle \gamma(t)\big) \\
&= \gamma(t) \triangle A
\end{aligned}$$

which proves Equation (2.27). To prove Equation (2.28), we may assume without loss of generality that $(s_i)_{i\in\mathbb{N}}$ with $s_i \to \inf I$ is decreasing. We then find that

$$\begin{aligned}
\bigcap_{i=1}^{\infty}\big(\gamma(s_i)\triangle A\big) &= \bigcap_{i=1}^{\infty}\Big(\underbrace{\gamma(s_i)\triangle B}_{\subseteq_\mu \bigcup_{j=1}^{\infty}(\gamma(s_j)\triangle B)}\triangle \bigcup_{j=1}^{\infty}\big(\gamma(s_j)\triangle B\big)\Big)\\
&= \bigcap_{i=1}^{\infty}\Big(\Big(\bigcup_{j=1}^{\infty}\big(\gamma(s_j)\triangle B\big)\Big)\cap\big(\gamma(s_i)\triangle B\big)^{\complement}\Big)\\
&= \bigcup_{j=1}^{\infty}\big(\gamma(s_j)\triangle B\big)\cap\Big(\bigcap_{i=1}^{\infty}\big(\gamma(s_i)\triangle B\big)\Big)^{\complement}\\
&= [\emptyset]_{\sim_\mu}.
\end{aligned}$$

Part 4 (Existence And Uniqueness). Proposition 2.3.11 shows that every geodesic has an origin point. For every origin point A of γ, $A\triangle \mathrm{TV}(\gamma)$ is a destination point of γ. Therefore, every geodesic has a destination point.

If B and B' are destination points of γ, then $B\triangle\mathrm{TV}(\gamma)$ and $B'\triangle\mathrm{TV}(\gamma)$ are origin points of γ. According to Proposition 2.3.11, the origin point of γ is unique if $I\neq\emptyset$. Therefore, $I\neq\emptyset$ implies

$$B\triangle\mathrm{TV}(\gamma) = B'\triangle\mathrm{TV}(\gamma).$$

By forming the symmetric difference with $\mathrm{TV}(\gamma)$, we obtain $B = B'$, proving that the destination point is unique if $I\neq\emptyset$. □

An interesting corollary of Propositions 2.3.11 and 2.3.12 is that a geodesic is canonical if and only if its origin point is $[\emptyset]_{\sim_\mu}$ and that if this is the case, then its destination point is equal to its total variation.

Corollary 2.3.13 (Strong Characterization of Canonical Geodesics). *Let (X, Σ, μ) be a measure space, let $I\subseteq\mathbb{R}$ be an interval, and let $\gamma\colon I\to \Sigma/_{\sim_\mu}$ be a geodesic. The following statements are equivalent:*

(1) γ is canonical,

(2) $[\emptyset]_{\sim_\mu}$ is an origin point of γ,

(3) $\mathrm{TV}(\gamma)$ is a destination point of γ. ◁

Proof. Part 1 ((1) ⟺ (2)). According to Proposition 2.3.11, $\gamma = \gamma\triangle[\emptyset]_{\sim_\mu}$ is canonical if and only if $[\emptyset]_{\sim_\mu}$ is an origin point of γ.

Part 2 ((2) ⟺ (3)). According to Proposition 2.3.12, $[\emptyset]_{\sim_\mu}$ is an origin point of γ if and only if $\mathrm{TV}(\gamma) = [\emptyset]_{\sim_\mu}\triangle\mathrm{TV}(\gamma)$ is a destination point of γ. □

We stress once more that Propositions 2.3.11 and 2.3.12 do not assume that the parameter interval I is bounded in either direction. Limit points remain well-defined for unbounded parameter intervals. In a technical sense, it could therefore be said that there exist geodesics that "connect" points that are infinitely far apart.

The existence and properties of limit points guarantee that we can continuously extend any geodesic to the boundaries of its parameter interval. Although we can technically even do this if the bounds are $-\infty$ or ∞, extending geodesics

to infinities leads to some theoretical complications and edge cases. We therefore restrict ourselves to the closure of the parameter interval within the real numbers.

Theorem 2.3.14 (Continuous Extension of Geodesics).
Let (X, Σ, μ) be a measure space, let $I \subseteq \mathbb{R}$ be an interval, let $\gamma\colon I \to \Sigma/_{\sim_\mu}$, let $A_\gamma \in \Sigma/_{\sim_\mu}$ be an origin point of γ, let $B_\gamma \in \Sigma/_{\sim_\mu}$ be a destination point of γ, and let $\overline{I} \subseteq \mathbb{R}$ be the closure of I within the real numbers. Then the map $\overline{\gamma}\colon \overline{I} \to \Sigma/_{\sim_\mu}$ with

$$\overline{\gamma}(t) \coloneqq \begin{cases} A_\gamma & \text{if } t = \inf I \\ B_\gamma & \text{if } t = \sup I \\ \gamma(t) & \text{if } \inf I < t < \sup I \end{cases} \qquad \forall t \in \overline{I}$$

is the unique geodesic on $\overline{I}$ that satisfies $\overline{\gamma}(t) = \gamma(t)$ for all $t \in I$. If $C \geq 0$ is a geodesic constant of γ, then C is also a geodesic constant of $\overline{\gamma}$. ◁

PROOF. PART 1 ($\overline{\gamma}(t) = \gamma(t)$ $\forall t \in I$). For $\inf I < t < \sup I$, this is true by definition. If $t \in I$ is such that $t = \inf I$ or $t = \sup I$. Then we can use the monotonic sequence $(t_i)_{i\in\mathbb{N}} \in I^{\mathbb{N}}$ with $t_i \coloneqq t$ for all $i \in \mathbb{N}$ to approximate t. Definition 2.3.8 then states that the symmetric difference between $\gamma(t_i)$ and the origin or destination point contracts to $[\emptyset]_{\sim_\mu}$ as $t_i \to t$, i.e.,

$$\bigcap_{i=1}^{\infty} \big(\gamma(t_i) \triangle A_\gamma\big) = [\emptyset]_{\sim_\mu} \quad \text{if } t = \inf I,$$
$$\bigcap_{i=1}^{\infty} \big(\gamma(t_i) \triangle B_\gamma\big) = [\emptyset]_{\sim_\mu} \quad \text{if } t = \sup I.$$

In either case, all elements in intersection on the left hand side are equal to $\gamma(t) \triangle A_\gamma$ or $\gamma(t) \triangle B_\gamma$, respectively, which means that

$$\gamma(t_0) \triangle A_\gamma = [\emptyset]_{\sim_\mu} \quad \text{if } t = \inf I,$$
$$\gamma(t_0) \triangle B_\gamma = [\emptyset]_{\sim_\mu} \quad \text{if } t = \sup I.$$

This proves that we have $\overline{\gamma}(t) = \gamma(t)$, even if $t \in I$ simultaneously satisfies $t_0 = \inf I$ or $t_0 = \sup I$.

PART 2 ($\overline{\gamma}$ IS A GEODESIC). Let $C \geq 0$ be a geodesic constant of γ. For $s, t \in I$, we know that $\overline{\gamma}(s) = \gamma(s)$ and $\overline{\gamma}(t) = \gamma(t)$. Therefore, we have

$$\mu\big(\overline{\gamma}(s) \triangle \overline{\gamma}(t)\big) = \mu\big(\gamma(s) \triangle \gamma(t)\big) = C \cdot |s - t|.$$

For cases where either s and t are in $\overline{I} \setminus I$, we have to make an approximation argument. We first note that, because the closure $\overline{I}$ explicitly does not contain infinities, this case can only occur if $I \neq \emptyset$.

We assume without loss of generality that $s \leq t$. Because of this, we can find sequences $(s_i)_{i\in\mathbb{N}} \in I^{\mathbb{N}}$ and $(t_i)_{i\in\mathbb{N}} \in I^{\mathbb{N}}$ such that $s \leq s_i \leq t_i \leq t$ for all $i \in \mathbb{N}$, $(s_i)_{i\in\mathbb{N}}$ is decreasing with $s_i \to s$ for $i \to \infty$, and $(t_i)_{i\in\mathbb{N}}$ is increasing with $t_i \to t$. We now find that

$$\begin{aligned} \overline{\gamma}(s) \triangle \overline{\gamma}(t) &= \big(\overline{\gamma}(s) \triangle \overline{\gamma}(s_i)\big) \triangle \big(\overline{\gamma}(s_i) \triangle \overline{\gamma}(t_i)\big) \triangle \big(\overline{\gamma}(t_i) \triangle \overline{\gamma}(t)\big) \\ &= \big(\overline{\gamma}(s) \triangle \gamma(s_i)\big) \triangle \big(\gamma(s_i) \triangle \gamma(t_i)\big) \triangle \big(\gamma(t_i) \triangle \overline{\gamma}(t)\big). \end{aligned}$$

Now, we have to make a minor case distinction. If $s \in I$, then $\overline{\gamma}(s) = \gamma(s)$ and therefore $\overline{\gamma}(s) \triangle \gamma(s_i) = \gamma(s) \triangle \gamma(s_i)$. In this case, we have $\overline{\gamma}(s) \triangle \gamma(s_i) \subseteq_\mu \overline{\gamma}(s) \triangle \gamma(t_i)$ because of the monotonicity of measure space geodesics. If $s \notin I$, then $s = \inf I$ and $\overline{\gamma}(s) = A_\gamma$ is an origin point of γ. In this case Equation (2.27) guarantees that $\overline{\gamma}(s) \triangle \gamma(s_i) \subseteq_\mu \overline{\gamma}(s) \triangle \gamma(t_i)$. By an analogous argument using Equation (2.29), we show that $\gamma(t_i) \triangle \overline{\gamma}(t) \subseteq_\mu \gamma(s_i) \triangle \overline{\gamma}(t)$. Because the symmetric difference with a subset is a set difference, we find that

$$\begin{aligned}\gamma(s_i) \triangle \gamma(t_i) &= \big(\overline{\gamma}(s) \triangle \gamma(t_i)\big) \setminus \big(\overline{\gamma}(s) \triangle \gamma(s_i)\big),\\ \gamma(s_i) \triangle \gamma(t_i) &= \big(\gamma(s_i) \triangle \overline{\gamma}(t)\big) \setminus \big(\gamma(t_i) \triangle \overline{\gamma}(t)\big),\end{aligned}$$

and that, therefore, $\gamma(s_i) \triangle \gamma(t_i)$ is essentially disjoint from both $\overline{\gamma}(s) \triangle \gamma(s_i)$ and $\gamma(t_i) \triangle \overline{\gamma}(t)$. We can then rewrite the measure of $\overline{\gamma}(s) \triangle \overline{\gamma}(t)$ as a disjoint union:

$$\begin{aligned}\overline{\gamma}(s) \triangle \overline{\gamma}(t) &= \big(\overline{\gamma}(s) \triangle \gamma(s_i)\big) \triangle \big(\gamma(s_i) \triangle \gamma(t_i)\big) \triangle \big(\gamma(t_i) \triangle \overline{\gamma}(t)\big)\\ &= \big(\gamma(s_i) \triangle \gamma(t_i)\big) \cup \Big(\big(\overline{\gamma}(s) \triangle \gamma(s_i)\big) \triangle \big(\gamma(t_i) \triangle \overline{\gamma}(t)\big)\Big)\\ &\subseteq_\mu \big(\gamma(s_i) \triangle \gamma(t_i)\big) \cup \big(\overline{\gamma}(s) \triangle \gamma(s_i)\big) \cup \big(\gamma(t_i) \triangle \overline{\gamma}(t)\big).\end{aligned}$$

Because of the continuity of the absolute value, it is evident that

$$\mu\big(\overline{\gamma}(s) \triangle \overline{\gamma}(t)\big) \geq \mu\big(\gamma(s_i) \triangle \gamma(t_i)\big) = C \cdot |s_i - t_i| \xrightarrow{i\to\infty} C \cdot |s - t|.$$

and therefore

$$\mu\big(\overline{\gamma}(s) \triangle \overline{\gamma}(t)\big) \geq C \cdot |s - t|.$$

Further, for each $i \in \mathbb{N}$, we have

$$\mu\big(\overline{\gamma}(s) \triangle \overline{\gamma}(t)\big) \leq \mu\big(\gamma(s_i) \triangle \gamma(t_i)\big) + \mu\big(\overline{\gamma}(s) \triangle \gamma(s_i)\big) + \mu\big(\gamma(t_i) \triangle \overline{\gamma}(t)\big),$$

where the latter two summands converge to zero because of either the geodesic property or Equations (2.28) and (2.30), depending on whether or not $s \in I$ or $t \in I$. In either case, we obtain

$$\mu\big(\overline{\gamma}(s) \triangle \overline{\gamma}(t)\big) \leq \lim_{i\to\infty} \mu\big(\gamma(s_i) \triangle \gamma(t_i)\big) = \lim_{i\to\infty} C \cdot |s_i - t_i| = C \cdot |s - t|.$$

This shows that

$$\mu\big(\overline{\gamma}(s) \triangle \overline{\gamma}(t)\big) = C \cdot |s - t| \quad \forall s, t \in \overline{I},$$

i.e., that $\overline{\gamma}$ is a geodesic. We note that, if $s = -\infty$ or $t = \infty$, then the last step in this line of reasoning breaks down because the measures of $\gamma(s_i) \triangle \overline{\gamma}(s)$ and $\gamma(t_i) \triangle \overline{\gamma}(t)$, respectively, do not converge to zero for $i \to \infty$.

Part 3 (Uniqueness). Let $\gamma' \colon \overline{I} \to \Sigma/_{\sim_\mu}$ be a geodesic with $\gamma'(t) = \gamma(t)$ for all $t \in I$. Let $C \geq 0$ be a geodesic constant of γ and let $C' \geq 0$ be a geodesic constant of γ'. Because $\overline{I}$ is the closure of I in the real numbers, for all $t \in \overline{I}$, we can find a sequence $(t_i)_{i\in\mathbb{N}} \in I^{\mathbb{N}}$ such that $t_i \xrightarrow{i\to\infty} t$. We then find that for all $i \in \mathbb{N}$,

$$\begin{aligned}\mu(\gamma'(t) \triangle \overline{\gamma}(t)) &\leq \mu(\gamma'(t) \triangle \gamma'(t_i)) + \mu(\gamma'(t_i) \triangle \overline{\gamma}(t))\\ &= C' \cdot |t - t_i| + \mu(\overline{\gamma}(t_i) \triangle \overline{\gamma}(t)\\ &= (C + C') \cdot |t - t_i|\\ &\xrightarrow{i\to\infty} 0.\end{aligned}$$

This implies $\gamma'(t) = \overline{\gamma}(t)$. Therefore, $\overline{\gamma}$ is the only geodesic defined on $\overline{I}$ for which $\overline{\gamma}(t) = \gamma(t)$ for all $t \in I$. □

This greatly simplifies the proof of some theorems because we no longer have to account for all constellations of intervals. If a parameter interval's infimum or supremum are finite, then we can simply assume them to be included in the parameter interval. Conversely, we may assume that if a parameter interval is open in one direction, then it is also unbounded in that same direction.

2.3.2 Geodesic Interpolation

Simply enumerating the properties of geodesics is of limited use if we do not have any concrete geodesics to work with. In this section, we lay the foundation of a set of methods that can be used to construct geodesics. The most straightforward way to construct a geodesic is to specify its output explicitly and verify the geodesic property by hand. This is prohibitively complex in many cases.

As we have established in the previous section, geodesics are continuous and monotonic, and allow for the determination of limits from either side by countable union of the variation. We can therefore use limits to "fill in" more sparsely specified geodesics.

We proceed in two steps. First, we interpolate a geodesic that is "densely specified" in the sense that its values are explicitly given on a dense subset of the parameter interval. This case is relatively straightforward. In the second step, we make use of atomlessness to fill in gaps of non-zero width in a "sparsely specified" geodesic. This second interpolation result makes use of the first by first "densifying" the set of support points and then invoking the dense interpolation result derived in the first step.

2.3.2.1 GEODESIC SUPPORT TUPLES

Before we begin, we define the format in which support points are given to our interpolation theorems. We call this format a *geodesic support tuple*.

Definition 2.3.15 (Geodesic Support Tuple).
Let $(X,\ \Sigma,\ \mu)$ be a measure space, let $I \subseteq \mathbb{R}$ be an interval, let $T \subseteq I$ be a finite or countably infinite set, and let $B\colon T \to \Sigma/_{\sim_\mu}$ be such that there exists a constant $C \geq 0$ with

$$\operatorname{conv}(T) \subseteq I \subseteq \overline{\operatorname{conv}(T)}; \tag{2.34}$$

$$\mu\big(B(s) \mathbin{\triangle} B(t)\big) = C \cdot |s - t| \quad \forall s, t \in T. \tag{2.35}$$

Then we refer to $(I,\ T,\ B)$ as a *geodesic support tuple* in $(X,\ \Sigma,\ \mu)$. We refer to the elements of T as the *support points* of the tuple and to the elements of $\big\{B(t)\,\big|\,t \in T\big\}$ as the *support values* of the tuple. We call a geodesic support tuple *dense* if T is dense in I. ◁

In principle, a support tuple is quite simple: we have a countable set of support points, each of which is associated with a corresponding similarity class. As was the case with geodesics themselves, Equation (2.35) implies a kind of monotonicity.

Lemma 2.3.16 (Monotonicity of Geodesic Support Tuples).
Let $(I,\ T,\ B)$ be a geodesic support tuple in a measure space $(X,\ \Sigma,\ \mu)$, and let $r, s, t \in T$ with $r \leq s \leq t$. Then

$$\big(B(r) \mathbin{\triangle} B(s)\big) \cap \big(B(s) \mathbin{\triangle} B(t)\big) = [\emptyset]_{\sim_\mu}. \qquad \triangleleft$$

PROOF. We prove this by contradiction. According to Definition 2.3.15, there exists a constant $C \geq 0$ such that $\mu\big(B(s) \triangle B(t)\big) = C \cdot |s-t|$ for all $s,t \in T$. If we were to assume that

$$\mu\Big(\big(B(r) \triangle B(s)\big) \cap \big(B(s) \triangle B(t)\big)\Big) > 0,$$

then we would have

$$\begin{aligned}
\mu\big(B(r) \triangle B(t)\big) &= \mu\Big(\big(B(r) \triangle B(s)\big) \triangle \big(B(s) \triangle B(t)\big)\Big) \\
&= \mu\Big(\big(B(r) \triangle B(s)\big) \cup \big(B(s) \triangle B(t)\big)\Big) - \mu\Big(\big(B(r) \triangle B(s)\big) \cap \big(B(s) \triangle B(t)\big)\Big) \\
&< \mu\Big(\big(B(r) \triangle B(s)\big) \cup \big(B(s) \triangle B(t)\big)\Big) \\
&\leq \mu\big(B(r) \triangle B(s)\big) + \mu\big(B(s) \triangle B(t)\big) \\
&= C \cdot |r-s| + C \cdot |s-t| \\
&= C \cdot (s-r+t-s) \\
&= C \cdot (t-r) \\
&= C \cdot |r-t|.
\end{aligned}$$

However, this would contradict Equation (2.35). Thus, our initial assumption that $B(r) \triangle B(s)$ and $B(s) \triangle B(t)$ are not essentially disjoint must be false. □

We note that this lemma implies that the symmetric difference

$$\big(B(r) \triangle B(s)\big) \triangle \big(B(s) \triangle B(t)\big)$$

is an essentially disjoint union, which means that

$$\begin{aligned}
B(r) \triangle B(s) &\subseteq_\mu B(r) \triangle B(t), \\
B(s) \triangle B(t) &\subseteq_\mu B(r) \triangle B(t).
\end{aligned}$$

It is to be expected that geodesic support tuples share this property of the geodesics that they approximate. For us, the monotonicity of geodesic support tuples provides a convenient way to avoid superfluous case distinctions. By approximating parameters in I with monotonic sequences and considering symmetric differences with respect to the first support value, we will see that the approximating sequence of similarity classes is always essentially increasing, irrespective of the monotonicity of B or the direction from which we approximate.

2.3.2.2 DENSE INTERPOLATION

If a support tuple $(I,\ T,\ B)$ is dense, then the process of finding the value associated with a parameter $t \in I \setminus T$ is straightforward. We just have to find the limit of the support values associated with the approximating sequence. The main challenge lies in proving that the result of this process does not depend on the choice of approximating sequence and that the geodesic property holds. For this, we first prove that the support tuple can be extended to any accumulation point of the support points in a well-defined manner.

Lemma 2.3.17 (Interpolation at Accumulation Points).
Let $(I,\ T,\ B)$ be a geodesic support tuple in a measure space $(X,\ \Sigma,\ \mu)$, and let $t^* \in I$ be an accumulation point of T. Then there exists a unique similarity class $B^* \in \Sigma/_{\sim_\mu}$ such that

$$B(t_i) \xrightarrow{i \to \infty} B^*$$

holds for every sequence $(t_i)_{i\in\mathbb{N}} \in T^{\mathbb{N}}$ *with* $t_i \to t^*$ *for* $i \to \infty$. *The tuple* $(I,\ T',\ B')$ *with* $T' \coloneqq T \cup \{t^*\}$ *and* $B' \colon T' \to \Sigma/_{\sim_\mu}$ *with*

$$B'(t) \coloneqq \begin{cases} B(t) & \text{if } t \in T, \\ B^* & \text{if } t = t^* \end{cases}$$

is also a geodesic support tuple. If $C \geq 0$ *is a constant for which* $(I,\ T,\ B)$ *satisfies Equation* (2.35)*, then* $(I,\ T',\ B')$ *satisfies Equation* (2.35) *with the same constant.* $\lhd$

PROOF. PART 1 (FINDING B^*). Let $t^* \in I$ be an accumulation point of T. By definition, there exists a sequence $(t_i)_{i\in\mathbb{N}} \in T^{\mathbb{N}}$ such that $t_i \to t^*$ for $i \to \infty$. We can now partition $\mathbb{N}$ into

$$N_{\leq} \coloneqq \{i \in \mathbb{N} \mid t_i \leq t^*\},$$
$$N_{>} \coloneqq \{i \in \mathbb{N} \mid t_i > t^*\}.$$

This is evidently a partition of $\mathbb{N}$. Because the union of $N_{\leq}$ and $N_{>}$ is infinite, at least one of the two parts is infinite. By restricting $(t_i)_{i\in\mathbb{N}}$ to the subsequence associated with that part, we may assume without loss of generality that either $t_i \leq t^*\ \forall i \in \mathbb{N}$ or $t_i > t^*\ \forall i \in \mathbb{N}$. Selecting a subsequence does not change the fact that $t_i \to t^*$ for $i \to \infty$.

If $t_i \leq t^*\ \forall i \in \mathbb{N}$, then $t_i \to t^*$ implies that we can select a monotonically increasing subsequence of $(t_i)_{i\in\mathbb{N}}$. If $t_i > t^*\ \forall i \in \mathbb{N}$, then we can select a monotonically decreasing subsequence. Without loss of generality, let $(t_i)_{i\in\mathbb{N}}$ be monotonic. Again, this does not affect convergence.

We now have a monotonic approximating sequence for t^* in T. Because the sequence is monotonic, Lemma 2.3.16 implies that the sequence $(D_i)_{i\in\mathbb{N}} \in (\Sigma/_{\sim_\mu})^{\mathbb{N}}$ with

$$D_i \coloneqq B(t_i) \triangle B(t_1) \quad \forall i \in \mathbb{N}$$

is monotonically increasing. Let

$$D^* \coloneqq \bigcup_{i=1}^{\infty} D_i,$$
$$B^* \coloneqq B(t_1) \triangle D^*.$$

For the subsequent parts of this proof, let $(t_i)_{i\in\mathbb{N}}$ denote the specific monotonic approximating sequence of t^* in T that was used to construct B^*.

PART 2 (EXTENDED SUPPORT TUPLE). In this part of the proof, we show that $(I,\ T',\ B')$ is a geodesic support tuple. One of the premises of this lemma is that $t^* \in I$. Therefore, adding t^* to T does not change the fact that $T \subseteq I$. In other words, we have $T' \subseteq I$. Because I is an interval, I is convex and therefore, we have

$$\operatorname{conv}(T') \subseteq I.$$

Because $T \subseteq T'$, we have $\operatorname{conv}(T) \subseteq \operatorname{conv}(T')$ and

$$I \subseteq \overline{\operatorname{conv}(T)} \subseteq \overline{\operatorname{conv}(T')}.$$

Together, these subset relationships prove that $(I,\ T',\ B')$ satisfies Equation (2.34). We therefore only need to prove Equation (2.35). More precisely, we need to prove that

$$\mu\big(B^* \triangle B(s)\big) = C \cdot |t^* - s| \quad \forall s \in T$$

where $C \geq 0$ is a constant for which $(I,\ T',\ B')$ satisfies Equation (2.35).

First, we note that, because $(D_i)_{i\in\mathbb{N}}$ is an essentially increasing sequence of similarity classes, we have

$$\begin{aligned}
\mu\big(B^* \triangle B(t_1)\big) &= \mu\Big(\bigcup_{i=1}^{\infty} D_i\Big) \\
&= \lim_{i\to\infty} \mu(D_i) \\
&= \lim_{i\to\infty} \mu\big(B(t_i) \triangle B(t_1)\big) \\
&= \lim_{i\to\infty} C \cdot |t_i - t_1| \\
&= C \cdot |t^* - t_1|.
\end{aligned}$$

Let $s \in T$. We have $t_i \in T$ for all $i \in \mathbb{N}$. This implies that

$$\mu\big(B(t_1) \triangle B(s)\big) = C \cdot |t_1 - s|.$$

We now make a case distinction based on the order relationship between s, t_1, and t^*.

Case 1 ($s = t^*$). In this case, t_i lies between s and t_1 for all $i \in \mathbb{N}$. Lemma 2.3.16 then demonstrates that we have

$$D_i = B(t_i) \triangle B(t_1) \subseteq_\mu B(s) \triangle B(t_1) \quad \forall i \in \mathbb{N}.$$

This implies $D^* \subseteq_\mu B(s) \triangle B(t_1)$ and therefore

$$\begin{aligned}
\mu\big(B^* \triangle B(s)\big) &= \mu\Big(\big(B^* \triangle B(t_1)\big) \triangle \big(B(s) \triangle B(t_1)\big)\Big) \\
&= \mu\Big(\big(B^* \triangle B(t_1)\big) \setminus \big(B(s) \triangle B(t_1)\big)\Big) \\
&= \mu\big(B^* \triangle B(t_1)\big) - \mu\big(B(s) \triangle B(t_1)\big) \\
&= C \cdot |t^* - t_1| - C \cdot |s - t_1| \\
&= 0 \\
&= C \cdot |t^* - s|. \qquad \triangleleft
\end{aligned}$$

Case 2 ($t_1 \leq s < t^*$). In this case, because $t_i \to t^* > s$, there exists $i_0 \in \mathbb{N}$ with $t_{i_0} \geq s$. Lemma 2.3.16 shows that

$$B(s) \triangle B(t_1) \subseteq_\mu B(t_{i_0}) \triangle B(t_1) = D_{i_0} \subseteq_\mu D^* = B^* \triangle B(t_1),$$

which implies that

$$\begin{aligned}
\mu\big(B^* \triangle B(s)\big) &= \mu\Big(\big(B^* \triangle B(t_1)\big) \triangle \big(B(s) \triangle B(t_1)\big)\Big) \\
&= \mu\Big(\big(B^* \triangle B(t_1)\big) \setminus \big(B(s) \triangle B(t_1)\big)\Big) \\
&= \mu\big(B^* \triangle B(t_1)\big) - \mu\big(B(s) \triangle B(t_1)\big) \\
&= C \cdot |t^* - t_1| - C \cdot |s - t_1| \\
&= C \cdot (t^* - t_1) - C \cdot (s - t_1) \\
&= C \cdot (t^* - s) \\
&= C \cdot |t^* - s|. \qquad \triangleleft
\end{aligned}$$

Case 3 ($s < t_1 \leq t^*$). Because $s < t_1 \leq t_i$ for all $i \in \mathbb{N}$. Lemma 2.3.16 states that $D_i = B(t_i) \triangle B(t_1)$ and $B(s) \triangle B(t_1)$ are essentially disjoint for all $i \in \mathbb{N}$. This implies that

$$\big(B^* \triangle B(t_1)\big) \cap \big(B(s) \triangle B(t_1)\big) = [\emptyset]_{\sim_\mu}.$$

We therefore have

$$\begin{aligned}
\mu\big(B^* \triangle B(s)\big) &= \mu\Big(\big(B^* \triangle B(t_1)\big) \triangle \big(B(s) \triangle B(t_1)\big)\Big) \\
&= \mu\Big(\big(B^* \triangle B(t_1)\big) \cup \big(B(s) \triangle B(t_1)\big)\Big) \\
&= \mu\big(B^* \triangle B(t_1)\big) + \mu\big(B(s) \triangle B(t_1)\big) \\
&= C \cdot |t^* - t_1| + C \cdot |s - t_1| \\
&= C \cdot (t^* - t_1) + C \cdot (t_1 - s) \\
&= C \cdot (t^* - s) \\
&= C \cdot |t^* - s|. \qquad \lhd
\end{aligned}$$

Case 4 ($s < t^* < t_1$). For all $i \in \mathbb{N}$, we have $s < t^* \leq t_i \leq t_1$. According to Lemma 2.3.16, we have

$$D_i = B(t_i) \triangle B(t_1) \subseteq_\mu B(s) \triangle B(t_1) \quad \forall i \in \mathbb{N}$$

and therefore $D^* = B^* \triangle B(t_1) \subseteq_\mu B(s) \triangle B(t_1)$. Therefore, we obtain

$$\begin{aligned}
\mu\big(B^* \triangle B(s)\big) &= \mu\Big(\big(B^* \triangle B(t_1)\big) \triangle \big(B(s) \triangle B(t_1)\big)\Big) \\
&= \mu\Big(\big(B(s) \triangle B(t_1)\big) \setminus \big(B^* \triangle B(t_1)\big)\Big) \\
&= \mu\big(B(s) \triangle B(t_1)\big) - \mu\big(B^* \triangle B(t_1)\big) \\
&= C \cdot |s - t_1| - C \cdot |t^* - t_1| \\
&= C \cdot (t_1 - s) - C \cdot (t_1 - t^*) \\
&= C \cdot (t^* - s) \\
&= C \cdot |t^* - s|. \qquad \lhd
\end{aligned}$$

Case 5 ($t^* < s \leq t_1$). This is analogous to $t_1 \leq s < t^*$. We have $B(s) \triangle B(t_1) \subseteq_\mu D_\infty$ and therefore

$$\begin{aligned}
\mu\big(B^* \triangle B(s)\big) &= \mu\big(B^* \triangle B(t_1)\big) - \mu\big(B(s) \triangle B(t_1)\big) \\
&= C \cdot |t^* - t_1| - C \cdot |s - t_1| \\
&= C \cdot (t_1 - t^*) - C \cdot (t_1 - s) \\
&= C \cdot (s - t^*) \\
&= C \cdot |t^* - s|. \qquad \lhd
\end{aligned}$$

Case 6 ($t^* \leq t_1 < s$). This case is symmetric to $s < t_1 \leq t^*$. We have

$$\big(B^* \triangle B(t_1)\big) \cap \big(B(s) \triangle B(t_1)\big) = [\emptyset]_{\sim_\mu}.$$

It then follows that

$$\begin{aligned}\mu\big(B^* \triangle B(s)\big) &= \mu\big(B^* \triangle B(t_1)\big) + \mu\big(B(s) \triangle B(t_1)\big)\\ &= C \cdot |t^* - t_1| + C \cdot |s - t_1|\\ &= C \cdot (t_1 - t^*) + C \cdot (s - t_1)\\ &= C \cdot (s - t^*)\\ &= C \cdot |t^* - s|.\end{aligned}$$ ◁

Case 7 ($t_1 < t^* < s$). This case is symmetric to $s < t^* < t_1$. We have

$$B^* \triangle B(t_1) \subseteq_\mu B(s) \triangle B(t_1).$$

We then obtain the equality

$$\begin{aligned}\mu\big(B^* \triangle B(s)\big) &= \mu\big(B(s) \triangle B(t_1)\big) - \mu\big(B^* \triangle B(t_1)\big)\\ &= C \cdot |s - t_1| - C \cdot |t^* - t_1|\\ &= C \cdot (s - t_1) - C \cdot (t^* - t_1)\\ &= C \cdot (s - t^*)\\ &= C \cdot |t^* - s|.\end{aligned}$$ ◁

With this, we have proven that

$$\mu\big(B^* \triangle B(s)\big) = C \cdot |t^* - s| \quad \forall s \in T.$$

This specifically implies that $B^* = B(t^*)$ if $t^* \in T$, thereby guaranteeing that B' is well defined. In conjunction with the evident facts that

$$\mu\big(B'(t^*) \triangle B'(t^*)\big) = \mu(B^* \triangle B^*) = 0 = C \cdot |t^* - t^*|$$

and that

$$\mu\big(B(s) \triangle B(t)\big) = C \cdot |s - t| \quad \forall s, t \in T,$$

this shows that $(I,\ T',\ B')$ satisfies Equation (2.35) with constant C. Together with Equation (2.34), which we had proven earlier, this proves that $(I,\ T',\ B')$ is a geodesic support tuple.

Part 3 (Well-Definedness and Uniqueness). Let $(s_i)_{i\in\mathbb{N}} \in T^{\mathbb{N}}$ be a sequence with $s_i \to t^*$ for $i \to \infty$. We have

$$\mu\big(B^* \triangle B(s_i)\big) = C \cdot |t^* - s_i| \quad \forall i \in \mathbb{N}.$$

Because $s_i \to t^*$ for $i \to \infty$, this means that

$$\mu\big(B^* \triangle B(s_i)\big) \xrightarrow{i\to\infty} 0,$$

which means that $B(s_i) \to B^*$ for $i \to \infty$. This shows that B^* does not depend on the choice of the approximating sequence $(t_i)_{i\in\mathbb{N}}$.

Uniqueness of B^* is quite trivial to prove. Let $B' \in \Sigma/_{\sim_\mu}$ and $(s_i)_{i\in\mathbb{N}} \in T^{\mathbb{N}}$ be such that

$$s_i \xrightarrow{i\to\infty} t^*, B(s_i) \xrightarrow{i\to\infty} B'.$$

Then we have

$$\mu\big(B' \triangle B(s_i)\big) \xrightarrow{i\to\infty} 0.$$

However, as we had previously demonstrated, $s_i \to t^*$ for $i \to \infty$ implies

$$\mu\big(B^* \triangle B(s_i)\big) \xrightarrow{i\to\infty} 0.$$

Together, this implies that

$$\mu(B^* \triangle B') \leq \mu\big(B^* \triangle B(s_i)\big) + \mu\big(B' \triangle B(s_i)\big) \xrightarrow{i\to\infty} 0$$

and therefore $B^* = B'$. □

Lemma 2.3.17 is not a dense interpolation theorem. The lemma only allows us to add points to the support tuple one by one. The result could never be defined on an uncountable parameter set. However, the dense interpolation theorem itself is a simple corollary of this lemma. The main advantage of breaking the primary effort behind the dense interpolation theorem off into a separate lemma is that, by doing it this way, we are able to reuse the lemma for sparse interpolation later.

Theorem 2.3.18 (Geodesic Interpolation With Dense Support).
Let $(I,\ T,\ B)$ be a dense geodesic support tuple in a measure space $(X,\ \Sigma,\ \mu)$. Then there exists a unique geodesic $\gamma\colon I \to {}^{\Sigma}\!/_{\sim_\mu}$ such that

$$\gamma(t) = B(t) \qquad \forall t \in T.$$

If $C \geq 0$ is a constant for which $(I,\ T,\ B)$ satisfies Equation (2.35). Then C is a geodesic constant of γ. ◁

PROOF. PART 1 (CONSTRUCTION OF γ). Let $t \in I$. Because T is dense in I, t is an accumulation point of T. According to Lemma 2.3.17, there exists a unique similarity class $B^*(t) \in {}^{\Sigma}\!/_{\sim_\mu}$ such that

$$B(s_i) \xrightarrow{i\to\infty} B^*(t)$$

for all sequences $(s_i)_{i\in\mathbb{N}} \in T^{\mathbb{N}}$. We define $\gamma\colon I \to {}^{\Sigma}\!/_{\sim_\mu}$ by

$$\gamma(t) \coloneqq B^*(t) \quad \forall i \in I.$$

Let $C \geq 0$ be a constant for which $(I,\ T,\ B)$ satisfies Equation (2.35). Let $s, t \in I$. By applying Lemma 2.3.17 twice, we can show that $(I,\ T',\ B')$ with $T' \coloneqq T \cup \{s,\ t\}$ and

$$B'(q) \coloneqq \left(\begin{cases} B(q) & \text{if } q \in T, \\ B^*(s) & \text{if } q = s, \\ B^*(t) & \text{if } q = t \end{cases}\right) \quad \forall q \in T'$$

is a well-defined geodesic support tuple that satisfies Equation (2.35) with constant C. This implies that

$$\mu\big(\gamma(s) \triangle \gamma(t)\big) = \mu\big(B^*(s) \triangle B^*(t)\big) = C \cdot |s - t|.$$

Because this holds for every valid choice of C, s, and t, and because at least one valid choice for C exists according to Definition 2.3.15, γ is a geodesic and C is a geodesic constant of γ.

PART 2 (REALIZATION OF SUPPORT). Let $t \in T$. We can approximate t with a constant approximating sequence $(t_i)_{i\in\mathbb{N}} \in T^{\mathbb{N}}$ with $t_i := t$ for all $i \in \mathbb{N}$. Then we have

$$B(t_i) \xrightarrow{i\to\infty} B^*(t) = \gamma(t).$$

Because $B(t_i) = B(t)$ is constant, this is only possible if

$$\gamma(t) = B(t).$$

PART 3 (UNIQUENESS). Finally, let $\gamma' \colon I \to \Sigma/_{\sim_\mu}$ be another geodesic with geodesic constant $C' \geq 0$ such that

$$\gamma'(t) = B(t) \quad \forall t \in T.$$

Let $t \in I$. Because T is dense in I, there exists a sequence $(t_i)_{i\in\mathbb{N}} \in T^{\mathbb{N}}$ with $t_i \to t$ for $i \to \infty$. For all $i \in \mathbb{N}$, we have

$$\begin{aligned}
\mu\big(\gamma(t) \triangle \gamma'(t)\big) &= \mu\Big(\big(\gamma(t) \triangle \gamma(t_i)\big) \triangle \big(\gamma'(t) \triangle \overbrace{\gamma(t_i)}^{=\gamma'(t_i)}\big)\Big) \\
&= \mu\Big(\big(\gamma(t) \triangle \gamma(t_i)\big) \triangle \big(\gamma'(t) \triangle \gamma'(t_i)\big)\Big) \\
&\leq \mu\Big(\big(\gamma(t) \triangle \gamma(t_i)\big) \cup \big(\gamma'(t) \triangle \gamma'(t_i)\big)\Big) \\
&\leq \mu\big(\gamma(t) \triangle \gamma(t_i)\big) + \mu\big(\gamma'(t) \triangle \gamma'(t_i)\big) \\
&= (C + C') \cdot |t - t_i|.
\end{aligned}$$

The fact that $|t - t_i| \to 0$ for $i \to \infty$ implies that $\gamma(t) = \gamma'(t)$. □

2.3.2.3 SPARSE INTERPOLATION

Sparse interpolation is a more involved procedure than dense interpolation. As the name suggests, sparse interpolation deals with geodesic support tuples that are not dense. Sparse support tuples can have arbitrarily large gaps in their support. While dense interpolation could arguably be referred to as "interpolation in name only" because the geodesic is already fully defined by the support tuple, sparse interpolation has to genuinely fill in gaps by interpolating between anchor points on either side of a gap.

Although sparse interpolation is distinct from dense interpolation, the latter is invoked as the final step of the former. The final geodesic is obtained by interpolating a dense support tuple which is derived from the sparse support tuple through a process that we refer to as "support densification."

Our chosen method of support densification is not quite what one might expect. Intuitively, one might think that we only need to fill in actual gaps in the support tuple. However, from a theoretical standpoint, it is much more convenient to define an entirely new set of supports that is dense on its own and consistent with the given sparse support. We can then form the union of both support tuples and use it for dense interpolation.

Of course, the way in which we fill in a gap is arbitrary. Therefore, we cannot reasonably expect the result of sparse interpolation to be unique. Before we begin, we first clearly define the anchor points that we will use for interpolation.

Definition 2.3.19 (Interpolation Anchors).
Let $(I,\ T,\ B)$ be a geodesic support tuple in a measure space $(X,\ \Sigma,\ \mu)$. Let $t \in \operatorname{conv}(T)$ and let

$$\check{T}(t,T) \coloneqq \sup\{s \in T \mid s \le t\},$$
$$\hat{T}(t,T) \coloneqq \inf\{s \in T \mid s \ge t\}.$$

Let $\check{B}(t,T,B) \in \Sigma/_{\sim_\mu}$ be such that

$$B(t_i) \xrightarrow{i\to\infty} \check{B}(t,T,B)$$

for all sequences $(t_i)_{i\in\mathbb{N}} \in T^{\mathbb{N}}$ with $t_i \to \check{T}(t,T)$ for $i \to \infty$. Let $\hat{B}(t,T,B) \in \Sigma/_{\sim_\mu}$ be such that

$$B(t_i) \xrightarrow{i\to\infty} \hat{B}(t,T,B)$$

for all sequences $(t_i)_{i\in\mathbb{N}} \in T^{\mathbb{N}}$ with $t_i \to \hat{T}(t,T)$ for $i \to \infty$. Then we refer to $\big(\check{T}(t,T),\ \check{B}(t,T,B)\big)$ and $\big(\hat{T}(t,T),\ \hat{B}(t,T,B)\big)$ as the *left* and *right interpolation anchor for t in* $(I,\ T,\ B)$, respectively. ◁

Definition 2.3.19 relies heavily on Lemma 2.3.17. For $t \in \operatorname{conv}(T)$, the sets $\{s \in T \mid s \le t\}$ and $\{s \in T \mid s \ge t\}$ are non-empty subsets of T that are bounded above or below, respectively, by t. Therefore, $\check{T}(t,T)$ and $\hat{T}(t,T)$ are finite accumulation points of subsets of T. This makes them accumulation points of T as well. This, in turn, guarantees the existence and uniqueness of $\check{B}(t,T,B)$ and $\hat{B}(t,T,B)$, which means that the interpolation anchors are well-defined.

Next, we formulate a lemma that demonstrates that, if the underlying measure space is atomless, then the support tuple can be extended to include a support value for any parameter within $\operatorname{conv}(T)$. This lemma forms the core of support densification.

Lemma 2.3.20 (Single-Point Support Fill-In).
Let $(I,\ T,\ B)$ be a geodesic support tuple in an atomless measure space $(X,\ \Sigma,\ \mu)$. Let $t^* \in \operatorname{conv}(T)$. Then there exists a similarity class $B^* \in \Sigma/_{\sim_\mu}$ such that the tuple $(I,\ T',\ B')$ with $T' \coloneqq T \cup \{t^*\}$ and $B' \colon T' \to \Sigma/_{\sim_\mu}$ such that

$$B'(t) \coloneqq \begin{cases} B(t) & \text{if } t \in T, \\ B^* & \text{if } t = t^* \end{cases} \qquad \forall t \in T'$$

is a geodesic support tuple. If $C \ge 0$ is a constant for which $(I,\ T,\ B)$ satisfies Equation (2.35), then $(I,\ T',\ B')$ satisfies Equation (2.35) with the same constant C. ◁

PROOF. PART 1 (ANCHORING). Let $C \ge 0$ be a constant for which $(I,\ T,\ B)$ satisfies Equation (2.35). We begin by finding the interpolation anchors from which we will determine B^*. Because $t^* \in \operatorname{conv}(T)$, we know that

$$T_\le \coloneqq \{t \in T \mid t \le t^*\},$$
$$T_\ge \coloneqq \{t \in T \mid t \ge t^*\}$$

are both non-empty. Because $T_\le$ is non-empty and bounded above by t^*, the left interpolation anchor's parameter satisfies

$$-\infty < \underbrace{\check{T}(t^*,T)}_{=\sup T_\le} \le t^*.$$

Similarly, the right interpolation anchor's parameter satisfies

$$t^* \leq \underbrace{\hat{T}(t^*, T)}_{=\inf T_\geq} < \infty$$

because $T_\geq$ is non-empty. According to Lemma 2.3.17, the respective similarity classes $\check{B}(t^*, T, B)$ and $\hat{B}(t^*, T, B)$ are uniquely defined and satisfy

$$B(t_i) \xrightarrow{i\to\infty} \check{B}(t^*, T, B),$$
$$B(t_i) \xrightarrow{i\to\infty} \hat{B}(t^*, T, B)$$

for all sequences $(t_i)_{i\in\mathbb{N}} \in T^{\mathbb{N}}$ with $t_i \to \check{T}(t^*, T)$ for $i \to \infty$ or $t_i \to \hat{T}(t^*, T)$ for $i \to \infty$, respectively. By applying Lemma 2.3.17 twice, we can show that $(I,\ \widetilde{T},\ \widetilde{B})$ with

$$\widetilde{T} \coloneqq T \cup \{\check{T}(t^*, T),\ \hat{T}(t^*, T)\},$$
$$\widetilde{B}(t) \coloneqq \left(\begin{cases} B(t) & \text{if } t \in T, \\ \check{B}(t^*, T, B) & \text{if } t = \check{T}(t^*, T), \\ \hat{B}(t^*, T, B) & \text{if } t = \hat{T}(t^*, T) \end{cases} \right) \quad \forall t \in \widetilde{T}$$

is a geodesic support tuple that satisfies Equation (2.35) with constant C.

Part 2 (Interpolation). For brevity, let

$$t_\leq \coloneqq \check{T}(t^*, T), \qquad t_\geq \coloneqq \hat{T}(t^*, T).$$

Because $-\infty < t_\leq \leq t^* \leq t_\geq < \infty$, we can write t^* as a convex combination of $t_\leq$ and $t_\geq$. Let $q \in [0,\ 1]$ be such that

$$t^* = t_\leq + q \cdot (t_\geq - t_\leq).$$

The symmetric difference between $\widetilde{B}(t_\leq)$ and $\widetilde{B}(t_\geq)$ satisfies

$$\mu\Big(\underbrace{\widetilde{B}(t_\leq) \triangle \widetilde{B}(t_\geq)}_{\coloneqq D}\Big) = C \cdot |t_\leq - t_\geq| = C \cdot (t_\geq - t_\leq).$$

Because $(X,\ \Sigma,\ \mu)$ is atomless, we can choose an essential subset $D^* \subseteq_\mu D$ such that $\mu(D^*) = C \cdot q \cdot (t_\geq - t_\leq) \leq \mu(D)$ (see, e.g., [Bog07, Cor. 1.12.10] in conjunction with Lemma 2.2.16). We then define

$$B^* \coloneqq \widetilde{B}(t_\leq) \triangle D^*.$$

Part 3 (Support Tuple Properties). We define $T' \coloneqq T \cup \{t^*\}$. Because $t^* \in \operatorname{conv}(T)$, we evidently have $\operatorname{conv}(T') = \operatorname{conv}(T)$. This implies

$$\operatorname{conv}(T') = \operatorname{conv}(T) \subseteq I \subseteq \overline{\operatorname{conv}(T)} = \overline{\operatorname{conv}(T')}$$

because $(I,\ T,\ B)$ satisfies Equation (2.34). We therefore only need to verify distances to prove that $(I,\ T',\ B')$ with

$$B'(t) \coloneqq \begin{cases} B(t) & \text{if } t \in T, \\ B^* & \text{if } t = t^* \end{cases} \quad \forall t \in T'$$

satisfies Definition 2.3.15. Let $s \in T$. We have to make a case distinction.

Case 1 ($s \le t^*$). By definition of $t_\le$, we have $s \le t_\le$. Because $(I,\ \widetilde{T},\ \widetilde{B})$ is a geodesic support tuple that satisfies Equation (2.35) with constant C, Lemma 2.3.16 implies that

$$D \cap \big(B(s) \triangle \widetilde{B}(t_\le)\big) = \big(\widetilde{B}(t_\le) \triangle \widetilde{B}(t_\ge)\big) \cap \big(\widetilde{B}(s) \triangle \widetilde{B}(t_\le)\big) = [\emptyset]_{\sim_\mu}$$

and we have the geodesic property

$$\mu\big(B(s) \triangle \widetilde{B}(t_\le)\big) = \mu\big(\widetilde{B}(s) \triangle \widetilde{B}(t_\le)\big) = C \cdot |s - t_\le| = C \cdot (t_\le - s).$$

Because D^* is an essential subset of D, D^* and $B(s) \triangle \widetilde{B}(t_\le)$ are essentially disjoint. Therefore, we have

$$\begin{aligned}
\mu\big(B(s) \triangle B^*\big) &= \mu\Big(\big(B(s) \triangle \widetilde{B}(t_\le)\big) \triangle \big(\widetilde{B}(t_\le) \triangle B^*\big)\Big) \\
&= \mu\Big(\big(B(s) \triangle \widetilde{B}(t_\le)\big) \triangle D^*\Big) \\
&= \mu\Big(\big(B(s) \triangle \widetilde{B}(t_\le)\big) \cup D^*\Big) \\
&= \mu\big(B(s) \triangle \widetilde{B}(t_\le)\big) + \mu(D^*) \\
&= C \cdot (t_\le - s) + C \cdot q \cdot (t_\ge - t_\le) \\
&= C \cdot \big(t_\le + q \cdot (t_\ge - t_\le) - s\big) \\
&= C \cdot (t^* - s) \\
&= C \cdot |s - t^*|.
\end{aligned}$$

◁

Case 2 ($s > t^*$). By definition of $t_\ge$, we have $s \ge t_\ge$. Similarly to the previous case, Lemma 2.3.16 implies

$$D \cap \big(B(s) \triangle \widetilde{B}(t_\ge)\big) = [\emptyset]_{\sim_\mu}$$

and we have the geodesic property

$$\begin{aligned}
\mu\big(B(s) \triangle \widetilde{B}(t_\ge)\big) &= C \cdot |s - t_\ge| = C \cdot (s - t_\ge), \\
\mu\big(D\big) &= C \cdot |t_\le - t_\ge| = C \cdot (t_\ge - t_\le).
\end{aligned}$$

The fractional step D^* is an essential subset of D and therefore essentially disjoint from $B(s) \triangle \widetilde{B}(t_\ge)$. Therefore, we have

$$\begin{aligned}
\mu\big(B(s) \triangle B^*\big) &= \mu\Big(\big(B(s) \triangle \widetilde{B}(t_\ge)\big) \triangle \big(\widetilde{B}(t_\ge) \triangle \widetilde{B}(t_\le)\big) \triangle \big(\widetilde{B}(t_\le) \triangle B^*\big)\Big) \\
&= \mu\Big(\big(B(s) \triangle \widetilde{B}(t_\ge)\big) \triangle D \triangle D^*\Big) \\
&= \mu\Big(\big(B(s) \triangle \widetilde{B}(t_\ge)\big) \cup D \setminus D^*\Big) \\
&= \mu\big(B(s) \triangle \widetilde{B}(t_\ge)\big) + \mu(D) - \mu(D^*) \\
&= C \cdot (s - t_\ge) + C \cdot (t_\ge - t_\le) - C \cdot q \cdot (t_\ge - t_\le) \\
&= C \cdot \big(s - t_\ge + (1 - q) \cdot (t_\ge - t_\le)\big) \\
&= C \cdot \big(s - t_\le - q \cdot (t_\ge - t_\le)\big) \\
&= C \cdot (s - t^*) \\
&= C \cdot |s - t^*|.
\end{aligned}$$

◁

In either case, we have

$$\mu\big(B(s) \triangle B^*\big) = C \cdot |s - t^*| \quad \forall s \in T.$$

This proves that B' is well-defined if $t^* \in T$. Along with the trivial cases $s, t \in T$ and $s = t = t^*$, this proves that

$$\mu\big(B'(s) \triangle B'(t)\big) = C \cdot |s - t| \quad \forall s, t \in T'.$$

Therefore, $(I,\ T',\ B')$ is a geodesic support tuple that satisfies Equation (2.35) with constant C. □

Lemma 2.3.20 allows us to add a single interpolated support point to a support tuple. We can iteratively apply the lemma to add any finite number of interpolated support points. However, in order to make a sparse support tuple dense, we need to add a countably infinite set of interpolated support points. The following lemma proves that we can take the limit of an infinite number of single-point interpolations.

Lemma 2.3.21 (Limit Support Tuples).
Let $(X,\ \Sigma,\ \mu)$ be a measure space, let $I \subseteq \mathbb{R}$ be an interval, let $(T_i)_{i\in\mathbb{N}} \in \big(2^I\big)^{\mathbb{N}}$ be a sequence of finite or countable subsets of I, and let $(B_i)_{i\in\mathbb{N}}$ be a sequence of mappings $B_i \colon T_i \to \Sigma/_{\sim_\mu}$ such that

(1) $T_i \subseteq T_{i+1}$ for all $i \in \mathbb{N}$;

(2) $B_{i+1}\big|_{T_i} = B_i$ for all $i \in \mathbb{N}$;

(3) $(I,\ T_i,\ B_i)$ is a geodesic support tuple for each $i \in \mathbb{N}$.

Then the tuple $(I,\ T_\infty,\ B_\infty)$ with

$$T_\infty := \bigcup_{i=1}^{\infty} T_i$$

and $B_\infty \colon T_\infty \to \Sigma/_{\sim_\mu}$ with

$$B_\infty(t) := \begin{cases} B_1(t) & \text{if } t \in T_1, \\ B_{i+1}(t) & \text{if } t \in T_{i+1} \setminus T_i \text{ for } i \in \mathbb{N} \end{cases} \qquad \forall t \in T_\infty$$

is a geodesic support tuple that satisfies

$$B_\infty(t) = B_i(t) \quad \forall i \in \mathbb{N},\ t \in T_i.$$

If $C \geq 0$ is a constant for which all $(I,\ T_i,\ B_i)$ satisfy Equation (2.35). Then $(I,\ T_\infty,\ B_\infty)$ satisfies Equation (2.35) with the same constant C. ◁

Proof. Because T_∞ is a countable union of finite or countably infinite sets, it is itself finite or countably infinite. For every $t \in T_\infty$, there must exist at least one $i \in N$ such that $t \in T_i$. Furthermore, because $(T_i)_{i\in\mathbb{N}}$ is monotonically increasing, there exists at most one $i \in \mathbb{N}$ for which $t \in T_i$ and $i = 1$ or $t \notin T_{i-1}$. This proves that B_∞ is well-defined.

We now prove that $B_\infty(t) = B_i(t)$ for all $i \in \mathbb{N}$. We do so by complete induction. For $i = 1$, we have $B_\infty(t) = B_i(t)$ for all $t \in T_i$ by definition. For the induction step, we work from the premise that $i \in \mathbb{N}$ is such that

$$B_\infty(t) = B_j(t) \quad \forall j \leq i,\ t \in T_j.$$

For all $t \in T_{i+1} \setminus T_i$, we have $B_\infty(t) = B_{i+1}(t)$ by definition of B_∞. For $t \in T_j$ for $j < i+1$, we have $B_\infty(t) = B_j(t)$ by our induction premise. We also have $t \in T_i$ and therefore

$$B_{i+1}(t) = B_i(t) = B_\infty(t)$$

according to the induction premise and Property 2.3.21 (2). By complete induction, this proves that

$$B_\infty(t) = B_i(t) \quad \forall i \in \mathbb{N},\ t \in T_i.$$

Next, we prove that $(I,\ T_\infty,\ B_\infty)$ is a geodesic support tuple. Because $T_i \subseteq I$ for all $i \in \mathbb{N}$, we have $T_\infty \subseteq I$ and therefore

$$\operatorname{conv}(T_\infty) \subseteq I.$$

This follows from the fact that I is its own convex hull and that the convex-hull operator is non-decreasing. We also know that $T_\infty \supseteq T_i$ for all $i \in \mathbb{N}$, which implies that

$$\overline{\operatorname{conv}(T_\infty)} \supseteq \overline{\operatorname{conv}(T_i)} \supseteq I,$$

which follows from the fact that $(I,\ T_i,\ B_i)$ is a geodesic support tuple for each $i \in \mathbb{N}$.

The only thing that remains to be proven is Equation (2.35). Let $C \geq 0$ be a constant such that all $(I,\ T_i,\ B_i)$ satisfy Equation (2.35) with constant C. To show that such a constant exists, we note that C is uniquely defined by the distance between two support values associated with distinct parameters. Therefore, unless T_i is empty or a singleton, C is uniquely defined.

If T_i is empty or a singleton for every $i \in \mathbb{N}$, then any choice of $C \geq 0$ fits for all $i \in \mathbb{N}$. Otherwise, there exists a minimal $i_0 \in \mathbb{N}$ such that $\#T_{i_0} > 1$. For this i_0, C is uniquely defined. For all subsequent $i > i_0$, $T_i \supseteq T_{i_0}$ and

$$B_i(t) = B_\infty(t) = B_{i_0}(t) \quad \forall t \in T_{i_0}$$

ensure that $(I,\ T_i,\ B_i)$ must satisfy Equation (2.35) with the same unique constant as $(I,\ T_{i_0},\ B_{i_0})$.

Let $s, t \in T_\infty$. There must exist $i, j \in \mathbb{N}$ such that $s \in T_i$ and $t \in T_j$. Without loss of generality, let $i \leq j$, which implies that $T_i \subseteq T_j$. We then find that

$$\mu\big(B_\infty(s) \mathbin{\triangle} B_\infty(t)\big) = \mu\big(B_j(s) \mathbin{\triangle} B_j(t)\big) = C \cdot |s - t|.$$

Because this holds for all $s, t \in T_\infty$, $(I,\ T_\infty,\ B_\infty)$ satisfies Equation (2.35) with constant C. □

With this result, the sparse interpolation theorem is relatively easy to prove: we find a countable dense subset $Q \subseteq I$, extend the support tuple by its interpolation at each point in Q, take the limit of that sequence of extensions, and perform a dense interpolation of the resulting support tuple. In our case, the dense set will be $Q \coloneqq \mathbb{Q} \cap \operatorname{conv}(T)$, which is a valid choice for any interval I that has a non-empty interior.

Theorem 2.3.22 (Geodesic Interpolation With Sparse Support).
Let $(I,\ T,\ B)$ be a geodesic support tuple in an atomless measure space $(X,\ \Sigma,\ \mu)$. Then there exists a geodesic $\gamma\colon I \to \Sigma/_{\sim_\mu}$ such that

$$\gamma(t) = B(t) \qquad \forall t \in T.$$

If $(I,\ T,\ B)$ satisfies Equation (2.35) for a constant $C \geq 0$, then C is a geodesic constant of γ. ◁

Proof. Part 1 (Edge Cases). We first deal with the edge cases in which $\operatorname{conv}(T)$ is empty or a singleton. In these cases, we have $\operatorname{conv}(T) = T$, which means that T is dense in

$$I \subseteq \overline{\operatorname{conv}(T)} = \overline{T}.$$

Therefore $(I,\ T,\ B)$ is a dense geodesic support tuple and we can invoke Theorem 2.3.18 directly. For all other cases, the interior of $\operatorname{conv}(T)$ is non-empty and $\mathbb{Q} \cap \operatorname{conv}(T)$ is dense in $I \subseteq \overline{\operatorname{conv}(T)}$.

Part 2 (Iterative Extension). Let $(q_i)_{i\in\mathbb{N}} \in \mathbb{Q}^{\mathbb{N}}$ be an enumeration of $\mathbb{Q}$. Let $C \geq 0$ be a constant for which $(I,\ T,\ B)$ satisfies Equation (2.35). Let $T_1 \coloneqq T$ and $B_1 \coloneqq B$. We now iteratively construct sequences $(T_i)_{i\in\mathbb{N}}$ and $(B_i)_{i\in\mathbb{N}}$ for use in Lemma 2.3.21. We simultaneously prove by complete induction that $(I,\ T_i,\ B_i)$ is a geodesic support tuple that satisfies Equation (2.35) with constant C, and that we have $T_i \subseteq T_{i+1}$, $B_{i+1}\big|_{T_i} = B_i$, as well as

$$q_j \in T_{i+1}\ \forall i \in \mathbb{N},\ j \leq i \colon q_j \in \operatorname{conv}(T).$$

We first consider $i = 1$. According to the theorem's premise $(I,\ T_i,\ B_i)$ is a geodesic support tuple. C is chosen such that $(I,\ T_i,\ B_i)$ satisfies Equation (2.35) with constant C.

For the construction and induction step, we work from the premise that $i \in \mathbb{N}$ is such that $(I,\ T_i,\ B_i)$ is a geodesic support tuple that satisfies Equation (2.35) with constant C as well as

$$q_j \in T_i\ \forall j < i \colon q_j \in \operatorname{conv}(T).$$

We now make a case distinction.

Case 1 ($q_i \notin \operatorname{conv}(T)$). In this case, we set $T_{i+1} \coloneqq T_i$ and $B_{i+1} \coloneqq B_i$. Evidently, this satisfies $T_i \subseteq T_{i+1}$ and $B_{i+1}\big|_{T_i} = B_i$. The fact that $(I,\ T_{i+1},\ B_{i+1})$ is a geodesic support tuple that satisfies Equation (2.35) with constant C as well as

$$q_j \in T_{i+1}\ \forall j \leq i \colon q_j \in I$$

follow directly from the induction premise. ◁

Case 2 ($q_i \in \operatorname{conv}(T)$). In this case, Lemma 2.3.20 allows us to choose $B_i^* \in \Sigma/_{\sim_\mu}$, such that $(I,\ T_{i+1},\ B_{i+1})$ with $T_{i+1} = T_i \cup \{q_i\}$ and $B_{i+1} \colon T_{i+1} \to \Sigma/_{\sim_\mu}$ such that

$$B_{i+1}(t) = \begin{cases} B_i(t) & \text{if } t \in T_i, \\ B_i^* & \text{if } t = q_i \end{cases} \qquad \forall t \in T_{i+1}$$

is a well-defined geodesic support tuple that satisfies Equation (2.35) with constant C. We have $T_i \subseteq T_{i+1}$ and $B_{i+1}\big|_{T_i} = B_i$ by definition. For $j < i+1$ with $q_j \in \operatorname{conv}(T)$, $q_j \in T_{i+1}$ follows from the induction premise, which states that $q_j \in T_i \subseteq T_{i+1}$. For $j = i+1$, we have $q_j \in T_{i+1}$ by definition. ◁

By complete induction, this proves that $\big((I,\ T_i,\ B_i)\big)_{i\in\mathbb{N}}$ is a sequence of geodesic support tuples that all satisfy Equation (2.35) with constant C that satisfies the premises of Lemma 2.3.21 as well as

$$q_j \in T_i\ \forall i \in \mathbb{N},\ j < i \colon q_j \in \operatorname{conv}(T).$$

PART 3 (DENSIFICATION AND INTERPOLATION). Because $\big((I,\, T_i,\, B_i)\big)_{i\in\mathbb{N}}$ is a sequence of support tuples that satisfies the premises of Lemma 2.3.21, we can apply that lemma to show that $(I,\, T_\infty,\, B_\infty)$ with

$$T_\infty = \bigcup_{i=1}^{\infty} T_i$$

and $B_\infty \colon T_\infty \to \Sigma/_{\sim_\mu}$ such that

$$B_\infty(t) = \begin{cases} B_1(t) & \text{if } t \in T_1, \\ B_{i+1}(t) & \text{if } t \in T_{i+1} \setminus T_i \text{ for } i \in \mathbb{N} \end{cases} \qquad \forall t \in T_\infty$$

is a well-defined geodesic support tuple that satisfies Equation (2.35) with constant C. Because $T_1 = T$ and $B_1 = B$, we have $B_\infty(t) = B(t)$ for all $t \in T$. Furthermore, because $q_j \in T_j \subseteq T_\infty$ for all $j \in \mathbb{N}$ with $q_j \in \operatorname{conv}(T)$, we know that $Q = \mathbb{Q} \cap \operatorname{conv}(T) \subseteq T_\infty$, which implies that T_∞ is dense in $\overline{\operatorname{conv}(T)} \supseteq I$.

We can therefore invoke Theorem 2.3.18 to obtain a geodesic $\gamma \colon I \to \Sigma/_{\sim_\mu}$ with geodesic constant C such that

$$\gamma(t) = B_\infty(t) \quad \forall t \in T_\infty.$$

Because $T \subseteq T_\infty$, this implies that

$$\gamma(t) = B_\infty(t) = B_1(t) = B(t) \quad \forall t \in T.$$

We note that, although γ is unique for a given dense support tuple $(I,\, T_\infty,\, B_\infty)$, there is flexibility in the fill-in process, which means that γ is not unique as an interpolator of $(I,\, T,\, B)$. □

The sparse interpolation theorem is very powerful. It allows us to arrange similarity classes along the parameter space according to their distance to one another without any prior regard as to whether or not the result is dense. We will use this in Sections 2.3.4 and 2.3.5 to arrange similarity classes whose measures are not known a priori. Another interesting application is in Section 2.3.6, where we will use it with just two distinct support points to construct a geodesic connecting two arbitrary similarity classes with each other. In this case, the sparse interpolation theorem essentially constructs the entire geodesic on its own.

2.3.3 Geodesic Level Set Functions

So far, one of the most relevant properties of geodesics is their monotonicity. Relative to its origin point, every geodesic is canonical and therefore monotonically increasing. In this section, we demonstrate that, if we were to stack the similarity classes of a canonical geodesic on top of one another along an additional coordinate axis, then they form the epigraph of a measurable function. We refer to this function as the "geodesic level set function" (GLSF) associated with the geodesic. Figure 2.4 on the next page illustrates this.

By showing this, we effectively establish an equivalence between a canonical geodesic and the preimage map of a GLSF. This equivalence is very valuable because it allows us to transfer some operations from measurable functions to

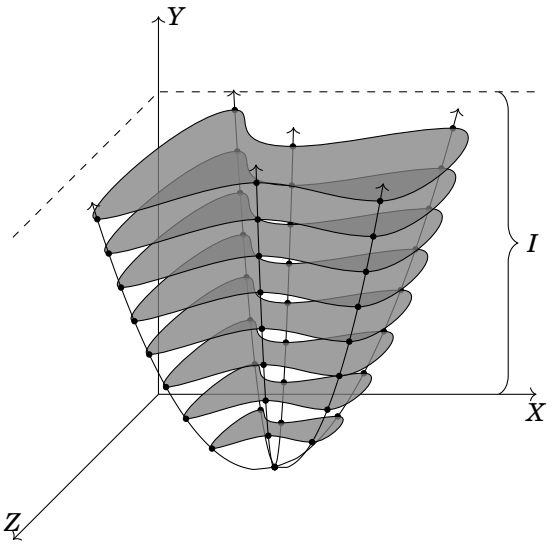


Figure 2.4: Illustration of a geodesic level set function. The parameter interval I is represented by a section of the Y axis. The sets of a canonical geodesic $\gamma\colon I \to \Sigma/_{\sim_\mu}$ form the sublevel sets of the level set function. We have marked six points on the boundary of the set to illustrate better how the growth rate of the set diameter decays to ensure steady growth in area.

measure space geodesics. One notable such operation is *composition*. Because the preimage map of a composed function $h = f \circ g$ is the reversed composition of the component preimage maps $h^{-1} = g^{-1} \circ f^{-1}$, the composition of two GLSFs can effectively be read as the GLSF of a composed geodesic. We refer to this operation as "rearrangement" because it effectively takes the changes made by one geodesic and arranges them in an order prescribed by the other. We discuss rearrangement later in this section.

As one might expect, not every measurable function encodes a geodesic. For instance, a constant function would encode a path that is discontinuous and could therefore not be a geodesic. The function whose sublevel sets form a geodesic must conform to the defining characteristics of a geodesic. We refer to the class of all measurable functions that satisfy these conditions as "geodesic level set functions." At first, this is regardless of whether these functions actually encode geodesics, though we will later see that they always do.

Definition 2.3.23 (Geodesic Level Set Function).
Let $(X,\ \Sigma,\ \mu)$ be a measure space, and let $g\colon X \to \mathbb{R} \cup \{\infty\}$ be a measurable function. We refer to g as a *geodesic level set function (GLSF) on* $(X,\ \Sigma,\ \mu)$ if and only if there exist an interval $I \subseteq \mathbb{R}$ and a constant $C \geq 0$ such that $g(X) \subseteq I \cup \{\infty\}$ and

$$\mu\Big(g^{-1}\big([s,\ t]\big)\Big) = C \cdot (t - s) \qquad \forall s, t \in I\colon s \leq t. \tag{2.36}$$

We refer to C as a *geodesic constant* of g and to I as the *parametric interval* of g. We denote the set of all geodesic level set functions on $(X,\ \Sigma,\ \mu)$ by

$$\mathcal{G}(X, \Sigma, \mu)$$

and the set of all geodesic level set functions on $(X,\ \Sigma,\ \mu)$ with a given parametric interval $I \subseteq \mathbb{R}$ by

$$\mathcal{G}(X, \Sigma, \mu, I). \qquad \triangleleft$$

We note that we explicitly allow for GLSFs to map to ∞. This is so that there is a value to which we can map points that are outside of the total variation of the encoded geodesic. Mapping a point to ∞ excludes it from all sublevel sets and therefore makes it so that the point is never included.

Because Equation (2.36) is almost exactly the geodesic property, it is straightforward to show that for every GLSF, there exists a corresponding canonical geodesic. It is substantially more complex to construct a GLSF from a geodesic. This is because geodesics yield similarity classes from which we then have to obtain suitable representatives.

Theorem 2.3.24 (Constructing Geodesics From GLSFs).
Let $(X,\ \Sigma,\ \mu_i)$ be a measure space, let $I \subseteq \mathbb{R}$ be an interval, and let $g \in \mathcal{G}(X, \Sigma, \mu, I)$. Then the map $\gamma\colon I \to \Sigma/_{\sim_\mu}$ with

$$\gamma(t) \coloneqq \left[\{g \leq t\}\right]_{\sim_\mu}$$

is a canonical geodesic with geodesic constant C and

$$\mathrm{TV}(\gamma) = \left[\{g < \infty\}\right]_{\sim_\mu}. \qquad \triangleleft$$

PROOF. Because g is measurable and because $\{g \leq t\} = g^{-1}\big((-\infty,\ t]\big)$ is the preimage of an interval (which is Borel-measurable) for $t \in \mathbb{R}$, $\{g \leq t\}$ is in Σ for all $t \in \mathbb{R}$. Let $C \geq 0$ be a geodesic constant of g.

We first show that γ is essentially increasing. Let $s, t \in I$ with $s \leq t$. Then we have

$$\gamma(s) = \big[\underbrace{\{g \leq s\}}_{\subseteq \{g \leq t\}}\big]_{\sim_\mu} \subseteq_\mu \left[\{g \leq t\}\right]_{\sim_\mu} = \gamma(t)$$

which proves that γ is essentially increasing.

Let $s, t \in I$. Without loss of generality, let $s \leq t$. We have

$$\begin{aligned}
\mu\big(\gamma(s) \triangle \gamma(t)\big) &= \mu\big(\gamma(t) \setminus \gamma(s)\big) \\
&= \mu\Big(\left[\{g \leq t\} \setminus \{g \leq s\}\right]_{\sim_\mu}\Big) \\
&= \mu\Big(g^{-1}\big((s,\ t]\big)\Big) \\
&= \mu\Big(g^{-1}\big([s,\ t]\big)\Big) - \mu\Big(g^{-1}\big([s,\ s]\big)\Big) \\
&= C \cdot (\underbrace{t - s}_{=|s-t|}) - C \cdot \underbrace{(s - s)}_{=0} \\
&= C \cdot |s - t|
\end{aligned}$$

which proves that γ is a geodesic with geodesic constant C. With this, we only have to prove the claim about the total variation.

If $I = \emptyset$, then we have $g(x) = \infty$ for all $x \in I$ according to Definition 2.3.23. In this case, it follows that $\mathrm{TV}(\gamma) = [\emptyset]_{\sim_\mu} = \left[\{g < \infty\}\right]_{\sim_\mu}$. If $I = \emptyset$, then γ is trivially canonical. We therefore subsequently only discuss the case in which $I \neq \emptyset$.

If $I \neq \emptyset$, then there exist sequences $(s_i)_{i\in\mathbb{N}} \subseteq I$ with $s_i \to \inf I$ for $i \to \infty$ and $(t_i)_{i\in\mathbb{N}} \subseteq I$ with $t_i \to \sup I$ for $i \to \infty$. Without loss of generality, let $s_i \leq t_i$ for all $i \in \mathbb{N}$. According to Proposition 2.3.5, we have

$$\begin{aligned}
\mathrm{TV}(\gamma) &= \bigcup_{i=1}^{\infty} \big(\gamma(s_i) \triangle \gamma(t_i)\big) \\
&= \bigcup_{i=1}^{\infty} \underbrace{\big[g^{-1}\big((s_i,\, t_i]\big)\big]_{\sim_\mu}}_{\subseteq_\mu [g^{-1}(I)]_{\sim_\mu}} \\
&\subseteq_\mu \big[g^{-1}(I)\big]_{\sim_\mu} \\
&= [\{g < \infty\}]_{\sim_\mu}.
\end{aligned}$$

This establishes that $\mathrm{TV}(\gamma) \subseteq_\mu \big[\{g < \infty\}\big]_{\sim_\mu}$. On the other hand, we have

$$\begin{aligned}
\big[g^{-1}(I)\big]_{\sim_\mu} \setminus \mathrm{TV}(\gamma) &= \Big[g^{-1}(I) \setminus \bigcup_{i=1}^{\infty} g^{-1}\big((s_i,\, t_i]\big)\Big]_{\sim_\mu} \\
&= \Big[g^{-1}\Big(I \setminus \bigcup_{i=1}^{\infty} (s_i,\, t_i]\Big)\Big]_{\sim_\mu} \\
&= \Big[g^{-1}\Big(\bigcap_{i=1}^{\infty} \big(I \setminus (s_i,\, t_i]\big)\Big)\Big]_{\sim_\mu}.
\end{aligned}$$

We now make a case distinction to find an upper bound on this similarity class.

Case 1 ($\inf I = -\infty$, $\sup I = \infty$). In this case, we have

$$\bigcup_{i=1}^{\infty} (s_i,\, t_i] = \mathbb{R}$$

and therefore

$$\big[g^{-1}(I)\big]_{\sim_\mu} \setminus \mathrm{TV}(\gamma) = \Big[g^{-1}\Big(I \setminus \bigcup_{i=1}^{\infty} (s_i,\, t_i]\Big)\Big]_{\sim_\mu} = [\emptyset]_{\sim_\mu}. \qquad \lhd$$

Case 2 ($\inf I > -\infty$, $\sup I = \infty$). Because $s_i \to \inf I$ for $i \to \infty$, we have

$$\bigcup_{i=1}^{\infty} (s_i,\, t_i] \supseteq (\inf I,\, \infty)$$

and therefore

$$\big[g^{-1}(I)\big]_{\sim_\mu} \setminus \mathrm{TV}(\gamma) = \Big[g^{-1}\Big(I \setminus \bigcup_{i=1}^{\infty} (s_i,\, t_i]\Big)\Big]_{\sim_\mu} \subseteq_\mu \big[g^{-1}(\{\inf I\})\big]_{\sim_\mu}. \qquad \lhd$$

Case 3 ($\inf I = -\infty$, $\sup I < \infty$). Because $t_i \to \sup I$ for $i \to \infty$, we have

$$\bigcup_{i=1}^{\infty} (s_i,\, t_i] \supseteq (-\infty,\, \sup I)$$

and therefore

$$\big[g^{-1}(I)\big]_{\sim_\mu} \setminus \mathrm{TV}(\gamma) = \Big[g^{-1}\Big(I \setminus \bigcup_{i=1}^{\infty} (s_i,\, t_i]\Big)\Big]_{\sim_\mu} \subseteq_\mu \big[g^{-1}(\{\sup I\})\big]_{\sim_\mu}. \qquad \lhd$$

Case 4 ($\inf I > -\infty$, $\sup I < \infty$). Because $s_i \to \inf I$ and $t_i \to \sup I$ for $i \to \infty$, we have

$$\bigcup_{i=1}^{\infty} (s_i,\ t_i] \supseteq (\inf I,\ \sup I)$$

and therefore

$$\left[g^{-1}(I)\right]_{\sim_\mu} \setminus \mathrm{TV}(\gamma) = \left[g^{-1}\left(I \setminus \bigcup_{i=1}^{\infty} (s_i,\ t_i]\right)\right]_{\sim_\mu} \subseteq_\mu \left[g^{-1}(\{\inf I,\ \sup I\})\right]_{\sim_\mu}. \qquad \triangleleft$$

In all four cases, we have

$$\left[g^{-1}(I)\right]_{\sim_\mu} \setminus \mathrm{TV}(\gamma) \subseteq_\mu \left[g^{-1}(\{\inf I,\ \sup I\})\right]_{\sim_\mu} = \left[g^{-1}(\{\inf I\})\right]_{\sim_\mu} \cup \left[g^{-1}(\{\sup I\})\right]_{\sim_\mu}$$

The geodesic property for GLSFs implies that the measures of both components of the union are nullsets, i.e.,

$$\mu\left(\left[g^{-1}(I)\right]_{\sim_\mu} \setminus \mathrm{TV}(\gamma)\right) \leq \mu\left(\left[g^{-1}(\{\inf I\})\right]_{\sim_\mu}\right) + \mu\left(\left[g^{-1}(\{\sup I\})\right]_{\sim_\mu}\right) = 0 + 0 = 0.$$

Therefore, we have

$$\left[g^{-1}(I)\right]_{\sim_\mu} \subseteq_\mu \mathrm{TV}(\gamma).$$

In summary, we find that

$$\mathrm{TV}(\gamma) = \left[g^{-1}(I)\right]_{\sim_\mu} = [\{g < \infty\}]_{\sim_\mu}.$$

Bearing in mind that the codomain of g is $I \cup \{\infty\}$, we can also conclude from this that

$$\begin{aligned} \gamma(t) &= \left[g^{-1}\left((-\infty,\ t]\right)\right]_{\sim_\mu} \\ &= \left[g^{-1}\left(I \cap (-\infty,\ t]\right)\right]_{\sim_\mu} \\ &\subseteq_\mu \left[g^{-1}(I)\right]_{\sim_\mu} \\ &= \mathrm{TV}(\gamma) \end{aligned}$$

holds for all $t \in I$. In conjunction with the fact that γ is μ-essentially increasing, this means that γ is canonical. $\square$

The converse transformation, i.e., transforming a canonical geodesic into a GLSF, is more complex. This is mainly because there is no straightforward way to define an expression for the GLSF in the same way in which we could define an explicit expression for the geodesic in Theorem 2.3.24. Instead, we have to construct the GLSF incrementally and apply a convergence theorem to prove the existence of the limit.

Our construction bears some resemblance to the proof of the sparse interpolation theorem (Theorem 2.3.22). In essence, we enumerate the rational numbers and adjust the sublevel set for each rational level to match the geodesic up to a nullset.

Instead of constructing the entire function at once, we construct the positive and negative part of the GLSF separately. This allows us to construct both

parts as pointwise suprema of monotonically increasing sequences, which are themselves measurable functions (see, e.g., [Bog07, Thm. 2.1.5]). We can then combine these functions into the final GLSF.

One important issue during this transition is that the geodesic only returns similarity classes. However, we require concrete sets to construct our function. These sets must be representatives of the respective similarity classes, but they must also be chosen such that they are truly non-decreasing. As before, Lemma 2.2.16 can be used to ensure this.

Because we have to perform the construction of the GLSF in two separate parts, it is convenient to first formulate a lemma that can be used to handle either part in isolation. Because we enumerate only the rational numbers and use limit arguments for the remainder of the real numbers, we must make a continuity argument. However, semicontinuity is sufficient for our purposes. As we will see later, requiring only semicontinuity instead of full continuity allows us to reuse the lemma for the construction of other functions at a later point.

Definition 2.3.25 (Monotonic Upper Semicontinuity).
Let (X, Σ, μ) be a measure space, let $U \subseteq \mathbb{R}$, and let $\gamma \colon U \to \Sigma/_{\sim_\mu}$ be essentially monotonically increasing. We call γ *upper semicontinuous in* $t \in U$ if

$$\gamma(t) = \bigcap_{i=1}^{\infty} \gamma(t_i)$$

for all sequences $(t_i)_{i\in\mathbb{N}} \in U^{\mathbb{N}}$ such that $t_i \geq t$ for all $i \in \mathbb{N}$ and $t_i \to t$ for $i \to \infty$. ◁

Definition 2.3.25 could be adapted to define *lower semicontinuity* by considering approximating sequences with $t_i \leq t$ and exchanging the intersection with a union. We could also extend the definition to monotonically decreasing mappings γ by exchanging the union and intersection operators in the definition. In the context of this thesis, these variants are irrelevant because we primarily intend to work with canonical geodesics, which are monotonically increasing. It is relatively easy to show that any monotonic mapping γ that is continuous is also both upper and lower semicontinuous.

Lemma 2.3.26 (Semicontinuity and Continuity).
Let (X, Σ, μ) be a measure space, let $U \subseteq \mathbb{R}$, and let $\gamma \colon U \to \Sigma/_{\sim_\mu}$ be monotonically increasing and continuous. Then γ is upper semicontinuous. ◁

Proof. Let $t \in U$ and let $(t_i)_{i\in\mathbb{N}} \in U^{\mathbb{N}}$ with $t_i \geq t$ for all $i \in \mathbb{N}$ and $t_i \to t$ for $i \to \infty$. Because γ is monotonically increasing, we have

$$\bigcap_{i=1}^{n} \gamma(t_i) = \gamma\Big(\min_{i\in[n]} t_i\Big) \quad \forall n \in \mathbb{N}.$$

Therefore, we can equivalently consider the sequence $(\widetilde{t}_i)_{i\in\mathbb{N}} \in U^{\mathbb{N}}$ with

$$\widetilde{t}_i \coloneqq \min_{j\in[i]} t_j \quad \forall i \in \mathbb{N}.$$

This is a monotonically decreasing sequence. Accordingly, the corresponding sequence $\big(\gamma(\widetilde{t}_i)\big)_{i\in\mathbb{N}}$ is also μ-essentially monotonically decreasing. Because γ is monotonically increasing and $\widetilde{t}_i \geq t$ for all $i \in \mathbb{N}$, we have

$$\gamma(t) \subseteq_\mu \gamma(\widetilde{t}_i) \quad \forall i \in \mathbb{N}.$$

Therefore, we have

$$\begin{aligned}
\mu\Big(\gamma(t) \vartriangle \bigcap_{i=1}^{\infty} \gamma(\widetilde{t}_i)\Big) &= \mu\Big(\Big(\bigcap_{i=1}^{\infty} \gamma(\widetilde{t}_i)\Big) \setminus \gamma(t)\Big) \\
&= \mu\Big(\bigcap_{i=1}^{\infty} \underbrace{\big(\gamma(\widetilde{t}_i) \setminus \gamma(t)\big)}_{\text{ess. decreasing}}\Big) \\
&= \lim_{i\to\infty} \mu\big(\gamma(\widetilde{t}_i) \setminus \gamma(t)\big) \\
&= \lim_{i\to\infty} \mu\big(\gamma(\widetilde{t}_i) \vartriangle \gamma(t)\big) \\
&= 0
\end{aligned}$$

because γ is continuous. This shows that

$$\gamma(t) = \bigcap_{i=1}^{\infty} \gamma(\widetilde{t}_i).$$

For the unmodified sequence, we have

$$\bigcap_{i=1}^{\infty} \gamma(t_i) = \bigcap_{i=1}^{\infty} \bigcap_{j=1}^{i} \gamma(t_j) = \bigcap_{i=1}^{\infty} \gamma(\widetilde{t}_i) = \gamma(t)$$

because of the way in which we had constructed the sequence $(\widetilde{t}_i)_{i\in\mathbb{N}}$ and because of the monotonicity of γ. □

With this, we can formulate the partial construction lemma for upper semicontinuous mappings and be assured that it still remains applicable to continuous essentially monotonically increasing maps such as canonical geodesics.

Lemma 2.3.27 (Partial GLSF Construction).
Let (X, Σ, μ) be a measure space, let $\gamma\colon \mathbb{R}_{\geq 0} \to {}^{\Sigma}\!/_{\sim_\mu}$ be essentially increasing and upper semicontinuous, let $B_0 \in \big(\gamma(0)\big)^{\complement}$, and let $B_\infty \in \Sigma$ with $B_\infty \subseteq B_0$ and $[B_\infty]_{\sim_\mu} \subseteq_\mu \big(\gamma(t)\big)^{\complement}$ for all $t \in \mathbb{R}_{\geq 0}$. Then there exists a measurable function $g\colon X \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ such that

(1) $\{g \leq t\} \in \gamma(t)$ for all $t \in \mathbb{R}_{\geq 0}$;

(2) $\{g > 0\} \subseteq B_0$;

(3) $B_\infty \subseteq \{g = \infty\}$. ◁

PROOF. Let $\mathbb{Q}_+ := \{q \in \mathbb{Q} \mid q > 0\}$ be the set of all strictly positive rational numbers. All numbers $q \in \mathbb{Q}_+$ satisfy $0 < q < \infty$. We also know that $\mathbb{Q}_+$ is dense in $[0, \infty)$. Evidently $\mathbb{Q}_+$ is infinite, but as a subset of $\mathbb{Q}$, it is still countable. Let $(q_i)_{i\in\mathbb{N}} \in \mathbb{Q}_+^{\mathbb{N}}$ be an enumeration of $\mathbb{Q}_+$. To simplify notation for this proof, we define

$$\begin{aligned}
\overline{\mathbb{Q}}_+ &:= \mathbb{Q}_+ \cup \{0, \infty\}, \\
\mathbb{Q}_{+i} &:= \{q_j \mid j \leq i\} \qquad \forall i \in \mathbb{N}_0, \\
\overline{\mathbb{Q}}_{+i} &:= \mathbb{Q}_{+i} \cup \{0, \infty\} \quad \forall i \in \mathbb{N}_0.
\end{aligned}$$

We define $\overline{\mathbb{Q}}_{+0} = \{0, \infty\}$ because it provides a convenient starting point for inductive arguments.

Part 1 (Defining $(B_q)_{q\in\mathbb{Q}_+}$). We first complement B_0 and B_∞ with measurable sets B_q for $q \in \mathbb{Q}_+$ such that

$$\begin{aligned} B_q &\supseteq B_r && \forall q,r \in \overline{\mathbb{Q}}_+ : q \le r, \\ B_q &\in \big(\gamma(q)\big)^{\complement} && \forall q \in \overline{\mathbb{Q}}_+ : q < \infty. \end{aligned}$$

We construct these B_q by iteration over the enumeration $(q_i)_{i\in\mathbb{N}}$. We simultaneously prove by complete induction that

$$B_q \supseteq B_r \qquad \forall i \in \mathbb{N}_0,\ (q,\ r) \in \overline{\mathbb{Q}}_{+i}^2 : q \le r, \tag{2.37}$$

$$B_q \in \big(\gamma(q)\big)^{\complement} \qquad \forall q \in \overline{\mathbb{Q}}_{+i} : q < \infty. \tag{2.38}$$

As we have indicated, $i = 0$ provides a simple starting point for our inductive argument. For $i = 0$, $q = 0$ and $r = \infty$ are the only two values in $\overline{\mathbb{Q}}_{+i}$ with $q \neq r$ and $q \le r$. In this case, $B_q \supseteq B_r$ is part of the premises of the lemma. Similarly, $B_0 \in \big(\gamma(0)\big)^{\complement}$ is a part of those premises.

We now come to the induction step. Let $i \in \mathbb{N}_0$ be such that Equations (2.37) and (2.38) hold. Because $\overline{\mathbb{Q}}_{+i}$ is a finite set, we can determine

$$\begin{aligned} q_{i+1}^+ &\coloneqq \min\{q \in \overline{\mathbb{Q}}_{+i} \,|\, q \ge q_{i+1}\}, \\ q_{i+1}^- &\coloneqq \max\{q \in \overline{\mathbb{Q}}_{+i} \,|\, q \le q_{i+1}\}. \end{aligned}$$

Let $\widetilde{B}_{q_{i+1}}$ be an arbitrary representative of $\big(\gamma(q_{i+1})\big)^{\complement}$. We define

$$B_{q_{i+1}} \coloneqq (\widetilde{B}_{q_{i+1}} \cup B_{q_{i+1}^+}) \cap B_{q_{i+1}^-}.$$

Because $q_{i+1}^- \le q_{i+1} \le q_{i+1}^+$, our induction premise states that

$$B_{q_{i+1}^-} \supseteq B_{q_{i+1}^+}.$$

This means that, aside from $B_{q_{i+1}} \subseteq B_{q_{i+1}^-}$, which follows by construction, we also have

$$\begin{aligned} B_{q_{i+1}} &= (\widetilde{B}_{q_{i+1}} \cup B_{q_{i+1}^+}) \cap B_{q_{i+1}^-} \\ &= (\widetilde{B}_{q_{i+1}} \cap B_{q_{i+1}^-}) \cup (B_{q_{i+1}^+} \cap B_{q_{i+1}^-}) \\ &= (\widetilde{B}_{q_{i+1}} \cap B_{q_{i+1}^-}) \cup B_{q_{i+1}^+} \\ &\supseteq B_{q_{i+1}^+}. \end{aligned}$$

For $r \in \overline{Q}_{+(i+1)}$ with $r < q_{i+1}$, $r \neq q_{i+1}$ implies $r \in \overline{Q}_{+i}$ and therefore $r \le q_{i+1}^-$. We can infer from the induction premise that

$$B_{q_{i+1}} \subseteq B_{q_{i+1}^-} \subseteq B_r.$$

Conversely, for $r \in \overline{Q}_{+(i+1)}$ with $r > q_{i+1}$ we have $r \in \overline{Q}_{+i}$ and $r \ge q_{i+1}^+$. We can then infer that

$$B_r \subseteq B_{q_{i+1}^+} \subseteq B_{q_{i+1}}.$$

Together, these relations show that Equation (2.37) holds up to $i+1$. To show that $B_{q_{i+1}} \in \big(\gamma(q_{i+1})\big)^{\complement}$, we have to make use of the monotonicity of γ. First, we note that $q_{i+1}^- \in \overline{\mathbb{Q}}_{+i}$ with $q_{i+1}^- \le q_{i+1} < \infty$ and therefore

$$B_{q_{i+1}^-} \in \big(\gamma(q_{i+1}^-)\big)^{\complement}$$

according to our induction premise. We then have

$$[B_{q_{i+1}^-}]_{\sim_\mu} = \big(\gamma(q_{i+1}^-)\big)^\complement \supseteq_\mu \big(\gamma(q_{i+1})\big)^\complement$$

because $q_{i+1}^- \leq q_{i+1}$ and because γ is monotonically increasing. For q_{i+1}^+, we have to distinguish two cases.

Case 1 ($q_{i+1}^+ < \infty$). In this case, our induction premise states that

$$B_{q_{i+1}^+} \in \big(\gamma(q_{i+1}^+)\big)^\complement$$

and the monotonicity of γ in conjunction with $q_{i+1} \leq q_{i+1}^+$ implies that

$$[B_{q_{i+1}^+}]_{\sim_\mu} = \big(\gamma(q_{i+1}^+)\big)^\complement \subseteq_\mu \big(\gamma(q_{i+1})\big)^\complement. \qquad \triangleleft$$

Case 2 ($q_{i+1}^+ = \infty$). In this case,

$$[B_{q_{i+1}^+}]_{\sim_\mu} = B_\infty^\complement \subseteq_\mu \big(\gamma(q_{i+1})\big)^\complement.$$

follows directly from the premises of the lemma. $\triangleleft$

In either case, we have

$$[B_{q_{i+1}^+}]_{\sim_\mu} \subseteq_\mu \big(\gamma(q_{i+1})\big)^\complement \subseteq_\mu [B_{q_{i+1}^-}]_{\sim_\mu}.$$

From this, we can infer that

$$\begin{aligned}
[B_{q_{i+1}}]_{\sim_\mu} &= \big([\widetilde{B}_{q_{i+1}}]_{\sim_\mu} \cup [B_{q_{i+1}^+}]_{\sim_\mu}\big) \cap [B_{q_{i+1}^-}]_{\sim_\mu} \\
&= \Big(\big(\gamma(q_{i+1})\big)^\complement \cup \underbrace{[B_{q_{i+1}^+}]_{\sim_\mu}}_{\subseteq_\mu (\gamma(q_{i+1}))^\complement}\Big) \cap [B_{q_{i+1}^-}]_{\sim_\mu} \\
&= \big(\gamma(q_{i+1})\big)^\complement \cap \underbrace{[B_{q_{i+1}^-}]_{\sim_\mu}}_{\supseteq_\mu (\gamma(q_{i+1}))^\complement} \\
&= \big(\gamma(q_{i+1})\big)^\complement,
\end{aligned}$$

which demonstrates that Equation (2.38) holds up to $i+1$. By complete induction, this proves that Equations (2.37) and (2.38) hold for all $i \in \mathbb{N}$.

Let $q, r \in \overline{\mathbb{Q}}_+$ be such that $q \leq r$. Because $\overline{\mathbb{Q}}_+ = \bigcup_{i=0}^\infty \overline{\mathbb{Q}}_{+i}$ and because $(\overline{\mathbb{Q}}_{+i})_{i\in\mathbb{N}}$ is monotonically increasing, there exists $i_0 \in \mathbb{N}_0$ such that $q, r \in \overline{\mathbb{Q}}_{+i_0}$. Then, $B_q \supseteq B_r$ follows from Equation (2.37) for $i = i_0$. Similarly, for $q \in \overline{\mathbb{Q}}_+$ with $q < \infty$, there exists $i_0 \in \mathbb{N}_0$ such that $q \in \overline{\mathbb{Q}}_{+i_0}$ and we obtain $B_q \in \big(\gamma(q)\big)^\complement$ by applying Equation (2.38) for $i = i_0$.

PART 2 (CONSTRUCTING g). We now have sets B_q for all $q \in \overline{\mathbb{Q}}_+$ that are monotonically decreasing in q. With these sets, we construct a sequence $(g_i)_{i\in\mathbb{N}_0}$ of measurable functions $g_i : X \to \overline{\mathbb{Q}}_{+i}$ that is pointwise monotonically increasing and satisfies

$$\{g_i > 0\} \subseteq B_0 \quad \forall i \in \mathbb{N}_0, \tag{2.39}$$

$$\{g_i \geq q\} = B_q \quad \forall i \in \mathbb{N}_0,\ q \in \overline{\mathbb{Q}}_{+i} : q > 0. \tag{2.40}$$

for all $i \in \mathbb{N}_0$. As before, we construct $(g_i)_{i\in\mathbb{N}_0}$ iteratively and prove the claim by total induction. For all $i \in \mathbb{N}$, let

$$q_i^+ \coloneqq \min\{q \in \overline{\mathbb{Q}}_{+(i-1)} \mid q \geq q_i\},$$
$$q_i^- \coloneqq \max\{q \in \overline{\mathbb{Q}}_{+(i-1)} \mid q \leq q_i\}$$

be the upper and lower proxima of q_i in $\overline{\mathbb{Q}}_{+(i-1)}$, respectively. We define the initial function $g_0 \colon X \to \overline{\mathbb{Q}}_{+0} = \{0, \infty\}$ with

$$g_0(x) \coloneqq \begin{cases} \infty & \text{if } x \in B_\infty, \\ 0 & \text{if } x \notin B_\infty \end{cases} \quad \forall x \in X.$$

For all $i \in \mathbb{N}$, we define

$$g_i \coloneqq g_{i-1} + (q_i - q_i^-) \cdot \chi_{B_{q_i} \setminus B_{q_i^+}}.$$

Evidently, we have $q_i^- \leq q_i$ and therefore $q_i - q_i^- \geq 0$, which guarantees that $(g_i)_{i\in\mathbb{N}_0}$ is pointwise monotonically increasing. We now prove the remaining properties by complete induction.

We start our induction at $i = 0$. By definition, we have

$$\{g_0 > 0\} = B_\infty \subseteq B_0.$$

There is only one $q \in \overline{\mathbb{Q}}_{+0}$ with $q > 0$, namely $q = \infty$. Here, we find that

$$\{g_0 \geq q\} = \{g_0 = \infty\} = B_\infty = B_q.$$

This proves that Equations (2.39) and (2.40) hold for $i = 0$.

Next, we proceed with the induction step. Let $i \in \mathbb{N}_0$ such that $g_i \colon X \to \overline{\mathbb{Q}}_{+i}$ satisfies Equations (2.39) and (2.40). By definition, $q_{i+1}^- \leq q_{i+1} \leq q_{i+1}^+$. With the monotonicity result from the previous part of the proof, this implies that $B_{q_{i+1}^+} \subseteq B_{q_{i+1}} \subseteq B_{q_{i+1}^-}$ and therefore

$$B_{q_{i+1}} \setminus B_{q_{i+1}^+} \subseteq B_{q_{i+1}^-} \setminus B_{q_{i+1}^+}.$$

According to the induction premise, this means that

$$B_{q_{i+1}} \setminus B_{q_{i+1}^+} \subseteq \{g_i \geq q_{i+1}^-\} \setminus \{g_i \geq q_{i+1}^+\} = \{q_{i+1}^- \leq g_i < q_{i+1}^+\}.$$

Because q_{i+1}^- and q_{i+1}^+ are defined as the largest number in $\overline{\mathbb{Q}}_{+i}$ that is no greater than q_{i+1} and the smallest number in $\overline{\mathbb{Q}}_{+i}$ that is no less than q_{i+1}, respectively, there is no element of $\overline{\mathbb{Q}}_{+i}$ that is strictly between q_{i+1}^- and q_{i+1}^+. Because $\overline{\mathbb{Q}}_{+i}$ is the codomain of g_i, this implies that

$$B_{q_{i+1}} \setminus B_{q_{i+1}^+} \subseteq \{g_i = q_{i+1}^-\}.$$

We therefore know that

$$\begin{aligned} g_{i+1}(x) &= g_i(x) + (q_{i+1} - q_{i+1}^-) \cdot \chi_{B_{q_{i+1}} \setminus B_{q_{i+1}^+}}(x) \\ &= \begin{cases} g_i(x) & \text{if } x \notin B_{q_{i+1}} \setminus B_{q_{i+1}^+}, \\ g_i(x) + q_{i+1} - q_{i+1}^- & \text{if } x \in B_{q_{i+1}} \setminus B_{q_{i+1}^+} \end{cases} \\ &= \begin{cases} g_i(x) & \text{if } x \notin B_{q_{i+1}} \setminus B_{q_{i+1}^+}, \\ q_{i+1}^- + q_{i+1} - q_{i+1}^- & \text{if } x \in B_{q_{i+1}} \setminus B_{q_{i+1}^+} \end{cases} \\ &= \begin{cases} g_i(x) & \text{if } x \notin B_{q_{i+1}} \setminus B_{q_{i+1}^+}, \\ q_{i+1} & \text{if } x \in B_{q_{i+1}} \setminus B_{q_{i+1}^+}. \end{cases} \end{aligned}$$

This proves that $g_{i+1}(x) \in \overline{\mathbb{Q}}_{+(i+1)} = \overline{\mathbb{Q}}_{+i} \cup \{q_{i+1}\}$ for all $x \in X$. Furthermore, because $g_{i+1}(x) \geq g_i(x)$ for all $x \in X$, we have

$$\{g_{i+1} \geq q\} = \{g_i \geq q\} = B_q \quad \forall q \in \overline{\mathbb{Q}}_{+i} \colon q < q_{i+1}.$$

For the support $\{g_{i+1} > 0\}$, we have

$$\{g_{i+1} > 0\} \subseteq \underbrace{\{g_i > 0\}}_{\subseteq B_0} \cup \underbrace{B_{q_{i+1}}}_{\subseteq B_0} \subseteq B_0.$$

For $q \in \overline{\mathbb{Q}}_{+i}$ with $q > q_{i+1}$, we also have

$$\{g_{i+1} \geq q\} = \{g_i \geq q\} = B_q$$

because the value of g_i is not changed to a value exceeding q_{i+1} anywhere. This only leaves the set $\{g_{i+1} \geq q_{i+1}\}$ to be verified. Because the codomain of g_{i+1} is $\overline{\mathbb{Q}}_{+(i+1)}$, we have

$$\{g_{i+1} \geq q_{i+1}\} = \{g_{i+1} \geq q_{i+1}^+\} \cup \{g_{i+1} = q_{i+1}\}.$$

Here, we have to make a minor case distinction. If $q_{i+1} \in \overline{\mathbb{Q}}_{+i}$, i.e., if q_{i+1} is a repeat entry in $(q_i)_{i \in \mathbb{N}}$, then we have $q_{i+1}^- = q_{i+1}^+ = q_{i+1}$ and therefore $g_i = g_{i+1}$, which means that

$$\{g_{i+1} \geq q_{i+1}\} = \{g_i \geq q_{i+1}^-\} = B_{q_{i+1}^-} = B_{q_{i+1}}.$$

If $q_{i+1} \notin \overline{\mathbb{Q}}_{+i}$, then we have $q_{i+1}^- < q_{i+1} < q_{i+1}^+$ and therefore

$$\begin{aligned} \{g_{i+1} \geq q_{i+1}\} &= \{g_{i+1} \geq q_{i+1}^+\} \cup \{g_{i+1} = q_{i+1}\} \\ &= \{g_i \geq q_{i+1}^+\} \cup \{g_{i+1} = q_{i+1}\} \\ &= B_{q_{i+1}^+} \cup (B_{q_{i+1}} \setminus B_{q_{i+1}^+}) \\ &= \underbrace{B_{q_{i+1}^+}}_{\subseteq B_{q_{i+1}}} \cup B_{q_{i+1}} \\ &= B_{q_{i+1}}. \end{aligned}$$

This proves that Equations (2.39) and (2.40) hold for $i+1$. By complete induction, this implies that the same equations hold for all $i \in \mathbb{N}_0$.

The sequence $(g_i)_{i\in\mathbb{N}}$ is a pointwise monotonically increasing sequence of measurable functions. We define $g\colon X \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ to be

$$g(x) := \sup_{i\in\mathbb{N}_0} g_i(x) \quad \forall x \in X.$$

The function g is a pointwise supremum of countably many measurable functions and is therefore also measurable. For the support of g, we find that

$$\{g > 0\} = \bigcup_{i=0}^{\infty} \underbrace{\{g_i > 0\}}_{\subseteq B_0} \subseteq B_0,$$

which proves Property 2.3.27 (2).

Due to the way in which the approximating functions g_i are constructed, we have $B_\infty = \{g_i = \infty\}$ for all $i \in \mathbb{N}$. Because g is the pointwise supremum over all g_i, we certainly have

$$\{g = \infty\} \supseteq \bigcup_{i=0}^{\infty} \underbrace{\{g_i = \infty\}}_{=B_\infty} = B_\infty$$

which proves Property 2.3.27 (3).

Part 3 ($\{g \leq t\} \in \gamma(t)\ \forall t \in \mathbb{R}_{\geq 0}$). Let $t \in \mathbb{R}_{\geq 0}$. Because g is the pointwise supremum over all g_i, we have $g(x) > t$ for any given $x \in X$ if and only if there exists any $i \in \mathbb{N}_0$ such that $g_i(x) > t$, i.e.,

$$\{g > t\} = \bigcup_{i=0}^{\infty} \{g_i > t\}$$

For each $i \in \mathbb{N}_0$, the codomain of g_i is $\overline{\mathbb{Q}}_{+i}$. We can therefore write

$$\{g > t\} = \bigcup_{i=0}^{\infty} \{g_i > t\} = \{g_0 = \infty\} \cup \bigcup_{i=1}^{\infty} \bigcup_{\substack{j=1 \\ q_j > t}}^{\infty} \{g_i \geq q_j\} = \{g_0 = \infty\} \cup \bigcup_{i=1}^{\infty} \bigcup_{\substack{j=1 \\ q_j > t}}^{i} \{g_i \geq q_j\}.$$

For every $j \in \mathbb{N}$, we have $\{g_i \geq q_j\} = B_q$ for all $i \geq j$. Because $i \mapsto g_i(x)$ is also monotonically increasing for all $x \in X$, $i \mapsto \{g_i \geq q_j\}$ is also monotonically increasing. The union of a set sequence that is monotonically increasing and assumes a known constant after a given point is that same constant. We obtain the simplified form

$$\{g > t\} = \{g_0 = \infty\} \cup \bigcup_{i=1}^{\infty} \bigcup_{\substack{j=1 \\ q_j > t}}^{i} \{g_i \geq q_j\} = \{g_0 = \infty\} \cup \bigcup_{\substack{i=1 \\ q_i > t}}^{\infty} \{g_i \geq q_i\},$$

which gives us

$$\begin{aligned} \{g > t\} &= \{g_0 = \infty\} \cup \bigcup_{\substack{i=1 \\ q_i > t}}^{\infty} \{g_i \geq q_i\} \\ &= \underbrace{B_\infty}_{\subseteq B_{q_i}\ \forall i \in \mathbb{N}} \cup \bigcup_{\substack{i=1 \\ q_i > t}}^{\infty} B_{q_i} \end{aligned}$$

$$
\begin{aligned}
&= \bigcup_{\substack{i=1 \\ q_i > t}}^{\infty} B_{q_i} \\
&\in \bigcup_{\substack{i=1 \\ q_i > t}}^{\infty} \big(\gamma(q_i)\big)^{\complement} \\
&= \Big(\bigcap_{\substack{q \in \mathbb{Q}_+ \\ q > t}}^{\infty} \gamma(q) \Big)^{\complement}.
\end{aligned}
$$

We can select a sequence $(q_i^+)_{i\in\mathbb{N}} \in \mathbb{Q}^+$ such that $q_i^+ > t$ for all $i \in \mathbb{N}$ and $q_i^+ \to t$ for $i \to \infty$. Because γ is essentially monotonically increasing and upper semicontinuous, we obtain

$$
\bigcap_{\substack{q \in \mathbb{Q}_+ \\ q > t}}^{\infty} \gamma(q) = \bigcap_{i=1}^{\infty} \gamma(q_i^+) = \gamma\big(\lim_{i\to\infty} q_i^+\big) = \gamma(t).
$$

Therefore, we have

$$
\{g > t\} \in \Big(\bigcap_{\substack{q \in \mathbb{Q}_+ \\ q > t}}^{\infty} \gamma(q) \Big)^{\complement} = \big(\gamma(t)\big)^{\complement}.
$$

This implies that

$$
\{g \le t\} = \{g > t\}^{\complement} \in \gamma(t),
$$

which proves Property 2.3.27 (1). □

Lemma 2.3.27 allows us to construct one half of the GLSF from a continuous extension of the geodesic. The main challenge is then to combine the two halves of the GLSF and properly confine the codomain so that the GLSF does not map to any finite values outside of the parameter interval I.

Theorem 2.3.28 (Constructing GLSFs from Geodesics).
Let $(X,\ \Sigma,\ \mu)$ be a measure space, let $I \subseteq \mathbb{R}$ be an interval, let $\gamma\colon I \to \Sigma/_{\sim_\mu}$ be a canonical geodesic with geodesic constant $C \ge 0$. Then there exists a unique GLSF $g \in \mathcal{G}(X,\Sigma,\mu,I)$ such that

$$
\big[g \le t\big]_{\sim_\mu} = \gamma(t) \quad \forall t \in I.
$$

The constant C is a geodesic constant of g. ◁

PROOF. PART 1 (EXTENDING γ). The parameter set I is an interval and therefore convex. We can decompose $I^{\complement}$ into

$$
\begin{aligned}
I_+^{\complement} &:= \big\{ t \in I^{\complement} \,\big|\, t > s \ \forall s \in I \big\}, \\
I_-^{\complement} &:= \big\{ t \in I^{\complement} \,\big|\, t < s \ \forall s \in I \big\}.
\end{aligned}
$$

In the edge case where $I = \emptyset$, we have $I_+^{\complement} = I_-^{\complement} = \mathbb{R}$. Otherwise, we can guarantee that $I_+^{\complement}$ and $I_-^{\complement}$ are disjoint. We now extend γ to $\overline{\gamma}\colon \mathbb{R} \to \Sigma/_{\sim_\mu}$ as follows:

$$
\overline{\gamma}(t) := \left(\begin{cases} \gamma(t) & \text{if } t \in I, \\ [\emptyset]_{\sim_\mu} & \text{if } t \in I_-^{\complement}, \\ \mathrm{TV}(\gamma) & \text{if } t \in I_+^{\complement} \end{cases} \right) \quad \forall t \in \mathbb{R}.
$$

As $\mathrm{TV}(\gamma) = [\emptyset]_{\sim_\mu}$ for $I = \emptyset$, $\overline{\gamma}$ is still well-defined if $I = \emptyset$. Because γ is canonical, we have $[\emptyset]_{\sim_\mu} \subseteq_\mu \gamma(t) \subseteq_\mu \mathrm{TV}(\gamma)$ for all $t \in I$, which guarantees that $\overline{\gamma}$ is monotonically increasing.

We can also verify that $\overline{\gamma}$ is still continuous. In the interiors of I, $I_+^\complement$, and $I_-^\complement$, this is evident. At the boundaries between those intervals, if they exist, continuity follows from the fact that γ is canonical and therefore has the origin point $[\emptyset]_{\sim_\mu}$ and the destination point $\mathrm{TV}(\gamma)$. As we have shown in Lemma 2.3.26, continuity and monotonicity imply upper semicontinuity.

Part 2 (Partial Construction). Because γ is canonical, we always have $\overline{\gamma}(0) \subseteq_\mu \mathrm{TV}(\gamma)$, irrespective of which of the three branches of the definition of $\overline{\gamma}$ $t = 0$ falls into. Let B_0 and B_∞ be representatives of $\overline{\gamma}(0)$ and $\mathrm{TV}(\gamma)$, respectively, such that $B_0 \subseteq B_\infty$.

Using Lemma 2.3.27, we construct a measurable function $g^+ \colon X \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ such that

$$\begin{aligned} \{g^+ > 0\} &\subseteq B_0, \\ \{g^+ = \infty\} &\supseteq B_\infty, \\ \{g^+ \leq t\} &\in \overline{\gamma}(t) \qquad \forall t \in \mathbb{R}_{\geq 0}. \end{aligned}$$

Similarly, we construct a measurable function $g^- \colon X \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ such that

$$\begin{aligned} \{g^- > 0\} &\subseteq B_0^\complement, \\ \{g^- = \infty\} &\supseteq \emptyset, \\ \{g^- \leq t\} &\in \left(\overline{\gamma}(-t)\right)^\complement \quad \forall t \in \mathbb{R}_{\geq 0}. \end{aligned}$$

We note that $t \mapsto \overline{\gamma}(-t)$ is a monotonically decreasing continuous map, which means that $t \mapsto \left(\overline{\gamma}(-t)\right)^\complement$ is monotonically increasing and continuous. This ensures that we can apply Lemma 2.3.27 to the construction of g^- as well.

Part 3 (Combination). Because both g^+ and g^- are non-negative functions, their supports are $\{g^+ > 0\} \subseteq B_0$ and $\{g^- > 0\} \subseteq B_0^\complement$. This guarantees that the two supports are disjoint and that specifically, there exists no $x \in X$ such that $g^+(x) = \infty$ and $g^-(x) = \infty$. This means that we can subtract both functions to obtain a well-defined measurable function $\widetilde{g} \coloneqq g^+ - g^- \colon X \to \mathbb{R} \cup \{\pm\infty\}$. This function is generally not a GLSF yet, because it can map to $-\infty$ as well as to values in $\mathbb{R} \setminus I$. We have to correct this by adjusting $\widetilde{g}$ on a subset of X. Before we do so, we establish some basic properties of $\widetilde{g}$. Because g^+ and g^- are non-negative functions with disjoint supports, we have

$$\begin{aligned} \{\widetilde{g} = \infty\} &= \{g^+ = \infty\}, \\ \{\widetilde{g} = -\infty\} &= \{g^- = \infty\}, \\ \{\widetilde{g} \leq t\} &= \{g^+ \leq t\} \qquad \forall t \geq 0, \\ \{\widetilde{g} \leq t\} &= \{g^- \geq -t\} \quad \forall t < 0. \end{aligned}$$

For $t \in \mathbb{R}_{\geq 0}$, this already gives us $\{\widetilde{g} \leq t\} \in \overline{\gamma}(t)$. For $t \in \mathbb{R}_{<0}$, this requires a little bit of additional work. Let $(t_i)_{i \in \mathbb{N}} \in \mathbb{R}^\mathbb{N}$ be a sequence with $t_i > t$ for all $i \in \mathbb{N}$ and $t_i \to t$ for $i \to \infty$. Because $t < 0$, we may assume without loss of generality that

$t_i < 0$ for all $i \in \mathbb{N}$ We then find that

$$\begin{aligned}\{\widetilde{g} \le t\} &= \{g^- \ge -t\} \\ &= \{g^- < -t\}^{\complement} \\ &= \Big(\bigcup_{i=1}^{\infty} \{g^- \le -t_i\}\Big)^{\complement} \\ &= \bigcap_{i=1}^{\infty} \{g^- \le -t_i\}^{\complement} \\ &\in \bigcap_{i=1}^{\infty} \overline{\gamma}(t_i).\end{aligned}$$

Because $\overline{\gamma}$ is essentially increasing and because t_i converges to its infimum t, we can exploit the continuity of $\overline{\gamma}$ to obtain

$$\begin{aligned}\mu\Big(\overline{\gamma}(t) \triangle \big[\{\widetilde{g} \le t\}\big]_{\sim_\mu}\Big) &= \mu\Big(\overline{\gamma}(t) \triangle \Big(\bigcap_{i=1}^{\infty} \underbrace{\overline{\gamma}(t_i)}_{\supseteq_\mu \overline{\gamma}(t)}\Big)\Big) \\ &= \mu\Big(\bigcap_{i=1}^{\infty} \big(\overline{\gamma}(t_i) \setminus \overline{\gamma}(t)\big)\Big) \\ &\le \lim_{i \in \mathbb{N}} \mu\big(\overline{\gamma}(t_i) \setminus \overline{\gamma}(t)\big) \\ &= \lim_{i \in \mathbb{N}} \mu\big(\overline{\gamma}(t_i) \triangle \overline{\gamma}(t)\big) \\ &\xrightarrow{i \to \infty} 0.\end{aligned}$$

Therefore, we have

$$\{\widetilde{g} \le t\} \in \overline{\gamma}(t) \quad \forall t < 0.$$

In summary, we have

$$\{\widetilde{g} \le t\} \in \overline{\gamma}(t) \quad \forall t \in \mathbb{R},$$

which implies $\{\widetilde{g} \le t\} \in \gamma(t)$ for all $t \in I$.

PART 4 (ADJUSTING IMAGE). The function $\widetilde{g} \colon X \to \mathbb{R} \cup \{\pm\infty\}$ constructed during the previous part is measurable and satisfies

$$\{\widetilde{g} \le t\} \in \overline{\gamma}(t) \quad \forall t \in \mathbb{R}.$$

By adjusting $\widetilde{g}$ on part of X, we now create $g \colon X \to I \cup \{\infty\}$ with the same property. First, let $t \in \mathbb{R}$ and let $(t_i)_{i \in \mathbb{N}} \in \mathbb{R}^{\mathbb{N}}$ be a monotonically increasing sequence with $t_i \to t$ for $i \to \infty$. We can once more exploit the continuity of $\overline{\gamma}$ to show that

$$\begin{aligned}\mu\big(\{\widetilde{g} = t\}\big) &= \mu\big(\{\widetilde{g} \le t\} \setminus \{\widetilde{g} < t\}\big) \\ &= \mu\Big(\{\widetilde{g} \le t\} \setminus \bigcup_{i=1}^{\infty} \{\widetilde{g} \le t_i\}\Big) \\ &= \mu\Big(\bigcup_{i=1}^{\infty} \underbrace{\big(\{\widetilde{g} \le t\} \setminus \{\widetilde{g} \le t_i\}\big)}_{\text{ess. decreasing}}\Big) \\ &= \lim_{i \to \infty} \mu\big(\{\widetilde{g} \le t\} \setminus \{\widetilde{g} \le t_i\}\big)\end{aligned}$$

$$\begin{aligned}
&= \lim_{i\to\infty} \mu\big(\overline{\gamma}(t) \setminus \underbrace{\overline{\gamma}(t_i)}_{\subseteq_\mu \overline{\gamma}(t)}\big) \\
&= \lim_{i\to\infty} \mu\big(\overline{\gamma}(t) \mathbin{\triangle} \overline{\gamma}(t_i)\big) \\
&= 0.
\end{aligned}$$

This demonstrates that all exact level sets of g are nullsets. We treat the edge case $I = \emptyset$ first. In this case, we can simply set $g(x) = \infty$ for all $x \in X$ and trivially have $g \colon X \to I \cup \{\infty\}$ with

$$\{g(x) \leq t\} \in \gamma(t) \quad \forall t \in I.$$

Outside of this edge case, we may assume that there exists $t_0 \in I$. We can therefore find a monotonically decreasing sequence $(s_i)_{i\in\mathbb{N}} \in I^{\mathbb{N}}$ with $s_i \to \inf I$ for $i \to \infty$. We have

$$N_- := \big\{x \in X \,\big|\, \widetilde{g}(x) < t \ \forall t \in I\big\} \subseteq \{\widetilde{g} \leq \inf I\} = \bigcap_{i=1}^{\infty} \{\widetilde{g} \leq s_i\} \in \bigcap_{i=1}^{\infty} \gamma(s_i).$$

According to the premises of the theorem, γ is canonical and therefore has $[\emptyset]_{\sim_\mu}$ as an origin point. This means that

$$\bigcap_{i=1}^{\infty} \gamma(s_i) = \bigcap_{i=1}^{\infty} \big(\gamma(s_i) \mathbin{\triangle} [\emptyset]_{\sim_\mu}\big) = [\emptyset]_{\sim_\mu},$$

which means that N_- is a μ-nullset. Similarly, we can find a monotonically increasing sequence $(t_i)_{i\in\mathbb{N}} \in I^{\mathbb{N}}$ with $t_i \to \sup I$ for $i \to \infty$. We then define

$$N_+ := \big\{x \in X \,\big|\, \widetilde{g}(x) > t \ \forall t \in I\big\} \subseteq \bigcap_{i=1}^{\infty} \{\widetilde{g} > t_i\} = \Big(\bigcup_{i=1}^{\infty} \{\widetilde{g} \leq t_i\}\Big)^{\complement} \in \Big(\bigcup_{i=1}^{\infty} \gamma(t_i)\Big)^{\complement}.$$

Because γ is canonical, $\mathrm{TV}(\gamma)$ is a destination point of γ. The fact that $t_i \to \sup I$ for $i \to \infty$ implies that

$$\mathrm{TV}(\gamma) \mathbin{\triangle} \bigcup_{i=1}^{\infty} \underbrace{\gamma(t_i)}_{\subseteq_\mu \mathrm{TV}(\gamma)} = \bigcup_{i=1}^{\infty} \big(\mathrm{TV}(\gamma) \setminus \gamma(t_i)\big) = \bigcup_{i=1}^{\infty} \big(\mathrm{TV}(\gamma) \mathbin{\triangle} \gamma(t_i)\big) = [\emptyset]_{\sim_\mu},$$

and therefore that $\bigcup_{i=1}^{\infty} \gamma(t_i) = \mathrm{TV}(\gamma)$. We therefore have

$$N_+ \in \big(\mathrm{TV}(\gamma)\big)^{\complement}.$$

We now define $g \colon X \to I \cup \{\infty\}$ by

$$g(x) := \begin{cases} \infty & \text{if } x \in N_- \cup N_+, \\ \widetilde{g}(x) & \text{if } x \notin N_- \cup N_+ \end{cases} \quad \forall x \in X.$$

The codomain of g is $I \cup \{\infty\}$ because $N_- \cup N_+$ encompasses the entire preimage of $I^{\complement}$ under $\widetilde{g}$. For all $t \in I$, we have

$$\begin{aligned}
\big[\{g \leq t\}\big]_{\sim_\mu} &= \big[\{\widetilde{g} \leq t\} \setminus (N_- \cup N_+)\big]_{\sim_\mu} \\
&= \underbrace{\big[\{\widetilde{g} \leq t\}\big]_{\sim_\mu}}_{=\overline{\gamma}(t)=\gamma(t)} \setminus \underbrace{\big[N_-\big]_{\sim_\mu}}_{=[\emptyset]_{\sim_\mu}} \setminus \underbrace{\big[N_+\big]_{\sim_\mu}}_{=(\mathrm{TV}(\gamma))^{\complement}} \\
&= \underbrace{\gamma(t)}_{\subseteq_\mu \mathrm{TV}(\gamma)} \setminus \big(\mathrm{TV}(\gamma)\big)^{\complement} \\
&= \gamma(t).
\end{aligned}$$

This proves that $\{g \leq t\} \in \gamma(t)$ for all $t \in I$. In either case, whether $I = \emptyset$ or $I \neq \emptyset$, we have thus proven this property.

We now prove that g is a GLSF with geodesic constant C. Let $s,t \in I$ with $s \leq t$. Let $(s_i)_{i\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}$ be a sequence such that $s_i < s$ for all $i \in \mathbb{N}$ and $s_i \to s$ for $i \to \infty$. Without loss of generality, let $(s_i)_{i\in\mathbb{N}}$ be monotonically increasing. We have

$$g^{-1}\big([s,\ t]\big) = \bigcap_{i=1}^{\infty} g^{-1}\big((s_i,\ t]\big) = \bigcap_{i=1}^{\infty} \big(\{g \leq t\} \setminus \{g \leq s_i\}\big).$$

Here, $\big(\{g \leq s_i\}\big)_{i\in\mathbb{N}}$ is monotonically increasing because $(s_i)_{i\in\mathbb{N}}$ is monotonically increasing. Therefore, the sequence $\big(\{g \leq t\} \setminus \{g \leq s_i\}\big)_{i\in\mathbb{N}}$ is monotonically decreasing and we have

$$\mu\Big(g^{-1}\big([s,\ t]\big)\Big) = \inf_{i\in\mathbb{N}} \mu\big(\{g \leq t\} \setminus \{g \leq s_i\}\big) = \lim_{i\to\infty} \mu\big(\{g \leq t\} \setminus \{g \leq s_i\}\big).$$

At this point, we have to make a minor case distinction based on whether the s_i are in I or not. If $s > \inf I$, then $s_i \to s$ implies that there exists $i_0 \in \mathbb{N}$ with $s_i \in I$ for all $i \geq i_0$. In this case, we may assume without loss of generality that $s_i \in I$ for all $i \in \mathbb{N}$. We also need to take into account that $s_i < s \leq t$ implies that $\{g \leq s_i\} \subseteq \{g \leq t\}$. We then have

$$\mu\big(\{g \leq t\} \setminus \{g \leq s_i\}\big) = \mu\big(\{g \leq t\} \triangle \{g \leq s_i\}\big) = \mu\big(\gamma(t) \triangle \gamma(s_i)\big) = C \cdot |t - s_i|$$

for all $i \in \mathbb{N}$, which implies that

$$\mu\Big(g^{-1}\big([s,\ t]\big)\Big) = \lim_{i\to\infty} \mu\big(\{g \leq t\} \setminus \{g \leq s_i\}\big) = \lim_{i\to\infty} C \cdot |t - s_i| = C \cdot |t - s|.$$

In the edge case of $s = \inf I$, $s_i < s$ implies $s_i < r$ for all $r \in I$ for all $i \in \mathbb{N}$. This then implies that $\{g \leq s_i\} = \emptyset$ for all $i \in \mathbb{N}$ because the codomain of g is $I \cup \{\infty\}$. In this case, we have to take into account that γ is canonical and therefore has $[\emptyset]_{\sim_\mu}$ as an origin point. Because $s \in I$, γ assumes this value in s, i.e., we have $\gamma(s) = [\emptyset]_{\sim_\mu}$. This then gives us

$$\mu\big(\{g \leq t\} \setminus \{g \leq s_i\}\big) = \mu\big(\{g \leq t\} \triangle \emptyset\big) = \mu\big(\gamma(t) \triangle [\emptyset]_{\sim_\mu}\big) = \mu\big(\gamma(t) \triangle \gamma(s)\big) = C \cdot |t - s|,$$

which directly yields

$$\mu\Big(g^{-1}\big([s,\ t]\big)\Big) = \lim_{i\to\infty} \mu\big(\{g \leq t\} \setminus \{g \leq s_i\}\big) = C \cdot |t - s|.$$

In either case, we have

$$\mu\Big(g^{-1}\big([s,\ t]\big)\Big) = C \cdot |t - s| \quad \forall s,t \in I,$$

which means that g is a GLSF with geodesic constant C.

PART 5 (UNIQUENESS). Let $g' : X \to I \cup \{\infty\}$ be a GLSF such that

$$\{g' \leq t\} \in \gamma(t) \quad \forall t \in I.$$

We prove the claim of uniqueness by contradiction. If we were to assume that the set $D := \{g - g' \neq 0\}$ is a non-nullset, then we could partition the set D into $D_+ := \{g - g' > 0\}$ and $D_- := \{g - g' < 0\}$. If $\mu(D) > 0$, then at least one of these

parts would have strictly positive measure. Without loss of generality, we assume that that part is D_+. The remaining case is analogous with the roles of g and g' exchanged.

We would then have $\mu(D_+) > 0$. Because

$$D_+ = \bigcup_{i=1}^{\infty} \underbrace{\left\{ g - g' \geq \frac{1}{i} \right\}}_{=:D_{+i}},$$

the σ-additivity of μ would imply that there exists at least one $i_0 \in \mathbb{N}$ such that $\mu(D_{+i_0}) > 0$. We note that, by definition, for all $x \in D_{+i_0}$, we would have

$$g(x) \geq g'(x) + \frac{1}{i_0}.$$

We could then further subdivide D_{+i_0} into bands of width $\frac{1}{i_0}$ according to the value of g. Let

$$D_{+i_0,j} \coloneqq D_{+i_0} \cap \left\{ \frac{j}{i_0} < g \leq \frac{j+1}{i_0} \right\} \quad \forall j \in \mathbb{Z}.$$

We would then have

$$D_{+i_0} = \bigcup_{j \in \mathbb{Z}} D_{+i_0,j},$$

which would imply that there exists at least one $j_0 \in \mathbb{Z}$ such that $\mu(D_{+i_0,j_0}) > 0$. For all $x \in D_{+i_0,j_0}$, we have

$$\begin{aligned} g(x) &> \frac{j_0}{i_0}, \\ g(x) &\leq \frac{j_0+1}{i_0}, \\ g'(x) &\leq g(x) - \frac{1}{i_0} \\ &\leq \frac{j_0+1}{i_0} - \frac{1}{i_0} \\ &= \frac{j_0}{i_0}. \end{aligned}$$

Let $t_0 \coloneqq \frac{j_0}{i_0}$. We would have

$$D_{+i_0,j_0} \subseteq \{g' \leq t_0\} \setminus \{g \leq t_0\} \subseteq \{g' \leq t_0\} \triangle \{g \leq t_0\},$$

which would then imply that

$$\mu\big(\{g' \leq t_0\} \triangle \{g \leq t_0\}\big) > 0$$

and therefore $\{g' \leq t_0\} \not\sim_\mu \{g \leq t_0\}$. However, this would contradict our initial assumption that

$$\big[\{g' \leq t_0\}\big]_{\sim_\mu} = \gamma(t_0) = \big[\{g \leq t_0\}\big]_{\sim_\mu}.$$

In order to avert this contradiction, our initial assumption that $\mu(D) > 0$ must be false, i.e., we must have $g = g'$ pointwise almost everywhere. □

Theorems 2.3.24 and 2.3.28 show that GLSFs and canonical geodesics can be used interchangeably. This is useful because some operations are easier to perform on GLSFs than they are on geodesics directly. Notably, this includes *rearrangement*, which we will discuss in Section 2.3.4.

Rearranging a geodesic is an operation that is easy to grasp on an intuitive level. A geodesic transitions gradually from one set to another. In order to do this gradually, the changes necessary to make the transition, which consists of "flipping" the set membership of almost all points in the symmetric difference, must be brought into an order where some changes are made early in the parameter interval while others are deferred until later. Intuitively, a rearrangement takes the order encoded in one geodesic and shuffles the parameter interval such that the changes are made at different points along the parameter interval.

Rearrangement is much harder to grasp in a rigorous theoretical sense. The concept of changes made over a given parameter interval is quite intuitive: they are given by the symmetric difference between the ends of the interval. However, this only helps us if we want to rearrange a geodesic in blocks corresponding to parameter intervals. GLSFs allow for a much broader concept of rearrangement.

Because GLSF are measurable functions, they assign a measurable preimage not only to intervals, but to every Borel measurable set. Using GLSFs, we can therefore define changesets associated with arbitrary Borel measurable subsets of the parameter interval. In Section 2.3.4, we will see that we can essentially "compose" two geodesics by taking the preimages of sets produced by one geodesic under a GLSF representing the other. The GLSF of the rearranged geodesic is then simply the composition of the GLSFs of the two component geodesics. For such a concatenation to yield another GLSF, we need to extend the geodesic property encoded in Equation (2.36) from closed intervals to arbitrary Borel sets.

This extension is a three step process. In the first step, we use the fact that half-open and open intervals can be approximated internally using closed intervals. Having extended the property to all intervals, we can then be sure that it extends to every intersection between the parameter interval I and another interval. In the second step, we extend the property to all open and closed sets by exploiting the fact that these can be rewritten as countable unions of intervals. Finally, we use a well-known approximation theorem for Borel measurable sets to extend the property to all Borel measurable subsets of I.

Lemma 2.3.29 (Geodesic Property on Intervals).
Let $(X,\ \Sigma,\ \mu)$ be a measure space, let $I \subseteq \mathbb{R}$ be an interval, let $g \in \mathcal{G}(X,\Sigma,\mu,I)$, and let $C \geq 0$ be a geodesic constant of g. Then for every interval $J \subseteq I$, we have

$$\mu\big(g^{-1}(J)\big) = C \cdot \lambda(J)$$

where λ is the Lebesgue measure on $\mathbb{R}$. ◁

PROOF. Let $\overline{s} := \inf J$ and $\overline{t} := \sup J$. We distinguish several cases based on whether $\overline{s} \in I$ and $\overline{t} \in I$. We note that $\overline{s} \leq \overline{t}$ in all cases except when $J = \emptyset$.

Case 1 ($J = \emptyset$). In this case, we have $g^{-1}(J) = g^{-1}(\emptyset) = \emptyset$. Therefore, we have $\mu\big(g^{-1}(J)\big) = \mu(\emptyset) = 0 = C \cdot \lambda(\emptyset) = C \cdot \lambda(J)$. ◁

Case 2 ($\overline{s} \in J \wedge \overline{t} \in J$). In this case, J is a closed interval and we use the original equation $\mu\big(g^{-1}(J)\big) = \mu\big(g^{-1}\big([\overline{s},\ \overline{t}]\big)\big) = C \cdot (\overline{t} - \overline{s}) = C \cdot \lambda\big([\overline{s},\ \overline{t}]\big) = C \cdot \lambda(J)$. ◁

Case 3 ($\bar{s} \in J \wedge \bar{t} \notin J$). Because $J \neq \emptyset$, there is a monotonically increasing sequence $(t_i)_{i\in\mathbb{N}} \in J^{\mathbb{N}}$ with $t_i \to \bar{t}$ for $i \to \infty$. We have

$$J = \bigcup_{i=1}^{\infty} [\bar{s},\ t_i]$$

and therefore

$$g^{-1}(J) = \bigcup_{i=1}^{\infty} g^{-1}\big([\bar{s},\ t_i]\big)$$

where both $\big([\bar{s},\ t_i]\big)_{i\in\mathbb{N}}$ and $\big(g^{-1}([\bar{s},\ t_i])\big)_{i\in\mathbb{N}}$ are monotonically increasing set sequences. This implies

$$\mu\big(g^{-1}(J)\big) = \lim_{i\to\infty} \mu\Big(g^{-1}\big([\bar{s},\ t_i]\big)\Big) = \lim_{i\to\infty} C\cdot(t_i - \bar{s}) = C\cdot(\bar{t}-\bar{s}) = C\cdot\lambda\big([\bar{s},\ \bar{t})\big) = C\cdot\lambda(J).$$

This argument still holds if $\bar{t} = \infty$. ◁

Case 4 ($J \neq \emptyset \wedge \bar{s} \notin J$). This case encompasses cases where $\bar{t} \in J$, cases where $\bar{t} \notin J$, cases where $\bar{t} = \infty$, and cases where $\bar{s} = -\infty$. Let $(s_i)_{i\in\mathbb{N}} \in J^{\mathbb{N}}$ be a monotonically decreasing sequence with $s_i \to \bar{s}$ for $i \to \infty$. We have

$$J = \bigcup_{i=1}^{\infty} \big(J \cap [s_i,\ \infty)\big)$$

where $J \cap [s_i,\ \infty)$ is the sub-interval of J that encompasses all numbers greater than or equal to s_i. Because s_i is monotonically decreasing, this is a monotonically increasing set sequence and so is $g^{-1}\big(J \cap [s_i,\ \infty)\big)$. We can apply Cases 2 and 3 to show that

$$\begin{aligned}
\mu\big(g^{-1}(J)\big) &= \lim_{i\to\infty} \mu\big(g^{-1}(J \cap [s_i,\ \infty))\big) \\
&= C\cdot \lim_{i\to\infty} (\bar{t} - s_i) \\
&= C\cdot(\bar{t}-\bar{s}) \\
&= C\cdot\lambda(J).
\end{aligned}$$

□

In the next step, we approximate general open sets in $\mathbb{R}$ with countable unions of disjoint intervals. In one dimension, this is possible without incurring any errors, including errors of nonzero measure. This is well-known. An equivalent result can be found, e.g., in [Coh13, Prop. C.4].

Lemma 2.3.30 (Interval Representation of Open Sets in $\mathbb{R}$).
Let $U \subseteq \mathbb{R}$ be an open set. Then there are sequences $(s_i)_{i\in\mathbb{N}} \in \big(\mathbb{R}\cup\{-\infty\}\big)^{\mathbb{N}}$ and $(t_i)_{i\in\mathbb{N}} \in \big(\mathbb{R}\cup\{\infty\}\big)^{\mathbb{N}}$ such that the open intervals $(s_i,\ t_i)$ are pairwise disjoint and satisfy

$$U = \bigcup_{i=1}^{\infty} (s_i,\ t_i).$$

◁

PROOF. For every $t \in U$, there exists a radius $\varepsilon > 0$ such that $B_\varepsilon(t) \subseteq U$. Because $\mathbb{Q}$ is dense in $\mathbb{R}$, we have $\mathbb{Q} \cap B_\varepsilon(t) \neq \emptyset$. This is significant because $\mathbb{Q}$ is countably infinite. We use this fact to cover U with a countable family of open intervals. Let

$$s_t \coloneqq \inf\{s \in (-\infty,\, t] \,|\, [s,\, t] \subseteq U\} \qquad \forall t \in \mathbb{Q},$$
$$u_t \coloneqq \sup\{u \in [t,\, \infty) \,|\, [t,\, u] \subseteq U\} \qquad \forall t \in \mathbb{Q}.$$

For $t \in \mathbb{Q}$ with $t \notin U$, these are infimum and supremum of two empty sets, which means that $s_t = \infty$ and $u_t = -\infty$.

By definition, for every $t \in \mathbb{Q}$, we have $(s_t,\, u_t) \subseteq U$. If there existed $t \in \mathbb{Q}$ such that $s_t \in U$, then there would exist $\varepsilon > 0$ such that $(s_t - \varepsilon,\, s_t] \subseteq U$. This would imply $(s_t - \varepsilon,\, t] = (s_t - \varepsilon,\, s_t] \cup (s_t,\, t] \subseteq U$ which would contradict the definition of s_t. Similarly, if $u_t \in U$, then $[t,\, u_t + \varepsilon) \subseteq U$ for suitable $\varepsilon > 0$ which would also contradict the definition of u_t. These contradictions show that $s_t \notin U$ and $u_t \notin U$ for all $t \in \mathbb{Q}$.

Let $t_1 \in \mathbb{Q} \cap U$, and $t_2 \in \mathbb{Q}$ such that $t_2 \in (s_{t_1},\, u_{t_1})$. As we have previously shown, this implies $t_2 \in U$ and therefore $s_{t_2} < t_2 < u_{t_2}$. If we had $s_{t_2} \geq t_1$, then we would have

$$s_{t_2} \in [t_1,\, t_2] \subseteq (s_{t_1},\, u_{t_1}) \subseteq U,$$

which would contradict $s_{t_2} \notin U$. Therefore, we must have $s_{t_2} < t_1$. If we had $s_{t_1} < s_{t_2}$, then we would have $s_{t_2} \in [s_{t_2},\, t_1] \subseteq U$ by definition of s_{t_1}, which would also contradict $s_{t_2} \notin U$. Therefore, we must have $s_{t_2} \leq s_{t_1}$.

We can make a similar argument on the other end of the interval. If $u_{t_2} \leq t_1$, then $u_{t_2} \in [t_2,\, t_1] \subseteq U$, which would contradict $u_{t_2} \notin U$. Therefore, we must have $u_{t_2} > t_1$. If $u_{t_2} < u_{t_1}$, then $u_{t_2} \in [t_1,\, u_{t_2}] \subseteq U$ by definition of u_{t_1}, which would contradict $u_{t_2} \notin U$. Therefore, we must have $u_{t_2} \geq u_{t_1}$.

By aggregating these two estimates, we obtain

$$(s_{t_1},\, u_{t_1}) \subseteq (s_{t_2},\, u_{t_2}).$$

By exchanging the roles of t_1 and t_2, we obtain the converse inclusion and therefore

$$(s_{t_1},\, u_{t_1}) = (s_{t_2},\, u_{t_2}).$$

This proves that the intervals $(s_t,\, u_t)$ are always either disjoint or identical. If $t_1, t_2 \in \mathbb{Q} \cap U$ with

$$(s_{t_1},\, u_{t_1}) \cap (s_{t_2},\, u_{t_2}) \neq \emptyset,$$

then we can choose $t_3 \in (s_{t_1},\, u_{t_1}) \cap (s_{t_2},\, u_{t_2})$ and have

$$(s_{t_1},\, u_{t_1}) = (s_{t_3},\, u_{t_3}) = (s_{t_2},\, u_{t_2}).$$

Let $\mathcal{U} \coloneqq \{(s_t,\, u_t) \mid t \in \mathbb{Q}\}$. We have already shown that $\mathcal{U}$ is pairwise disjoint. The family $\mathcal{U}$ is also at most as large as $\mathbb{Q}$, which is countably infinite. We can therefore choose an enumeration $(U_i)_{i \in \mathbb{N}} \in \mathcal{U}^{\mathbb{N}}$ of $\mathcal{U}$. If $\mathcal{U}$ is finite, we can extend an enumeration of $\mathcal{U}$ with instances of $\emptyset$, which is an interval. We then have

$$\bigcup_{i=1}^{\infty} U_i = \bigcup_{I \in \mathcal{U}} I.$$

Let $t \in U$. Because U is open, there exists $\varepsilon > 0$ such that $B_\varepsilon(t) \subseteq U$, and because $\mathbb{Q}$ is dense in $\mathbb{R}$, there exists $t' \in B_\varepsilon(t) \cap \mathbb{Q}$. The fact that $B_\varepsilon(t) \subseteq U$ with $t' \in B_\varepsilon(t)$

means that $B_\varepsilon(t) \subseteq (s_{t'},\ u_{t'})$, which means that there is an element of $\mathscr{U}$ that contains t, i.e.,

$$t \in \bigcup_{I \in \mathscr{U}} I = \bigcup_{i=1}^{\infty} U_i.$$

We therefore have

$$U \subseteq \bigcup_{i=1}^{\infty} U_i.$$

Conversely, let $t \in U_i$ for some $i \in \mathbb{N}$. Because $t \in U_i$, we have $U_i \neq \emptyset$ and therefore $U_i \in \mathscr{U}$, which implies that there exists $t' \in \mathbb{Q}$ such that

$$t \in (s_{t'},\ u_{t'}) \subseteq U.$$

This implies

$$\bigcup_{i=1}^{\infty} U_i \subseteq U.$$

□

We can use this fact to extend the geodesic property of GLSFs to the preimages of both relatively open and relatively closed sets. In order to extend it to closed sets, we exploit the fact that $\big(\mathbb{R},\ \mathscr{B}(\mathbb{R}),\ \lambda\big)$ is σ-finite, i.e., that $\mathbb{R}$ can be written as a countable union of sets of finite measure.

Lemma 2.3.31 (Geodesic Property on Relatively Open or Closed Sets). *Let $(X,\ \Sigma,\ \mu)$ be a measure space, let $I \subseteq \mathbb{R}$ be an interval, let $g \in \mathscr{G}(X, \Sigma, \mu, I)$, and let $C \geq 0$ be a geodesic constant of g. For every set $A \subseteq \mathbb{R}$ that is either open or closed, we have*

$$\mu\big(g^{-1}(A \cap I)\big) = C \cdot \lambda(A \cap I).$$ ◁

Proof. Case 1 (A is an open set). According to Lemma 2.3.30, there exists a sequence of pairwise disjoint open intervals $(A_i)_{i \in \mathbb{N}}$ such that

$$A = \bigcup_{i=1}^{\infty} A_i.$$

For every $i \in \mathbb{N}$, $A_i \cap I$ is also an interval. These restricted intervals are also pairwise disjoint. Because disjointness is preserved by the preimage map, we can use the σ-additivity of μ and λ to show that

$$\begin{aligned}
\mu\big(g^{-1}(A \cap I)\big) &= \mu\Big(\bigcup_{i=1}^{\infty} g^{-1}(A_i \cap I)\Big) \\
&= \sum_{i=1}^{\infty} \mu\big(g^{-1}(A_i \cap I)\big) \\
&= \sum_{i=1}^{\infty} \big(C \cdot \lambda(A_i \cap I)\big) \\
&= C \cdot \sum_{i=1}^{\infty} \lambda(A_i \cap I) \\
&= C \cdot \lambda\Big(\bigcup_{i=1}^{\infty} (A_i \cap I)\Big) \\
&= C \cdot \lambda(A \cap I).
\end{aligned}$$ ◁

Case 2 (A is a closed set). If A is closed, then $A^{\complement}$ is open by definition. However, the formula $\mu(A \cap I) = \mu(I) - \mu(A^{\complement} \cap I)$ only holds if I is of finite measure. Therefore, we need to exploit the σ-finiteness of $\mathbb{R}$. Let $(s_i)_{i\in\mathbb{N}} \in \big(\mathbb{R} \cup \{-\infty\}\big)^{\mathbb{N}}$ and $(t_i)_{i\in\mathbb{N}} \in \big(\mathbb{R} \cup \{\infty\}\big)^{\mathbb{N}}$ be sequences such that

$$A^{\complement} = \bigcup_{i=1}^{\infty} (s_i,\ t_i).$$

Let further

$$I_j := I \cap \begin{cases} \{0\} & \text{if } j = 0, \\ (j-1,\ j] & \text{if } j > 0, \\ [j,\ j+1) & \text{if } j < 0 \end{cases} \qquad \forall j \in \mathbb{Z}.$$

Evidently, each I_j is the intersection of two intervals and therefore itself an interval. Moreover, the I_j are pairwise disjoint, each I_j has finite measure, and we have

$$I = \bigcup_{j\in\mathbb{Z}} I_j.$$

We then have

$$\begin{aligned}
\mu\big(g^{-1}(A \cap I)\big) = \mu\Big(g^{-1}\big(I \setminus (A^{\complement} \cap I)\big)\Big) &= \mu\Big(\bigcup_{j\in\mathbb{Z}} g^{-1}\big(I_j \setminus (A^{\complement} \cap I_j)\big)\Big) \\
&= \sum_{j\in\mathbb{Z}} \mu\Big(g^{-1}\big(I_j \setminus (A^{\complement} \cap I_j)\big)\Big) \\
&= \sum_{j\in\mathbb{Z}} \Big(\mu\big(g^{-1}(I_j)\big) - \mu\big(g^{-1}(A^{\complement} \cap I_j)\big)\Big) \\
&= \sum_{j\in\mathbb{Z}} \Big(\mu\big(g^{-1}(I_j)\big) - \sum_{i=1}^{\infty} \mu\Big(g^{-1}\big((s_i,\ t_i) \cap I_j\big)\Big)\Big) \\
&= \sum_{j\in\mathbb{Z}} \Big(C \cdot \lambda(I_j) - C \cdot \sum_{i=1}^{\infty} \lambda\big((s_i,\ t_i) \cap I_j\big)\Big) \\
&= C \cdot \sum_{j\in\mathbb{Z}} \big(\lambda(I_j) - \lambda(A^{\complement} \cap I_j)\big) \\
&= C \cdot \sum_{j\in\mathbb{Z}} \lambda(A \cap I_j) \\
&= C \cdot \lambda(A \cap I). \qquad \triangleleft
\end{aligned}$$

The last step is to transfer the geodesic property from open and closed sets to general Borel sets. This is possible because every Borel set can be approximated from inside by closed sets and from outside by open sets. For finite sets, this is somewhat technical to prove and we refer, for instance, to [Coh13, Prop. 1.4.1] for more detail. For infinite sets, we need to do some additional work.

Lemma 2.3.32 (Approximation of Borel Sets).
Let $B \in \mathscr{B}(\mathbb{R})$, let λ be the Lebesgue measure on $\mathscr{B}(\mathbb{R})$, and let $\varepsilon > 0$. Then there exist an open set $U_\varepsilon \subseteq \mathbb{R}$ and a closed set $F_\varepsilon \subseteq \mathbb{R}$ such that $F_\varepsilon \subseteq B \subseteq U_\varepsilon$ and

$$\lambda(U_\varepsilon \setminus F_\varepsilon) \leq \varepsilon. \qquad \triangleleft$$

Proof. Part 1 ($\lambda(B) < \infty$). We invoke [Coh13, Prop. 1.4.1] which states that

$$\lambda(B) = \inf\{\lambda(U) \mid B \subseteq U,\ U \subseteq \mathbb{R} \text{ open}\},$$
$$\lambda(B) = \sup\{\lambda(F) \mid F \subseteq B,\ F \subseteq \mathbb{R} \text{ closed}\}.$$

This implies that there exist an open set $U_\varepsilon \subseteq \mathbb{R}$ with $B \subseteq U_\varepsilon$ and $\lambda(U_\varepsilon) \leq \lambda(B) + \frac{\varepsilon}{2}$ and a closed set $F_\varepsilon \subseteq B$ with $\lambda(F_\varepsilon) \geq \lambda(B) - \frac{\varepsilon}{2}$. We then have $F_\varepsilon \subseteq B \subseteq U_\varepsilon$ and

$$\lambda(U_\varepsilon \setminus F_\varepsilon) = \lambda(U_\varepsilon \setminus B) + \lambda(B \setminus F_\varepsilon) = \underbrace{\lambda(U_\varepsilon) - \lambda(B)}_{\leq \frac{\varepsilon}{2}} + \underbrace{\lambda(B) - \lambda(F_\varepsilon)}_{\leq \frac{\varepsilon}{2}} \leq \varepsilon.$$

Part 2 ($\lambda(B) = \infty$). For each $i \in \mathbb{N}$, let $B_i \coloneqq B \cap (-i,\ i)$. Because B_i has finite measure, we can apply the first part of the proof to find an open set $U_{i,\varepsilon}$ and a closed set $F_{i,\varepsilon}$ such that $F_{i,\varepsilon} \subseteq B_i \subseteq U_{i,\varepsilon}$ and

$$\lambda(U_{i,\varepsilon} \setminus F_{i,\varepsilon}) \leq \frac{\varepsilon}{2^{i+1}}.$$

Because $U_{i,\varepsilon} \setminus B_i \subseteq U_{i,\varepsilon} \setminus F_{i,\varepsilon}$, this implies $\lambda(U_{i,\varepsilon} \setminus B_i) \leq \frac{\varepsilon}{2^{i+1}}$. The set

$$U_\varepsilon \coloneqq \bigcup_{i=1}^{\infty} U_{i,\varepsilon}$$

is a union of open sets and therefore itself open, and satisfies both

$$B = \bigcup_{i=1}^{\infty} \underbrace{B_i}_{\subseteq U_{i,\varepsilon}} \subseteq \bigcup_{i=1}^{\infty} U_{i,\varepsilon} = U_\varepsilon$$

and

$$\lambda(U_\varepsilon \setminus B) = \sum_{i=1}^{\infty} \lambda(\underbrace{U_{i,\varepsilon} \setminus B}_{\subseteq U_{i,\varepsilon} \setminus B_i}) \leq \sum_{i=1}^{\infty} \lambda(U_{i,\varepsilon} \setminus B_i) \leq \sum_{i=1}^{\infty} \frac{\varepsilon}{2^{i+1}} = \frac{\varepsilon}{2}.$$

Similarly, we can construct an open set $\overline{U}_\varepsilon \subseteq \mathbb{R}$ such that we have $B^\complement \subseteq \overline{U}_\varepsilon$ and $\lambda(\overline{U}_\varepsilon \setminus B^\complement) \leq \frac{\varepsilon}{2}$. As the complement of an open superset of $B^\complement$,

$$F_\varepsilon \coloneqq \overline{U}_\varepsilon^\complement$$

is a closed subset of B. It satisfies

$$\lambda(B \setminus F_\varepsilon) = \lambda(B \cap F_\varepsilon^\complement) = \lambda(F_\varepsilon^\complement \setminus B^\complement) = \lambda(\overline{U}_\varepsilon \setminus B^\complement) \leq \frac{\varepsilon}{2}.$$

We therefore have $F_\varepsilon \subseteq B \subseteq U_\varepsilon$ and

$$\lambda(U_\varepsilon \setminus F_\varepsilon) = \lambda(U_\varepsilon \setminus B) + \lambda(B \setminus F_\varepsilon) \leq \varepsilon. \qquad \square$$

This last approximation result allows us to transfer the geodesic property to arbitrary Borel sets. Because this is the largest σ-algebra such that the preimages under a GLSF are guaranteed to be measurable, this result effectively cannot be improved outside of special cases where the preimages of a larger σ-algebra are guaranteed to be measurable.

Theorem 2.3.33 (Geodesic Property on General Borel Sets).
Let $(X,\ \Sigma,\ \mu)$ be a measure space, let $I \subseteq \mathbb{R}$ be an interval, let $g \in \mathcal{G}(X,\Sigma,\mu,I)$, and let $C \geq 0$ be a geodesic constant of g. Then we have

$$\mu\big(g^{-1}(B)\big) = C \cdot \lambda(B) \quad \forall B \in \mathcal{B}(I). \qquad \triangleleft$$

PROOF. We first note that $\mathcal{B}(I) = \big\{B \cap I \,\big|\, B \in \mathcal{B}(\mathbb{R})\big\}$. It is therefore sufficient to show that

$$\mu\big(g^{-1}(B \cap I)\big) = C \cdot \lambda(B \cap I) \qquad \forall B \in \mathcal{B}(\mathbb{R}).$$

If $B \in \mathcal{B}(\mathbb{R})$, then $B \cap I \in \mathcal{B}(\mathbb{R})$ because $I \in \mathcal{B}(\mathbb{R})$. According to Lemma 2.3.32, for every $\varepsilon > 0$, there exist an open set $U_\varepsilon \subseteq \mathbb{R}$ and a closed set $F_\varepsilon \subseteq \mathbb{R}$ such that $F_\varepsilon \subseteq B \cap I \subseteq U_\varepsilon$ and $\lambda(U_\varepsilon \setminus F_\varepsilon) \leq \varepsilon$.

This notably implies that $\lambda\big((B \cap I) \setminus F_\varepsilon\big) \leq \lambda(U_\varepsilon \setminus F_\varepsilon) \leq \varepsilon$ for all $\varepsilon > 0$. It also implies that $\lambda\big((U_\varepsilon \cap I) \setminus (B \cap I)\big) \leq \lambda\big(U_\varepsilon \setminus (B \cap I)\big) \leq \lambda(U_\varepsilon \setminus F_\varepsilon) \leq \varepsilon$ for all $\varepsilon > 0$. Regardless of whether $\lambda(B \cap I) < \infty$ or not, we therefore have

$$\mu\big(g^{-1}(B \cap I)\big) \geq \mu\big(g^{-1}(F_\varepsilon)\big) = C \cdot \lambda(F_\varepsilon) \geq C \cdot \big(\lambda(B \cap I) - \varepsilon\big) \quad \forall \varepsilon > 0$$

and

$$\mu\big(g^{-1}(B \cap I)\big) \leq \mu\big(g^{-1}(U_\varepsilon \cap I)\big) = C \cdot \lambda(U_\varepsilon \cap I) \leq C \cdot (\lambda(B \cap I) + \varepsilon) \quad \forall \varepsilon > 0.$$

We therefore have

$$\mu\big(g^{-1}(B \cap I)\big) \in \big[C \cdot (\lambda(B \cap I) - \varepsilon),\ C \cdot (\lambda(B \cap I) + \varepsilon)\big] \quad \forall \varepsilon > 0$$

which implies

$$\mu\big(g^{-1}(B \cap I)\big) = C \cdot \lambda(B \cap I).$$

We note that this argument holds for all cases with $C = 0$ or $\lambda(B \cap I) = \infty$ as well. □

The transfer of the geodesic property to general Borel sets is significant because it allows us to chain the geodesic properties of two GLSFs together to show that the composition of two GLSFs is itself a GLSF.

Theorem 2.3.34 (Composition of GLSFs).
Let $(X,\ \Sigma,\ \mu)$ be a measure space, let $I, J \subseteq \mathbb{R}$ be intervals, let $g \in \mathcal{G}(X,\Sigma,\mu,I)$, and let $f \in \mathcal{G}(I,\mathcal{B}(I),\lambda,J)$. Then the map $f \odot g \colon X \to J \cup \{\infty\}$ with

$$(f \odot g)(x) \coloneqq \begin{cases} f\big(g(x)\big) & \text{if } g(x) < \infty, \\ \infty & \text{if } g(x) = \infty \end{cases} \qquad \forall x \in X$$

satisfies $f \odot g \in \mathcal{G}(X,\Sigma,\mu,J)$.

Let further $C_f \geq 0$ and $C_g \geq 0$ be geodesic constants of f and g, respectively. Then $f \odot g$ satisfies

$$\lambda\Big((f \odot g)\big([s,\ t]\big)\Big) = C_f C_g \cdot (t - s) \qquad \forall s,t \in J : s \leq t. \qquad \triangleleft$$

PROOF. We first have to show that $f \odot g$ is well defined. If $g(x) < \infty$, then $g \in \mathcal{G}(X, \Sigma, \mu, I)$ implies $g(x) \in I$. Therefore, $f\big(g(x)\big) \in J \cup \{\infty\}$ is well-defined. Next, we show that $f \odot g$ is measurable. Let $B \subseteq J \cup \{\infty\}$ be Borel measurable. We define

$$B' \coloneqq B \setminus \{\infty\}.$$

We note that, due to the way that we define Borel measurable sets that contain infinities, we have $B' \in \mathcal{B}(J)$. Because f is measurable, we have

$$\begin{aligned} f^{-1}(B) &= f^{-1}(B') \cup \begin{cases} f^{-1}\big(\{\infty\}\big) & \text{if } \infty \in B, \\ \emptyset & \text{if } \infty \notin B \end{cases} \\ &\in \{\underbrace{f^{-1}(B')}_{\in \mathcal{B}(I)},\ f^{-1}(B') \cup \underbrace{f^{-1}(\{\infty\})}_{\in \mathcal{B}(I)}\} \\ &\in \mathcal{B}(I). \end{aligned}$$

We then exploit the measurability of g, which gives us

$$\begin{aligned} (f \odot g)^{-1}(B) &= g^{-1}\big(f^{-1}(B)\big) \cup \begin{cases} g^{-1}\big(\{\infty\}\big) & \text{if } \infty \in B, \\ \emptyset & \text{if } \infty \notin B \end{cases} \\ &\in \Big\{\underbrace{g^{-1}\big(f^{-1}(B)\big)}_{\in \Sigma},\ g^{-1}\big(f^{-1}(B)\big) \cup \underbrace{g^{-1}\big(\{\infty\}\big)}_{\in \Sigma}\Big\} \\ &\in \Sigma. \end{aligned}$$

This shows that the preimages of Borel-measurable sets under $f \odot g$ are measurable, i.e., that $f \odot g$ is a measurable function.

Next, we show the geodesic property. Let $C_f \geq 0$ and $C_g \geq 0$ be geodesic constants for f and g, respectively. For all $s, t \in J$ with $s \leq t$, we have shown that $f^{-1}\big([s,\ t]\big) \in \mathcal{B}(I)$. The geodesic property for f states that

$$\lambda\Big(f^{-1}\big([s,\ t]\big)\Big) = C_f \cdot (t - s).$$

Because $f^{-1}\big([s,\ t]\big) \in \mathcal{B}(I)$, we have $(f \odot g)^{-1}\big([s,\ t]\big) = g^{-1}\big(f^{-1}\big([s,\ t]\big)\big) \in \Sigma$ and Theorem 2.3.33 states that

$$\mu\Big((f \odot g)^{-1}\big([s,\ t]\big)\Big) = \mu\Big(g^{-1}\Big(f^{-1}\big([s,\ t]\big)\Big)\Big) = C_g \cdot \lambda\Big(f^{-1}\big([s,\ t]\big)\Big) = C_f C_g \cdot (t - s).$$

Thus, $f \odot g$ is measurable and satisfies the geodesic property with geodesic constant $C_f C_g > 0$. We therefore have $f \odot g \in \mathcal{G}(X, \Sigma, \mu, J)$. □

We use the symbol "$\odot$" instead of "$\circ$" because $f \odot g$ is not a strict composition of f and g. The function g can map some points to ∞ and $f(\infty)$ is not defined. Therefore, we have to handle these points differently. Fortunately, the remedy of mapping ∞ to itself results in fairly intuitive behavior if we think of the composition in terms of the geodesics associated with the functions f, g, and $f \odot g$. Let those geodesics be $\phi\colon J \to \mathcal{B}(I)/_{\sim_\lambda}$, $\gamma\colon I \to \Sigma/_{\sim_\mu}$, and $\theta\colon J \to \Sigma/_{\sim_\mu}$, respectively.

The points that g maps to ∞ form a representative of the complement of $\mathrm{TV}(\gamma)$. We can think of them as points whose set membership is unaffected by γ. If we think of the composition θ as the geodesic that applies the set changes that

γ associates with the Borel measurable parameter sets produced by the geodesic ϕ, then it is evident that θ can only ever affect points that are also affected by γ. It is therefore appropriate that points unaffected by γ, i.e., those that are mapped to ∞ by g, would also be mapped to ∞ by $f \odot g$. Of course, it is not technically correct to speak of the effect that a geodesic has on a single point. Geodesics do not affect the set membership of nullsets in a well-defined way. However, it is helpful to think of this issue in those terms.

As we have already hinted at, the GLSF $f \odot g$ has an associated geodesic to which we referred as θ. This geodesic is a composition of the geodesic ϕ and γ. Through GLSFs, we can conceptualize the image of a general Borel set under γ. We can therefore reparameterize γ with an arbitrary geodesic ϕ in its parameter space. We introduce a general notation for this composition to simplify its use later on.

Definition 2.3.35 (Composition of Geodesics).
Let (X, Σ, μ) be a measure space, let $I, J \subseteq \mathbb{R}$ be intervals, and let $\gamma\colon I \to \Sigma/_{\sim_\mu}$ and $\phi\colon J \to \mathscr{B}(I)/_{\sim_\lambda}$ be geodesics. Let further $g \in \mathscr{G}(X, \Sigma, \mu, I)$ and $f \in \mathscr{G}(I, \mathscr{B}(I), \lambda, J)$ such that

$$\begin{aligned} \{g \leq t\} &\in \gamma(t) \qquad \forall t \in I, \\ \{f \leq t\} &\in \phi(t) \qquad \forall t \in J, \end{aligned}$$

and let $f \odot g \in \mathscr{G}(X, \Sigma, \mu, J)$ be as defined in Theorem 2.3.34. Then we refer to the geodesic $\gamma \odot \phi\colon J \to \Sigma/_{\sim_\mu}$ with

$$(\gamma \odot \phi)(t) := \big[\{f \odot g \leq t\}\big]_{\sim_\mu} \qquad t \in J$$

as the *composition of γ and ϕ* or the *parameterization of γ by ϕ*. ◁

We stress that for geodesics, the notation of the "$\odot$" operator is exactly reversed. This is because the composition of γ and ϕ is more intuitively thought of as a mapping $t \mapsto \gamma\big(\phi(t)\big)$. For GLSFs, this order of composition is reversed because geodesics are essentially the preimage maps of GLSFs, which behave in many ways like generalized inverse mappings.

Theorem 2.3.34 and Definition 2.3.35 are major theoretical achievements of this chapter. We had initially conceptualized a geodesic as an order in which changes to a set are performed. We had then formulated the intuition that there should be many orders in which a given set of changes can be applied. The composition of geodesics formalizes this intuition. Theorem 2.3.34 shows that a given geodesics can be reordered in at least as many ways as there are geodesics connecting the empty set with its full parameter interval. In Section 2.3.4, we will see that this is an exact characterization and that every rearrangement of a geodesic can be written in this way.

Composition is our first way to "modify" an existing geodesic. We will investigate other ways to modify geodesics Section 2.3.4. The ability to modify geodesics is intriguing because it suggests that geodesics could be constructed by iterative modification of a simple starting geodesic. We will address this possibility in Chapter 5. There, we will discuss the convergence issues that arise with such construction methods.

2.3.3.1 GENERATED SIMILARITY SPACE

Every canonical geodesic can be equivalently described by a GLSF, which is a measurable function. For measurable functions, there exists the concept of a *generated* or *induced* σ-algebra, which is the smallest σ-algebra that contains all preimages of measurable sets. For a GLSF, where the preimages of Borel sets reflect the changes made by a geodesic over that parameter set, that σ-algebra encompasses all changesets that we could achieve by some rearrangement of the original geodesic.

Knowing the generated σ-algebra of a GLSF is very valuable, because it tells us what we can achieve by rearranging a geodesic. However, there is a major problem when we apply information about the GLSF's generated σ-algebra to the corresponding geodesic. Geodesics never distinguish between sets that are equal up to a nullset, but a loss of nullsets can lead to a smaller σ-algebra. Notably, the Lebesgue and Borel σ-algebras on $\mathbb{R}^n$ are exactly equal up to nullset differences, but are generally held to be distinct σ-algebras. If we completely disregard nullsets, then two GLSFs representing the same geodesic could potentially generate different σ-algebras.

In this section, we want to demonstrate that these differences on nullsets do not affect the similarity space associated with a σ-algebra generated by a GLSF. If the similiarity space is stable with respect to such differences, then we can still use it. Fortunately, we can show that two families of sets that are equal except for differences on nullsets always generate σ-algebras that are equal except for nullset differences.

Lemma 2.3.36 (Similarity of Induced σ-Algebras).
Let (X, Σ, μ) be a measure space, and let $\mathscr{F}, \mathscr{G} \subseteq \Sigma$ be such that

$$\forall F \in \mathscr{F}\ \exists G \in \mathscr{G} \colon F \sim_\mu G.$$

Then we have

$$\forall F \in \sigma(\mathscr{F})\ \exists G \in \sigma(\mathscr{G}) \colon F \sim_\mu G. \qquad \triangleleft$$

PROOF. We prove the claim using the good set principle. The good set principle is a well-known proof technique for claims about generated σ-algebras: If a σ-algebra $\mathscr{S}$ contains $\mathscr{F}$, then the fact that $\sigma(\mathscr{F})$ is the smallest σ-algebra containing $\mathscr{F}$ implies that $\sigma(\mathscr{F}) \subseteq \mathscr{S}$.

In our case, that outer σ-algebra $\mathscr{S}$ has the form

$$\mathscr{S} := \left\{ F \in \sigma(\mathscr{F}) \,\middle|\, \exists G \in \sigma(\mathscr{G}) \colon F \sim_\mu G \right\}.$$

According to the premise of the lemma, for every $F \in \mathscr{F}$, there exists $G \in \mathscr{G}$ with $F \sim_\mu G$. Because $\mathscr{F} \subseteq \sigma(\mathscr{F})$ and $\mathscr{G} \subseteq \sigma(\mathscr{G})$, this means that $F \subseteq \mathscr{S}$. Because of the good set principle, it is sufficient to show that $\mathscr{S}$ is a σ-algebra. This then implies that $\sigma(\mathscr{F}) \subseteq \mathscr{S}$, which implies the conclusion of the theorem.

Because $\sigma(\mathscr{F})$ and $\sigma(\mathscr{G})$ are σ-algebras, they both contain $\emptyset$, which shows that $\emptyset \in \mathscr{S}$. Next, we show that $\mathscr{S}$ is closed under complementation. Let $F \in \mathscr{S}$. Because $\sigma(\mathscr{F})$ is a σ-algebra, we have $F^\complement \in \sigma(\mathscr{F})$. Because $F \in \mathscr{S}$, there exists $G \in \sigma(\mathscr{G})$ such that $F \sim_\mu G$. Because $\sigma(\mathscr{G})$ is a σ-algebra, we have $G^\complement \in \sigma(\mathscr{G})$. As we had shown earlier when we proved that complementation is well-defined on similarity spaces, $F \sim_\mu G$ implies $F^\complement \sim_\mu G^\complement \in \sigma(\mathscr{G})$ and therefore $F^\complement \in \mathscr{S}$.

Let $(F_i)_{i\in\mathbb{N}} \in \mathscr{S}^{\mathbb{N}}$. By definition of $\mathscr{S}$, there exist a sequence $(G_i)_{i\in\mathbb{N}} \in \big(\sigma(\mathscr{G})\big)^{\mathbb{N}}$ such that $F_i \sim_\mu G_i$ for all $i \in \mathbb{N}$. As we had shown earlier when we proved that countable union is well-defined on similarity spaces, this implies that

$$\bigcup_{i=1}^{\infty} F_i \sim_\mu \bigcup_{i=1}^{\infty} G_i.$$

Because

$$\bigcup_{i=1}^{\infty} F_i \in \sigma(\mathscr{F}), \qquad \bigcup_{i=1}^{\infty} G_i \in \sigma(\mathscr{G}),$$

this implies that

$$\bigcup_{i=1}^{\infty} F_i \in \mathscr{S}.$$

Overall, we find that $\mathscr{S}$ is a σ-algebra with $\mathscr{F} \subseteq \mathscr{S}$. Therefore, $\sigma(\mathscr{F}) \subseteq \mathscr{S}$, which means that

$$\forall F \in \sigma(\mathscr{F})\ \exists G \in \sigma(\mathscr{G})\colon F \sim_\mu G. \qquad \square$$

As a corollary of Lemma 2.3.36, we can show that the similarity spaces associated with the σ-algebras generated by two GLSFs representing the same geodesic are equal.

Theorem 2.3.37 (Uniqueness of the Generated Similarity Space).
Let $(X,\ \Sigma,\ \mu)$ be a measure space, let $I \subseteq \mathbb{R}$ be an interval, let $\gamma\colon I \to {}^{\Sigma}\!/_{\sim_\mu}$ be a canonical geodesic, and let $f, g \in \mathscr{G}(X,\Sigma,\mu,I)$ be such that

$$\{f \le t\} \in \gamma(t) \quad \forall t \in I,$$
$$\{g \le t\} \in \gamma(t) \quad \forall t \in I.$$

Then we have ${}^{\sigma(f)}\!/_{\sim_\mu} = {}^{\sigma(g)}\!/_{\sim_\mu}$. $\triangleleft$

PROOF. As we had shown in Theorem 2.3.28, $f = g$ up to a nullset, which implies that $f^{-1}(\{\infty\}) \sim_\mu g^{-1}(\{\infty\})$. Let

$$\mathscr{F} := \big\{\{f \le t\} \;\big|\; t \in I\big\} \cup \big\{f^{-1}(\{\infty\})\big\},$$
$$\mathscr{G} := \big\{\{g \le t\} \;\big|\; t \in I\big\} \cup \big\{g^{-1}(\{\infty\})\big\}.$$

We recall that $\sigma(f) = \sigma(\mathscr{F})$ and $\sigma(g) = \sigma(\mathscr{G})$. Since $f^{-1}(\{\infty\}) \sim_\mu g^{-1}(\{\infty\})$ and $\{f \le t\} \sim_\mu \{g \le t\}$ for all $t \in I$, we can apply Lemma 2.3.36 to show that

$$\forall F \in \sigma(f)\ \exists G \in \sigma(g)\colon F \sim_\mu G,$$
$$\forall G \in \sigma(f)\ \exists F \in \sigma(g)\colon F \sim_\mu G.$$

This implies ${}^{\sigma(f)}\!/_{\sim_\mu} \subseteq {}^{\sigma(g)}\!/_{\sim_\mu}$ and ${}^{\sigma(g)}\!/_{\sim_\mu} \subseteq {}^{\sigma(f)}\!/_{\sim_\mu}$, respectively. $\square$

This means that, given any geodesic γ, the similarity space associated with the generated σ-algebra of a GLSF that represents γ is the same, regardless of which GLSF we choose. We refer to this similarity space as the *generated similarity space* of the geodesic γ.

Definition 2.3.38 (Generated Similarity Space).
Let $(X,\ \Sigma,\ \mu)$ be a measure space, let $I \subseteq \mathbb{R}$ be an interval, let $\gamma\colon I \to {}^{\Sigma}\!/_{\sim_\mu}$ be a geodesic, and let $g \in \mathscr{G}(X,\Sigma,\mu,I)$ be such that

$$\{g \leq t\} \in \check{\gamma}(t) \quad \forall t \in I,$$

then we refer to

$$\sigma(\gamma) \coloneqq {}^{\sigma(g)}\!/_{\sim_\mu}$$

as the *similarity space generated by* γ. ◁

2.3.3.2 PUSHFORWARD AND PULLBACK

Our intention is to ultimately use geodesics as search directions in an iterative optimization scheme. This means that we have to accumulate the effects of derivatives along a geodesic. This is not necessarily a problem. Given a geodesic $\gamma\colon I \to {}^{\Sigma}\!/_{\sim_\mu}$ over a measure space $(X,\ \Sigma,\ \mu)$ and a μ-integrable function $f : X \to \mathbb{R}$, the mapping

$$t \mapsto \int_{\gamma(t)} f \,\mathrm{d}\mu$$

is a relatively benign mapping. It is evident that, because γ is continuous and because of the absolute continuity of the Lebesgue integral, the composition of both is still continuous. The integral over f is therefore something that can be considered "along a geodesic" with relative ease.

However, the composite mapping of a geodesic and an integral, although continuous, is not always differentiable. This means that there is not necessarily always a function $g\colon I \to \mathbb{R}$ such that

$$\int_{I\cap(-\infty,\ t]} g(s)\mathrm{d}s = \int_{\gamma(t)} f \,\mathrm{d}\mu \quad \forall t \in I.$$

In other words, although we can always consider the *integral of* f along a geodesic γ, the function f itself cannot necessarily always be considered as an integrable function along a geodesic. In this section, we investigate the circumstances under which this is possible. Transforming a function defined on a multi-dimensional domain into an equivalent function on a one-dimensional domain brings with it many benefits, some of which we will discuss in later sections.

We note that a geodesic can never resolve differences between individual points. Therefore, the transition from multi-dimensional to one-dimensional space should not be thought of as preserving the pointwise values. Instead, it can be thought of more accurately in the spirit of the Lebesgue differentiation theorem (see, e.g., [BC09, Thm. 8.2.4]), where a mean value over a sequence of decreasing balls is used to approximate the function value at the point to which they decrease. A decreasing sequence of parameter intervals passed through a geodesic can similarly give us a sequence of decreasing encompassing volumes such that the limit of the mean values over those volumes still accurately represents the value at the point to which those volumes contract.

Conceptually, this is easier to understand if we introduce the concepts of *push-forward* and *pullback*, both of which are well-established terms from measure theory and analysis. The following definition is adapted from [Bog07, Sec. 3.6].

Definition 2.3.39 (Pushforward Measure [Bog07, Sec. 3.6]).
Let $(X,\ \Sigma,\ \mu)$ be a measure space, let $(Y,\ \Pi)$ be a measurable space, and let $f\colon X \to Y$ be $(\Sigma,\ \Pi)$ measurable. The measure $f_*(\mu)\colon \Pi \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ with

$$f_*(\mu)(P) \coloneqq \mu\big(f^{-1}(P)\big) \qquad \forall P \in \Pi$$

is known as the *image of the measure μ under the mapping f*. We also refer to it as the *pushforward of μ* wherever the function f is evident. ◁

Because pushforward makes use of the preimage of the function f, it is evident that this definition relates closely with GLSFs. Let $I \subseteq \mathbb{R}$ be an interval and let $g \in \mathcal{G}(X, \Sigma, \mu, I)$. Then we have seen in Theorem 2.3.33 that the pushforward measure of μ satisfies

$$g_*(\mu)(B) = C \cdot \lambda(B) \qquad \forall B \in \mathscr{B}(I)$$

for an appropriately chosen constant $C \geq 0$. This is interesting if we take into account the change of variables formula.

Lemma 2.3.40 (Change of Variables).
Let $(X,\ \Sigma,\ \mu)$ be a measure space, let $(Y,\ \Pi)$ be a measurable space, let $g\colon X \to Y$ be $(\Sigma,\ \Pi)$ measurable. Then a function $f\colon Y \to \mathbb{R} \cup \{\pm\infty\}$ is $g_*(\mu)$-integrable if and only if $f \circ g\colon X \to \mathbb{R} \cup \{\pm\infty\}$ is μ-integrable and we have

$$\int_B f \,\mathrm{d}g_*(\mu) = \int_{g^{-1}(B)} f \circ g \,\mathrm{d}\mu \qquad \forall B \in \Pi. \qquad \triangleleft$$

PROOF. See [Coh13, Prop. 2.6.8] or [Bog07, Thm. 3.6.1]. □

This suggests the concept of what we will refer to as a *pullback function*. The word "pullback" here is meant to signify that the function transformation is converse to the pushforward transformation of the measure. The pushforward measure of μ is a measure on the parameter space of the geodesic, whereas the pullback transformation is a function in the underlying measure space $(X,\ \Sigma,\ \mu)$.

We note that this is, at first glance, precisely inverse to what we would expect pushforward and pullback to do. Intuitively, "pushing something forward" through a geodesic should transform a mapping on its parameter space into a mapping on the underlying measure space, while a pullback should work the other way around. The reversal results from the fact that the terms are derived from operations on the GLSF to which the geodesic is essentially the preimage function. We choose this wording so that our terminology aligns more closely with the traditional definition of "pushforward" measures (see, e.g., [Bog07, Sec. 3.6]), which is stated with respect to measurable functions as opposed to geodesics.

Theorem 2.3.41 (Pullback Function).
Let $(X,\ \Sigma,\ \mu)$ be a measure space, let $I \subseteq \mathbb{R}$ be an interval, let $g \in \mathcal{G}(X, \Sigma, \mu, I)$ have a strictly positive geodesic constant $C > 0$. For every measurable function $f\colon I \to \mathbb{R} \cup \{\pm\infty\}$, the map $f \otimes g\colon X \to \mathbb{R} \cup \{\pm\infty\}$ with

$$(f \otimes g)(x) \coloneqq \begin{cases} 0 & \text{if } g(x) = \infty \\ \frac{1}{C} f\big(g(x)\big) & \text{if } g(x) < \infty \end{cases} \qquad \forall x \in X$$

is measurable with $\sigma(f \otimes g) \subseteq \sigma(g)$ *and satisfies,*

$$\int_B |f| \, \mathrm{d}\lambda = \int_{g^{-1}(B)} |f \otimes g| \, \mathrm{d}\mu \quad \forall B \in \mathscr{B}(I), \tag{2.41}$$

$$\int_B f \, \mathrm{d}\lambda = \int_{g^{-1}(B)} (f \otimes g) \, \mathrm{d}\mu \quad \forall B \in \mathscr{B}(I) \colon \int_B |f| \, \mathrm{d}\lambda < \infty. \tag{2.42}$$

We will subsequently refer to $f \otimes g$ *as the* pullback function of f through g. $\triangleleft$

PROOF. We begin by proving that $f \otimes g$ is a measurable function. Let $B \in \overline{\mathscr{B}}(\mathbb{R})$. Then the preimage $(f \otimes g)^{-1}(B)$ can be derived by well-known rules for compositions and is given by

$$(f \otimes g)^{-1}(B) = g^{-1}\big(f^{-1}(C \cdot B)\big) \cup \begin{cases} g^{-1}\big(\{\infty\}\big) & \text{if } 0 \in B, \\ \emptyset & \text{if } 0 \notin B. \end{cases}$$

We note that $\pm\infty$ are invariant under multiplication with strictly positive real constants and $\mathscr{B}(\mathbb{R})$ is closed under linear transformation. Therefore, we have $C \cdot B \in \overline{\mathscr{B}}(\mathbb{R})$. Measurability of f and g implies $f^{-1}(C \cdot B) \in \mathscr{B}(I)$ and $g^{-1}\big(f^{-1}(C \cdot B)\big) \in \Sigma$ as well as $g^{-1}\big(\{\infty\}\big) \in \Sigma$. Overall, this implies $(f \otimes g)(B) \in \Sigma$. Because this holds for all $B \in \overline{\mathscr{B}}(\mathbb{R})$, $f \otimes g$ is measurable. Furthermore, because $(f \otimes g)^{-1}(B)$ is a union of preimages under g for every $B \in \overline{\mathscr{B}}(\mathbb{R})$, we have $\sigma(f \otimes g) \subseteq \sigma(g)$.

Having thus established that $f \otimes g$ is measurable, we proceed to show Equations (2.41) and (2.42). Let $B \in \mathscr{B}(I)$. Because $B \subseteq I$, we know that $g(x) < \infty$ for all $x \in g^{-1}(B)$, which implies that $(f \otimes g)(x) = \frac{1}{C} f\big(g(x)\big)$ for all $x \in g^{-1}(B)$. As we had discussed previously, the fact that g is a GLSF ensures that $g_*(\mu) = C \cdot \lambda$ on $\mathscr{B}(I)$. Therefore, we have

$$\int_B |f| \, \mathrm{d}\lambda = \frac{1}{C} \int_B |f| \, \mathrm{d}g_*(\mu).$$

If $\int_B |f| \, \mathrm{d}\lambda = \infty$, then Lemma 2.3.40 implies that $f \circ g$ is also not integrable on $g^{-1}(B)$. We therefore have

$$\int_{g^{-1}(B)} \underbrace{|f \otimes g|}_{= \frac{1}{C}(f \circ g)} \mathrm{d}\mu = \frac{1}{C} \int_{g^{-1}(B)} |f \circ g| \, \mathrm{d}\mu = \infty = \int_B |f| \, \mathrm{d}\lambda.$$

If $\int_B |f| \, \mathrm{d}\lambda < \infty$, then Lemma 2.3.40 implies

$$\begin{aligned} \int_{g^{-1}(B)} |f \otimes g| \, \mathrm{d}\mu &= \frac{1}{C} \int_{g^{-1}(B)} |f \circ g| \, \mathrm{d}\mu \\ &= \frac{1}{C} \int_{g^{-1}(B)} |f| \circ g \, \mathrm{d}\mu \\ &= \frac{1}{C} \int_B |f| \, \mathrm{d}g_*(\mu) \\ &= \int_B |f| \, \mathrm{d}\lambda \end{aligned}$$

which proves Equation (2.41).

For Equation (2.42), we assume that $\int_B |f| \, \mathrm{d}\lambda < \infty$. As shown, this implies $\int_{g^{-1}(B)} |f \otimes g| \, \mathrm{d}\mu < \infty$. Thus, both integrals are well-defined and we have

$$\int_{g^{-1}(B)} (f \otimes g) \, \mathrm{d}\mu = \frac{1}{C} \int_{g^{-1}(B)} (f \circ g) \, \mathrm{d}\mu = \frac{1}{C} \int_B f \, \mathrm{d}g_*(\mu) = \int_B f \, \mathrm{d}\lambda$$

which proves Equation (2.42). □

Theorem 2.3.41 proves that we can pull an integrable function defined on a geodesic's parameter space through the geodesic to obtain a corresponding integrable function in the measure space in which the geodesic is defined. We note that this is possible for every geodesic. A logical next question to ask is whether there is a converse operation where we take an integrable function in the underlying measure space and "push it forward" through the geodesic to obtain a corresponding function in the geodesic's parameter space.

A pushforward operation is harder to define because it is generally not possible to write it down as a composition. This is due to the fact that geodesic level set functions are not generally invertible. However, we can still perform a pushforward if the geodesic is chosen such that it can resolve all details of the integrable function to be pushed forward in the sense that it generates a similarity space that is a superset of that generated by the function.

Theorem 2.3.42 (Pushforward Function).
Let (X, Σ, μ) be a measure space, let $I \subseteq \mathbb{R}$ be an interval, let $f : X \to \mathbb{R} \cup \{\pm\infty\}$ be a measurable function, and let $g \in \mathcal{G}(X, \Sigma, \mu, I)$ be such that

$$\begin{aligned} &\sigma(f)/_{\sim_\mu} \subseteq \sigma(g)/_{\sim_\mu}, \\ &f(x) = 0 \qquad \text{for a.a. } x \in X \text{ with } g(x) = \infty, \end{aligned}$$

and such that there exists a strictly positive geodesic constant $C > 0$ of g. Then there is a measurable function $h : I \to \mathbb{R} \cup \{\pm\infty\}$ such that

$$(h \otimes g)(x) = C \cdot f(x) \quad \text{for a.a. } x \in X. \tag{2.43}$$

The function h is unique up to differences on nullsets and satisfies

$$\int_B |h| \, \mathrm{d}\lambda = \int_{g^{-1}(B)} |f| \, \mathrm{d}\mu \quad \forall B \in \mathcal{B}(I), \tag{2.44}$$

$$\int_B h \, \mathrm{d}\lambda = \int_{g^{-1}(B)} f \, \mathrm{d}\mu \quad \forall B \in \mathcal{B}(I) : \int_{g^{-1}(B)} |f| \, \mathrm{d}\mu < \infty. \tag{2.45}$$

We write $f \otimes g^{-1} := h$ and refer to h as the pushforward function of f through g. ◁

PROOF. Let subsequently $C > 0$ be a geodesic constant of the GLSF g. Due to Theorem 2.3.33, we know that

$$\mu\big(g^{-1}(B)\big) = C \cdot \lambda(B) \quad \forall B \in \mathcal{B}(I).$$

Our goal is to use the partial construction lemma for GLSFs (see Lemma 2.3.27) to construct the function h. To that end, we first have to construct a suitable essentially monotonic upper semicontinuous mapping γ to generate the desired non-strict sublevel sets for the positive and negative parts of h.

PART 1 (CONSTRUCTION OF h FOR $f \geq 0$). If $C > 0$, then we want to invoke Lemma 2.3.27. First, we note the following well-known identity for the generated σ-algebra of the measurable functions f and g:

$$\sigma(f) = \{f^{-1}(B) \,|\, B \in \overline{\mathscr{B}}(\mathbb{R})\},$$
$$\sigma(g) = \{g^{-1}(B) \,|\, B \in \overline{\mathscr{B}}(I)\}.$$

This identity stems from the fact that the right hand side is both the family that induces the generated σ-algebra and also a σ-algebra itself. This means that it is trivially the smallest σ-algebra containing itself. The second identity of note is that

$$\overline{\mathscr{B}}(\mathbb{R}) = \{A \cup B \,|\, A \in \mathscr{B}(\mathbb{R}),\ B \subseteq \{\pm\infty\}\},$$
$$\overline{\mathscr{B}}(I) = \{A \cup B \,|\, A \in \mathscr{B}(I),\ B \subseteq \{\pm\infty\}\}.$$

This means that we can always decompose sets in $\sigma(f)$ and $\sigma(g)$ the preimages of a Borel set, $\{\infty\}$, and $\{-\infty\}$ under the respective function. For g, this is further simplified by the fact that GLSFs never map to $-\infty$ and therefore $g^{-1}(\{-\infty\}) = \emptyset$.

With these identities in hand, let $t \in \mathbb{R}$. Because ${}^{\sigma(f)}\!/\!_{\sim_\mu} \subseteq {}^{\sigma(g)}\!/\!_{\sim_\mu}$, there exists $B_t \in \overline{\mathscr{B}}(I)$ such that

$$g^{-1}(B_t) \sim_\mu f^{-1}\big([-\infty,\ t]\big) = \{f \leq t\}.$$

Because g is a GLSF with geodesic constant $C > 0$, we can show that the Borel measurable part of B_t is essentially unique. Let $\widetilde{B}_t \in \overline{\mathscr{B}}(I)$ be another set such that

$$g^{-1}(\widetilde{B}_t) \sim_\mu \{f \leq t\}.$$

We may assume without loss of generality that $-\infty$ is contained in neither B_t nor $\widetilde{B}_t$ because g never maps to $-\infty$. Let

$$B'_t \coloneqq B_t \setminus \{\infty\}, \qquad \widetilde{B}'_t \coloneqq \widetilde{B}_t \setminus \{\infty\}.$$

Then we have

$$\begin{aligned}
\lambda(B'_t \triangle \widetilde{B}'_t) &= \frac{1}{C} \cdot \mu\big(g^{-1}(B'_t \triangle \widetilde{B}'_t)\big) \\
&= \frac{1}{C} \cdot \mu\Big(g^{-1}\big((B'_t \setminus \widetilde{B}'_t) \cup (B'_t \setminus \widetilde{B}'_t)\big)\Big) \\
&= \frac{1}{C} \cdot \mu\big(g^{-1}(B'_t \setminus \widetilde{B}'_t) \cup g^{-1}(B'_t \setminus \widetilde{B}'_t)\big) \\
&= \frac{1}{C} \cdot \mu\Big(\big(g^{-1}(B'_t) \setminus g^{-1}(\widetilde{B}'_t)\big) \cup \big(g^{-1}(B'_t) \setminus g^{-1}(\widetilde{B}'_t)\big)\Big) \\
&= \frac{1}{C} \cdot \mu\big(g^{-1}(B'_t) \triangle g^{-1}(\widetilde{B}'_t)\big).
\end{aligned}$$

We have

$$\begin{aligned}
g^{-1}(B'_t) &= g^{-1}\big(B_t \setminus \{\infty\}\big) \\
&= g^{-1}(B_t) \setminus g^{-1}\big(\{\infty\}\big) \\
&\sim_\mu \{f \leq t\} \setminus g^{-1}\big(\{\infty\}\big)
\end{aligned}$$

and similarly $g^{-1}(\widetilde{B}'_t) \sim_\mu \{f \le t\} \setminus g^{-1}(\{\infty\})$. Therefore, we have

$$g^{-1}(B'_t) \sim_\mu g^{-1}(\widetilde{B}'_t),$$

which implies that

$$\lambda(B'_t \triangle \widetilde{B}'_t) = \frac{1}{C} \cdot \mu\big(g^{-1}(B'_t) \triangle g^{-1}(\widetilde{B}'_t)\big) = \frac{1}{C} \cdot 0 = 0.$$

Thus, B'_t is λ-essentially unique for all $t \in \mathbb{R}$ and the mapping $\gamma\colon \mathbb{R} \to \mathscr{B}(I)/_{\sim_\lambda}$ with

$$\gamma(t) \coloneqq \big[B'_{\frac{1}{C}\cdot t}\big]_{\sim_\lambda} \quad \forall t \in \mathbb{R}$$

is well-defined. Let $s, t \in \mathbb{R}$ with $s \le t$. We have

$$\begin{aligned}
\lambda\big(\gamma(s) \setminus \gamma(t)\big) &= \lambda(B'_{s/C} \setminus B'_{t/C}) \\
&= \frac{1}{C} \cdot \mu\big(g^{-1}(B'_{s/C} \setminus B'_{t/C})\big) \\
&= \frac{1}{C} \cdot \mu\big(g^{-1}(B'_{s/C}) \setminus g^{-1}(B'_{t/C})\big) \\
&= \frac{1}{C} \cdot \mu\Big(\big(\{f \le s/C\} \setminus g^{-1}(\{\infty\})\big) \setminus \big(\{f \le t/C\} \setminus g^{-1}(\{\infty\})\big)\Big) \\
&= \frac{1}{C} \cdot \mu\big(\{f \le s/C\} \setminus \{f \le t/C\} \setminus g^{-1}(\{\infty\})\big) \\
&\le \frac{1}{C} \cdot \mu\big(\underbrace{\{f \le s/C\}}_{\subseteq \{f \le t/C\}} \setminus \{f \le t/C\}\big) \\
&= 0,
\end{aligned}$$

which means that $\gamma(s) \subseteq_\lambda \gamma(t)$. Therefore, γ is λ-essentially monotonically increasing. As a consequence, so is $t \mapsto (\gamma(-t))^{\complement}$. Let $t \in \mathbb{R}$ and let $(t_i)_{i\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}$ be a sequence such that $t_i \ge t$ for all $i \in \mathbb{N}$ and $t_i \to t$ for $i \to \infty$. Because γ is λ-essentially monotonically increasing, we have

$$\begin{aligned}
\lambda\Big(\gamma(t) \triangle \bigcap_{i=1}^{\infty} \gamma(t_i)\Big) &= \lambda\Big(\Big(\bigcap_{i=1}^{\infty} \gamma(t_i)\Big) \setminus \gamma(t)\Big) \\
&= \lambda\Big(\bigcap_{i=1}^{\infty} \big(\gamma(t_i) \setminus \gamma(t)\big)\Big) \\
&= \frac{1}{C} \cdot \mu\Big(\bigcap_{i=1}^{\infty} \big(\{f \le t_i/C\} \setminus \{f \le t/C\} \setminus g^{-1}(\{\infty\})\big)\Big) \\
&\le \frac{1}{C} \cdot \mu\Big(\bigcap_{i=1}^{\infty} \{t/C < f \le t_i/C\}\Big) \\
&= \frac{1}{C} \cdot \mu\Big(\big\{t/C < f \le \inf_{i\in\mathbb{N}}(t_i/C)\big\}\Big) \\
&= \frac{1}{C} \cdot \mu\big(\{t/C < f \le t/C\}\big) \\
&= 0.
\end{aligned}$$

This demonstrates that γ is upper semicontinuous in the sense of Definition 2.3.25.

We now choose $B_0 \in (\gamma(0))^{\complement}$ and $B_\infty := \emptyset$ and invoke Lemma 2.3.27 to obtain a measurable function $h : I \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ such that

$$\begin{aligned} &\{h \leq t\} \in \gamma(t) \quad \forall t \geq 0, \\ &\{h > 0\} \subseteq B_0. \end{aligned}$$

Next, we show that $(h \otimes g)(x) = f(x)$ almost everywhere. Let $t \in \mathbb{R}$. We have

$$\begin{aligned} \{h \otimes g \leq t\} &= g^{-1}\Big(\underbrace{\{h \leq C \cdot t\}}_{\in \gamma(Ct) = [B'_t]_{\sim_\mu}}\Big) \cup \begin{cases} g^{-1}(\{\infty\}) & \text{if } t \geq 0, \\ \emptyset & \text{if } t < 0 \end{cases} \\ &\overset{(*)}{\sim}_\mu \big(\{f \leq t\} \setminus \underbrace{g^{-1}(\{\infty\})}_{\subseteq_\mu \{f=0\}}\big) \cup \begin{cases} g^{-1}(\{\infty\}) & \text{if } t \geq 0, \\ \emptyset & \text{if } t < 0 \end{cases} \\ &\sim_\mu \big(\{f \leq t\} \setminus g^{-1}(\{\infty\})\big) \cup \big(\{f \leq t\} \cap g^{-1}(\{\infty\})\big) \\ &\sim_\mu \{f \leq t\}. \end{aligned}$$

This shows that the non-strict sublevel sets for every level in $\mathbb{R}$ are similar between f and $h \otimes g$. We can now make use of the fact that $f(x) \neq (h \otimes g)(x)$ implies that there exists $q \in \mathbb{Q}$ with $f(x) \leq q < (h \otimes g)(x)$ or $(h \otimes g)(x) \leq q < g(x)$. Because $\mathbb{Q}$ is countable, we can estimate that

$$\begin{aligned} \mu\big(\{f \neq h \otimes g\}\big) &= \mu\Big(\bigcup_{q \in \mathbb{Q}} \Big(\big(\{f \leq q\} \cap \{h \otimes g > q\}\big) \cup \big(\{h \otimes g \leq q\} \cap \{f > q\}\big)\Big)\Big) \\ &\leq \sum_{q \in \mathbb{Q}} \mu\Big(\big(\{f \leq q\} \setminus \{h \otimes g \leq q\}\big) \cup \big(\{h \otimes g \leq q\} \setminus \{f \leq q\}\big)\Big) \\ &= \sum_{q \in \mathbb{Q}} \underbrace{\mu\big(\{f \leq q\} \triangle \{h \otimes g \leq q\}\big)}_{=0} \\ &= 0. \end{aligned}$$

This proves that $h \otimes g = f$ pointwise almost everywhere.

Part 2 (Construction for General f). Let $N_1 \in \Sigma$ be a μ-nullset such that

$$f(x) = 0 \quad \forall x \in g^{-1}(\{\infty\}) \setminus N_1.$$

A suitable set N_1 exists according to the premises of the theorem. Because f is a measurable function, we can decompose $f : X \to \mathbb{R} \cup \{\pm\infty\}$ into two functions $f^+, f^- : X \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ via

$$f^+ := \max\{0,\, f\}, \qquad f^- := -\min\{0,\, f\}.$$

It is important to note that $F^+ := \operatorname{supp} f^+ \in \sigma(f)$ and $F^- := \operatorname{supp} f^- \in \sigma(f)$ are disjoint.

We invoke Part 1 to construct measurable functions $h^+, h^- : I \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ such that there exist μ-nullsets $N_2, N_3 \in \Sigma$ with

$$\begin{aligned} &(h^+ \otimes g)(x) = f^+(x) \qquad \forall x \in X \setminus N_2, \\ &(h^- \otimes g)(x) = f^-(x) \qquad \forall x \in X \setminus N_3. \end{aligned}$$

We now show that $N_4 := g^{-1}(\operatorname{supp} h^+ \cap \operatorname{supp} h^-) \subseteq N_2 \cup N_3$. Due to the fact that $\operatorname{supp} h^+ \subseteq I$ and $\operatorname{supp} h^- \subseteq I$, we certainly have $g(x) < \infty$ for all $x \in N_4$. This then implies that

$$g^{-1}(\operatorname{supp} h^+) = \{x \in X \mid h^+(g(x)) > 0\} = \{x \in X \mid (h^+ \otimes g)(x) > 0\} \subseteq F^+ \cup N_2$$

and similarly

$$g^{-1}(\operatorname{supp} h^-) \subseteq F^- \cup N_3.$$

We therefore have

$$\begin{aligned} N_4 &= g^{-1}(\operatorname{supp} h^+) \cap g^{-1}(\operatorname{supp} h^-) \\ &\subseteq (F^+ \cup N_2) \cap (F^- \cup N_3) \\ &= \underbrace{(F^+ \cap F^-)}_{=\emptyset} \cup \underbrace{(N_2 \cap F^-) \cup (N_3 \cap F^+) \cup (N_2 \cap N_3)}_{\subseteq N_2 \cup N_3} \\ &\subseteq N_2 \cup N_3. \end{aligned}$$

Thus, N_4 is also a μ-nullset. We now define $N := N_1 \cup N_2 \cup N_3$ and $h : I \to \mathbb{R} \cup \{\pm\infty\}$ via

$$h(t) := \begin{cases} h^+(t) - h^-(t) & \text{if } h^+(t) = 0 \text{ or } h^-(t) = 0, \\ 0 & \text{otherwise.} \end{cases}$$

Let $x \in X \setminus N$. If $g(x) = \infty$, then

$$(h \otimes g)(x) = 0 = f(x)$$

because $x \notin N_1$. If $g(x) < \infty$, then $g(x) \notin \operatorname{supp} h^+ \cap \operatorname{supp} h^-$ because $N_4 \subseteq N_2 \cup N_3$ and $x \notin N_2 \cup N_3$. Therefore, we have

$$\begin{aligned} (h \otimes g)(x) &= \frac{1}{C} h(g(x)) \\ &= \frac{1}{C} \big(h^+(g(x)) - h^-(g(x))\big) \\ &= (h^+ \otimes g)(x) - (h^- \otimes g)(x) \\ &= f^+(x) - f^-(x) \\ &= f(x). \end{aligned}$$

This shows Equation (2.43).

PART 3 (h IS UNIQUE). Let $h' : I \to \mathbb{R} \cup \{\pm\infty\}$ be measurable with $h' \otimes g = f$ almost everywhere. If we were to assume that there was a non-nullset $B \in \mathscr{B}(I)$ with $h'(x) \neq h(x)$ for all $x \in B$, then the fact that g is a GLSF would imply

$$\mu(g^{-1}(B)) = C \cdot \lambda(B) > 0.$$

Simultaneously, we would have $g(x) < \infty$ and

$$(h \otimes g)(x) = \frac{1}{C} h(g(x)) \neq \frac{1}{C} h'(g(x)) = (h' \otimes g)(x)$$

for all $x \in g^{-1}(B)$ which would contradict the fact that $h' \otimes g = f = h \otimes g$ almost everywhere. The contradiction shows that h is unique up to differences on nullsets.

Part 4 ((2.43) $\implies$ (2.44) $\wedge$ (2.45)). According to Theorem 2.3.41, we have

$$\int_B |h| \, \mathrm{d}\lambda = \int_{g^{-1}(B)} |h \otimes g| \, \mathrm{d}\mu = \int_{g^{-1}(B)} |f| \, \mathrm{d}\mu \qquad \forall B \in \mathscr{B}(I)$$

which shows Equation (2.44). For all $B \in \mathscr{B}(I)$ with $\int_B |h| \, \mathrm{d}\lambda < \infty$, Theorem 2.3.41 shows that

$$\int_B h \, \mathrm{d}\lambda = \int_{g^{-1}(B)} (h \otimes g) \, \mathrm{d}\mu = \int_{g^{-1}(B)} f \, \mathrm{d}\mu.$$

Since $\int_{g^{-1}(B)} |f| \, \mathrm{d}\mu = \int_B |h| \, \mathrm{d}\lambda$ for all $B \in \mathscr{B}(I)$, this proves Equation (2.45). □

In Section 2.3.4, we show that Theorems 2.3.41 and 2.3.42 can be used to show exactly when one geodesic can be made out of another through composition. We close this section by showing that for truly integrable functions, pushforward and pullback can be seen as inverse to one another.

Definition 2.3.43 (Pushforward And Pullback Operators).
Let (X, Σ, μ) be a measure space, let $I \subseteq \mathbb{R}$ be an interval, and let $g \in \mathscr{G}(X, \Sigma, \mu, I)$ have a strictly positive geodesic constant. Let

$$\begin{aligned} F_g &:= L^1(\mathscr{B}(I), \lambda), \\ F_{g^{-1}} &:= \left\{ f \in L^1(\Sigma, \mu) \;\middle|\; {}^{\sigma(f)}\!/_{\sim_\mu} \subseteq {}^{\sigma(g)}\!/_{\sim_\mu},\ f(x) = 0 \text{ for a.a. } x \in X \text{ with } g(x) = \infty \right\}. \end{aligned}$$

Then we refer to the operator $\cdot \otimes g \colon F_g \to F_{g^{-1}}$ as the *pullback operator through* g and to $\cdot \otimes g^{-1} \colon F_{g^{-1}} \to F_g$ as the *pushforward operator through* g. ◁

Theorem 2.3.44 (Relationship Between Pushforward And Pullback).
Let (X, Σ, μ) be a measure space, let $I \subseteq \mathbb{R}$ be an interval, and let $g \in \mathscr{G}(X, \Sigma, \mu, I)$ have a strictly positive geodesic constant. Let F_g, $F_{g^{-1}}$, $\cdot \otimes g$, and $\cdot \otimes g^{-1}$ be defined as in Definition 2.3.43. Then $\cdot \otimes g$ and $\cdot \otimes g^{-1}$ are bounded linear operators whose operator norms satisfy

$$\| \cdot \otimes g \|_{\mathscr{L}(F_g, F_{g^{-1}})} = \| \cdot \otimes g^{-1} \|_{\mathscr{L}(F_{g^{-1}}, F_g)} = 1.$$

Furthermore, they are inverse to one another. ◁

Proof. Let subsequently $C > 0$ be a strictly positive geodesic constant of g.

Part 1 ($\cdot \otimes g$). Let $\alpha, \beta \in \mathbb{R}$, let $f, h \in F_g$. By definition of $\cdot \otimes g$, we have

$$\begin{aligned} \big((\alpha f + \beta h) \otimes g\big)(x) &= \begin{cases} 0 & \text{if } g(x) = \infty \\ \frac{1}{C}\big(\alpha \cdot f(g(x)) + \beta h(g(x))\big) & \text{if } g(x) < \infty \end{cases} \\ &= \alpha \cdot (f \otimes g)(x) + \beta \cdot (h \otimes g)(x) \end{aligned}$$

for all $x \in X$.

To show that $\cdot \otimes g$ is bounded, we need only refer to Equation (2.41). For every $f \in F_g$, we have

$$\int_{g^{-1}(I)} |f \otimes g| \, \mathrm{d}\mu = \int_I |f| \, \mathrm{d}\lambda = \| f \|_{L^1(I, \mathscr{B}(I), \lambda)} < \infty.$$

and therefore

$$\|f \otimes g\|_{L^1(\Sigma,\mu)} = \int_X |f \otimes g| \, \mathrm{d}\mu = \int_{g^{-1}(I)} |f \otimes g| \, \mathrm{d}\mu + \int_{g^{-1}(\{\infty\})} \underbrace{|f \otimes g|}_{=0} \mathrm{d}\mu = \|f\|_{L^1(\mathscr{B}(I),\lambda)}.$$

This implies that $\cdot \otimes g \in \mathscr{L}(F_g, F_{g^{-1}})$ with $\|\cdot \otimes g\|_{\mathscr{L}(F_g, F_{g^{-1}})} = 1$.

PART 2 ($\cdot \otimes g^{-1}$). We use the uniqueness statement for the pushforward function that we had stated in Theorem 2.3.42. Let $f, h \in F_{g^{-1}}$, and let $\alpha, \beta \in \mathbb{R}$. Let $F \coloneqq \alpha \cdot (f \otimes g^{-1}) + \beta \cdot (h \otimes g^{-1}) \in F_g$.

According to the previous part of this proof, we have

$$F \otimes g = \alpha \cdot \big((f \otimes g^{-1}) \otimes g\big) + \beta \cdot \big((h \otimes g^{-1}) \otimes g\big) = \alpha f + \beta h.$$

Thus, F satisfies $F = (\alpha f + \beta h) \otimes g^{-1}$ and is pointwise unique almost everywhere.

To show that $\cdot \otimes g^{-1}$ is bounded, we once more use the equality of absolute integrals stated in Equation (2.44). For every $f \in F_{g^{-1}}$, we have

$$\|f \otimes g^{-1}\|_{L^1(\mathscr{B}(I),\lambda)} = \int_I |f \otimes g^{-1}| \, \mathrm{d}\lambda = \int_{g^{-1}(I)} |f| \, \mathrm{d}\mu \leq \int_X |f| \, \mathrm{d}\mu = \|f\|_{L^1(\Sigma,\mu)}$$

with exact equality being realized for all f that assume the value 0 outside of $g^{-1}(I)$. Since the indicator function $\chi_{g^{-1}(I)} \in F_{g^{-1}}$, we have

$$\|\cdot \otimes g^{-1}\|_{\mathscr{L}(F_{g^{-1}}, F_g)} = 1.$$

PART 3 (INVERSE RELATIONSHIP). As shown in Theorem 2.3.42, we have

$$\big((f \otimes g^{-1}) \otimes g\big)(x) = f(x) \qquad \text{for a.a. } x \in X$$

for all $f \in F_{g^{-1}}$.

Let $f \in F_g$. Then $f \otimes g \in F_{g^{-1}}$. Let $h \coloneqq (f \otimes g) \otimes g^{-1}$. We have

$$\begin{aligned} \int_I |f - h| \, \mathrm{d}\lambda &= \int_{g^{-1}(I)} |(f - h) \otimes g| \, \mathrm{d}\mu \\ &= \int_{g^{-1}(I)} \big|(f \otimes g) - \underbrace{(h \otimes g)}_{= f \otimes g}\big| \, \mathrm{d}\mu \\ &= 0 \end{aligned}$$

which proves that $f = h$ almost everywhere. □

2.3.4 Modifying Geodesics

In this section, we discuss four ways of modifying existing geodesics. We discuss modification before construction because some of these modifications are relevant to the construction methods discussed in Section 2.3.5.

2.3.4.1 REARRANGEMENT

We have already briefly addressed the possibility of rearranging a canonical geodesic in Section 2.3.3. In Theorem 2.3.34, we have shown that canonical geodesics can be composed with geodesics in their parameter space to produce rearranged versions of themselves. We now answer the question whether any given canonical geodesic can be rearranged into any other canonical geodesic.

It seems intuitively unlikely that there would be no limit whatsoever on this type of rearrangement. If we imagine a geodesic in $\mathbb{R}^2$ that produces only similarity classes that are essentially symmetric with respect to the x_1 axis, for instance, we would not expect there to be a rearrangement of that geodesic that breaks that symmetry.

In this section, we give a formal proof of this intuition by using pushforward functions and the concept of generated similarity spaces. Specifically, we show that the act of rearranging a canonical geodesic can only restrict its generated similarity space and can never expand it. However, we also show that this is the only limitation on rearrangement. If the generated similarity space of one geodesic is the same or smaller than another, then there always exists a parameter geodesic with which that other geodesic can be composed to produce it.

We prove this constructively. Let $\gamma\colon I \to {}^{\Sigma}\!/_{\sim_\mu}$ and $\phi\colon J \to {}^{\Sigma}\!/_{\sim_\mu}$ be canonical geodesics with $\sigma(\phi) \subseteq \sigma(\gamma)$ and let $g\colon X \to I \cup \{\infty\}$ and $f\colon X \to J \cup \{\infty\}$ be their respective GLSFs. We construct a GLSF for the parameter geodesic by modifying the pushforward function $p \coloneqq f \otimes g^{-1}$. The modification is very simple. We only need to redefine the value of p for points in which $g(x) = \infty$ because the pushforward function maps such points to 0, while a GLSF must map them to ∞. After this modification, p is a GLSF corresponding to a canonical geodesic $\rho\colon J \to {}^{\mathscr{B}(I)}\!/_{\sim_\lambda}$ such that $\gamma \odot \rho = \phi$.

For this type of construction to be possible, it is essential that the total variation of ϕ not be larger than that of γ, because no rearrangement of γ can ever affect sets of points not affected by γ itself. It appears counterintuitive that this would follow from $\sigma(\phi) \subseteq \sigma(\gamma)$. However, it is straightforward if we take into account that every subset of the total variation can be subdivided into arbitrarily small parts using the parameterization of the geodesic. If we take any Borel set $B \in \mathscr{B}(I)$ such that its preimage $g^{-1}(B)$ under g has strictly positive measure, then we can construct a Borel measurable subset of B whose preimage under g has strictly smaller but positive measure. However, we cannot do this for $\mathrm{TV}(\gamma)^{\complement}$ because $\mathrm{TV}(\gamma)^{\complement}$ is entirely mapped to the value ∞. It is therefore either a nullset or an atom.

Lemma 2.3.45 (Atomicity of $(\mathrm{TV}(\gamma))^{\complement}$).
Let (X, Σ, μ) be a measure space, let $I \subseteq \mathbb{R}$ be an interval, and let $g \in \mathscr{G}(X, \Sigma, \mu, I)$. The generated σ-algebra $\sigma(g)$ contains no non-empty true subsets of $g^{-1}(\{\infty\})$. For $A \in \sigma(g)$ with $\mu(A) > 0$, A is a μ-atom in $\sigma(g)$ if and only if $A \sim_\mu g^{-1}(\{\infty\})$. ◁

PROOF. Let $C \geq 0$ be a geodesic constant of g.

PART 1 ($\emptyset$ IS THE ONLY TRUE SUBSET OF $g^{-1}(\{\infty\})$ IN $\sigma(g)$). Since $\{\infty\}$ is a singleton, there exists no $B \subset \{\infty\}$ with $B \neq \emptyset$. This means that for all $B \in \overline{\mathscr{B}}(I)$ with $B \subseteq \{\infty\}$, we either have $g^{-1}(B) = g^{-1}(\{\infty\})$ or $g^{-1}(B) = g^{-1}(\emptyset) = \emptyset$. Since

$$\sigma(g) = \{g^{-1}(B) \mid B \in \overline{\mathscr{B}}(I)\},$$

every $A \in \sigma(g)$ has a corresponding $B \in \overline{\mathscr{B}}(I)$ such that $A = g^{-1}(B)$. If $A \in \sigma(g)$ satisfies $A \subseteq g^{-1}(\{\infty\})$, then its corresponding set B satisfies $B \subseteq \{\infty\}$ which implies either $A = \emptyset$ or $A = g^{-1}(\{\infty\})$.

PART 2 ($A \sim_\mu g^{-1}(\{\infty\}) \implies A$ IS μ-ATOM). Let $A \in \sigma(g)$ satisfy $\mu(A) > 0$. We partition A into $A_1 := A \setminus g^{-1}(\{\infty\})$ and $A_2 := A \cap g^{-1}(\{\infty\})$. Because $\sigma(g)$ contains no non-empty true subsets of $g^{-1}(\{\infty\})$, we have $A_2 \in \{\emptyset,\ g^{-1}(\{\infty\})\}$.

By definition, $A \sim_\mu g^{-1}(\{\infty\})$ implies

$$\mu(A_1) = \mu\big(\underbrace{A \setminus g^{-1}(\{\infty\})}_{\subseteq A \triangle g^{-1}(\{\infty\})}\big) \leq \mu\big(A \triangle g^{-1}(\{\infty\})\big) = 0$$

which implies $\mu(A_1) = 0$. Because $\mu(A_1) + \mu(A_2) = \mu(A) > 0$, this implies $\mu(A_2) > 0$ and therefore $A_2 = g^{-1}(\{\infty\})$.

Let $B \in \sigma(g)$ with $B \subseteq A$ and $\mu(B) > 0$. We partition B into $B_1 := B \setminus g^{-1}(\{\infty\})$ and $B_2 := B \cap g^{-1}(\{\infty\})$. $B_1 \subseteq A_1$ implies $\mu(B_1) \leq \mu(A_1) = 0$ and thus $\mu(B_1) = 0$. This then implies $B_2 = g^{-1}(\{\infty\})$ and

$$\mu(B) = \mu(B_1) + \mu(B_2) = \mu(g^{-1}(\{\infty\})) = \mu(A).$$

Because this holds for all subsets in $\sigma(g)$ that have non-zero measure, A is a μ-atom in $\sigma(g)$.

PART 3 (A IS μ-ATOM $\implies A \sim_\mu g^{-1}(\{\infty\})$). Let $A \in \sigma(g)$ with $\mu(A) > 0$. We partition A into $A_1 := A \setminus g^{-1}(\{\infty\})$ and $A_2 := A \cap g^{-1}(\{\infty\})$. Because $\sigma(g)$ contains no non-empty true subsets of $g^{-1}(\{\infty\})$, we have $A_2 \in \{\emptyset,\ g^{-1}(\{\infty\})\}$.

If $\mu(A_1) = 0$, then $\mu(A_1 \cup A_2) = \mu(A) > 0$ implies $\mu(A_2) > 0$ and thus $A_2 \neq \emptyset$. In this case, we have

$$\mu(A \triangle \underbrace{g^{-1}(\{\infty\})}_{=A_2}) = \mu\big((A_1 \cup A_2) \triangle A_2\big) = \mu(A_1) = 0$$

and therefore $A \sim_\mu A_2 = g^{-1}(\{\infty\})$.

If $\mu(A_1) > 0$, then we find $B \in \mathscr{B}(I)$ such that $A_1 = g^{-1}(B)$. We note that $0 < \mu(A_1) = \mu(g^{-1}(B)) = C \cdot \lambda(B)$ implies $C > 0$. We also know that $\infty \notin B$ because $A_1 \cap g^{-1}(\{\infty\}) = \emptyset$ by definition of A_1. We now define the map $\varphi\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ with

$$\varphi(r) := \lambda\big(B \cap (-r,\ r)\big) \qquad \forall r \in \mathbb{R}_{\geq 0}.$$

Because λ is continuous from both above and below, φ is continuous. It is evident that $\varphi(0) = 0$. If there existed no $r_1 > 0$ with $\varphi(r_1) > 0$, then we would have

$$\mu(A_1) = C \cdot \lambda(B) = C \cdot \lambda\Big(\bigcup_{i=1}^{\infty}\big(B \cap (-i,\ i)\big)\Big) = C \cdot \lim_{i \to \infty} \varphi(i) = 0,$$

which would contradict our prior assumption that $\mu(A_1) > 0$. There must therefore exist $r_1 > 0$ such that $\varphi(r_1) > 0$. If $\varphi(r_1) < \lambda(B)$, then we define $r_2 := r_1$. Otherwise, the intermediate value theorem dictates that there must exist $r_2 \in (0,\ r_1)$ such that $\varphi(r_2) = \frac{\varphi(r_1)}{2}$. Let

$$A' := g^{-1}\big(B \cap (-r_2,\ r_2)\big).$$

Since $B \cap (-r_2,\ r_2) \in \mathscr{B}(I)$, we have $A' \in \sigma(g)$. We also have

$$A' = g^{-1}(B \cap (-r_2,\ r_2)) \subseteq g^{-1}(B) = A_1 \subseteq A.$$

However, by invoking the geodesic property, we have

$$\mu(A') = C \cdot \lambda(B \cap (-r_2,\ r_2)) = C \cdot \varphi(r_2) < C \cdot \lambda(B) = \mu(A_1) \leq \mu(A).$$

Thus, A is not a μ-atom. Thus, if $\mu(A_1) > 0$, then A is not a μ-atom. Conversely, if A is a μ-atom, then $\mu(A_1) = 0$, which then implies that $A \sim_\mu g^{-1}(\{\infty\})$. □

This type of intermediate-value argument may have further application beyond Lemma 2.3.45. However, the result that $g^{-1}(\{\infty\})$ is always either an atom or a nullset for every GLSF g is sufficient for our purposes. If $f^{-1}(\{\infty\})$ was not an essential superset of $g^{-1}(\{\infty\})$, then the set $g^{-1}(\{\infty\}) \setminus f^{-1}(\{\infty\})$ would have subsets of non-zero measure in $\sigma(f)$ that would not be similar to any set in $\sigma(g)$. This would then imply $\sigma(\phi) \not\subseteq \sigma(\gamma)$.

Lemma 2.3.46 (Subset Relation of Total Variations).
Let $(X,\ \Sigma,\ \mu)$ be a measure space, let $I, J \subseteq \mathbb{R}$ be intervals, and let $\gamma\colon I \to {}^{\Sigma}\!/_{\sim_\mu}$ and $\phi\colon J \to {}^{\Sigma}\!/_{\sim_\mu}$ be canonical geodesics. If $\sigma(\phi) \subseteq \sigma(\gamma)$, then $\mathrm{TV}(\phi) \subseteq_\mu \mathrm{TV}(\gamma)$. ◁

Proof. Let $g \in \mathscr{G}(X, \Sigma, \mu, I)$ and $f \in \mathscr{G}(X, \Sigma, \mu, J)$ be GLSFs representing γ and ϕ, respectively. As we have previously argued, we have $g^{-1}(\{\infty\}) \in \mathrm{TV}(\gamma)^{\complement}$ and $f^{-1}(\{\infty\}) \in \mathrm{TV}(\phi)^{\complement}$. If $\mathrm{TV}(\phi) \not\subseteq_\mu \mathrm{TV}(\gamma)$, then

$$A \coloneqq \underbrace{g^{-1}(\{\infty\})}_{\in \mathrm{TV}(\gamma)^{\complement}} \setminus \underbrace{f^{-1}(\{\infty\})}_{\in \mathrm{TV}(\phi)^{\complement}}$$

has strictly positive measure. For all $x \in A$, we have $g(x) = \infty$ and $f(x) \in J$. Because $\mu(A) > 0$ and A because is disjoint from $f^{-1}(\{\infty\})$, Lemma 2.3.45 guarantees that A is not a μ-atom in $\sigma(f)$. There therefore exists $A' \in \sigma(f)$ with $A' \subseteq A$ and $\mu(A') \in (0,\ \mu(A))$.

For all $B \in \sigma(g)$ with $B \subseteq_\mu A \subseteq g^{-1}(\{\infty\})$ and $\mu(B) > 0$, Lemma 2.3.45 dictates that $g^{-1}(\{\infty\}) \subseteq B$ because $g^{-1}(\{\infty\})$ has no non-empty true subsets in $\sigma(g)$. This then implies $\mu(B) \geq \mu(g^{-1}(\{\infty\})) \geq \mu(A)$. Because $\mu(A) > \mu(A')$, this means that $B \not\sim_\mu A'$.

For all $B \in \sigma(g)$ with $B \sim_\mu A'$, we would have $\mu(B) = \mu(A') > 0$ and $B \subseteq_\mu A$, because $A' \subseteq A$. The fact that all such B satisfy $B \not\sim_\mu A'$ implies that there exists no $B \in \sigma(g)$ with $B \sim_\mu A'$. Thus, we have $\sigma(\phi) \not\subseteq \sigma(\gamma)$. This indirectly proves that if $\sigma(\phi) \subseteq \sigma(\gamma)$, then $\mathrm{TV}(\phi) \subseteq_\mu \mathrm{TV}(\gamma)$. □

Having thus established that a subset relationship between generated similarity spaces implies an essential subset relationship between total variations, we can guarantee that the indicated modification of the pushforward function on $g^{-1}(\{\infty\})$ essentially only affects points where $f(x) = \infty$ and therefore does not destroy any information substantial to the geodesic. We can proceed with proving the rearrangeability theorem.

Theorem 2.3.47 (Rearrangeability of Canonical Geodesics).
Let $(X,\ \Sigma,\ \mu)$ be a measure space, let $I, J \subseteq \mathbb{R}$ be intervals, let $\gamma\colon I \to {}^{\Sigma}\!/_{\sim_\mu}$ and $\phi\colon J \to {}^{\Sigma}\!/_{\sim_\mu}$ be canonical geodesics. There is a canonical geodesic $\rho\colon J \to {}^{\mathscr{B}(I)}\!/_{\sim_\lambda}$ such that $\phi = \gamma \odot \rho$ if and only if $\sigma(\phi) \subseteq \sigma(\gamma)$. ◁

PROOF. PART 1 ($\Longrightarrow$). In this part, we assume that there exists a canonical geodesic $\rho\colon J \to \mathscr{B}(I)/_{\sim_\lambda}$ such that $\phi = \gamma \odot \rho$. By definition, there exist GLSFs $g \in \mathscr{G}(X,\Sigma,\mu,I)$, $f \in \mathscr{G}(X,\Sigma,\mu,J)$, and $p \in \mathscr{G}(I,\mathscr{B}(I),\lambda,J)$ representing γ, ϕ, and ρ, respectively, and satisfying $f = p \odot g$ where the "$\odot$" operator is defined in Theorem 2.3.34.

By definition, we have $\sigma(\phi) = \sigma(f)/_{\sim_\mu}$ and $\sigma(\gamma) = \sigma(g)/_{\sim_\mu}$. It is therefore sufficient to show that for every $B \in \overline{\mathscr{B}}(J)$, there exists $B' \in \overline{\mathscr{B}}(I)$ with $g^{-1}(B') \sim_\mu f^{-1}(B)$.

Let $B \in \overline{\mathscr{B}}(J)$. We have

$$f^{-1}(B) \sim_\mu (p \odot g)^{-1}(B) = \underbrace{g^{-1}\big(\overbrace{p^{-1}(B)}^{\subseteq I}\big) \cup \begin{cases} g^{-1}(\{\infty\}) & \text{if } \infty \in B, \\ \emptyset & \text{if } \infty \notin B. \end{cases}}_{=:G}$$

Since $p^{-1}(B) \in \mathscr{B}(I)$, we have $g^{-1}\big(p^{-1}(B)\big) \in \sigma(g)$. In addition, we certainly have $g^{-1}(\{\infty\}) \in \sigma(g)$. Together, these imply

$$f^{-1}(B) \sim_\mu G \in \sigma(g) = \big\{g^{-1}(B') \,\big|\, B' \in \overline{\mathscr{B}}(I)\big\}.$$

This demonstrates that $\sigma(\phi) \subseteq \sigma(\gamma)$.

PART 2 ($\Longleftarrow$). In this part, we assume that $\sigma(\phi) \subseteq \sigma(\gamma)$. Let $f \in \mathscr{G}(X,\Sigma,\mu,J)$ and $g \in \mathscr{G}(X,\Sigma,\mu,I)$ be GLSFs representing ϕ and γ respectively.

We first address the edge cases in which g does not have a strictly positive geodesic constant. The preimage of the associated canonical geodesic's parameter interval under any GLSF that has 0 as a geodesic constant must be a nullset. Therefore, such GLSFs must map almost all points in X to ∞. Essentially, if g does not have a strictly positive geodesic constant, then

$$\mathrm{TV}(\gamma) = \big[g^{-1}(I)\big]_{\sim_\mu} = [\emptyset]_{\sim_\mu}.$$

However, we have already established in Lemma 2.3.46 that $\sigma(\phi) \subseteq \sigma(\gamma)$ implies $\mathrm{TV}(\phi) \subseteq_\mu \mathrm{TV}(\gamma)$. It would therefore follow that $\mathrm{TV}(\phi) = [\emptyset]_{\sim_\mu}$. Since both γ and ϕ are canonical, we would have $\gamma(t) = \phi(u) = [\emptyset]_{\sim_\mu}$ for all $t \in I$ and $u \in J$. We could then simply choose $\rho\colon J \to \mathscr{B}(I)/_{\sim_\lambda}$ with

$$\rho(u) \coloneqq [\emptyset]_{\sim_\lambda} \quad \forall u \in J$$

which would satisfy $\phi = \gamma \odot \rho$. Having handled this edge case, we will subsequently assume that g has a strictly positive geodesic constant $C_g > 0$.

According to Lemma 2.3.46, $\sigma(\phi) \subseteq \sigma(\gamma)$ implies $\mathrm{TV}(\phi) \subseteq_\mu \mathrm{TV}(\gamma)$. Because $f^{-1}(\{\infty\}) \in \mathrm{TV}(\phi)^{\complement}$ and $g^{-1}(\{\infty\}) \in \mathrm{TV}(\gamma)^{\complement}$, we have $g^{-1}(\{\infty\}) \subseteq_\mu f^{-1}(\{\infty\})$. Therefore,

$$N_\infty \coloneqq g^{-1}(\{\infty\}) \setminus f^{-1}(\{\infty\})$$

is a μ-nullset. Let $\widetilde{f}\colon X \to \mathbb{R} \cup \{\infty\}$ with

$$\widetilde{f}(x) \coloneqq \begin{cases} f(x) & \text{if } g(x) < \infty \\ 0 & \text{if } g(x) = \infty \end{cases} \qquad \forall x \in X.$$

We note that because $g^{-1}(\{\infty\})$ is measurable, the function $\widetilde{f}$ remains measurable. By definition, we have $\widetilde{f}(x) = 0$ for all $x \in X$ with $g(x) = \infty$. Let $B \in \overline{\mathscr{B}}(\mathbb{R})$. Since

$\sigma(f)/_{\sim_\mu} \subseteq \sigma(g)/_{\sim_\mu}$, there exists $Y \in \sigma(G)$ with $Y \sim_\mu f^{-1}(B)$. We define

$$\widetilde{Y} \coloneqq \begin{cases} Y \setminus g^{-1}(\{\infty\}) & \text{if } 0 \notin B, \\ Y \cup g^{-1}(\{\infty\}) & \text{if } 0 \in B. \end{cases}$$

It is evident that $\widetilde{Y} \in \sigma(g)$ in either case. We also have

$$\begin{aligned} \widetilde{f}^{-1}(B) &= \begin{cases} f^{-1}(B) \setminus g^{-1}(\{\infty\}) & \text{if } 0 \notin B \\ f^{-1}(B) \cup g^{-1}(\{\infty\}) & \text{if } 0 \in B \end{cases} \\ &\sim_\mu \widetilde{Y}. \end{aligned}$$

Therefore, for every, $B \in \overline{\mathscr{B}}(\mathbb{R})$, there exists $\widetilde{Y} \in \sigma(g)$ such that $\widetilde{Y} \sim_\mu \widetilde{f}^{-1}(B)$. This means that $\sigma(\widetilde{f})/_{\sim_\mu} \subseteq \sigma(g)/_{\sim_\mu}$.

Because $\sigma(\widetilde{f})/_{\sim_\mu} \subseteq \sigma(g)/_{\sim_\mu}$ and $\widetilde{f}(x) = 0$ for all $x \in X$ with $g(x) = \infty$, we can invoke Theorem 2.3.42 to show that there exists an essentially unique measurable pushforward function $\widetilde{p} \coloneqq \widetilde{f} \otimes g^{-1} \colon I \to \mathbb{R} \cup \{\pm\infty\}$ with

$$(\widetilde{p} \otimes g)(x) = \widetilde{f}(x) \quad \text{for a.a. } x \in X.$$

We now construct a GLSF from $\widetilde{p}$. Let

$$N \coloneqq \widetilde{p}^{-1}\Big(\big(\mathbb{R} \cup \{\pm\infty\}\big) \setminus \Big(\Big(\frac{1}{C_g} \cdot J\Big) \cup \{\infty\}\Big)\Big).$$

In simple terms, N is the set of all points where $C_g \cdot \widetilde{p}$ is outside of $J \cup \{\infty\}$, which is the desired codomain of the GLSF p representing the parameter geodesic ρ.

Because $(\widetilde{p} \otimes g)(x) = \widetilde{f}(x)$ almost everywhere and $\widetilde{f}(x) = f(x) \in J \cup \{\infty\}$ outside of $g^{-1}(\{\infty\})$, $N' \coloneqq N \setminus g^{-1}(\{\infty\})$ is a μ-nullset. We define $p \colon I \to J \cup \{\infty\}$ by

$$p(t) \coloneqq \begin{cases} C_g \cdot \widetilde{p}(t) & \text{if } t \notin N \\ \infty & \text{if } t \in N \end{cases} \quad \forall t \in I.$$

Because N is a measurable set, p is a measurable function. We include the factor C_g to compensate for the factor $\frac{1}{C_g}$ that is introduced in the definition of the pullback function to ensure equality of integrals. To prove that p is a GLSF, let $B \in \mathscr{B}(J)$. We have

$$p^{-1}(B) = \widetilde{p}^{-1}\Big(\underbrace{\frac{1}{C_g} \cdot B}_{\subseteq \frac{1}{C_g} \cdot J}\Big) \setminus N = \widetilde{p}^{-1}\Big(\frac{1}{C_g} \cdot B\Big).$$

Here, we make use of the fact that N is disjoint from the preimage of $\frac{1}{C_g} \cdot J$ under $\widetilde{p}$. Because $p^{-1}(B) \in \mathscr{B}(I)$, we can apply the geodesic property of g to obtain

$$\mu\Big(g^{-1}\big(p^{-1}(B)\big)\Big) = C_g \cdot \lambda\big(p^{-1}(B)\big).$$

Simultaneously, we have

$$\begin{aligned} g^{-1}\big(p^{-1}(B)\big) &= g^{-1}\Big(\widetilde{p}^{-1}\Big(\frac{1}{C_g} \cdot B\Big)\Big) \\ &= (\widetilde{p} \otimes g)^{-1}(B) \setminus g^{-1}\big(\{\infty\}\big) \\ &\sim_\mu \widetilde{f}^{-1}(B) \setminus g^{-1}\big(\{\infty\}\big). \end{aligned}$$

As we had noted, $\widetilde{f}^{-1}(B)$ equals either $f^{-1}(B) \setminus g^{-1}(\{\infty\})$ or $f^{-1}(B) \cup g^{-1}(\{\infty\})$ depending on whether $0 \in B$ or not. If we take into account Lemma 2.3.46, however, we know that $g^{-1}(\{\infty\}) \subseteq_\mu f^{-1}(\{\infty\})$. Since $\infty \notin B$, $f^{-1}(B)$ is μ-essentially disjoint from $g^{-1}(\{\infty\})$ and we have

$$g^{-1}\big(p^{-1}(B)\big) \sim_\mu \widetilde{f}^{-1}(B) \setminus g^{-1}\big(\{\infty\}\big) \sim_\mu f^{-1}(B).$$

Let $C_f \geq 0$ be a geodesic constant of f. Due to the geodesic property of f, we have

$$C_f \cdot \lambda(B) = \mu\big(f^{-1}(B)\big) = \mu\Big(g^{-1}\big(p^{-1}(B)\big)\Big) = C_g \cdot \lambda\big(p^{-1}(B)\big)$$

and therefore $\lambda\big(p^{-1}(B)\big) = \frac{C_f}{C_g} \cdot \lambda(B)$. Thus, $p \in \mathscr{G}(I, \mathscr{B}(I), \lambda, J)$ and the geodesic constant associated with p is $\frac{C_f}{C_g}$.

Finally, we show that the composition GLSF $p \odot g$ that we had defined in Theorem 2.3.34 satisfies $p \odot g = f$ almost everywhere. According to Theorem 2.2.20, it is sufficient to show $f^{-1}(B) \sim_\mu (p \odot g)^{-1}(B)$ for all $B \in \overline{\mathscr{B}}(\mathbb{R})$. Since f and $p \odot g$ both map to $J \cup \{\infty\}$, it is sufficient to do so for all $B \in \mathscr{B}(J)$ and $B = \{\infty\}$. We have already shown that

$$(p \odot g)^{-1}(B) = g^{-1}\big(p^{-1}(B)\big) \sim_\mu f^{-1}(B) \quad \forall B \in \mathscr{B}(J).$$

For $B = \{\infty\}$, we have

$$\begin{aligned}
(p \odot g)^{-1}(B) &= g^{-1}\Big(p^{-1}\big(\{\infty\}\big)\Big) \cup g^{-1}\big(\{\infty\}\big) \\
&= g^{-1}\Big(\widetilde{p}^{-1}\big(\{\infty\}\big)\Big) \cup g^{-1}(N) \cup g^{-1}\big(\{\infty\}\big) \\
&= g^{-1}\Big(\widetilde{p}^{-1}\big(\{\infty\}\big)\Big) \cup g^{-1}(N') \cup g^{-1}\big(\{\infty\}\big) \\
&\sim_\mu g^{-1}\Big(\widetilde{p}^{-1}\big(\{\infty\}\big)\Big) \cup g^{-1}\big(\{\infty\}\big).
\end{aligned}$$

Because $(\widetilde{p} \otimes g)(x) = \widetilde{f}(x)$ almost everywhere, $\widetilde{f}(x) = f(x)$ everywhere outside of $g^{-1}\big(\{\infty\}\big)$, $\widetilde{f}(x) \neq \infty$ for all $x \in g^{-1}\big(\{\infty\}\big)$, and $f(x) = \infty$ for almost all $x \in g^{-1}\big(\{\infty\}\big)$, we have

$$g^{-1}\Big(\widetilde{p}^{-1}\big(\{\infty\}\big)\Big) \cup g^{-1}\big(\{\infty\}\big) = (\widetilde{p} \otimes g)^{-1}\big(\{\infty\}\big) \cup g^{-1}\big(\{\infty\}\big) \sim_\mu f^{-1}\big(\{\infty\}\big).$$

and thus $(p \odot g)^{-1}(\{\infty\}) \sim_\mu f^{-1}(\{\infty\})$. Theorem 2.2.20 then shows that $p \odot g = f$ almost everywhere.

Let $\rho\colon I \to J \cup \{\infty\}$ be the canonical geodesic associated with p. Then $p \odot g$ is the GLSF associated with the composite geodesic $\gamma \odot \rho$. Since $p \odot g = f$ pointwise almost everywhere, we have

$$\phi(t) = \big[\{f \leq t\}\big]_{\sim_\mu} = \big[\{p \odot g \leq t\}\big]_{\sim_\mu} = (\gamma \odot \rho)(t) \qquad \forall t \in J. \qquad \square$$

2.3.4.2 REPARAMETERIZATION AND JUNCTION

Some reparameterizations of geodesics are themselves geodesics. This is notably the case for affine linear reparameterizations.

Theorem 2.3.48 (Reparameterization Of Geodesics).
Let $(X,\ \Sigma,\ \mu)$ be a measure space, let $I,J\subseteq\mathbb{R}$ be intervals, let $\gamma\colon I\to {}^{\Sigma}\!/_{\sim_\mu}$ be a geodesic, and let $\rho\colon J\to I$ be an affine linear map. Then $\gamma\circ\rho\colon J\to {}^{\Sigma}\!/_{\sim_\mu}$ is a geodesic.

Let $C\geq 0$ be a geodesic constant of γ and let $a,b\in\mathbb{R}$ be such that $\rho(t)=a\cdot t+b$ for all $t\in J$. Then $C\cdot|a|$ is a geodesic constant associated with $\gamma\circ\rho$. ◁

Proof. Because ρ is affine linear, there exist $a,b\in\mathbb{R}$ such that $\rho(t)=a\cdot t+b$ for all $t\in J$. Let $C\geq 0$ be a geodesic constant of γ. We have

$$\mu\Big(\gamma\big(\rho(s)\big)\mathbin{\triangle}\gamma\big(\rho(t)\big)\Big)=C\cdot\big|\rho(s)-\rho(t)\big|=C\cdot\big|a\cdot(s-t)\big|=C\cdot|a|\cdot|s-t|\quad\forall s,t\in J.$$

Therefore $\gamma\circ\rho$ is a geodesic with geodesic constant $C\cdot|a|$. □

It is important to note that Theorem 2.3.48 is not limited to either canonical geodesics or particular affine linear maps. The map need not be either surjective or injective for the theorem to hold. Using non-surjective maps allows us to restrict ourselves to certain subsections of the geodesic γ. By contrast, violations of injectivity only allow us to produce constant geodesics, which is unlikely to be useful.

Reparameterizations are also the only universally applicable way of reversing an infinitely long geodesic. This is generally not possible using only translation and rearrangement. Imagine a canonical, non-constant geodesic $\gamma\colon[0,\ \infty)\to {}^{\Sigma}\!/_{\sim_\mu}$. We can translate γ by $\mathrm{TV}(\gamma)$ to exchange its origin and destination points. However, this does not change the order in which changes are applied by γ. In order to rearrange γ to apply changes in reverse order, we would have to start at the beginning of the parameter interval with the changes previously applied at its end. However, we cannot select changes that are applied at $\pm\infty$. Even with reparameterization, we are limited to mapping one infinity to another.

Having the flexibility to restrict and transform the parameter intervals of geodesics allows us to join geodesics together. With the word *junction*, we will generally refer to end-to-end junctions, although we note that through reparameterization, this result can be extended to allow for junctions at parameter values other than 0 or concatenations. By continuously extending a geodesic to the boundary of its parameter interval, we can also apply the junction theorem to join geodesics on open ends of their parameter interval.

Theorem 2.3.49 (Junction of Geodesics).
Let $(X,\ \Sigma,\ \mu)$ be a measure space, let $I,J\subseteq\mathbb{R}$ be intervals with $\sup I=\inf J=0$ and $0\in I\cap J$, and let $\gamma\colon I\to {}^{\Sigma}\!/_{\sim_\mu}$ and $\phi\colon J\to {}^{\Sigma}\!/_{\sim_\mu}$ be geodesics that share a geodesic constant $C\geq 0$ such that $\mathrm{TV}(\gamma)$ and $\mathrm{TV}(\phi)$ are essentially disjoint. Let further $\gamma(0)=\phi(0)$. Then the map $\gamma\,|\,\phi\colon I\cup J\cup\{0\}$ with

$$(\gamma\,|\,\phi)(t):=\begin{cases}\gamma(t) & \text{if } t\leq 0,\\ \phi(t) & \text{if } t>0\end{cases}\qquad\forall t\in I\cup J$$

is a geodesic with geodesic constant C. ◁

Proof. We note that because $0=\sup I$ and $0=\inf J$, we have $I\cap J=\{0\}\neq\emptyset$. This implies that $I\cup J$ is an interval. Let $s,t\in I\cup J$. Without loss of generality, we have $s\leq t$. If $s\leq 0$ and $t\leq 0$, or if $s>0$ and $t>0$, then the geodesic property simply follows from the geodesic property of γ or ϕ, respectively.

We therefore only consider the case where $s \leq 0$ and $t > 0$. In this case, we have

$$\begin{aligned}
\mu\big((\gamma \mid \phi)(s) \mathbin{\triangle} (\gamma \mid \phi)(t)\big) &= \mu\big(\gamma(s) \mathbin{\triangle} \phi(t)\big) \\
&= \mu\big(\gamma(s) \mathbin{\triangle} \gamma(0) \mathbin{\triangle} \gamma(0) \mathbin{\triangle} \phi(t)\big) \\
&= \mu\big(\gamma(s) \mathbin{\triangle} \gamma(0) \mathbin{\triangle} \phi(0) \mathbin{\triangle} \phi(t)\big) \\
&= \mu\big(\gamma(s) \mathbin{\triangle} \gamma(0)\big) + \mu\big(\phi(0) \mathbin{\triangle} \phi(t)\big) \\
&\quad - 2 \cdot \mu\big(\underbrace{(\gamma(s) \mathbin{\triangle} \gamma(0))}_{\subseteq_\mu \mathrm{TV}(\gamma)} \cap \underbrace{(\phi(0) \mathbin{\triangle} \phi(t))}_{\subseteq_\mu \mathrm{TV}(\phi)}\big) \\
&= C \cdot (0 - s + t - 0) - 2 \cdot 0 \\
&= C \cdot (t - s) \\
&= C \cdot |s - t|. \qquad \square
\end{aligned}$$

2.3.4.3 RESTRICTION IN IMAGE

In the previous section, we have shown that affine linear reparameterizations of geodesics are themselves geodesics. Unsurprisingly, this includes any restriction of a geodesic to a sub-interval of its original parameter interval. In this section, we discuss a second, more obscure way of restricting a geodesics: *restriction in image*.

On several occasions, we have described geodesics as a way to impose an order on the changes necessary to gradually transform one set into another. The basic idea of a restriction in image is to preserve the order of changes while only considering changes to a specific subset of the universal set.

The best way of encoding the pure order of changes performed by a geodesic is to consider its canonical form. Let $(X,\ \Sigma,\ \mu)$ be a measure space, let $I \subseteq \mathbb{R}$ be an interval, let $\gamma\colon I \to \Sigma/_{\sim_\mu}$ be a geodesic, and let $R \in \Sigma/_{\sim_\mu}$. Imagine that we want to perform only the changes done by γ in the same order but simultaneously want to limit ourselves to only changes affecting essential subsets of R. The most straightforward way to achieve this would be to use the map $\gamma'\colon I \to \Sigma/_{\sim_\mu}$ with

$$\gamma'(t) \coloneqq \gamma(t) \cap R \qquad t \in I.$$

However, because the changes in R need not be equally distributed over the parameter interval I, γ' is not always a geodesic. We can remedy this by choosing a nonlinear reparameterization of γ' that passes more quickly over segments of the parameter interval where less changes occur in R.

The first step in designing such a parameterization is to determine how quickly the measure of the localized changes made by the geodesic rises. The measure of this restricted changeset rises continuously and monotonically. We refer to this measure as the *localized measure of variation*.

Lemma 2.3.50 (Localized Measure of Variation).
Let $(X,\ \Sigma,\ \mu)$ be a measure space, let either $I = [0,\ T]$ for some $T \geq 0$ or $I = [0,\ \infty)$, let $\gamma\colon I \to \Sigma/_{\sim_\mu}$ be a canonical geodesic, and let $R \in \Sigma/_{\sim_\mu}$. Then $\mu_{\gamma,R}\colon I \to \mathbb{R}_{\geq 0}$ with

$$\mu_{\gamma,R}(t) \coloneqq \mu\big(\gamma(t) \cap R\big) \qquad \forall t \in I$$

is uniformly Lipschitz-continuous and monotonically increasing. Accordingly, $\mu_{\gamma,R}(I) \subseteq \mathbb{R}_{\geq 0}$ with $\inf \mu_{\gamma,R}(I) = 0 \in \mu_{\gamma,R}(I)$. If $I = [0,\ T]$ for some $T \geq 0$, then $\mu_{\gamma,R}(I)$ is a closed interval. $\triangleleft$

PROOF. We note that $\mu_{\gamma,R}(t)$ is evidently well-defined for all $t \in I$ and satisfies $\mu_{\gamma,R}(t) \geq 0$ for all $t \in [a,\ b]$. Because $0 \in I$ and γ is canonical, $\gamma(0)$ must be the origin point of γ which is $[\varnothing]_{\sim_\mu}$. We therefore have

$$\mu_{\gamma,R}(0) = \mu(\gamma(0) \cap R) = \mu(\varnothing) = 0$$

which proves $\inf \mu_{\gamma,R}(I) = 0 \in \mu_{\gamma,R}(I)$.

To show that $\mu_{\gamma,R}(t)$ is monotonically increasing, let $s,t \in I$ with $s \leq t$. Because γ is canonical, we have $\gamma(s) \subseteq_\mu \gamma(t)$ and therefore $\gamma(s) \cap R \subseteq_\mu \gamma(t) \cap R$. It follows that

$$\mu_{\gamma,R}(s) = \mu\big(\gamma(s) \cap R\big) \leq \mu\big(\gamma(t) \cap R\big) = \mu_{\gamma,R}(t) \quad \forall s,t \in I \colon s \leq t.$$

To show that $\mu_{\gamma,R}$ is uniformly Lipschitz-continuous, let $C \geq 0$ be a geodesic constant for γ. Let $s,t \in I$. Without loss of generality, let $s \leq t$. Because we have already shown $\mu_{\gamma,R}$ to be monotonically increasing, we can then make the reformulation

$$\begin{aligned}
\big|\mu_{\gamma,R}(s) - \mu_{\gamma,R}(t)\big| &= \mu_{\gamma,R}(t) - \mu_{\gamma,R}(s) \\
&= \mu\big(\gamma(t) \cap R\big) - \mu\big(\underbrace{\gamma(s) \cap R}_{\subseteq_\mu \gamma(t) \cap R}\big) \\
&= \mu\Big(\big(\gamma(t) \cap R\big) \setminus \big(\gamma(s) \cap R\big)\Big) \\
&= \mu\Big(\big(\gamma(t) \cap R\big) \triangle \big(\gamma(s) \cap R\big)\Big) \\
&= \mu\Big(\big(\gamma(t) \triangle \gamma(s)\big) \cap R\Big) \\
&\leq \mu\big(\gamma(t) \triangle \gamma(s)\big) \\
&= C \cdot |s - t|.
\end{aligned}$$

If $s > t$, then we can exchange the roles of s and t to arrive at the same result.

To show that $J \coloneqq \mu_{\gamma,R}(I) \subseteq \mathbb{R}_{\geq 0}$ is an interval, we show that J is convex. Let $u,v \in J$. Without loss of generality, let $u \leq v$. By definition, there exist $s,t \in I$ such that $u = \mu_{\gamma,R}(s)$ and $v = \mu_{\gamma,R}(t)$. Because $\mu_{\gamma,R}$ is monotonically increasing, we have $s \leq t$. Furthermore, the continuity of $\mu_{\gamma,R}$ allows us to use the intermediate value theorem which states that

$$\forall w \in (u,v)\ \exists \tau \in \underbrace{(s,t)}_{\subseteq I} \colon \mu_{\gamma,R}(\tau) = w.$$

This implies that $w \in J$ for all $w \in (u,v)$. Therefore, J is convex.

Because $\mu_{\gamma,I}$ is uniformly continuous, if I is compact, then J is also compact. We have already proven that J is an interval. Therefore, J would be a closed interval in this case. □

The localized measure of variation is continuous and monotonically increasing. However, it is not straightforwardly invertible. This is because there may be parts of the parameter interval where no changes of strictly positive measure are made within the restriction set R. In these parts, $\mu_{\gamma,R}$ would plateau and there would be multiple parameter values with the same localized measure of variation.

This does not present an issue because the set $\gamma(t) \cap R$ remains essentially unchanged over such segments of I. Therefore, it is irrelevant which point from

the parameter segment we choose, since all of them yield the same similarity class.

Theorem 2.3.51 (Restriction in Image).
Let (X, Σ, μ) be a measure space, let either $I = [0, T]$ for $T \geq 0$ or $I = [0, \infty)$, let $\gamma\colon I \to \Sigma/_{\sim_\mu}$ be a canonical geodesic with geodesic constant $C \geq 0$, and let $R \in \Sigma/_{\sim_\mu}$. Let $\mu_{\gamma,R}$ be the localized measure of variation as defined in Lemma 2.3.50, let $J \coloneqq \mu_{\gamma,R}(I)$, and let $\tau_{\gamma,R}\colon J \to I$ be given by

$$\tau_{\gamma,R}(p) \coloneqq \inf\{t \in I \mid \mu_{\gamma,R}(t) \geq p\} \quad \forall p \in J.$$

Then the map $\gamma\|_R\colon J \to \Sigma/_{\sim_\mu}$ with

$$\gamma\|_R(p) \coloneqq \gamma(\tau_{\gamma,R}(p)) \cap R \quad \forall p \in J.$$

is a canonical geodesic with geodesic constant 1 and $\mathrm{TV}(\gamma\|_R) = \mathrm{TV}(\gamma) \cap R$. ◁

PROOF. PART 1 (PROPERTIES OF $\tau_{\gamma,R}$). For each $p \in J = \mu_{\gamma,R}(I)$, there exists $t_0 \in I$ such that $\mu_{\gamma,R}(t_0) = p$. Therefore, we have

$$\tau_{\gamma,R}(p) \in [\inf I,\ t_0] \subseteq I \quad \forall p \in J.$$

Because the infimum of a well-defined real set is well-defined, so is $\tau_{\gamma,R}$. Let $s, t \in J$ with $s \leq t$. We have $\mu_{\gamma,R}(p) < s < t$ for all $p < \tau_{\gamma,R}(s)$. Therefore $\tau_{\gamma,R}(t)$ is the infimum over a subset of the set that $\tau_{\gamma,R}(s)$ is the infimum of. This implies that $\tau_{\gamma,R}(t) \geq \tau_{\gamma,R}(s)$.

PART 2 ($\gamma\|_R$ IS AN INCREASING GEODESIC). Let $s, t \in J$. We assume without loss of generality that $s \leq t$, which implies $\tau_{\gamma,R}(s) \leq \tau_{\gamma,R}(t)$.

Because $\mu_{\gamma,R}$ is continuous and monotonically increasing, we always find that $\mu_{\gamma,R}(\tau_{\gamma,R}(s)) = s$ and $\mu_{\gamma,R}(\tau_{\gamma,R}(t)) = t$. We now exploit the fact that γ is canonical and $\tau_{\gamma,R}(s) \leq \tau_{\gamma,R}(t)$ to obtain

$$\gamma\|_R(s) = \underbrace{\gamma\big(\tau_{\gamma,R}(s)\big)}_{\subseteq_\mu \gamma(\tau_{\gamma,R}(t)}\cap R \subseteq_\mu \gamma\big(\tau_{\gamma,R}(t)\big) \cap R = \gamma\|_R(t).$$

The fact that $\gamma\|_R$ is μ-essentially increasing allows us to make the reformulation

$$\begin{aligned}
\mu\big(\gamma\|_R(s) \triangle \gamma\|_R(t)\big) &= \mu\big(\gamma\|_R(t) \setminus \gamma\|_R(s)\big) \\
&= \mu\Big(\gamma\big(\tau_{\gamma,R}(t)\big) \cap R\Big) - \mu\Big(\gamma\big(\tau_{\gamma,R}(s)\big) \cap R\Big) \\
&= \mu_{\gamma,R}\big(\tau_{\gamma,R}(t)\big) - \mu_{\gamma,R}\big(\tau_{\gamma,R}(s)\big) \\
&= t - s \\
&= |s - t|.
\end{aligned}$$

Thus, the geodesic property holds with geodesic constant 1.

PART 3 ($\gamma\|_R$ IS CANONICAL). We show that the origin point of $\gamma\|_R$ is $[\emptyset]_{\sim_\mu}$. This is straightforward because $\inf J = 0 \in J$. The origin point is therefore

$$\begin{aligned}
\gamma\|_R(0) &= \gamma\big(\tau_{\gamma,R}(0)\big) \cap R \\
&= \gamma\left(\inf\Big\{t \in I \,\Big|\, \mu\big(\gamma(t) \cap R\big) \geq 0\Big\}\right) \cap R \\
&= \gamma(\inf I) \cap R \\
&= \gamma(0) \cap R \\
&= [\emptyset]_{\sim_\mu}.
\end{aligned}$$

Therefore, the origin point of $\gamma\|_R$ is $[\emptyset]_{\sim_\mu}$. In conjunction with the fact that $\gamma\|_R$ is essentially monotonically increasing, this means that $\gamma\|_R$ is canonical.

Part 4 ($\mathbf{TV(\gamma\|_R) = TV(\gamma) \cap R}$). By definition of J, we have $\mu_{\gamma,R}(t) \in J$ for all $t \in I$. Let $p := \mu_{\gamma,R}(t) \in J$ and $s := \tau_{\gamma,R}(p) \in I$. Because s is an infimum of a set that includes t, we have $s \leq t$. The continuity of $\mu_{\gamma,R}$ further implies that

$$\mu\big(\gamma(s) \cap R\big) = p = \mu\big(\gamma(t) \cap R\big).$$

Because γ is canonical, we have $\gamma(s) \subseteq_\mu \gamma(t)$ and therefore

$$\mu\Big(\big(\gamma(s) \cap R\big) \triangle \big(\gamma(t) \cap R\big)\Big) = \mu\big(\gamma(t) \cap R\big) - \mu\big(\gamma(s) \cap R\big) = p - p = 0.$$

This implies that for every $t \in I$, there exists $p \in J$ with

$$\gamma\|_R(p) = \gamma\big(\tau_{\gamma,R}(p)\big) \cap R = \gamma(t) \cap R.$$

Let $(t_i)_{i\in\mathbb{N}} \subseteq I$ with $t_i \to \sup I$ for $i \to \infty$. Let $(p_i)_{i\in\mathbb{N}} \subseteq J$ be a sequence such that $\gamma\|_R(p_i) = \gamma(t_i) \cap R$ for all $i \in \mathbb{N}$. Because γ is canonical, we have

$$\mathrm{TV}(\gamma) \cap R = \bigcup_{i=1}^{\infty} \big(\gamma(t_i) \cap R\big) = \bigcup_{i=1}^{\infty} \gamma\|_R(p_i) \subseteq_\mu \mathrm{TV}(\gamma\|_R).$$

Conversely, let $(p_i)_{i\in\mathbb{N}} \subseteq J$ with $p_i \to \sup J$ for $i \to \infty$, then

$$\mathrm{TV}(\gamma\|_R) = \bigcup_{i=1}^{\infty} \gamma\|_R(p_i) = \bigcup_{i=1}^{\infty} \Big(\gamma\big(\tau_{\gamma,R}(p_i)\big) \cap R\Big) \subseteq_\mu \mathrm{TV}(\gamma) \cap R.$$

Together, this shows that $\mathrm{TV}(\gamma\|_R) = \mathrm{TV}(\gamma) \cap R$. □

2.3.4.4 INTERLEAVING

The final type of geodesic modification that we will discuss here is interleaving. In interleaving, we merge two geodesics into one that applies the changes from both at the same time.

Definition 2.3.52 (Interleaving Geodesics).
Let (X, Σ, μ) be a measure space, let $I \subseteq \mathbb{R}$ be an interval, let $\gamma\colon I \to \Sigma/_{\sim_\mu}$ and $\phi\colon I \to \Sigma/_{\sim_\mu}$ be canonical geodesics such that $\mathrm{TV}(\gamma)$ and $\mathrm{TV}(\phi)$ are essentially disjoint. Then we refer to the map $\gamma \parallel \phi\colon I \to \Sigma/_{\sim_\mu}$ with

$$(\gamma \parallel \phi)(t) := \gamma(t) \cup \phi(t) \qquad \forall t \in I$$

as the geodesic produced by *interleaving γ and ϕ*. ◁

It is fairly easy to prove that the geodesic produced by interleaving two canonical geodesics with disjoint total variations is a geodesic.

Theorem 2.3.53 (Interleaving Geodesics).
Let (X, Σ, μ) be a measure space, let $I \subseteq \mathbb{R}$ be a geodesic, let $\gamma\colon I \to \Sigma/_{\sim_\mu}$ and $\phi\colon I \to \Sigma/_{\sim_\mu}$ be canonical geodesics such that $\mathrm{TV}(\gamma)$ and $\mathrm{TV}(\phi)$ are essentially disjoint. Let $C_\gamma \geq 0$ and $C_\phi \geq 0$ be geodesic constants of γ and ϕ, respectively. Then the map $\gamma \parallel \phi\colon I \to \Sigma/_{\sim_\mu}$ produced by interleaving γ and ϕ is a canonical geodesic with geodesic constant $C_\gamma + C_\phi$ and

$$\mathrm{TV}(\gamma \parallel \phi) = \mathrm{TV}(\gamma) \cup \mathrm{TV}(\phi).$$ ◁

PROOF. Because γ and ϕ are canonical, we have $\gamma(t) \subseteq_\mu \mathrm{TV}(\gamma)$ and $\phi(t) \subseteq_\mu \mathrm{TV}(\phi)$ for all $t \in I$. Since $\mathrm{TV}(\gamma)$ and $\mathrm{TV}(\phi)$ are essentially disjoint, this means that we have $\gamma(s) \cap \phi(t) = [\emptyset]_{\sim_\mu}$ for all $s,t \in I$. From this, it follows that

$$\begin{aligned}\big(\gamma(s) \cup \phi(s)\big) \cap \big(\gamma(t) \cup \phi(t)\big) &= \big(\gamma(s) \cap \gamma(t)\big) \cup \underbrace{\big(\gamma(s) \cap \phi(t)\big)}_{=[\emptyset]_{\sim_\mu}} \cup \underbrace{\big(\phi(s) \cap \gamma(t)\big)}_{=[\emptyset]_{\sim_\mu}} \cup \big(\phi(s) \cap \phi(t)\big) \\ &= \big(\gamma(s) \cap \gamma(t)\big) \cup \big(\phi(s) \cap \phi(t)\big)\end{aligned}$$

for all $s,t \in I$. This is an essentially disjoint union because the total variations of γ and ϕ are essentially disjoint. We use this to verify the geodesic property. Let $s,t \in I$. We have

$$\begin{aligned}\mu\big((\gamma \parallel \phi)(s) \triangle (\gamma \parallel \phi)(t)\big) &= \mu\big((\gamma(s) \cup \phi(s)) \triangle (\gamma(t) \cup \phi(t))\big) \\ &= \mu\Big(\big(\gamma(s) \cup \gamma(t) \cup \phi(s) \cup \phi(t)\big) \setminus \big(\big(\gamma(s) \cup \phi(s)\big) \cap \big(\gamma(t) \cup \phi(t)\big)\big)\Big) \\ &= \mu\Big(\big(\gamma(s) \cup \gamma(t) \cup \phi(s) \cup \phi(t)\big) \setminus \big(\big(\gamma(s) \cap \gamma(t)\big) \cup \big(\phi(s) \cap \phi(t)\big)\big)\Big) \\ &= \mu\Big(\big(\big(\gamma(s) \cup \gamma(t)\big) \setminus \big(\gamma(s) \cap \gamma(t)\big) \setminus \big(\phi(s) \cap \phi(t)\big)\big) \\ &\qquad \cup \big(\big(\phi(s) \cup \phi(t)\big) \setminus \big(\phi(s) \cap \phi(t)\big) \setminus \big(\gamma(s) \cap \gamma(t)\big)\big)\Big) \\ &= \mu\Big(\big(\underbrace{\big(\gamma(s) \triangle \gamma(t)\big)}_{\subseteq_\mu \mathrm{TV}(\gamma)} \setminus \underbrace{\big(\phi(s) \cap \phi(t)\big)}_{\subseteq_\mu \mathrm{TV}(\phi)}\big) \\ &\qquad \cup \big(\underbrace{\big(\phi(s) \triangle \phi(t)\big)}_{\subseteq_\mu \mathrm{TV}(\phi)} \setminus \underbrace{\big(\gamma(s) \cap \gamma(t)\big)}_{\subseteq_\mu \mathrm{TV}(\gamma)}\big)\Big) \\ &= \mu\Big(\underbrace{\big(\gamma(s) \triangle \gamma(t)\big)}_{\subseteq_\mu \mathrm{TV}(\gamma)} \cup \underbrace{\big(\phi(s) \triangle \phi(t)\big)}_{\subseteq_\mu \mathrm{TV}(\phi)}\Big) \\ &= \mu\big(\gamma(s) \triangle \gamma(t)\big) + \mu\big(\phi(s) \triangle \phi(t)\big) \\ &= (C_\gamma + C_\phi) \cdot |s-t|.\end{aligned}$$

Therefore $\gamma \parallel \phi$ is a geodesic with geodesic constant $C_\gamma + C_\phi$.

For $s,t \in I$ with $s \le t$, we have

$$(\gamma \parallel \phi)(s) = \gamma(s) \cup \phi(s) \subseteq_\mu \underbrace{\gamma(t) \cup \phi(t)}_{=(\gamma \parallel \phi)(t)} \subseteq_\mu \mathrm{TV}(\gamma) \cup \mathrm{TV}(\phi)$$

because γ and ϕ are canonical. For $\gamma \parallel \phi$ to be canonical, we need only show that $\mathrm{TV}(\gamma \parallel \phi) = \mathrm{TV}(\gamma) \cup \mathrm{TV}(\phi)$. For $I = \emptyset$, we have

$$\mathrm{TV}(\gamma \parallel \phi) = [\emptyset]_{\sim_\mu} = \mathrm{TV}(\gamma) \cup \mathrm{TV}(\phi).$$

For $I \neq \emptyset$, we can construct a decreasing sequence $(s_i)_{i \in \mathbb{N}} \subseteq I$ with $s_i \to \inf I$ and

an increasing sequence $(t_i)_{i\in\mathbb{N}} \subseteq I$ with $t_i \to \sup I$. We then have

$$\begin{aligned}
\mathrm{TV}(\gamma \parallel \phi) &= \bigcup_{i=1}^{\infty} \big((\gamma \parallel \phi)(s_i) \triangle (\gamma \parallel \phi)(t_i)\big) \\
&= \bigcup_{i=1}^{\infty} \Big(\big(\gamma(s_i) \triangle \gamma(t_i)\big) \cup \big(\phi(s_i) \triangle \phi(t_i)\big)\Big) \\
&= \Big(\bigcup_{i=1}^{\infty} \big(\gamma(s_i) \triangle \gamma(t_i)\big)\Big) \cup \Big(\bigcup_{i=1}^{\infty} \big(\phi(s_i) \triangle \phi(t_i)\big)\Big) \\
&= \mathrm{TV}(\gamma) \cup \mathrm{TV}(\phi). \qquad \square
\end{aligned}$$

The preconditions for interleaving may seem restrictive at first. However, they can be established with relative ease. If the parameter intervals of γ and ϕ do not agree, we can perform reparameterization to bring them into agreement. If one interval is finite and the other is not, interleaving can be applied on a sub-interval.

Having thus brought the parameter intervals into agreement, the remaining concern is what to do if $\mathrm{TV}(\gamma)$ and $\mathrm{TV}(\phi)$ are not essentially disjoint. In this case, restriction in image restricts both geodesics to $\mathrm{TV}(\gamma) \triangle \mathrm{TV}(\phi)$. This effectively omits all changes where γ and ϕ would interfere with one another and would make the total variations disjoint.

2.3.5 Special Geodesics

With Theorem 2.3.22, we have a very general construction method for geodesics. In this section, we apply the knowledge gathered thus far to define a few practically useful types of geodesics.

There are two main ways of constructing geodesics without directly defining its value for each individual parameter. The first way is to construct a countable family of sets to be used as support points for the geodesic. This allows us to apply the sparse interpolation theorem.

The second method is to use the preimages of measurable sets under measurable functions to generate a geodesic. This is similar to the way that we derive geodesics from GLSFs. However, we can also apply this method in cases where the level set function does not satisfy a geodesic property.

2.3.5.1 MINIMAL MEAN GEODESICS

The minimal mean geodesic is a geodesic derived from a measurable function. Let $(X,\ \Sigma,\ \mu)$ be the underlying measure space, and let $f : X \to \mathbb{R} \cup \{\pm\infty\}$ be a measurable function. A minimal mean geodesic corresponding to the function f would be a geodesic $\phi : I \to {}^{\Sigma}\!/_{\sim_\mu}$ such that

$$\int_{\gamma(t)} f \,\mathrm{d}\mu = \min_{\substack{A\in\Sigma \\ \mu(A)=\mu(\gamma(t))}} \int_A f \,\mathrm{d}\mu \quad \forall t \in I.$$

For convenience, we select the interval $I \subseteq \mathbb{R}$ and geodesic γ such that γ is minimizing (i.e., has geodesic constant 1) and the infimum of the interval I is 0.

Although we will argue this in greater detail, it is evident that we can obtain a minimal mean geodesic from the sublevel sets of f. This is similar to the way in which we generate geodesics from GLSFs. However, we need to take into account the fact that f is not a GLSF.

Theorem 2.3.54 (Minimal Mean Geodesics).
Let $(X,\ \Sigma,\ \mu)$ be a finite measure space, let $f\colon X \to \mathbb{R}\cup\{\pm\infty\}$ be a measurable function such that $\mu\big(\{f=-\infty\}\big)=0$, let $I \coloneqq \big[0,\ \mu(X)\big]$, and let $\gamma\colon I \to {}^{\Sigma}\!/_{\sim_\mu}$ be a canonical geodesic with $\mathrm{TV}(\gamma)=[X]_{\sim_\mu}$ and geodesic constant $C_\gamma = 1$. Let further $T \coloneqq \mu\big(\{f<\infty\}\big)$, let $J \coloneqq (0,\ T]$, and let

$$\eta^*(t) \coloneqq \sup\Big\{\eta\in\mathbb{R}\ \Big|\ \mu\big(\{f\le\eta\}\big)<t\Big\} \quad \forall t\in J.$$

Then the geodesic $\phi\colon J \to {}^{\Sigma}\!/_{\sim_\mu}$ with

$$\phi(t) \coloneqq \big[\{f<\eta^*(t)\}\big]_{\sim_\mu} \cup (\gamma\|_{\{f=\eta^*(t)\}})\big(t-\mu(\{f<\eta^*(t)\})\big) \quad \forall t\in J$$

is canonical with $\mathrm{TV}(\phi)=[\{f<\infty\}]_{\sim_\mu}$, has geodesic constant $C_\phi = 1$, and satisfies

$$\int_{\phi(t)} f\,\mathrm{d}\mu \le \int_A f\,\mathrm{d}\mu \quad \forall t\in J,\ A\in{}^{\Sigma}\!/_{\sim_\mu}\colon \mu(A)=\mu\big(\phi(t)\big).$$

◁

PROOF. PART 1 (ϕ IS WELL-DEFINED). We first prove the well-definedness of $\eta^*: J\to\mathbb{R}\cup\{\infty\}$. Let $(\eta_i)_{i\in\mathbb{N}}\in\mathbb{R}^{\mathbb{N}}$ be a decreasing sequence with $\eta_i\to-\infty$ for $i\to\infty$. Then the corresponding sequence $\big(\{f\le\eta_i\}\big)_{i\in\mathbb{N}}\in\Sigma^{\mathbb{N}}$ of sublevel sets is decreasing and we have

$$\lim_{i\to\infty}\mu\big(\{f\le\eta_i\}\big)=\mu\Big(\bigcap_{i=1}^{\infty}\{f\le\eta_i\}\Big)=\mu(\{f=-\infty\})=0.$$

From this, it follows that for every $t>0$, there exists $\eta\in\mathbb{R}$ with $\mu\big(\{f\le\eta\}\big)\le t$. This ensures that $\eta^*(t)>-\infty$ for all $t\in J$. Conversely, for every increasing sequence $(\eta_i)_{i\in\mathbb{N}}\in\mathbb{R}^{\mathbb{N}}$ with $\eta_i\to\infty$ for $i\to\infty$, the corresponding sequence of sublevel sets $\big(\{f\le\eta_i\}\big)_{i\in\mathbb{N}}\in\Sigma^{\mathbb{N}}$ is increasing and satisfies

$$\lim_{i\to\infty}\mu\big(\{f\le\eta_i\}\big)=\mu\Big(\bigcup_{i=0}^{\infty}\{f\le\eta_i\}\Big)=\mu\big(\{f<\infty\}\big)=T.$$

Therefore, for every $t<T$, there exists $\eta\in\mathbb{R}$ such that $\mu\big(\{f\le\eta\}\big)>t$. This ensures that $\eta^*(t)\in\mathbb{R}$ for all $t\in(0,\ T)$.

As the supremum of a bounded non-empty set of real numbers, $\eta^*(t)$ is well-defined for every $t\in(0,\ T]$. Let $t_1,t_2\in(0,\ T)$ with $t_1\le t_2$. Then we have

$$\begin{aligned}\eta^*(t_1)&=\sup\Big\{\eta\in\mathbb{R}\ \Big|\ \mu\big(\{f\le\eta\}\big)\le\underbrace{t_1}_{\le t_2}\Big\}\\&\le\sup\Big\{\eta\in\mathbb{R}\ \Big|\ \mu\big(\{f\le\eta\}\big)\le t_2\Big\}\\&=\eta^*(t_2)\end{aligned}$$

because $\mu\big(\{f\le\eta\}\big)\le t_2$ for every $\eta\in\mathbb{R}$ with $\mu\big(\{f\le\eta\}\big)\le t_1$ and the supremum over a superset is always greater than or equal to that of its subset. This shows that η^* is monotonically increasing.

We now show that ϕ is well-defined. Let $t\in(0,\ T]$. Since $\eta^*(t)\in\mathbb{R}\cup\{\infty\}$, $\big\{f<\eta^*(t)\big\}$ and $\big\{f=\eta^*(t)\big\}$ are well-defined measurable sets. These sets are by definition disjoint and form a partition of the set $\big\{f\le\eta^*(t)\big\}$.

If $\eta^*(t) = \infty$, then we have $t = T$ and therefore $\mu\big(\{f < \eta^*(t)\}\big) = \mu\big(\{f < \infty\}\big) = T$. This implies that

$$\phi(t) = \big[\{f < \infty\}\big)\big]_{\sim_\mu} \cup \underbrace{(\gamma\|_{\{f=\infty\}})(0)}_{=[\emptyset]_{\sim_\mu}} = \big[\{f < \infty\}\big]_{\sim_\mu}.$$

We subsequently discuss the case in which $\eta^*(t) < \infty$.

Let $(\eta_i^-)_{i\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}$ be an increasing sequence with $\eta_i^- < \eta^*(t)$ for all $i \in \mathbb{N}$ and $\eta_i^- \to \eta^*(t)$ for $i \to \infty$. Let $(\eta_i^+)_{i\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}$ be a decreasing sequence with $\eta_i^+ > \eta^*(t)$ for all $i \in \mathbb{N}$ and $\eta_i^+ \to \eta^*(t)$ for $i \to \infty$. We have

$$\begin{aligned}
\mu\Big(\big\{f < \eta^*(t)\big\}\Big) &= \mu\Big(\bigcup_{i=1}^{\infty} \{f \le \eta_i^-\}\Big) \\
&= \lim_{i\to\infty} \underbrace{\mu\big(\{f \le \eta_i^-\}\big)}_{<t} \\
&\le t.
\end{aligned}$$

$$\begin{aligned}
\mu\Big(\big\{f \le \eta^*(t)\big\}\Big) &= \mu\Big(\bigcap_{i=1}^{\infty} \{f \le \eta_i^+\}\Big) \\
&= \lim_{i\to\infty} \underbrace{\mu\big(\{f \le \eta_i^+\}\big)}_{\ge t} \\
&\ge t.
\end{aligned}$$

The latter inequality ensures that

$$\mu\Big(\big\{f = \eta^*(t)\big\}\Big) = \mu\Big(\big\{f \le \eta^*(t)\big\}\Big) - \mu\Big(\big\{f < \eta^*(t)\big\}\Big) \ge t - \mu\Big(\big\{f < \eta^*(t)\big\}\Big).$$

The restriction in image $\gamma\|_{\{f=\eta^*(t)\}}$ is a minimizing canonical geodesic with total variation $\mathrm{TV}(\gamma) \cap \big[\{f = \eta^*(t)\}\big]_{\sim_\mu} = \big[\{f = \eta^*(t)\}\big]_{\sim_\mu}$. We note that $\mu\big(\{f = \eta^*(t)\}\big) < \infty$ because $\mu(X) < \infty$. This ensures that the parameter interval of $\gamma\|_{\{f=\eta^*(t)\}}$ is the closed interval $\big[0,\ \mu(\{f = \eta^*(t)\})\big]$ and therefore contains $t - \mu\big(\{f < \eta^*(t)\}\big)$.

Part 2 (ϕ is a canonical geodesic). We first show that ϕ is μ-essentially increasing. Let $s, t \in I$ with $s \le t$. Because η^* is increasing, we have $\eta^*(s) \le \eta^*(t)$. If $\eta^*(s) < \eta^*(t)$, then we have

$$\begin{aligned}
\phi(s) &= \Big[\big\{f < \eta^*(s)\big\}\Big]_{\sim_\mu} \cup (\gamma\|_{\{f=\eta^*(s)\}})\Big(s - \mu\Big(\big\{f < \eta^*(s)\big\}\Big)\Big) \\
&\subseteq_\mu \Big[\big\{f \le \eta^*(s)\big\}\Big]_{\sim_\mu} \\
&\subseteq_\mu \Big[\big\{f < \eta^*(t)\big\}\Big]_{\sim_\mu} \\
&\subseteq_\mu \phi(t).
\end{aligned}$$

If $\eta^*(s) = \eta^*(t)$, then we have

$$\begin{aligned}
\phi(s) &= \left[\{f < \eta^*(s)\}\right]_{\sim_\mu} \cup (\gamma\|_{\{f=\eta^*(s)\}})\Big(s - \mu\Big(\{f < \eta^*(s)\}\Big)\Big) \\
&= \left[\{f < \eta^*(t)\}\right]_{\sim_\mu} \cup (\gamma\|_{\{f=\eta^*(s)\}})\Big(s - \mu\Big(\{f < \eta^*(t)\}\Big)\Big) \\
&\subseteq_\mu \left[\{f < \eta^*(t)\}\right]_{\sim_\mu} \cup (\gamma\|_{\{f=\eta^*(s)\}})\Big(t - \mu\Big(\{f < \eta^*(t)\}\Big)\Big) \\
&= \phi(t).
\end{aligned}$$

Next, we show that the measure of $\phi(t)$ is equal to t for all $t \in (0,\ T]$. For all $t \in (0,\ T]$, we have

$$\begin{aligned}
\mu\big(\phi(t)\big) &= \mu\Bigg(\Big[\{f < \eta^*(t)\}\Big]_{\sim_\mu} \cup \underbrace{(\gamma\|_{\{f=\eta^*(s)\}})\Big(s - \mu\big(\{f < \eta^*(t)\}\big)\Big)}_{\subseteq_\mu [\{f=\eta^*(t)\}]_{\sim_\mu}}\Bigg) \\
&= \underbrace{\mu\Big(\Big[\{f < \eta^*(t)\}\Big]_{\sim_\mu}\Big)}_{\leq t} + \underbrace{\mu\big((\gamma\|_{\{f=\eta^*(s)\}})(s - \mu(\{f < \eta^*(t)\}))\big)}_{=t-\mu(\{f<\eta^*(t)\})} \\
&= t.
\end{aligned}$$

In conjunction with the fact that ϕ is μ-essentially increasing, this means that

$$\mu\big(\phi(s) \triangle \phi(t)\big) = \mu\Big(\phi\big(\max\{s,\ t\}\big) \setminus \phi\big(\min\{s,\ t\}\big)\Big) = \max\{s,\ t\} - \min\{s,\ t\} = |s - t|$$

and that $\mu\big(\phi(t)\big) \to 0$ for $t \to 0$. This implies that ϕ is a canonical geodesic and that $C_\phi = 1$ is a geodesic constant of ϕ.

PART 3 (TV(ϕ) = $[\{f < \infty\}]_{\sim_\mu}$). For $T = 0$, $\left[\{f < \infty\}\right]_{\sim_\mu} = [X]_{\sim_\mu} \setminus \left[\{f = \infty\}\right]_{\sim_\mu}$ implies

$$\mu\big(X \setminus \{f = \infty\}\big) = \mu\big(\{f < \infty\}\big) = T = 0,$$

i.e., $f = \infty$ almost everywhere. Thus, we have

$$\left[\{f < \infty\}\right]_{\sim_\mu} = [\emptyset]_{\sim_\mu} = \mathrm{TV}(\phi)$$

because ϕ has an empty parameter interval.

For $T > 0$, we use the fact that ϕ is canonical. This implies that TV(ϕ) is the destination point of ϕ, which is equal to $\phi(T)$. As we have shown before, $\mu\big(\phi(T)\big) = T$. If $\eta^*(T) = \infty$, then $\left[\{f < \eta^*(T)\}\right]_{\sim_\mu} \subseteq_\mu \phi(T)$ implies that we have $\phi(T) = \left[\{f < \infty\}\right]_{\sim_\mu}$. If $\eta^*(T) < \infty$, then we have

$$\phi(T) \subseteq_\mu \left[\{f \leq \eta^*(T)\}\right]_{\sim_\mu} \subseteq_\mu \left[\{f < \infty\}\right]_{\sim_\mu}$$

which implies that

$$\mu\Big(\phi(T) \triangle \left[\{f < \infty\}\right]_{\sim_\mu}\Big) = \mu\big(\{f < \infty\}\big) - \mu\big(\phi(T)\big) = T - T = 0.$$

We can rewrite the measure of the set difference as a difference of measures because $\mu(X) < \infty$. In either case, we have $\mathrm{TV}(\phi) = \phi(T) = \left[\{f < \infty\}\right]_{\sim_\mu}$.

PART 4 (MINIMALITY OF INTEGRAL). Let $t \in J$. By definition, we have

$$\left[\{f < \eta^*(t)\}\right]_{\sim_\mu} \subseteq_\mu \phi(t) \subseteq_\mu \left[\{f \leq \eta^*(t)\}\right]_{\sim_\mu}$$

Let $A \in \Sigma/_{\sim_\mu}$ with $\mu(A) = \mu\big(\phi(t)\big)$. We have

$$A \setminus \phi(t) \subseteq_\mu \left[\{f \geq \eta^*(t)\}\right]_{\sim_\mu},$$
$$\phi(t) \setminus A \subseteq_\mu \left[\{f \leq \eta^*(t)\}\right]_{\sim_\mu}.$$

Therefore, we have

$$\begin{aligned}\int_{\phi(t)} f \,\mathrm{d}\mu - \int_A f \,\mathrm{d}\mu &= \int_{\phi(t)\setminus A} \underbrace{f}_{\leq \eta^*(t)} \,\mathrm{d}\mu - \int_{A\setminus\phi(t)} \underbrace{f}_{\geq \eta^*(t)} \,\mathrm{d}\mu \\ &\leq \eta^*(t) \cdot \Big(\mu\big(\phi(t)\setminus A\big) - \mu\big(A \setminus \phi(t)\big)\Big).\end{aligned}$$

Since $\mu(A) = \mu\big(\phi(t)\big)$, we have $\mu\big(\phi(t)\setminus A\big) = \mu\big(A \setminus \phi(t)\big)$ and therefore

$$\int_{\phi(t)} f \,\mathrm{d}\mu - \int_A f \,\mathrm{d}\mu \leq 0.$$

□

The way in which we construct a minimal mean geodesic in Theorem 2.3.54 is very complicated. It requires a prior geodesic γ that is used to break ties between points. We note that, if the measure space (X, Σ, μ) is assumed to be both finite and atomless, then this construction is much simpler because we can invoke the sparse interpolation theorem (see Theorem 2.3.22). The complicated variant is useful because it allows us to specify the tie breaking geodesic. However, it does not generalize the theorem because the existence of a tie-breaking geodesic γ implies atomlessness.

The simplified construction method consists of determining the sublevel set $\{f \leq q\}$ for each $q \in \mathbb{Q}$ and associating the similarity classes of those sublevel sets with their respective measures in a support tuple. Because $\mathbb{Q}$ is countable, this yields a support tuple suitable for use in the sparse interpolation theorem. We can then use the theorem to quickly show the existence of ϕ.

Fundamentally, this does not change the underlying construction process. The sparse interpolation theorem implicitly constructs the relevant segments of the restricted tie-breaking geodesic $\gamma\|_{\{f=t\}}$ as part of its constructive proof. In Section 2.3.6, we will show that the conjunction of finiteness and atomlessness is precisely the condition under which a tie-breaking geodesic can always be constructed. There, we will use the implicit tie-breaking capability of the sparse interpolation theorem to construct geodesics connecting arbitrary similarity classes by interpolating support tuples that do not include any information besides origin and destination point (see Theorem 2.3.71).

2.3.5.2 CONSTANT MEAN GEODESICS

As we begin to move from pure geodesic theory to optimization theory, we rely heavily on two analogies that hold in most circumstances:

- geodesics behave like affine linear paths; and

- signed measures behave like linear forms.

In other words, whenever vector space optimization theory uses linear forms, e.g., for derivatives, we use signed measures. Whenever vector space optimization theory would use a vector to represent, e.g., a descent direction, we would use a canonical geodesic. In this way, we can transfer a lot of optimization theory from vector spaces to measure spaces.

There is one notable breakdown in this analogy. Usually, we would expect the composition of a linear form with an affine linear path to yield an affine linear function. This is not the case with geodesics and signed measures.

Example 2.3.55 (Composing Geodesic and Signed Measure).
We consider the signed measure $\mu\colon \mathscr{B}(I) \to \mathbb{R}$ with $I := [0,\ 1]$ and

$$\mu(A) := \int_A \left(x - \frac{1}{3}\right) \mathrm{d}x \quad \forall A \in \mathscr{B}(I).$$

Let $\gamma\colon I \to {}^{\mathscr{B}(I)}\!/_{\sim_\lambda}$ with

$$\gamma(t) := \left[[0,\ t]\right]_{\sim_\lambda} \quad \forall t \in I.$$

We recall that λ denotes the Lebesgue measure. It is evident that γ is a geodesic and that μ is a signed measure with $\mu \ll \lambda$. The composition of both is $f := \mu \circ \gamma\colon [0,\ 1] \to \mathbb{R}$, which is well-defined because the value of μ is the same for all representatives of a similarity class. We have

$$\begin{aligned} f(t) &= \mu\big([0,\ t]\big) \\ &= \int_0^t \left(x - \frac{1}{3}\right) \mathrm{d}x \\ &= \frac{1}{2}t^2 - \frac{1}{3}t \\ &= \frac{1}{2}t \cdot \left(t - \frac{2}{3}\right) \end{aligned}$$

for all $t \in [0,\ 1]$.

Thus, rather than being affine linear, f is a quadratic function that starts at $f(0) = 0$, decreases down to its minimal value $f(1/3) = -1/18$ and then begins to increase again, becoming non-negative once more at $t = 2/3$ and finally ending at $f(1) = 1/6$. Figure 2.5 on the next page illustrates this example. ◁

Example 2.3.55 proves that concatenations of signed measures and geodesics are not always affine linear. An interesting question that arises when we think about this problem is whether we can construct *specific* geodesics for which this is the case.

We can, in fact, construct such geodesics. We will refer to them as *constant mean geodesics*. This name makes more sense once we take into account that in a measure space $(X,\ \Sigma,\ \mu)$, any signed measure $\phi\colon \Sigma \to \mathbb{R} \cup \{\pm\infty\}$ with $\phi \ll \mu$ has a density function, i.e., a measurable function $f\colon X \to \mathbb{R}$ such that

$$\phi(A) = \int_A f \,\mathrm{d}\mu.$$

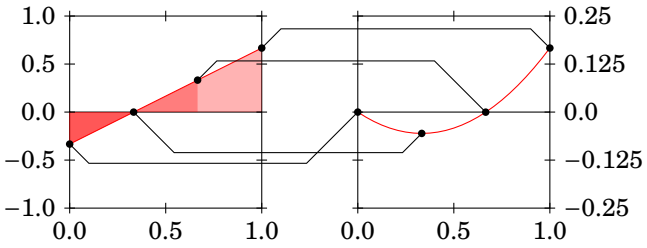


Figure 2.5: Illustration of Example 2.3.55. On the left side, we see the density function of the measure μ with the noted parameter points $t = 0$, $t = 1/3$, $t = 2/3$, and $t = 1$ marked and associated with their corresponding points on the graph of f, which is shown on the right side.

Let $I \subseteq \mathbb{R}$ be an interval and let $\gamma\colon I \to \Sigma/_{\sim_\mu}$ be a geodesic with geodesic constant $C_\gamma \geq 0$. For $s,t \in I$, we have

$$\begin{aligned}\phi\big(\gamma(s)\big) - \phi\big(\gamma(t)\big) &= \int_{\gamma(s)} f \,\mathrm{d}\mu - \int_{\gamma(t)} f \,\mathrm{d}\mu \\ &= \int_{\gamma(s)\setminus\gamma(t)} f \,\mathrm{d}\mu - \int_{\gamma(t)\setminus\gamma(s)} f \,\mathrm{d}\mu.\end{aligned}$$

If γ is essentially increasing, then $\gamma(s)\setminus\gamma(t) = [\emptyset]_{\sim_\mu}$ for $s \leq t$ and $\gamma(t)\setminus\gamma(s) = [\emptyset]_{\sim_\mu}$ for $s > t$. In this case we obtain

$$\begin{aligned}\phi\big(\gamma(s)\big) - \phi\big(\gamma(t)\big) &= \int_{\gamma(s)\setminus\gamma(t)} f \,\mathrm{d}\mu - \int_{\gamma(t)\setminus\gamma(s)} f \,\mathrm{d}\mu \\ &= \underbrace{\frac{C_\gamma \cdot (s-t)}{\mu(\gamma(s) \triangle \gamma(t))}}_{=\operatorname{sgn}(s-t)} \cdot \int_{\gamma(s)\triangle\gamma(t)} f \,\mathrm{d}\mu \\ &= C_\gamma \cdot \left(\frac{1}{\mu(\gamma(s) \triangle \gamma(t))} \cdot \int_{\gamma(s)\triangle\gamma(t)} f \,\mathrm{d}\mu \right) \cdot (s-t).\end{aligned}$$

Therefore, for an essentially increasing geodesic γ, $\phi \circ \gamma$ is affine linear if and only if

$$\frac{1}{\mu(\gamma(s) \triangle \gamma(t))} \int_{\gamma(s)\triangle\gamma(t)} f \,\mathrm{d}\mu$$

is constant for all $s,t \in I$, i.e., if the mean value of the density function over all $\gamma(s) \triangle \gamma(t)$ is constant. The name *constant mean geodesic* stems from this fact.

Definition 2.3.56 (Constant Mean Geodesic).
Let $(X,\ \Sigma,\ \mu)$ be a measure space, let $I \subseteq \mathbb{R}$ be an interval, let $\gamma\colon I \to \Sigma/_{\sim_\mu}$ be a canonical geodesic with $\mu\big(\mathrm{TV}(\gamma)\big) < \infty$, and let $f\colon X \to \mathbb{R} \cup \{\pm\infty\}$ be measurable with

$$\int_{\mathrm{TV}(\gamma)} |f| \,\mathrm{d}\mu < \infty.$$

We refer to γ as a *constant mean geodesic of* f if and only if there exists a constant $M \in \mathbb{R}$ such that

$$\frac{1}{\mu(\gamma(s) \triangle \gamma(t))} \cdot \int_{\gamma(s)\triangle\gamma(t)} f \,\mathrm{d}\mu = M \quad \forall s,t \in I.$$

◁

We can construct a constant mean geodesic by a special type of interleaving. If we have a canonical geodesic γ_+ with $f \geq M$ in $\mathrm{TV}(\gamma_+)$ and a canonical geodesic γ_- with $f \leq M$ in $\mathrm{TV}(\gamma_-)$ such that the total variations of both geodesics are essentially disjoint, then we can find reparameterizations of both geodesics such that their union becomes a constant mean geodesic.

Lemma 2.3.57.
Let (X, Σ, μ) be a measure space, let $R \in \Sigma/_{\sim_\mu}$ be such that $\mu(R) \in (0, \infty)$, let $\gamma\colon [0, \mu(R)] \to \Sigma/_{\sim_\mu}$ be a canonical geodesic with $\mathrm{TV}(\gamma) = R$, and let the function $f\colon X \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ be measurable with

$$M := \int_R f \,\mathrm{d}\mu = \int_R |f| \,\mathrm{d}\mu < \infty.$$

Then the function $f_\gamma\colon [0, \mu(R)] \to [0, M]$ with

$$f_\gamma(t) := \int_{\gamma(t)} f \,\mathrm{d}\mu$$

is uniformly continuous, monotonically increasing, and surjective. If $f > 0$ almost everywhere in R, then f_γ is strictly monotonic and therefore bijective. ◁

PROOF. We note that $\gamma(t) \subseteq_\mu \mathrm{TV}(\gamma) = R$ for all $t \in [0, \mu(R)]$ and therefore

$$0 \leq \int_{\gamma(t)} f \,\mathrm{d}\mu \leq \int_R |f| \,\mathrm{d}\mu = M \quad \forall t \in [0, \mu(R)].$$

Let $\widetilde{R} \in R \subseteq \Sigma$. Due to the absolute continuity of the Lebesgue integral and the fact that f is integrable on $\widetilde{R}$, for every $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\left| \int_B f \,\mathrm{d}\mu \right| < \varepsilon \quad \forall B \in \Sigma\colon B \subseteq \widetilde{R} \wedge \mu(B) < \delta.$$

This estimate transfers to all representatives of essential subsets of R. Since $\mathrm{TV}(\gamma) = R$, 1 is a geodesic constant of γ. For all $s, t \in [0, \mu(R)]$ with $|s - t| < \delta$, we can assume without loss of generality that $s \leq t$ and then obtain

$$\begin{aligned} |f_\gamma(s) - f_\gamma(t)| &= \left| \int_{\gamma(t)} f \,\mathrm{d}\mu - \int_{\gamma(s)} f \,\mathrm{d}\mu \right| \\ &= \left| \int_{\gamma(t) \setminus \gamma(s)} f \,\mathrm{d}\mu \right| \\ &\leq \varepsilon \end{aligned}$$

because $\mu(\gamma(t) \setminus \gamma(s)) = |t - s| \leq \delta$. This shows that f_γ is uniformly continuous.

For $s, t \in [0, \mu(R)]$ with $s \leq t$, we have

$$f_\gamma(t) - f_\gamma(s) = \int_{\gamma(t)} f \,\mathrm{d}\mu - \int_{\gamma(s)} f \,\mathrm{d}\mu = \int_{\gamma(t) \setminus \gamma(s)} f \,\mathrm{d}\mu \geq 0$$

which proves that f_γ is monotonically increasing. If $f > 0$ almost everywhere in R and $s < t$, then $\gamma(t) \setminus \gamma(s) = \gamma(s) \triangle \gamma(t)$ has strictly positive measure and we have

$$f_\gamma(t) - f_\gamma(s) = \int_{\gamma(t) \setminus \gamma(s)} f \,\mathrm{d}\mu > 0.$$

Since f_γ is continuous and $f_\gamma(0) = 0$, it is sufficient to show that $f_\gamma\big(\mu(R)\big) = M$ to show that f_γ is surjective. Since $\gamma\big(\mu(R)\big)$ is the destination point of γ and γ is canonical, we have

$$f_\gamma\big(\mu(R)\big) = \int_{\gamma(\mu(R))} f \,\mathrm{d}\mu = \int_{\mathrm{TV}(\gamma)} f \,\mathrm{d}\mu = M.$$

Therefore, f_γ assumes all values between 0 and M and is therefore surjective. If f_γ is strictly monotonic, then f_γ is also injective and therefore bijective. □

The bijectivity of f_γ as defined in Lemma 2.3.57 for strictly positive functions f significantly simplifies the construction of constant mean geodesics. If we choose γ_+ such that $f(x) \geq M$ for almost all $x \in \mathrm{TV}(\gamma_+)$ and γ_- such that $f(x) < M$ for almost all $x \in \mathrm{TV}(\gamma_-)$, then f_{γ_-} will be bijective. This means that for every parameter t of γ_+, the corresponding parameter for γ_- is given by $f_{\gamma_-}^{-1}\big(f_{\gamma_+}(t)\big)$.

Theorem 2.3.58 (Constant Mean Geodesic Construction).
Let (X, Σ, μ) be an atomless measure space, let $R \in \Sigma/_{\sim_\mu}$ with $\mu(R) < \infty$, and let $f : X \to \mathbb{R} \cup \{\pm\infty\}$ be measurable with

$$\int_R |f| \,\mathrm{d}\mu < \infty.$$

Let

$$M := \frac{1}{\mu(R)} \cdot \int_R f \,\mathrm{d}\mu \in \mathbb{R}.$$

Then we can partition R into $R_- := R \cap \big[\{f < M\}\big]_{\sim_\mu}$ and $R_+ := R \cap \big[\{f \geq M\}\big]_{\sim_\mu}$ and there exist canonical geodesics $\gamma_- : \big[0,\ \mu(R_-)\big] \to \Sigma/_{\sim_\mu}$ with $\mathrm{TV}(\gamma_-) = R_-$ and $\gamma_+ : \big[0,\ \mu(R_+)\big] \to \Sigma/_{\sim_\mu}$ with $\mathrm{TV}(\gamma_+) = R_+$.

Let any such pair of canonical geodesics γ_+, γ_- be given, let

$$V_+ := \int_{R_+} (f - M) \,\mathrm{d}x,$$
$$V_- := \int_{R_-} (M - f) \,\mathrm{d}x,$$

and let $f_{\gamma_+} : \big[0,\ \mu(R_+)\big] \to [0,\ V_+]$ and $f_{\gamma_-} : \big[0,\ \mu(R_-)\big] \to [0,\ V_-]$ with

$$f_{\gamma_+}(t) := \int_{\gamma_+(t)} (f - M) \,\mathrm{d}\mu \quad \forall t \in \big[0,\ \mu(R_+)\big],$$
$$f_{\gamma_-}(t) := \int_{\gamma_-(t)} (M - f) \,\mathrm{d}\mu \quad \forall t \in \big[0,\ \mu(R_-)\big].$$

Then we have $V_+ = V_-$, f_{γ_-} is bijective, and the map $\rho : \big[0,\ \mu(R_+)\big] \to \big[0,\ \mu(R)\big]$ with

$$\rho(t) := t + f_{\gamma_-}^{-1}\big(f_{\gamma_+}(t)\big) \quad \forall t \in \big[0,\ \mu(R_+)\big]$$

is continuous, strictly monotonically increasing, and bijective. Furthermore, the map $\gamma : \big[0,\ \mu(R)\big] \to \Sigma/_{\sim_\mu}$ with

$$\gamma(t) := \gamma_+\big(\rho^{-1}(t)\big) \cup \gamma_-\big(t - \rho^{-1}(t)\big) \quad \forall t \in \big[0,\ \mu(R)\big]$$

is a constant mean geodesic of f with geodesic constant 1 and $\mathrm{TV}(\gamma) = R$. ◁

PROOF. PART 1 (EXISTENCE OF $\gamma_\pm$). The existence of $\gamma_\pm$ follows from Theorem 2.3.22. Let

$$\begin{aligned} I_+ &:= [0,\ \mu(R_+)], \quad & T_+ &:= \{0,\ \mu(R_+)\}, \quad & B_+(0) &:= [\varnothing]_{\sim_\mu}, \quad & B_+\big(\mu(R_+)\big) &:= R_+, \\ I_- &:= [0,\ \mu(R_-)], \quad & T_- &:= \{0,\ \mu(R_-)\}, \quad & B_-(0) &:= [\varnothing]_{\sim_\mu}, \quad & B_-\big(\mu(R_-)\big) &:= R_-. \end{aligned}$$

Then $(I_+,\ T_+,\ B_+)$ and $(I_-,\ T_-,\ B_-)$ are geodesic support tuples according to Definition 2.3.15. Because $(X,\ \Sigma,\ \mu)$ is atomless, we can invoke Theorem 2.3.22 to prove that there are geodesics $\gamma_+ \colon I_+ \to \Sigma/_{\sim_\mu}$ and $\gamma_- \colon I_- \to \Sigma/_{\sim_\mu}$ such that

$$\begin{aligned} \gamma_+(0) &= [\varnothing]_{\sim_\mu}, & \gamma_+\big(\mu(R_+)\big) &= R_+, \\ \gamma_-(0) &= [\varnothing]_{\sim_\mu}, & \gamma_-\big(\mu(R_-)\big) &= R_-. \end{aligned}$$

The infimum of both I_+ and I_- is 0 and we have $\gamma_+(0) = \gamma_-(0) = [\varnothing]_{\sim_\mu}$. According to Corollary 2.3.13, this means that both γ_- and γ_+ are canonical. Furthermore, we have

$$\begin{aligned} \mathrm{TV}(\gamma_+) &= \gamma_+(0) \triangle \gamma_+\big(\mu(R_+)\big) = R_+, \\ \mathrm{TV}(\gamma_-) &= \gamma_-(0) \triangle \gamma_-\big(\mu(R_-)\big) = R_-. \end{aligned}$$

Let subsequently $\gamma_+ \colon I_+ \to \Sigma/_{\sim_\mu}$ and $\gamma_- \colon I_- \to \Sigma/_{\sim_\mu}$ be canonical geodesics with $\mathrm{TV}(\gamma_+) = R_+$ and $\mathrm{TV}(\gamma_-) = R_-$.

PART 2 (PROPERTIES OF f_{γ_+}, f_{γ_-}, ρ). Lemma 2.3.57 states that f_{γ_+} is uniformly continuous, monotonically increasing, and surjective. It further states that f_{γ_-} is uniformly continous, strictly monotonically increasing, and bijective. Concerning the constants V_+ and V_-, we have

$$V_+ - V_- = \int_{R_+} (f - M)\,\mathrm{d}\mu - \int_{R_-} (M - f)\,\mathrm{d}\mu = \int_R f\,\mathrm{d}\mu - M \cdot \mu(R) = 0$$

and therefore $V_+ = V_-$. This means that $f_{\gamma_-}^{-1} \circ f_{\gamma_+} \colon [0,\ \mu(R_+)] \to [0,\ \mu(R_-)]$ is well-defined. We have

$$\rho(t) = \underbrace{t}_{\leq \mu(R_+)} + \underbrace{f_{\gamma_-}^{-1}\big(f_{\gamma_+}(t)\big)}_{\leq \mu(R_-)} \in [0,\ \mu(R_+) + \mu(R_-)] = [0,\ \mu(R)].$$

Therefore, ρ is well-defined. The continuity of ρ is evident if $f_{\gamma_-}^{-1} \circ f_{\gamma_+}$ is continuous. Let $C \subseteq \mathbb{R}$ be a closed set. Since $C \cap [0,\ \mu(R_-)]$ is bounded and closed, it is compact and therefore

$$C' := (f_{\gamma_-}^{-1})^{-1}(C) = (f_{\gamma_-}^{-1})^{-1}\Big(C \cap [0,\ \mu(R_-)]\Big) = f_{\gamma_-}\Big(C \cap [0,\ \mu(R_-)]\Big) \subseteq [0,\ V_-]$$

is the image of a compact set under a continuous map, which is compact and thus closed. Since f_{γ_+} is also continuous,

$$C'' := (f_{\gamma_-}^{-1} \circ f_{\gamma_+})^{-1}(C) = f_{\gamma_+}^{-1}\big((f_{\gamma_-}^{-1})^{-1}(C)\big) = f_{\gamma_+}^{-1}(C') \subseteq [0,\ \mu(R^+)]$$

is also closed. Because the preimage of every closed set under $f_{\gamma_-}^{-1} \circ f_{\gamma_+}$ is closed, $f_{\gamma_-}^{-1} \circ f_{\gamma_+}$ is continuous. Therefore, so is ρ.

The strict monotonicity of ρ follows from the strict monotonicity of $t \mapsto t$ and the non-strict monotonicity of $f_{\gamma_-}^{-1} \circ f_{\gamma_+}$. The latter follows from the strict

monotonicity of $f_{\gamma_-}^{-1}$, which follows from the strict monotonicity of f_{γ_-}, and from the non-strict monotonicity of f_{γ_+}.

Since ρ is strictly monotonically increasing, ρ is injective. Surjectivity follows from continuity if we can show that $\rho(0) = 0$ and $\rho\big(\mu(R_+)\big) = \mu(R)$. We note that $f_{\gamma_+}(0) = f_{\gamma_-}(0) = 0$, $f_{\gamma_+}\big(\mu(R_+)\big) = V_+$, and $f_{\gamma_-}\big(\mu(R_-)\big) = V_- = V_+$, which implies that

$$\begin{aligned}\rho(0) &= 0 + f_{\gamma_-}^{-1}\big(f_{\gamma_+}(0)\big)\\ &= 0 + f_{\gamma_-}^{-1}(0)\\ &= 0,\end{aligned}$$

$$\begin{aligned}\rho\big(\mu(R_+)\big) &= \mu(R_+) + f_{\gamma_-}^{-1}\Big(f_{\gamma_+}\big(\mu(R_+)\big)\Big)\\ &= \mu(R_+) + f_{\gamma_-}^{-1}(V_+)\\ &= \mu(R_+) + \mu(R_-)\\ &= \mu(R_+ \cup R_-)\\ &= \mu(R).\end{aligned}$$

Because ρ is continuous, this implies that ρ is surjective. In conjunction with the injectivity of ρ, this proves that ρ is bijective.

Part 3 (Properties of γ). For all $t \in \big[0,\ \mu(R)\big]$ we have $\rho^{-1}(t) \in \big[0,\ \mu(R_+)\big]$. Thus, $\gamma_+\big(\rho^{-1}(t)\big)$ is well-defined. Since $\rho(\tau) \geq t$ for all $\tau \in \big[0,\ \mu(R_+)\big]$, we also have $\rho^{-1}(t) \leq t$ for all $t \in \big[0,\ \mu(R)\big]$. We therefore have $t - \rho^{-1}(t) \geq 0$.

We can most easily show by contradiction that $t - \rho^{-1}(t) \leq \mu(R_-)$. If there existed $t \in [0,\ \mu(R)]$ with $t - \rho^{-1}(t) > \mu(R_-)$ or $\rho^{-1}(t) < t - \mu(R_-)$. Because ρ is strictly monotonically increasing, this would imply that

$$\begin{aligned}t &< \rho(t - \mu(R_-))\\ &= t - \mu(R_-) + \underbrace{f_{\gamma_-}^{-1}\Big(f_{\gamma_+}\big(t - \mu(R_-)\big)\Big)}_{\leq \mu(R_-)}\\ &\leq t\end{aligned}$$

which is a clearly impossible. Therefore, we have $t - \rho^{-1}(t) \in \big[0,\ \mu(R_-)\big]$ for all $t \in \big[0,\ \mu(R)\big]$. This shows that $\gamma_-\big(t - \rho^{-1}(t)\big)$ is well-defined. Therefore, γ as a whole is well-defined.

Next, we show that γ is a geodesic. We note that γ_+ and γ_- are canonical geodesics with geodesic constant 1 and that their total variations $\mathrm{TV}(\gamma_+) = R_+$ and $\mathrm{TV}(\gamma_-) = R_-$ are essentially disjoint. Since ρ is strictly increasing, so is ρ^{-1}. For $s, t \in \big[0,\ \mu(R)\big]$ with $s \leq t$, we have $\rho^{-1}(s) \leq \rho^{-1}(t)$. The non-strict monotonicity of $f_{\gamma_-}^{-1} \circ f_{\gamma_+}$ yields

$$\underbrace{f_{\gamma_-}^{-1}\Big(f_{\gamma_+}\big(\rho^{-1}(s)\big)\Big)}_{=\rho(\rho^{-1}(s)) - \rho^{-1}(s)} \leq \underbrace{f_{\gamma_-}^{-1}\Big(f_{\gamma_+}\big(\rho^{-1}(t)\big)\Big)}_{=\rho(\rho^{-1}(t)) - \rho^{-1}(t)}$$

and therefore

$$s - \rho^{-1}(s) \leq t - \rho^{-1}(t) \quad \forall s, t \in \big[0,\ \mu(R)\big] : s \leq t.$$

This means that $t \mapsto t - \rho^{-1}(t)$ is also monotonically increasing. We now consider $s,t \in [0,\ \mu(R)]$. Without loss of generality, let $s \leq t$. Because γ_+ and γ_- have essentially disjoint total variation, we have

$$\begin{aligned}\mu\big(\gamma(s) \triangle \gamma(t)\big) &= \mu\Big(\gamma_+\big(\rho^{-1}(s)\big) \triangle \gamma_+\big(\rho^{-1}(t)\big)\Big) + \mu\Big(\gamma_-\big(s - \rho^{-1}(s)\big) \triangle \gamma_-\big(t - \rho^{-1}(t)\big)\Big) \\ &= \big(\rho^{-1}(t) - \rho^{-1}(s)\big) + \big(t - \rho^{-1}(t) - s + \rho^{-1}(s)\big) \\ &= t - s \\ &= |s - t|.\end{aligned}$$

which proves that γ is a geodesic with geodesic constant 1. Because we have $0 = \rho^{-1}(0) = 0 - \rho^{-1}(0)$, we can infer that

$$\gamma(0) = \gamma_+(0) \cup \gamma_-(0) = [\emptyset]_{\sim_\mu}.$$

Therefore, γ is canonical and we have

$$\begin{aligned}\mathrm{TV}(\gamma) &= \gamma\big(\mu(R)\big) \\ &= \gamma_+\Big(\rho^{-1}\big(\mu(R)\big)\Big) \cup \gamma_-\Big(\mu(R) - \rho^{-1}\big(\mu(R)\big)\Big) \\ &= \gamma_+\big(\mu(R_+)\big) \cup \gamma_-\big(\mu(R_-)\big) \\ &= \mathrm{TV}(\gamma_+) \cup \mathrm{TV}(\gamma_-) \\ &= R_+ \cup R_- \\ &= R.\end{aligned}$$

To verify the constant mean property, we examine a single $t \in [0,\ \mu(R)]$. Let $t_+ \coloneqq \rho^{-1}(t) \in [0,\ \mu(R_+)]$ and $t_- \coloneqq f_{\gamma_-}^{-1}(f_{\gamma_+}(t_+)) \in [0,\ \mu(R_-)]$. By definition, we have

$$t = \rho(t_+) = t_+ + f_{\gamma_-}^{-1}\big(f_{\gamma_+}(t_+)\big) = t_+ + t_-.$$

This implies that $t_- = t - t_+ = t - \rho^{-1}(t)$ and therefore

$$\gamma(t) = \gamma_+(t_+) \cup \gamma_-(t_-).$$

Since $\gamma_+(t_+)$ and $\gamma_-(t_-)$ are essentially disjoint, we have

$$\begin{aligned}\int_{\gamma(t)} (f - M)\,\mathrm{d}\mu &= \int_{\gamma_+(t_+)} (f - M)\,\mathrm{d}\mu - \int_{\gamma_-(t_-)} (M - f)\,\mathrm{d}\mu \\ &= f_{\gamma_+}(t_+) - f_{\gamma_-}(t_-) \\ &= f_{\gamma_+}(t_+) - f_{\gamma_-}\Big(f_{\gamma_-}^{-1}\big(f_{\gamma_+}(t_+)\big)\Big) \\ &= f_{\gamma_+}(t_+) - f_{\gamma_+}(t_+) \\ &= 0.\end{aligned}$$

This implies that

$$\int_{\gamma(t)} f\,\mathrm{d}\mu = M \cdot t = M \cdot \mu\big(\gamma(t)\big) \quad \forall t \in [0,\ \mu(R)].$$

Let $s, t \in [0,\ \mu(R)]$. Without loss of generality, let $s \leq t$. Then we have

$$\begin{aligned}\int_{\gamma(s)\triangle\gamma(t)} f \,\mathrm{d}\mu &= \int_{\gamma(t)\setminus\gamma(s)} f \,\mathrm{d}\mu \\ &= \int_{\gamma(t)} f \,\mathrm{d}\mu - \int_{\gamma(s)} f \,\mathrm{d}\mu \\ &= M \cdot \Big(\mu\big(\gamma(t)\big) - \mu\big(\gamma(s)\big)\Big) \\ &= M \cdot \mu\big(\gamma(t) \setminus \gamma(s)\big) \\ &= M \cdot \mu\big(\gamma(s) \triangle \gamma(t)\big)\end{aligned}$$

and therefore

$$\frac{1}{\mu(\gamma(s) \triangle \gamma(t))} \cdot \int_{\gamma(s)\triangle\gamma(t)} f \,\mathrm{d}\mu = M$$

which proves that γ is a constant mean geodesic of f. □

The proof described in Theorem 2.3.58 is constructive, but is notably unspecific in its choice of γ_+ and γ_-. It only states that in atomless measure spaces, such geodesics exist. Indeed, any suitable choice will produce a constant mean geodesics and there is no "preferred" or "canonical" way to construct a constant mean geodesic. An obvious choice is to select γ_+ and γ_- as minimal mean geodesics of $f - M$ and $M - f$, respectively. We refer to constant mean geodesics constructed in this way as *barycenter-focused constant mean geodesics*.

Definition 2.3.59 (Barycenter-Focused Constant Mean Geodesics).
Let $(X,\ \Sigma,\ \mu)$ be an atomless measure space, let $R \in \Sigma/_{\sim_\mu}$ with $\mu(R) < 0$, and let $f : X \to \mathbb{R} \cup \{\pm\infty\}$ be measurable with

$$\int_R |f| \,\mathrm{d}\mu < \infty.$$

Let further

$$\begin{aligned}M &\coloneqq \int_R f \,\mathrm{d}\mu, \\ R_+ &\coloneqq R \cap \{f \geq M\}, \\ R_- &\coloneqq R \cap \{f < M\}.\end{aligned}$$

Let $\gamma_+ : [0,\ \mu(R_+)] \to \Sigma/_{\sim_\mu}$ and $\gamma_- : [0,\ \mu(R_-)] \to \Sigma/_{\sim_\mu}$ be minimal mean geodesics for $(f - M)|_{R_+}$ and $(M - f)|_{R_-}$, respectively. Then we refer to the geodesic γ described in Theorem 2.3.58 as a *barycenter-focused constant mean geodesic of* f. ◁

2.3.5.3 GENERATOR GEODESICS

In this section, we discuss a class of geodesics that we will subsequently refer to as *generator geodesic*. The distinctive property of a generator geodesic is that its generated similarity space encompasses the similarity space of the entire σ-algebra. This makes generator geodesics particularly interesting in the context of rearrangement and pushforward functions because they can perfectly resolve all functions.

Definition 2.3.60 (Generator Geodesic).
Let (X, Σ, μ) be a measure space, let $I \subseteq \mathbb{R}$ be an interval. We refer to a canonical geodesic $\gamma\colon I \to {}^{\Sigma}\!/_{\sim_\mu}$ as a *generator geodesic* if and only if its generated similarity space $\sigma(\gamma)$ satisfies $\sigma(\gamma) = {}^{\Sigma}\!/_{\sim_\mu}$ and $\mathrm{TV}(\gamma) = [X]_{\sim_\mu}$. ◁

Let $\gamma\colon I \to {}^{\Sigma}\!/_{\sim_\mu}$ be a generator geodesic. As the similarity space $\sigma(\phi)$ generated by *any* canonical geodesic $\phi\colon J \to {}^{\Sigma}\!/_{\sim_\mu}$ is a subspace of ${}^{\Sigma}\!/_{\sim_\mu}$, Theorem 2.3.47 states that any canonical geodesic can be written as a rearrangement of γ. Similarly, because the generated similarity space ${}^{\sigma(f)}\!/_{\sim_\mu}$ of any measurable function $f\colon X \to \mathbb{R} \cup \{\pm\infty\}$ is a subspace of $\sigma(\gamma)$ and since $\mathrm{TV}(\gamma) = [X]_{\sim_\mu}$, Theorem 2.3.42 guarantees that we can push any such function forward through γ.

At first, one might question the requirement that $\mathrm{TV}(\gamma) = [X]_{\sim_\mu}$. After all, one would assume that this is implied by $\sigma(\gamma) = {}^{\Sigma}\!/_{\sim_\mu}$. However, this is not the case. As $\sigma(\gamma)$ is defined as the similarity space induced by the σ-algebra generated by GLSFs of γ, $\sigma(\gamma)$ also contains $\mathrm{TV}(\gamma)^{\complement}$ which is the similarity class of the preimage of $\{\infty\}$. Therefore, as long as $\mathrm{TV}(\gamma)^{\complement}$ is a μ-atom, a function must have value 0 on $\mathrm{TV}(\gamma)^{\complement}$ in order to be pushed forward through γ, which restricts the set of functions that can be pushed forward through γ.

We can safely omit this requirement if the measure space is atomless. This is because in this case $\sigma(\gamma) = {}^{\Sigma}\!/_{\sim_\mu}$ implies $\mathrm{TV}(\gamma) = [X]_{\sim_\mu}$. This is relevant because we will later show that all geodesic and weakly geodesic measure spaces are atomless. Therefore, all measure spaces in which we can perform iterative nonlinear optimization using the methods described in Chapter 3 are atomless measure spaces.

Lemma 2.3.61 (Generator Geodesics in Atomless Spaces).
Let (X, Σ, μ) be an atomless measure space, let $I \subseteq \mathbb{R}$ be an interval. A canonical geodesic $\gamma\colon I \to {}^{\Sigma}\!/_{\sim_\mu}$ is a generator geodesic if and only if $\sigma(\gamma) = {}^{\Sigma}\!/_{\sim_\mu}$. ◁

PROOF. If γ is a generator geodesic then $\sigma(\gamma) = {}^{\Sigma}\!/_{\sim_\mu}$ by definition. We therefore only need to address the inverse implication. More precisely, we only need to show that $\sigma(\gamma) = {}^{\Sigma}\!/_{\sim_\mu}$ implies $\mathrm{TV}(\gamma) = [X]_{\sim_\mu}$. We prove this indirectly.

If $\mathrm{TV}(\gamma) \neq [X]_{\sim_\mu}$, then we have $[X]_{\sim_\mu} \neq [\emptyset]_{\sim_\mu}$ because $\mathrm{TV}(\gamma) \subseteq [X]_{\sim_\mu}$. Therefore, we have $\big(\mathrm{TV}(\gamma)\big)^{\complement} = [X]_{\sim_\mu} \setminus \mathrm{TV}(\gamma) \neq [\emptyset]_{\sim_\mu}$. Let $g \in \mathcal{G}(X, \Sigma, \mu, I)$ be a GLSF corresponding to γ. Then $g^{-1}\big(\{\infty\}\big) \in \big(\mathrm{TV}(\gamma)\big)^{\complement}$ and thus $\mu\big(g^{-1}(\{\infty\})\big) > 0$. According to Lemma 2.3.45, $g^{-1}\big(\{\infty\}\big)$ is a μ-atom in $\sigma(g)$.

Because (X, Σ, μ) is atomless, there exists a subset $A \subseteq g^{-1}\big(\{\infty\}\big)$ with $A \in \Sigma$ and $0 < \mu(A) < \mu\big(g^{-1}(\{\infty\})\big)$. If there existed $B \in \sigma(g)$ with $B \sim_\mu A$, then the set $C := B \cap g^{-1}\big(\{\infty\}\big) \in \sigma(g)$ would be a subset of $g^{-1}\big(\{\infty\}\big)$ with $\mu(C) = \mu(A)$. This would contradict the fact that $g^{-1}\big(\{\infty\}\big)$ is a μ-atom in $\sigma(g)$.

There therefore exists a set $A \in \Sigma$ such that there exists no $B \in \sigma(g)$ with $B \sim_\mu A$. This implies $\sigma(\gamma) = {}^{\sigma(g)}\!/_{\sim_\mu} \neq {}^{\Sigma}\!/_{\sim_\mu}$. This indirectly shows that $\sigma(\gamma) = {}^{\Sigma}\!/_{\sim_\mu}$ implies $\mathrm{TV}(\gamma) = [X]_{\sim_\mu}$. Therefore, γ, which is assumed to be canonical, is a generator geodesic if $\sigma(\gamma) = {}^{\Sigma}\!/_{\sim_\mu}$. □

As the name suggests, generator geodesics can be constructed by interpolating between support points that are derived from a generator of the underlying σ-algebra. To do so, we have to assume that the underlying measure space is atomless. Because a generator geodesic is a canonical geodesic whose destination point is $[X]_{\sim_\mu}$, the proof of Lemma 2.3.68 will later show that the existence of a generator geodesic implies the atomlessness of the underlying measure space.

Theorem 2.3.62 (Existence of Generator Geodesics).
Let $(X,\ \Sigma,\ \mu)$ be a countably generated atomless measure space, let $(G_i)_{i\in\mathbb{N}} \subseteq \Sigma$ be a countable generator of Σ with $\mu(G_i) < \infty$ for all $i \in \mathbb{N}$. Then there exist an interval $I \subseteq \mathbb{R}$ with either $I = [0,\ \mu(X)]$ or $I = [0,\ \mu(X))$ and a generator geodesic $\gamma\colon I \to {}^{\Sigma}/_{\sim_\mu}$. ◁

Proof. Part 1 (Support Tuple). We make use of the fact that $\mathbb{N}_0$ is countable. This allows us to define the support tuple through a process of gradual refinement. Subsequently, each support point will be identified by an index tuple $(i,\ j) \in \mathbb{N}_0^2$ where j identifies the refinement level of a support point and i identifies the index of the support point within that level. To simplify notation, we include duplicates of support points from prior levels in each subsequent level.

The idea is to gradually build a countable union of all sets in the generator on level $j = 0$ and refine the steps on each subsequent level $j > 0$ by decomposing each step into its intersection with G_j and its set difference from G_j. We first introduce the "step sets" $D_{i,j}$. Let

$$D_{i,0} \coloneqq G_i \setminus \Big(\bigcup_{k=1}^{i-1} G_k\Big) \qquad \forall i \in \mathbb{N},$$

$$D_{i,j} \coloneqq \begin{cases} D_{\frac{i}{2},j-1} \cap G_j & \text{if } i \equiv 0 \pmod 2, \\ D_{\frac{i+1}{2},j-1} \setminus G_j & \text{if } i \equiv 1 \pmod 2 \end{cases} \qquad \forall j \in \mathbb{N},\ i \in \mathbb{N}.$$

It is evident that the entries of $(D_{i,0})_{i\in\mathbb{N}}$ are pairwise disjoint by construction. From this, it follows inductively that the entries $(D_{i,j})_{i\in\mathbb{N}}$ are pairwise disjoint for all levels $j \in \mathbb{N}_0$. We define

$$B_{i,j} \coloneqq \bigcup_{k=1}^{i} D_{k,j} \quad \forall (i,\ j) \in \mathbb{N}_0^2,$$
$$t_{i,j} \coloneqq \mu(B_{i,j}) \quad \forall (i,\ j) \in \mathbb{N}_0^2.$$

We note that, because $\mu(G_j) < \infty$ for all $j \in \mathbb{N}$, we have $t_{i,j} < \infty$ for all $(i,\ j) \in \mathbb{N}_0^2$. The three components of our support tuple are $T \coloneqq \{t_{i,j} \,|\, (i,\ j) \in \mathbb{N}_0^2\}$, $I \coloneqq \operatorname{conv}(T)$, and $B\colon T \to {}^{\Sigma}/_{\sim_\mu}$ with

$$B(t_{i,j}) \coloneqq \big[B_{i,j}\big]_{\sim_\mu} \quad \forall (i,\ j) \in \mathbb{N}_0.$$

The sequences $(B_{i,j})_{i\in\mathbb{N}}$ and $(t_{i,j})_{i\in\mathbb{N}}$ are straightforwardly monotonically increasing for all levels $j \in \mathbb{N}_0$. In order to show the required inclusion relations, we have to establish a relationship between different refinement levels. For all $(i,\ j) \in \mathbb{N}_0^2$, we have

$$\begin{aligned} B_{2i,j+1} &= \bigcup_{k=1}^{2i} D_{k,j+1} \\ &= \bigcup_{k=1}^{i} (D_{2k-1,j+1} \cup D_{2k,j+1}) \\ &= \bigcup_{k=1}^{i} \big((D_{k,j} \setminus G_{j+1}) \cup (D_{k,j} \cap G_{j+1})\big) \\ &= \bigcup_{k=1}^{i} D_{k,j} \\ &= B_{i,j}. \end{aligned}$$

By applying this identity inductively over $k \in \mathbb{N}_0$, we obtain

$$B_{2^k i, j+k} = B_{i,j} \quad \forall (i,\ j,\ k) \in \mathbb{N}_0^3.$$

We note that this implies $t_{2^k i, j+k} = t_{i,j}$ because of the definition of t. Let $(i,\ j)$ and $(k,\ l)$ be index tuples in $\mathbb{N}_0^2$. Without loss of generality, let $l \leq j$, which means that we have

$$\begin{aligned} t_{k,l} &= t_{2^{j-l}k,j}, \\ B_{k,l} &= B_{2^{j-l}k,j}. \end{aligned}$$

We now distinguish two cases. If $2^{j-l}k \leq i$, then the monotonicity of $i \mapsto t_{i,j}$ and $i \mapsto B_{i,j}$ guarantees that $t_{2^{j-l}k,j} \leq t_{i,j}$ and $B_{2^{j-l}k,j} \subseteq B_{i,j}$. In this case, we have

$$\begin{aligned} \mu(B_{i,j} \triangle B_{k,l}) &= \mu(B_{i,j} \triangle B_{2^{j-l}k,j}) \\ &= \mu(B_{i,j} \setminus B_{2^{j-l}k,j}) \\ &= \mu(B_{i,j}) - \mu(B_{2^{j-l}k,j}) \\ &= t_{i,j} - t_{2^{j-l}k,j} \\ &= |t_{i,j} - t_{2^{j-l}k,j}| \\ &= |t_{i,j} - t_{k,l}|. \end{aligned}$$

Conversely, if $2^{j-l}k > i$, then $t_{2^{j-l}k,j} \geq t_{i,j}$ and $B_{2^{j-l}k,j} \supseteq B_{i,j}$, which means that we have

$$\begin{aligned} \mu(B_{i,j} \triangle B_{k,l}) &= \mu(B_{i,j} \triangle B_{2^{j-l}k,j}) \\ &= \mu(B_{2^{j-l}k,j} \setminus B_{i,j}) \\ &= \mu(B_{2^{j-l}k,j}) - \mu(B_{i,j}) \\ &= t_{2^{j-l}k,j} - t_{i,j} \\ &= |t_{i,j} - t_{2^{j-l}k,j}| \\ &= |t_{i,j} - t_{k,l}|. \end{aligned}$$

This means that $t_{i,j} = t_{k,l}$ always implies $[B_{i,j}]_{\sim_\mu} = [B_{k,l}]_{\sim_\mu}$. Therefore, the mapping B is well-defined. In all cases, we have

$$\mu\big(B(t_{i,j}) \triangle B(t_{k,l})\big) = |t_{i,j} - t_{k,l}| \quad \forall (i,\ j,\ k,\ l) \in \mathbb{N}_0^4.$$

In conjunction with $I = \operatorname{conv}(T)$, this means that $(I,\ T,\ B)$ is a geodesic support tuple according to Definition 2.3.15.

For all $i, j \in \mathbb{N}_0$, we have $t_{i,j} = \mu(B_{i,j}) \geq 0$. This lower bound is also realized by

$$t_{0,0} = \mu(B_{0,0}) = \mu(\emptyset) = 0.$$

Therefore, I includes its infimum 0. This means that I is either a closed or a half-open interval with included infimum 0. Because $(G_i)_{i \in \mathbb{N}}$ is an enumeration of a countable generator of Σ, we have

$$\sup_{i,j \in \mathbb{N}_0} t_{i,j} = \sup_{i,j \in \mathbb{N}_0} \mu\Big(\bigcup_{k=1}^{i} D_{k,j}\Big) = \sup_{j \in \mathbb{N}_0} \mu\Big(\bigcup_{k=1}^{\infty} D_{k,j}\Big) = \mu\Big(\bigcup_{k=1}^{\infty} G_k\Big) = \mu(X).$$

We therefore have either $I = \big[0,\ \mu(X)\big]$ or $I = \big[0,\ \mu(X)\big)$. If $\mu(X) < \infty$, then the final geodesic can be continuously extended to $\mu(X)$.

Part 2 (Density of Parameters). In order to use the dense interpolation theorem (Theorem 2.3.18), we have to ensure that $(I,\ T,\ B)$ is a dense support tuple. To this end, we first have to prove that the σ-algebra generated by the step sets $D_{i,j}$ is equal to Σ. To simplify notation, let

$$\begin{aligned}\mathscr{D} &:= \{D_{i,j} \mid i,j \in \mathbb{N}\},\\ \mathscr{G} &:= \{G_i \mid i \in \mathbb{N}\}.\end{aligned}$$

Because the sets in $\mathscr{D}$ are constructed from the sets of the generator $\mathscr{G}$ entirely by finite union and intersection, we have $\mathscr{D} \subseteq \sigma(\mathscr{G})$ and therefore

$$\sigma(\mathscr{D}) \subseteq \sigma(\mathscr{G}) = \Sigma.$$

If we can show that all elements of $\mathscr{G}$ can be constructed by countable union from sets in $\mathscr{D}$, then that would prove the converse inclusion. For this, we can make use of the fact that the j-th step refinement level consists precisely of the intersections with and differences of prior steps with G_j.

Let $j \in \mathbb{N}$. We have

$$\begin{aligned}X &= \bigcup_{i=1}^{\infty} G_i\\ &= \bigcup_{i=1}^{\infty} B_{i,0}\\ &= \bigcup_{i=1}^{\infty} B_{2^j i,j}\\ &= \bigcup_{i=1}^{\infty} B_{i,j}\\ &= \bigcup_{i=1}^{\infty} D_{i,j}.\end{aligned}$$

Because $D_{2i-1,j} \cap G_j = \emptyset$ for all $i \in \mathbb{N}$ by definition, we have

$$\begin{aligned}G_j &= G_j \cap X\\ &= \bigcup_{i=1}^{\infty} (D_{i,j} \cap G_j)\\ &= \bigcup_{i=1}^{\infty} (\underbrace{D_{2i,j}}_{\subseteq G_j} \cap G_j)\\ &= \bigcup_{i=1}^{\infty} D_{2i,j}\end{aligned}$$

and therefore $G_j \in \sigma(\mathscr{D})$. Thus, we have $\mathscr{G} \subseteq \sigma(\mathscr{D})$ and therefore

$$\Sigma = \sigma(\mathscr{G}) \subseteq \sigma(\mathscr{D}).$$

By mutual inclusion, we have therefore proven that $\sigma(\mathscr{D}) = \Sigma$. This is significant because $(X,\ \Sigma,\ \mu)$ is atomless, which allows us to construct a proof by contradiction.

If we were to assume that T is not dense in I, then there would exist $t^* \in I$ such that t^* is not an accumulation point of T. This would imply that there is a neighborhood of t^* that does not intersect T. Let $R > 0$ be such that

$$|t^* - t_{i,j}| > R \quad \forall i, j \in \mathbb{N}_0.$$

Because $t_{0,0} = 0$, we would know that $t^* > 0$. Because t^* would not be an accumulation point of T, we would know that $t^* < \sup I$. Finally, because $i \mapsto t_{i,j}$ would be monotonically increasing for all $j \in \mathbb{N}_0$ and because $t_{k,l} = t_{2^{j-l}k,j}$ for all $(j,\ k,\ l) \in \mathbb{N}_0^3$ with $l \leq j$, we would know that $\operatorname{conv}\{t_{i,j} \mid i \in \mathbb{N}_0\} = I$ and $t_{i,j} \to \sup I$ for $i \to \infty$ for all $j \in \mathbb{N}_0$. Let

$$i_j^* := \max \underbrace{\{i \in \mathbb{N}_0 \mid t_{i,j} \leq t^*\}}_{\text{finite set}} \quad \forall j \in \mathbb{N}_0.$$

Because t^* would not be in T, we would have $t_{i_j^*,j} < t^*$ for all $j \in \mathbb{N}_0$. Furthermore, because support points are replicated on subsequent levels, $j \mapsto t_{i_j^*,j}$ would be monotonically increasing. Similarly, due to the monotonicity of $i \mapsto t_{i,j}$, $j \mapsto t_{i_j^*+1,j}$ would be monotonically decreasing with $t_{i_j^*+1,j} > t^*$ for all $j \in \mathbb{N}_0$. Due to the monotonicity theorem for geodesic support tuples (Lemma 2.3.16) the sequence

$$j \mapsto \left(B(t_{i_j^*,j}) \triangle B(t_{i_j^*+1,j})\right)$$

would essentially monotonically decreasing. We would then define

$$\begin{aligned} Q &:= \bigcap_{j=0}^{\infty} \left(B(t_{i_j^*,j}) \triangle B(t_{i_j^*+1,j})\right) \\ &= \bigcap_{j=0}^{\infty} (B_{i_j^*,j} \triangle B_{i_j^*+1,j}) \\ &= \bigcap_{j=0}^{\infty} (B_{i_j^*+1,j} \setminus B_{i_j^*,j}) \\ &= \bigcap_{j=0}^{\infty} D_{i_j^*+1,j}. \end{aligned}$$

Evidently, we have $Q \in \Sigma$. For $j \in \mathbb{N}$, depending on whether $i_j^* \equiv 1 \pmod 2$ or $i_j^* \equiv 0 \pmod 2$, we would have

$$Q \subseteq D_{i_j^*+1,j} \subseteq \begin{cases} G_j & \text{if } i_j^* \equiv 1 \pmod 2, \\ G_j^{\complement} & \text{if } i_j^* \equiv 0 \pmod 2. \end{cases}$$

For the measure of Q, we would find that

$$\begin{aligned} \mu(Q) &= \mu\Big(\bigcap_{j=0}^{\infty} (B_{i_j^*,j} \triangle B_{i_j^*+1,j})\Big) \\ &= \inf_{j \in \mathbb{N}_0} \mu(B_{i_j^*,j} \triangle B_{i_j^*+1,j}) \\ &= \inf_{j \in \mathbb{N}_0} \underbrace{|t_{i_j^*,j} - t_{i_j^*+1,j}|}_{\geq 2R} \\ &\geq 2R \\ &> 0. \end{aligned}$$

In order to show that Q would be a μ-atom, let

$$\mathscr{F} := \left(\Sigma \cap Q^{\complement}\right) \cup \left\{A \cup Q \,\middle|\, A \in \Sigma \cap Q^{\complement}\right\} \cup \left\{A \cup Q^{\complement} \,\middle|\, A \in \Sigma \cap Q^{\complement}\right\} \subseteq \Sigma$$

where $\Sigma \cap Q^{\complement}$ is the σ-algebra consisting of the intersections of sets in Σ with $Q^{\complement}$. $\mathscr{F}$ is specifically designed such that it contains no true subsets of Q other than $\emptyset$. Furthermore, $\mathscr{F}$ trivially contains $\emptyset$ and is closed under complementation and countable union. Thus, $\mathscr{F}$ is a σ-algebra.

For each $j \in \mathbb{N}$, we would have either $Q \subseteq G_j$, in which case

$$G_j = \underbrace{(G_j \setminus Q)}_{\in \Sigma \cap Q^{\complement}} \cup Q \in \mathscr{F},$$

or we would have $Q \subseteq G_j^{\complement}$, in which case

$$G_j = \underbrace{G_j \setminus Q}_{\in \Sigma \cap Q^{\complement}} \in \Sigma \cap Q^{\complement} \subseteq \mathscr{F}.$$

Therefore $\mathscr{F}$ would contain all members of $\mathscr{G}$. We would then have $\Sigma = \sigma(\mathscr{G}) \subseteq \mathscr{F}$. In conjunction with $\mathscr{F} \subseteq \Sigma$, this would mean that $\mathscr{F} = \Sigma$. However, because $\mathscr{F}$ contains no non-empty true subsets of Q, this would imply that Q would be a μ-atom in Σ, which would contradict the premise that $(X,\ \Sigma,\ \mu)$ is atomless.

To avoid this contradiction, our initial assumption that T is not dense in I must be false. Therefore T must be dense in I and $(I,\ T,\ B)$ must be a dense geodesic support tuple.

Part 3 (Construction of γ). Because $(I,\ T,\ B)$ is a dense geodesic support tuple, Theorem 2.3.18 shows that there is a unique geodesic $\gamma\colon I \to {}^{\Sigma}\!/_{\sim_\mu}$ such that

$$\gamma(t_{i,j}) = B_{i,j} \quad \forall i,j \in \mathbb{N}_0.$$

Because $\gamma(\inf I) = \gamma(t_{0,0}) = [\emptyset]_{\sim_\mu}$ is an origin point of γ, γ is canonical. Let $g \in \mathscr{G}(X,\Sigma,\mu,I)$ be a GLSF corresponding to γ. The similarity space generated by γ is given by $\sigma(\gamma) = {}^{\sigma(g)}\!/_{\sim_\mu}$. Bearing in mind that $\left(\mathrm{TV}(\gamma)\right)^{\complement}$ would have to be a μ-atom in $\sigma(g)$ if it had non-zero measure (see Lemma 2.3.45) and that $(X,\ \Sigma,\ \mu)$ is atomless, it is sufficient to show that ${}^{\Sigma}\!/_{\sim_\mu} \subseteq {}^{\sigma(g)}\!/_{\sim_\mu}$ to show that γ is a generator geodesic. $\mathrm{TV}(\gamma) = [X]_{\sim_\mu}$ then follows from the fact that the complement of $\mathrm{TV}(\gamma)$ has to be a nullset. The converse inclusion follows from the fact that g is measurable.

We have already shown that $\Sigma = \sigma(\mathscr{G}) = \sigma(\mathscr{D})$. We invoke Lemma 2.3.36 to show that ${}^{\sigma(D)}\!/_{\sim_\mu} \subseteq {}^{\sigma(g)}\!/_{\sim_\mu}$. According to that lemma, if for each $(i,\ j) \in \mathbb{N}^2$, there exists a Borel set $C_{i,j} \in \mathscr{B}(I)$ such that $g^{-1}(C_{i,j}) \sim_\mu D_{i,j}$, then for each measurable set $A \in \sigma(\mathscr{D}) = \Sigma$, there exists $B \in \sigma(g)$ with $A \sim_\mu B$.

Let $i,j \in \mathbb{N}$. We have

$$\begin{aligned} D_{i,j} &= B_{i,j} \setminus B_{i-1,j} \\ &\in B(t_{i,j}) \setminus B(t_{i-1,j}) \\ &= \gamma(t_{i,j}) \setminus \gamma(t_{i-1,j}) \\ &= \left[g^{-1}\left((t_{i-1,j},\ t_{i,j}]\right)\right]_{\sim_\mu}, \end{aligned}$$

which means that $D_{i,j} \sim_\mu g^{-1}\big((t_{i-1,j},\ t_{i,j}]\big) \in \sigma(g)$. According to Lemma 2.3.36, this means that for every $A \in \sigma(\mathscr{D}) = \Sigma$, there exists $B \in \sigma(g)$ with $A \sim_\mu B$. This implies

$$\Sigma/_{\sim_\mu} \subseteq \sigma(g)/_{\sim_\mu}.$$

As indicated, the converse inclusion follows from the measurability of g. Thus,

$$\sigma(\gamma) = \sigma(g)/_{\sim_\mu} = \Sigma/_{\sim_\mu}.$$

Because of this identity, the fact that $(X,\ \Sigma,\ \mu)$ is atomless implies that $\big(\mathrm{TV}(\gamma)\big)^{\complement}$ is a nullset. Therefore, we have $\mathrm{TV}(\gamma) = [X]_{\sim_\mu}$. Together with $\sigma(\gamma) = \Sigma/_{\sim_\mu}$ and the fact that γ is canonical, this implies that γ is a generator geodesic. □

Generator geodesics have great value in the context of relaxation methods with adaptive grid choice, where they can be used to transfer one-dimensional rounding methods to multi-dimensional problems. They could also potentially be used to restore the generated σ-algebra of a geodesic after it has been degraded through rearrangement, though exactly how one would achieve this is unclear.

2.3.6 Characterizing Geodesic Measure Spaces

In this section, we provide strong characterizations of when the similarity space associated with a measure space is geodesic. A metric space is called *geodesic* if and only if any two points within it can be connected using a geodesic.

Definition 2.3.63 (Geodesic Measure Space).
We refer to a measure space $(X,\ \Sigma,\ \mu)$ as being *geodesic* if and only if for any two similarity classes $A, B \in \Sigma/_{\sim_\mu}$, there exist an interval $I \subseteq \mathbb{R}$, parameters $s, t \in I$, and a geodesic $\gamma\colon I \to \Sigma/_{\sim_\mu}$ such that $A = \gamma(s)$ and $B = \gamma(t)$. ◁

This definition is very restrictive. In a geodesic measure spaces, the distance between $\emptyset$ and the universal set X must be realized by some finite parameter interval of a geodesic. Because parameter differences are proportional to distance for geodesics, it is evident that a geodesic measure space must be finite.

Lemma 2.3.64 (Finiteness of Geodesic Measure Spaces).
Let $(X,\ \Sigma,\ \mu)$ be a geodesic measure space. Then $\mu(X) < \infty$. ◁

PROOF. The σ-algebra Σ contains both $\emptyset$ and X. Because $(X,\ \Sigma,\ \mu)$ is geodesic, there exist an interval $I \subseteq \mathbb{R}$, parameters $s, t \in I$, and a geodesic $\gamma\colon I \to \Sigma/_{\sim_\mu}$ such that $\gamma(s) = [\emptyset]_{\sim_\mu}$ and $\gamma(t) = [X]_{\sim_\mu}$. Let $C_\gamma \geq 0$ be a geodesic constant of γ. We have

$$\mu(X) = \mu(X \mathbin{\triangle} \emptyset) = \mu\big(\gamma(t) \mathbin{\triangle} \gamma(s)\big) = C_\gamma \cdot |s - t| < \infty.$$ □

For our purposes, this restriction is not problematic because nonlinear optimization is generally only performed in bounded search spaces. However, we can mitigate the restriction somewhat by weakening the definition of geodesicity. This allows us to make use of the fact that even infinitely long measure space geodesics have well-defined origin and destination points.

Definition 2.3.65 (Weakly Geodesic Measure Space).
We refer to a measure space $(X,\ \Sigma,\ \mu)$ as *weakly geodesic* if and only if for any two similarity classes $A, B \in \Sigma/_{\sim_\mu}$, there exist an interval $I \subseteq \mathbb{R}$ and a geodesic $\gamma\colon I \to \Sigma/_{\sim_\mu}$ such that A is an origin point of γ and B is a destination point of γ. ◁

Weak geodesicity is indeed strictly weaker than strong geodesicity. To demonstrate this, we first show that every geodesic measure space is also weakly geodesic.

Lemma 2.3.66 (Strong and Weak Geodesicity).
Let (X, Σ, μ) *be a geodesic measure space. Then* (X, Σ, μ) *is weakly geodesic.* ◁

Proof. Let $A, B \in \Sigma/_{\sim_\mu}$. Because (X, Σ, μ) is geodesic, there exist an interval $I \subseteq \mathbb{R}$, parameters $s, t \in I$, and a geodesic $\gamma\colon I \to \Sigma/_{\sim_\mu}$ such that $\gamma(s) = A$ and $\gamma(t) = B$. Without loss of generality, let $s \leq t$. Otherwise we reparameterize γ with $t \mapsto -t$. The map $\gamma|_{[s,\,t]}\colon [s,\,t] \to \Sigma/_{\sim_\mu}$ is a geodesic with origin point A and destination point B. □

The fact that weak geodesicity is strictly weaker than strong geodesicity stems from the fact that it allows for connections between infinitely distant points. This means that we are no longer restricted to finite measure spaces. However, there is still a notable finiteness requirement on the underlying measure space: in order for $\emptyset$ and X to be weakly connected by a geodesic, the symmetric difference between them has to be similar to a countable union of step differences along the geodesic, all of which have to have finite measure. Therefore, X must be similar to a countable union of sets of finite measure. This means that the underlying measure space must be at least σ-finite (see Definition 2.1.2).

Lemma 2.3.67 (Weak Geodesicity and σ-Finiteness).
Let (X, Σ, μ) *be a weakly geodesic measure space. Then* (X, Σ, μ) *is* σ*-finite, i.e., there exists a sequence* $(A_i)_{i\in\mathbb{N}_0} \subseteq \Sigma$ *such that* $\mu(A_i) < \infty$ *for all* $i \in \mathbb{N}_0$ *and*

$$X = \bigcup_{i=0}^{\infty} A_i.$$

◁

Proof. Since $\emptyset \in \Sigma$ and $X \in \Sigma$, there exist an interval $I \subseteq \mathbb{R}$ and a geodesic $\gamma\colon I \to \gamma/_{\sim_\mu}$ such that $[\emptyset]_{\sim_\mu}$ is an origin point of γ and $[X]_{\sim_\mu}$ is a destination point of γ. Because $[\emptyset]_{\sim_\mu}$ is an origin point of γ, γ is canonical. Let $C_\gamma \geq 0$ be a geodesic constant of γ.

If $I = \emptyset$, then $\mu(X) = 0$ and we can set $A_i \coloneqq X$ for all $i \in \mathbb{N}_0$. If $I \neq \emptyset$, then we can find sequences $(s_i)_{i\in\mathbb{N}} \subseteq I$ and $(t_i)_{i\in\mathbb{N}} \subseteq I$ with $s_i \to \inf I$ and $t_i \to \sup I$ for $i \to \infty$, respectively. We define $\widetilde{A}_i \coloneqq \gamma(s_i) \mathbin{\triangle} \gamma(t_i)$ for all $i \in \mathbb{N}$. Then we have

$$\bigcup_{i=1}^{\infty} \widetilde{A}_i = \mathrm{TV}(\gamma) = [X]_{\sim_\mu}$$

with $\widetilde{A}_i \subseteq_\mu [X]_{\sim_\mu}$ for all $i \in \mathbb{N}$ because γ is canonical. Furthermore, we have

$$\mu(\widetilde{A}_i) = C_\gamma \cdot |s_i - t_i| < \infty \quad \forall i \in \mathbb{N}.$$

For each $i \in \mathbb{N}$, let $A_i \in \widetilde{A}_i$ be a representative of $\widetilde{A}_i$ such that $A_i \subseteq X$. We have

$$\bigcup_{i=1}^{\infty} A_i \sim_\mu X.$$

Let

$$A_0 \coloneqq X \mathbin{\triangle} \bigcup_{i=1}^{\infty} A_i = X \setminus \bigcup_{i=1}^{\infty} A_i.$$

Because $X \sim_\mu \bigcup_{i=1}^{\infty} A_i$, we have $\mu(A_0) = 0$. We also have

$$\bigcup_{i=0}^{\infty} A_i = A_0 \cup \bigcup_{i=1}^{\infty} A_i = X$$

which proves that (X, Σ, μ) is σ-finite. □

On its own, σ-finiteness is not yet a sufficient condition for weak geodesicity. There is a second necessary requirement for geodesicity that complements the finiteness requirement: atomlessness. Together, atomlessness and σ-finiteness will turn out to be sufficient to imply weak geodesicity. We can already see the atomlessness requirement being foreshadowed in Lemma 2.3.45. Geodesicity requires the existence of geodesics whose total variation is represented by the universal set X. Therefore, μ-atoms must be confined to the complement of $[X]_{\sim_\mu}$ which consists entirely of nullsets. Since nullsets cannot be atoms, this means that weakly geodesic measure spaces must be atomless.

Lemma 2.3.68 (Weak Geodesicity and Atomlessness).
Let (X, Σ, μ) be a weakly geodesic measure space. Then (X, Σ, μ) is atomless. ◁

PROOF. We prove the claim by contradiction. We assume that there exists a μ-atom $A \in \Sigma$. Then there would exist an interval $I \subseteq \mathbb{R}$ and a geodesic $\gamma\colon I \to \Sigma/_{\sim_\mu}$ such that $[\emptyset]_{\sim_\mu}$ is an origin point of γ and $[A]_{\sim_\mu}$ is a destination point of γ. Because $[\emptyset]_{\sim_\mu}$ is an origin point of γ, γ is canonical. Since $\mu([A]_{\sim_\mu}) = \mu(A) > 0$, γ has a strictly positive geodesic constant $C_\gamma > 0$ and I has nonzero length.

Because I has nonzero length, we can select $t_0 \in I$ from the interior of I. The fact that γ is canonical means that we have $\gamma(t_0) \subseteq_\mu \mathrm{TV}(\gamma) = [A]_{\sim_\mu}$. Let $B \in \gamma(t_0)$ be a representative of $\gamma(t_0)$ such that $B \subseteq A$.

Because t_0 is in the interior of I, there exists $t_1 \in I$ with $t_1 > t_0$. We have

$$\mu(B) = \mu\big(\gamma(t_0)\big) = C_\gamma \cdot t_0 > 0$$

and

$$\mu\big(\gamma(t_1)\big) - \mu(B) = \mu\big(\gamma(t_1)\big) - \mu\big(\underbrace{\gamma(t_0)}_{\subseteq_\mu \gamma(t_1)}\big) = \mu\big(\gamma(t_1) \triangle \gamma(t_0)\big) = C_\gamma \cdot |t_1 - t_0| > 0.$$

The first estimate proves that $\mu(B) > 0$. The second estimate demonstrates that $\mu(B) < \mu\big(\gamma(t_1)\big) \leq \mu(A)$. In conjunction with the fact that B is a measurable set with $B \subseteq A$, this contradicts the assumption that A is a μ-atom. The contradiction shows that (X, Σ, μ) must be atomless. □

We now know that σ-finiteness and atomlessness are necessary conditions for weak geodesicity. Similarly, because strong geodesicity implies weak geodesicity, we know that finiteness and atomlessness are necessary conditions for strong geodesicity. We now demonstrate the converse implication.

Lemma 2.3.69 (Weakly Connecting Geodesics).
Let (X, Σ, μ) be an atomless, σ-finite measure space, and let $A, B \in \Sigma/_{\sim_\mu}$. Then there exist an interval $I \subseteq \mathbb{R}$ and a geodesic $\gamma\colon I \to \Sigma/_{\sim_\mu}$ such that A is an origin point of γ and B is a destination point of γ. ◁

PROOF. We proceed in three stages. First, we show that there exists a canonical geodesic $\gamma_1 \colon J \to \Sigma/_{\sim_\mu}$ with destination point $[\Sigma]_{\sim_\mu}$. We use restriction in image to construct a canonical geodesic $\gamma_2 \colon I \to \Sigma/_{\sim_\mu}$ with destination point $A \triangle B$. Finally, we translate γ_2 by A to obtain a geodesic with origin point A and destination point B.

PART 1 (CONNECTING $[\emptyset]_{\sim_\mu}$ AND $[X]_{\sim_\mu}$). Because (X, Σ, μ) is σ-finite, there exists a sequence $(A_i)_{i\in\mathbb{N}} \subseteq \Sigma$ with $\mu(A_i) < \infty$ for all $i \in \mathbb{N}$ and

$$\bigcup_{i=1}^{\infty} A_i = X.$$

Let

$$B_i \coloneqq \Big[\bigcup_{j=1}^{i} A_j\Big]_{\sim_\mu} \in \Sigma/_{\sim_\mu} \quad \forall i \in \mathbb{N}_0.$$

We note that $B_0 = [\emptyset]_{\sim_\mu}$. The sequence $i \mapsto B_i$ is essentially increasing and satisfies

$$\mu(B_{i,0}) \leq \sum_{j=1}^{i} \mu(A_i) < \infty \quad \forall i \in \mathbb{N}_0.$$

Because $i \mapsto B_i$ is essentially increasing, so is $i \mapsto t_i$ with

$$t_i \coloneqq \mu(B_i) \quad \forall i \in \mathbb{N}_0.$$

Let $i, j \in \mathbb{N}_0$. Without loss of generality, let $i \leq j$. Then we have $B_i \subseteq_\mu B_j$ and $t_i \leq t_j$. We further have

$$\mu(B_i \triangle B_j) = \mu(B_j \setminus B_i) = \mu(B_j) - \mu(B_i) = t_j - t_i = |t_j - t_i|.$$

This means that for $T \coloneqq \{t_i \mid i \in \mathbb{N}_0\}$, $I \coloneqq \operatorname{conv}(T)$, and $B \colon T \to \Sigma/_{\sim_\mu}$ with

$$B(t_i) \coloneqq B_i \quad \forall i \in \mathbb{N}_0,$$

the tuple (I, T, B) is a geodesic support tuple according to Definition 2.3.15. Because (X, Σ, μ) is atomless, according to the sparse interpolation theorem (Theorem 2.3.22), there exists a geodesic $\gamma_1 \colon I \to \Sigma/_{\sim_\mu}$ with $\gamma_1(t_i) = B_i$ for all $i \in \mathbb{N}_0$. Because $t_0 = 0 = \inf I$, the origin point of γ_1 is $[\emptyset]_{\sim_\mu}$, which means that γ_1 is canonical. Because $i \to t_i$ is monotonically increasing, we have $t_i \to \sup I$ for $i \to \infty$. Therefore, we have

$$\begin{aligned}
\operatorname{TV}(\gamma_1) &= \bigcup_{i=0}^{\infty} \big(\underbrace{\gamma(0)}_{=[\emptyset]_{\sim_\mu}} \triangle \gamma(t_i) \big) \\
&= \bigcup_{i=0}^{\infty} \gamma(t_i) \\
&= \bigcup_{i=0}^{\infty} B_i \\
&= \Big[\bigcup_{i=0}^{\infty} \bigcup_{j=1}^{i} A_j\Big]_{\sim_\mu} \\
&= \Big[\bigcup_{i=1}^{\infty} A_i\Big]_{\sim_\mu} \\
&= [X]_{\sim_\mu}.
\end{aligned}$$

Because γ_1 is canonical, $\mathrm{TV}(\gamma_1)$ is the destination point of γ_1. Thus, γ_1 weakly connects $[\emptyset]_{\sim_\mu}$ and $[X]_{\sim_\mu}$.

PART 2 (CONNECTING $[\emptyset]_{\sim_\mu}$ AND $A \triangle B$). Let γ_1 be the geodesic constructed in the previous part. If $\mu(X) < \infty$, then we invoke Theorem 2.3.14 to extend γ_1 to the parameter interval $[0,\ \mu(X)]$ without changing its total variation. Let $J \subseteq \mathbb{R}$ be the resulting parameter interval which is either closed or equal to the half open interval $[0,\ \infty)$.

Let $A, B \in \Sigma/_{\sim_\mu}$. Let $\mu_{\gamma_1, A \triangle B}$ be the localized measure of variation according to Lemma 2.3.50, and let $I \coloneqq \mu_{\gamma_1,\ A \triangle B}(J)$. Let $\gamma_2 \coloneqq \gamma_1 \| _{A \triangle B} : I \to \Sigma/_{\sim_\mu}$ be the restriction in image of γ_1 to $A \triangle B$. Then γ_2 is a canonical geodesic with origin point $\gamma_2(0) = [\emptyset]_{\sim_\mu}$ and destination point

$$\mathrm{TV}(\gamma_2) = \mathrm{TV}(\gamma_1) \cap (A \triangle B) = [X]_{\sim_\mu} \cap (A \triangle B) = A \triangle B$$

according to Theorem 2.3.51.

PART 3 (CONNECTING A AND B). Let γ_2 be the geodesic connecting $[\emptyset]_{\sim_\mu}$ with $A \triangle B$ that we have constructed in the previous part. By translating γ_2 we obtain $\gamma \coloneqq \gamma_2 \triangle A : I \to \Sigma/_{\sim_\mu}$ with origin point

$$[\emptyset]_{\sim_\mu} \triangle A = A$$

and destination point

$$\mathrm{TV}(\gamma_2) \triangle A = (A \triangle B) \triangle A = B. \qquad \square$$

With this, we have both necessary and sufficient conditions for weak geodesicity, which means that we have an "if and only if" characterization of weakly geodesic measure spaces.

Theorem 2.3.70 (Strong Characterization of Weak Geodesicity).
A measure space $(X,\ \Sigma,\ \mu)$ is weakly geodesic if and only if it is both atomless and σ-finite. $\lhd$

PROOF. PART 1 ($\Longrightarrow$). For this part of the proof, we invoke Lemmas 2.3.67 and 2.3.68.

PART 2 ($\Longleftarrow$). For this part of the proof, we invoke Lemma 2.3.69. $\square$

As one might expect, the strong characterization of strong geodesicity is similar, except that it requires finiteness instead of σ-finiteness.

Theorem 2.3.71 (Strong Characterization of Geodesicity).
A measure space $(X,\ \Sigma,\ \mu)$ is geodesic if and only if it is both atomless and finite. $\lhd$

PROOF. PART 1 ($\Longrightarrow$). We invoke Lemma 2.3.66 and Theorem 2.3.70 to show that $(X,\ \Sigma,\ \mu)$ is atomless. We invoke Lemma 2.3.64 to show that $(X,\ \Sigma,\ \mu)$ is finite.

PART 2 ($\Longleftarrow$). Since (X, Σ, μ) is finite, it is also σ-finite. Theorem 2.3.70 states that (X, Σ, μ) is weakly geodesic.

Let $A, B \in \Sigma/_{\sim_\mu}$. If $A = B$, then A and B can be connected using the geodesic $\gamma\colon [0, 0] \to \Sigma/_{\sim_\mu}$ with $\gamma(0) := A = B$. Let subsequently $A \neq B$. Because (X, Σ, μ) is weakly geodesic, there exist an interval $I \subseteq \mathbb{R}$ and a geodesic $\gamma\colon I \to \Sigma/_{\sim_\mu}$ with origin point A and destination point B.

Because $A \neq B$, γ must have a strictly positive geodesic constant $C_\gamma > 0$. The length of the interval I is bounded by

$$\lambda(I) \leq \frac{\mu(A \triangle B)}{C_\gamma} \leq \frac{\mu(X)}{C_\gamma} < \infty.$$

Thus, I must be bounded. According to Theorem 2.3.14, we can extend γ to the closure $\overline{I}$ of I using its limit points. We then have $\gamma(\inf I) = A$ and $\gamma(\sup I) = B$.□

This closes our discussion of measure space geodesics. We now have a firm grasp of what measure space geodesics are, how to work with them, and what kinds of requirements a space must satisfy to have them. With this, we can now focus on the theory of optimization in similarity spaces.

2.4 DIFFERENTIABLE FUNCTIONS IN SIMILARITY SPACES

Our discussion is limited to optimization problems where all involved functionals are differentiable. In this section, we discuss the meaning of "differentiability." Differentiability is a concept that is usually discussed in vector space contexts. Since some concepts from vector spaces are not transferrable to similarity spaces, we define differentiability slightly differently in our context.

Throughout this section, we assume that the underlying measure space (X, Σ, μ) is finite and atomless. We discuss set functionals of the form

$$F\colon \Sigma/_{\sim_\mu} \to \mathbb{R}$$

which satisfy a differentiability criterion analogous to Taylor's theorem.

We introduce this criterion in Section 2.4.1 and show necessary conditions for local optimality as well as a curvature-based sub-optimality estimate. To show that our derivative concept is practically useful, we show in Section 2.4.2 that suitable derivatives can be derived from Banach space derivatives. Sections 2.4.3 and 2.4.4 apply this approach to reduced functions whose evaluation requires the solution of an ODE or a sufficiently simple PDE.

2.4.1 Taylor Criterion

Definition 2.4.1 and Proposition 2.4.7 were first formulated in [HLS22]. The concept of benign differentiability was added in this thesis. The curvature criterion in Proposition 2.4.8 was stated in a weaker form that did not require geodesics but required additional assumptions about the functional.

We base our differentiability concept on a truncated first order Taylor expansion. In vector spaces, the first order derivative is generally expressed as a linear form. The analogue of a linear form in a similarity space is a signed measure. To ensure that this measure is well-defined on the similarity space, we require that it is absolutely continuous with respect to the measure of the underlying measure space because this allows us to invoke Proposition 2.2.13.

Definition 2.4.1 (Differentiability).
Let $(X,\ \Sigma,\ \mu)$ be a finite atomless measure space. We refer to a function $F\colon \Sigma/_{\sim_\mu} \to \mathbb{R}$ as *differentiable in* $U \in \Sigma/_{\sim_\mu}$ if and only if there exists a finite signed measure

$$\phi(U)\colon \Sigma \to \mathbb{R}$$

with $\phi(U) \ll \mu$ such that

$$F(U \triangle D) - F(U) = \phi(U)(D) + o\big(\mu(D)\big) \quad \forall D \in \Sigma/_{\sim_\mu}.$$

In this case, we refer to $\phi(U)$ as the *derivative* or *gradient* of F in U. We write $\nabla F(U) \coloneqq \phi$. We refer to F as being *differentiable* if and only if F is differentiable in every $U \in \Sigma/_{\sim_\mu}$. ◁

Because part of our definition of differentiability is that the signed measure $\phi(U)$ must be absolutely continuous with respect to μ, we can invoke the Radon-Nikodym theorem to show that the signed gradient measure has a density function.

Lemma 2.4.2 (Gradient Density Function).
Let $(X,\ \Sigma,\ \mu)$ *be a finite atomless measure space, let* $U \in \Sigma/_{\sim_\mu}$*, and let* $F\colon \Sigma/_{\sim_\mu} \to \mathbb{R}$ *be differentiable in* U*. Then there exists an integrable function* $g(U) \in L^1(\Sigma, \mu)$ *such that*

$$\nabla F(U)(D) = \int_D g(U)\,\mathrm{d}\mu \quad \forall D \in \Sigma.$$

We refer to $g(U)$ *as the* gradient density function of F in U. ◁

PROOF. According to Definition 2.4.1, we have $\nabla F(U) \ll \mu$. According to the Radon-Nikodym theorem (Theorem 2.1.12), there exists a μ-integrable function $g(U)\colon X \to \mathbb{R}$ such that

$$\nabla F(U)(D) = \int_D g(U)\,\mathrm{d}\mu \quad \forall D \in \Sigma.$$

□

An interesting phenomenon arises because, in similarity spaces, every class is its own additive inverse. This means that, even for small steps $U \triangle V$, the gradient ∇F has to invert sign on the similarity class $U \triangle V$. For continuously differentiable functionals, we would therefore expect that

$$\big(\nabla F(U) - \nabla F(V)\big)(U \triangle V) \approx 2 \cdot \nabla F(U)(U \triangle V).$$

This is potentially a concern because it shows that, even for "continuously" differentiable set functionals, where the gradient measures in two points should converge as the distance between the points decreases, there will always be at least a small set where the average value of the gradient measure changes to a value close to the additive inverse.

To compensate for this effect, we define the "distance" between derivatives slightly differently. When we compare $\nabla F(U)$ with $\nabla F(V)$, we use the variation of their differences as we normally would. However, inside of the step $U \triangle V$, we use the variation of the sum instead, because we expect the sign to be inverted. We refer to the resulting positive measure as the "locally inverted difference variation."

Definition 2.4.3 (Locally Inverted Difference Variation).
Let $(X,\ \Sigma,\ \mu)$ be a measure space, let $R \in \Sigma/_{\sim_\mu}$, let $S(\Sigma,\mu)$ be the set of all finite signed measures $\varphi\colon \Sigma \to \mathbb{R}$ with $\varphi \ll \mu$, and let $P(\Sigma,\mu)$ be the set of all finite positive measures $\varphi\colon \Sigma \to \mathbb{R}_{\geq 0}$ with $\varphi \ll \mu$. We refer to the map $(\cdot \underset{R}{\ominus} \cdot)\colon S(\Sigma,\mu) \times S(\Sigma,\mu) \to P(\Sigma,\mu)$ with

$$(\varphi \underset{R}{\ominus} \nu)(D) := |\varphi - \nu|(D \setminus R) + |\varphi + \nu|(D \cap R) \quad \forall(\varphi,\ \nu,\ D) \in \big(S(\Sigma,\mu)\big)^2 \times \Sigma$$

as the *R-locally inverted difference variation*. ◁

We then define continuous differentiability to be the property that for sufficiently small steps $U \triangle V$, the $(U \triangle V)$-locally inverted difference variation between $\nabla F(U)$ and $\nabla F(V)$ disappears.

Definition 2.4.4 (Continuous Differentiability).
Let $(X,\ \Sigma,\ \mu)$ be a finite, atomless measure spaces. We refer to a differentiable set functional $F\colon \Sigma/_{\sim_\mu} \to \mathbb{R}$ as

(1) *continuously differentiable in $U \in \Sigma/_{\sim_\mu}$* if and only if for every $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\big(\nabla F(U) \underset{U \triangle V}{\ominus} \nabla F(V)\big)(D) \leq \varepsilon \cdot \mu(D) \quad \forall V, D \in \Sigma/_{\sim_\mu}\colon \mu(U \triangle V) \leq \delta;$$

(2) *continuously differentiable* if and only if F is continuously differentiable in every $U \in \Sigma/_{\sim_\mu}$;

(3) *uniformly continuously differentiable* if and only if for every $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\big(\nabla F(U) \underset{U \triangle V}{\ominus} \nabla F(V)\big)(D) \leq \varepsilon \cdot \mu(D) \quad \forall U, V, D \in \Sigma/_{\sim_\mu}\colon \mu(U \triangle V) \leq \delta;$$

(4) *locally Lipschitz-continuously differentiable* if and only if for every $U \in \Sigma/_{\sim_\mu}$, there exists a neighborhood $\mathcal{N}$ of U and constant $L \geq 0$ such that

$$\big(\nabla F(V) \underset{V \triangle W}{\ominus} \nabla F(W)\big)(D) \leq L \cdot \mu(V \triangle W) \cdot \mu(D) \quad \forall(V,\ W) \in \mathcal{N}^2,\ D \in \Sigma/_{\sim_\mu}.$$

In this case, we refer to the constant L as a *Lipschitz constant* of the derivative on $\mathcal{N}$;

(5) *Lipschitz-continuously differentiable* if and only if there exists a constant $L \geq 0$ such that

$$\big(\nabla F(U) \underset{U \triangle V}{\ominus} \nabla F(V)\big)(D) \leq L \cdot \mu(U \triangle V) \cdot \mu(D) \quad \forall U, V, D \in \Sigma/_{\sim_\mu}.$$

In this case, we refer to the constant L as a *Lipschitz constant* of the derivative. ◁

Because the underlying measure space is finite, it is evident that Lipschitz continuous differentiability implies uniform continous differentiability, which in turn implies continuous differentiability.

The definitions of continuous and Lipschitz-continuous differentiability appear very restrictive because they require that the locally inverted difference

variation be bounded above by a multiple of the "step measure" $\mu(D)$, which would appear to imply some sort of L^∞ continuity of the gradient measure's density function. This type of continuity is essential when estimating changes along a geodesic as we shall see in Proposition 2.4.8. We show in Theorem 2.4.13 that these types of continuity can be derived from corresponding continuity of Fréchet derivatives.

We note that, even if the derivative is continuous in some way, the integrable density function can still assume arbitrarily high local average values as long as the set on which they are assumed is sufficiently small. This is because the absolute continuity $\nabla F(U) \ll \mu$ implies $\nabla F(U) \to 0$ for $\mu(U) \to 0$, but not necessarily $\nabla F(U) = o\big(\mu(U)\big)$.

For unconstrained optimization, the latter is only required for suboptimality estimation. However, we will see that it is required to formulate a chain rule for compositions where a differentiable real function is applied to the result of a diffentiable set functional. This is essential for penalty or barrier methods where the output of a set functional is used as an argument to a nonlinear function. We will also see that the way in which a functional changes *along a geodesic* becomes much more significant in the context of constrained optimization.

Definition 2.4.5 (Benign Differentiality).
Let $(X,\ \Sigma,\ \mu)$ be a finite, atomless measure space. We refer to a differentiable set functional $F\colon \Sigma/_{\sim_\mu} \to \mathbb{R}$ as *benignly differentiable in* $U \in \Sigma/_{\sim_\mu}$ whenever the density function g of $\nabla F(U)$ is essentially bounded, i.e., $g \in L^\infty(\Sigma, \mu)$. We refer to signed measures with essentially bounded density functions as *benign*.

We call F *benignly differentiable* if it is benignly differentiable everywhere. Similarly, we allow the adjective *benign* or its adverb to be combined with the various types of continous differentiability given in Definition 2.4.4. ◁

For benignly differentiable set functionals, the aforementioned chain rule follows straightforwardly.

Theorem 2.4.6 (Chain Rule for Benignly Differentiable Functionals).
Let $(X,\ \Sigma,\ \mu)$ be a finite, atomless measure space. Let $F\colon \Sigma/_{\sim_\mu} \to \mathbb{R}$ be benignly differentiable in $U \in \Sigma/_{\sim_\mu}$ and let $h\colon F(\Sigma/_{\sim_\mu}) \to \mathbb{R}$ be differentiable in $F(U)$.

Then the set functional $h \circ F\colon \Sigma/_{\sim_\mu} \to \mathbb{R}$ obtained by applying h to the result of F is benignly differentiable in U and its derivative satisfies

$$\nabla(h \circ F)(U) = h'\big(F(U)\big) \cdot \nabla F(U). \qquad \triangleleft$$

PROOF. Let $V \in \Sigma/_{\sim_\mu}$. We have

$$\begin{aligned} h\big(F(V)\big) - h\big(F(U)\big) &= h'\big(F(U)\big) \cdot \big(F(V) - F(U)\big) + o\Big(\big|F(V) - F(U)\big|\Big) \\ &= h'\big(F(U)\big) \cdot \Big(\nabla F(U)(U \mathbin{\triangle} V) + o\big(\mu(U \mathbin{\triangle} V)\big)\Big) + o\Big(\big|F(V) - F(U)\big|\Big) \\ &= h'\big(F(U)\big) \cdot \nabla F(U)(U \mathbin{\triangle} V) + o\big(\mu(U \mathbin{\triangle} V)\big) + o\Big(\big|F(V) - F(U)\big|\Big). \end{aligned}$$

Let $g \in L^1(\Sigma, \mu) \cap L^\infty(\Sigma, \mu)$ be the density function of $\nabla F(U)$. Because the density function is μ-essentially bounded, there exists a constant $C < \infty$ such that $|g| \leq C$

μ-almost everywhere in X. We therefore have

$$\begin{aligned}\big|F(V)-F(U)\big| &= \Big|\nabla F(U)(U \triangle V) + o\big(\mu(U \triangle V)\big)\Big| \\ &= \left|\int_{U\triangle V} g \,\mathrm{d}\mu + o\big(\mu(U \triangle V)\big)\right| \\ &\leq \int_{U\triangle V} |g| \,\mathrm{d}\mu + o\big(\mu(U \triangle V)\big) \\ &\leq C \cdot \mu(U \triangle V) + o\big(\mu(U \triangle V)\big).\end{aligned}$$

Therefore, we have

$$\begin{aligned}h\big(F(V)\big) - h\big(F(U)\big) &= h'\big(F(U)\big) \cdot \nabla F(U)(U \triangle V) + o\big(\mu(U \triangle V)\big) + o\Big(\big|F(V)-F(U)\big|\Big) \\ &= h'\big(F(U)\big) \cdot \nabla F(U)(U \triangle V) + o\big(\mu(U \triangle V)\big) + o\Big(o\big(\mu(U \triangle V)\big)\Big) \\ &= h'\big(F(U)\big) \cdot \nabla F(U)(U \triangle V) + o\big(\mu(U \triangle V)\big),\end{aligned}$$

which means that $h \circ F$ is differentiable in U with the derivative measure $\nabla(h \circ F) = h'\big(F(U)\big) \cdot \nabla F(U)$. □

We will make extensive use of this chain rule when we discuss penalty methods in Section 3.2. For now, we turn our attention to the formulation of first-order necessary optimality criteria.

We can relatively easily show that if the point U around which the derivative is developed is a local optimum, then the gradient measures are positive measures and the gradient density functions are non-negative almost everywhere.

Proposition 2.4.7 (Necessary Optimality Criterion).
Let (X, Σ, μ) be a finite atomless measure space, and let $F : \Sigma/_{\sim_\mu} \to \mathbb{R}$ be differentiable in $U \in \Sigma/_{\sim_\mu}$ such that there exists a neighborhood $\mathcal{N} \subseteq \Sigma/_{\sim_\mu}$ of U with

$$F(V) \geq F(U) \quad \forall V \in \mathcal{N}.$$

Then we have

$$\nabla F(U)(D) \geq 0 \quad \forall D \in \Sigma.$$

If $g(U)$ is the gradient density function of F in U, then this is equivalent to

$$\int_X \min\{g(U),\, 0\} \,\mathrm{d}\mu = 0 \quad \textit{for a.a. } x \in X.$$

◁

Proof. We prove the claim indirectly. Let $D^* \in \Sigma$ be such that

$$\nabla F(U)(D^*) < 0.$$

According to the Hahn decomposition theorem (Theorem 2.1.5), we can partition D^* into $D^*_+, D^*_- \in \Sigma$ such that

$$\begin{aligned}\nabla F(U)(D) &\geq 0 \quad \forall D \in \Sigma : D \subseteq D^*_+, \\ \nabla F(U)(D) &\leq 0 \quad \forall D \in \Sigma : D \subseteq D^*_-.\end{aligned}$$

We then have

$$\nabla F(U)(D_-^*) \leq \underbrace{\nabla F(U)(D_+^*)}_{\geq 0} + \nabla F(U)(D_-^*) = \nabla F(U)(D^*) < 0$$

and therefore $\nabla F(U)(D_-^*) < 0$. Since $\nabla F(U) \ll \mu$, $\nabla F(U)(D_-^*) \neq 0$ implies that $\mu(D_-^*) > 0$.

Let $R_0 > 0$ be fixed such that $V \in \mathcal{N}$ for all $V \in \Sigma/_{\sim_\mu}$ with $\mu(U \triangle V) \leq R_0$. Such an R_0 exists because $\mathcal{N}$ is a neighborhood of U. According to Definition 2.4.1, there exists $R_1 > 0$ such that

$$\left|F(U) - F(V) - \nabla F(U)(U \triangle V)\right| \leq \frac{\left|\nabla F(U)(D_-^*)\right|}{2\mu(D_-^*)} \cdot \mu(U \triangle V) \quad \forall V \in \Sigma/_{\sim_\mu} \colon \mu(U \triangle V) \leq R_1.$$

Let $R \coloneqq \min\{R_0, R_1\}$. Since (X, Σ, μ) is finite and atomless, [Bog07, Thm. 1.12.9] demonstrates that there exists a finite partition $(D_{-i}^*)_{i \in [n]} \in \Sigma^n$ of D_-^* such that $\mu(D_{-i}^*) \leq R$ for all $i \in [n]$. Without loss of generality, let $\mu(D_{-i}^*) > 0$ for all $i \in [n]$. This is always achievable because $\mu(D_-^*) > 0$, which implies that there exists at least one $i \in [n]$ with $\mu(D_{-i}^*) > 0$ to which nullsets can be attached.

It is evident by contradiction that there must exist $i^* \in [n]$ with

$$\nabla F(U)(D_{-i^*}^*) \leq \nabla F(U)(D_-^*) \cdot \frac{\mu(D_{-i^*}^*)}{\mu(D_-^*)}.$$

This is evident because the negation would imply that

$$\begin{aligned}
\nabla F(U)(D_-^*) &= \sum_{i=1}^n \nabla F(U)(D_{-i}^*) \\
&= \sum_{i=1}^n \nabla F(U)(D_{-i}^*) \cdot \frac{\mu(D_{-i}^*)}{\mu(D_{-i}^*)} \\
&= \sum_{i=1}^n \underbrace{\mu(D_{-i}^*)}_{>0} \cdot \underbrace{\frac{\nabla F(U)(D_{-i}^*)}{\mu(D_{-i}^*)}}_{> \frac{\nabla F(U)(D_-^*)}{\mu(D_-^*)}} \\
&> \frac{\nabla F(U)(D_-^*)}{\mu(D_-^*)} \cdot \underbrace{\sum_{i=1}^n \mu(D_{-i}^*)}_{= \mu(D_-^*) > 0} \\
&= \nabla F(U)(D_-^*).
\end{aligned}$$

Let subsequently $D \coloneqq D_{-i^*}^*$ and let $V \coloneqq U \triangle D$. We have

$$\mu(U \triangle V) = \mu(D) \leq R \leq \min\{R_0, R_1\}$$

and therefore $V \in \mathcal{N}$ as well as

$$\begin{aligned} F(V) &\leq F(U) + \nabla F(U)(U \triangle V) + \frac{\overbrace{\left|\nabla F(U)(D_-^*)\right|}^{<0}}{2\mu(D_-^*)} \cdot \underbrace{\mu(U \triangle V)}_{=\mu(D)} \\ &= F(U) + \underbrace{\nabla F(U)(D)}_{\leq \frac{\nabla F(U)(D_-^*)}{\mu(D_-^*)} \cdot \mu(D)} - \frac{1}{2} \frac{\mu(D)}{\mu(D_-^*)} \cdot \nabla F(U)(D_-^*) \\ &\leq F(U) + \frac{1}{2} \underbrace{\frac{\mu(D)}{\mu(D_-^*)}}_{>0} \cdot \underbrace{\nabla F(U)(D_-^*)}_{<0} \\ &< F(U). \end{aligned}$$

We have thus proven that, if $\nabla F(U)(D) < 0$ for any $D \in \Sigma$, then every neighborhood $\mathcal{N}$ of U contains a point $V \in \mathcal{N}$ with $F(V) < F(U)$. Conversely, $F(U) \leq F(V)$ for all V within some neighborhood $\mathcal{N}$ of U implies $\nabla F(U)(D) \geq 0$ for all $D \in \Sigma$.

Let $g(U)$ be the gradient density function of F in U. We have

$$\int_D g \, \mathrm{d}\mu = \nabla F(U)(D) \geq 0 \quad \forall D \in \Sigma.$$

By choosing $D := \{g(U) < 0\}$, we obtain

$$\int_{\{g(U)<0\}} g(U) \, \mathrm{d}\mu = \nabla F(U)\Big(\big\{g(U) < 0\big\}\Big) \geq 0.$$

Since $\mu\big(\{g(U) < 0\}\big) > 0$ would imply that the integral is strictly negative, we have $\mu\big(\{g(U) < 0\}\big) = 0$ and therefore

$$g(U)(x) \geq 0 \text{ almost everywhere.} \qquad \square$$

Proposition 2.4.7 can be thought of as a necessary condition for local optimality in unconstrained problems. Such results are only of limited use. Generally speaking, similarity spaces are not compact. Therefore, it is often difficult to prove convergence within them. Thus, a sequence of feasible points such that the negative parts of $\nabla F(U)$ converge to zero need not converge to an optimum.

The situation is slightly different if one moves along a geodesic. Because geodesics prohibit backtracking, they always converge to their destination point. Generally, however, we cannot ensure that our optimization algorithms move along geodesics. Therefore, we must work around the possibility that an optimum of a given function may not exist.

Even if the optimum does not exist, that does not mean that we cannot reach approximate optimality. To assess approximate optimality, we have to find underestimators for the function. The first such underestimator applies if the derivative is Lipschitz continuous.

Proposition 2.4.8 (Curvature Limit Underestimator).
Let (X, Σ, μ) be a finite atomless measure space, and let $F : \Sigma/_{\sim_\mu} \to \mathbb{R}$ be Lipschitz continuously differentiable with Lipschitz constant $L \geq 0$. Then we have

$$\left|F(V) - F(U) - \nabla F(U)(U \triangle V)\right| \leq \frac{L}{2}\big(\mu(U \triangle V)\big)^2 \quad \forall U, V \in \Sigma/_{\sim_\mu}.$$

Let $U \in \Sigma/_{\sim_\mu}$, and let $g(U)$ be the gradient density function of $\nabla F(U)$. For all $R > 0$, we have

$$F(V) \geq F(U) + \left(\int_{U \triangle V} \min\{g(U),\ 0\} \, \mathrm{d}\mu \right) - \frac{L}{2} \cdot R^2$$

for all $V \in \Sigma/_{\sim_\mu}$ with $\mu(U \triangle V) \leq R$. ◁

PROOF. Let $U, V \in \Sigma/_{\sim_\mu}$. According to Theorem 2.3.71, there is an interval $I \subseteq \mathbb{R}$, parameters $s, t \in I$, and a geodesic $\gamma \colon I \to \Sigma/_{\sim_\mu}$ with $\gamma(s) = U$ and $\gamma(t) = V$. After reparameterization and restriction, we may assume without loss of generality that $I = [0,\ \mu(U \triangle V)]$, $\gamma(0) = U$, and $\gamma\big(\mu(U \triangle V)\big) = V$. This then implies that $C = 1$ is a geodesic constant of γ. Let $\check{\gamma} = \gamma \triangle U$ be the canonical form of γ.

For each parameter $t \in I$, F is differentiable around $\gamma(t)$. Therefore, for every $j \in \mathbb{N}$ and every $t \in I$, there exists $R_j(t) > 0$ such that

$$\Big| F(W) - F\big(\gamma(t)\big) - \nabla F\big(\gamma(t)\big)\big(W \triangle \gamma(t)\big) \Big| \leq \frac{1}{2^j} \cdot \mu\big(W \triangle \gamma(t)\big)$$

holds for all $W \in \Sigma/_{\sim_\mu}$ with $\mu\big(W \triangle \gamma(t)\big) \leq R_j(t)$. Without loss of generality, let $R_j(t) \leq \frac{1}{2^{j+1}}$ for all $j \in \mathbb{N}$ and $t \in I$. For each $j \in \mathbb{N}$ and $t \in I$, let $B_{R_j(t)}(t)$ be the open ball with radius $R_j(t)$ around t.

Because $B_{R_j(t)}(t)$ includes at least the point t, it is obvious that for every $j \in \mathbb{N}$, the family

$$\big\{ B_{R_j(t)}(t) \,\big|\, t \in I \big\}$$

is an open cover of the parameter interval I. Because I is a compact interval, the Heine-Borel theorem states that for each $j \in \mathbb{N}$, there exists a tuple $(t_{i,j})_{i \in [N_j]} \subseteq I$ with $N_j \in \mathbb{N}$ such that

$$\big\{ B_{R_j(t_{i,j})}(t_{i,j}) \,\big|\, i \in [N_j] \big\}$$

is an open cover of I for every $j \in \mathbb{N}$. To simplify notation in the next steps, we define $R_{i,j} \coloneqq R_j(t_{i,j})$ and $B_{i,j} \coloneqq B_{R_{i,j}}(t_{i,j})$ for all $j \in \mathbb{N}$ and $i \in [N_j]$. Without loss of generality, let $(t_{i,j})_{i \in [N_j]}$ be strictly increasing.

We may further assume without loss of generality that we have

$$t_{i,j} + R_{i,j} > t_{i',j} + R_{i',j} \quad \forall j \in \mathbb{N},\ (i,\ i') \in [N_j]^2 \colon i > i', \tag{2.46}$$

$$t_{i,j} - R_{i,j} < t_{i',j} - R_{i',j} \quad \forall j \in \mathbb{N},\ (i,\ i') \in [N_j]^2 \colon i < i'. \tag{2.47}$$

Let $j \in \mathbb{N}$. If there existed $(i,\ i') \in [N_j]^2$ with $i > i'$ and $t_{i,j} + R_{i,j} \leq t_{i',j} + R_{i',j}$, then we would have

$$t_{i,j} - R_{i,j} \geq \underbrace{2t_{i,j}}_{> 2t_{i',j}} - t_{i',j} - R_{i',j} > t_{i',j} - R_{i',j},$$

which would imply that $B_{i,j} \subset B_{i',j}$. We could therefore omit $t_{i,j}$ from the tuple and still have a finite open cover of I.

If there existed $(i,\ i')$ with $i < i'$ and $t_{i,j} - R_{i,j} \geq t_{i',j} - R_{i',j}$, then we could similarly argue that

$$t_{i',j} + R_{i',j} \geq \underbrace{2t_{i',j}}_{> 2t_{i,j}} - t_{i,j} + R_{i,j} > t_{i,j} + R_{i,j}.$$

which would again imply that we could omit $t_{i,j}$ from the tuple and still have a finite open cover of I.

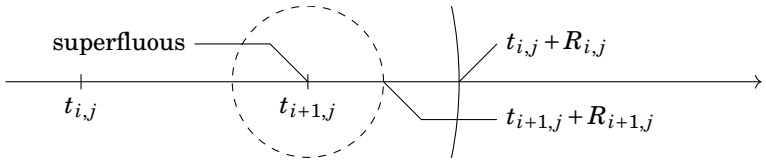


Figure 2.6: Illustration of the argument for Equations (2.46) and (2.47): If $t_{i,j} + R_{i,j} \geq t_{i+1,j} + R_{i+1,j}$, then $B_{i+1,j}$ is a subset of $B_{i,j}$ and $t_{i+1,j}$ is not required as a support point for the open cover.

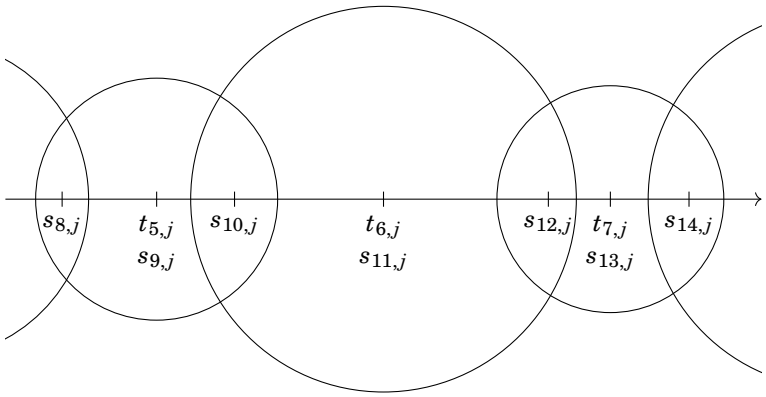


Figure 2.7: Illustration of support point selection. By introducing intermediate support points in the intersection between adjacent balls, we ensure that every segment of the parameter interval is fully covered by one of the covering balls. The behavior of the function can then be approximated by using the Taylor approximation from one of the segment's support points, which is only guaranteed to be sufficiently accurate inside of a single ball.

With Equations (2.46) and (2.47) guaranteed, we can prove by contradiction that the neighborhoods of adjacent support points must overlap. If this was not the case, then there would exist $j \in \mathbb{N}$ and $i \in [N_j]$ such that $B_{i,j} \cap B_{i+1,j} = \emptyset$. Because $t_{i,j} < t_{i+1,j}$, this would imply that $t_{i,j} + R_{i,j} < t_{i+1,j} - R_{i+1,j} < t_{i+1,j}$. If we take into account that I is convex with $t_{i,j} \in I$ and $t_{i+1,j} \in I$, then this would mean that $\tilde{t} \coloneqq t_{i,j} + R_{i,j} \in I$. However, we would have $\tilde{t} \notin B_{i,j}$ and $\tilde{t} \notin B_{i+1,j}$.

For all $i' \in [N_j]$ with $i' < i$, we would have $t_{i',j} + R_{i',j} < t_{i,j} + R_{i,j} = \tilde{t}$ and thus $\tilde{t} \notin B_{i',j}$. For all $i' > i+1$, we would find $t_{i',j} - R_{i',j} > t_{i+1,j} - R_{i+1,j} > \tilde{t}$, which would also mean that $\tilde{t} \notin B_{i',j}$. This would contradict the fact that $(B_{i,j})_{i \in [N_j]}$ enumerates an open cover of I. Thus, we must have $B_{i,j} \cap B_{i+1,j} \neq \emptyset$ for all $j \in \mathbb{N}$ and $i \in [N_j - 1]$.

Because $0 \in I$, for each $j \in \mathbb{N}$, there must exist $i \in [N_j]$ such that $0 \in B_{i,j}$. By a similar argument as before, we can then argue that $0 \in B_{1,j}$ for all $j \in \mathbb{N}$. Similarly, we find that $\mu(U \triangle V) \in B_{N_j,j}$ for all $j \in \mathbb{N}$.

We define support points $(s_{i,j})_{i\in[2N_j]_0} \subseteq I$ such that

$$\begin{aligned} s_{0,j} &\coloneqq 0 && \forall j \in \mathbb{N}, \\ s_{2N_j,j} &\coloneqq \mu(U \triangle V) && \forall j \in \mathbb{N}, \\ s_{2i-1,j} &\coloneqq t_{i,j} && \forall j \in \mathbb{N},\ i \in [N_j], \\ s_{2i,j} &\in (t_{i,j},\ t_{i+1,j}) \cap B_{i,j} \cap B_{i+1,\,j} && \forall j \in \mathbb{N},\ i \in [N_j - 1]. \end{aligned}$$

Suitable choices for $s_{2i,j}$ always exists because $B_{i,j}$ and $B_{i+1,j}$ are overlapping intervals in $\mathbb{R}$ with $t_{i,j} \in B_{i,j}$ and $t_{i+1,j} \in B_{i+1,j}$ and because we had assumed that $t_{i,j} < t_{i+1,j}$. This definition ensures that $(s_{i,j})_{i\in[2N_j]_0}$ is increasing for all $j \in \mathbb{N}$. We now define

$$\begin{aligned} S_{i,j} &\coloneqq \gamma(s_{i,j}) && \forall j \in \mathbb{N},\ i \in [2N_j]_0, \\ D_{i,j} &\coloneqq S_{i,j} \triangle S_{i-1,j} && \forall j \in \mathbb{N},\ i \in [2N_j]. \end{aligned}$$

We note that because $(s_{i,j})_{i\in[2N_j]_0}$ is increasing and γ is a geodesic, the step sets $(D_{i,j})_{i\in[2N_j]}$ are pairwise essentially disjoint for all $j \in \mathbb{N}$. For each $j \in \mathbb{N}$ and $i \in [2N_j]_0$, we have

$$S_{i,j} \triangle U = S_{i,j} \triangle S_{0,j} = \bigcup_{k=1}^{i} D_{k,j}.$$

We first look at steps from oddly indexed to evenly indexed support points. For each $j \in \mathbb{N}$ and $i \in [N_j]$, we have

$$\mu(D_{2i,j}) = |s_{2i,j} - s_{2i-1,j}| = |s_{2i,j} - t_{i,j}| < R_{i,j}.$$

This means that

$$\left|F(S_{2i,j}) - F(S_{2i-1,j}) - \nabla F(S_{2i-1,j})(D_{2i,j})\right| \leq \frac{1}{2^j} \cdot \mu(D_{2i,j}) \quad \forall j \in \mathbb{N},\ i \in [N_j].$$

Because $S_{2i-1,j} \triangle U = \bigcup_{k=1}^{2i-1} D_{k,j}$, $S_{2i-1,j} \triangle U$ is essentially disjoint from $D_{2i,j}$. We also have

$$\begin{aligned} \left|\nabla F(S_{2i-1,j})(D_{2i,j}) - \nabla F(U)(D_{2i,j})\right| &= \left|\nabla F(S_{2i-1,j}) - \nabla F(U)\right|(D_{2i,j}) \\ &= \left(\nabla F(S_{2i-1,j}) \underset{S_{2i-1,j}\triangle U}{\ominus} \nabla F(U)\right)(D_{2i,j}) \\ &\leq L \cdot \mu(S_{2i-1,j} \triangle U) \cdot \mu(D_{2i,j}) \end{aligned}$$

and therefore

$$\left|F(S_{2i,j}) - F(S_{2i-1,j}) - \nabla F(U)(D_{2i,j})\right| \leq \left(\frac{1}{2^j} + L \cdot \mu(S_{2i-1,j} \triangle U)\right) \cdot \mu(D_{2i,j})$$

for all $j \in \mathbb{N}$ and $i \in [N_j]$.

We can make a similar argument for steps from evenly indexed to oddly indexed support points. The difference here is that, because we have to rely on derivatives around the oddly indexed support points, we have to rely on the predicted reverse change from end to start point. Here, the locally inverted difference variation becomes useful. We have

$$\mu(D_{2i-1,j}) = |s_{2i-1,j} - s_{2i-2}| = |t_{i,j} - s_{2i-2}| < R_{i,j} \quad \forall j \in \mathbb{N},\ i \in [N_j]$$

and therefore

$$\left|F(S_{2i-2,j})-F(S_{2i-1,j})-\nabla F(S_{2i-1,j})(D_{2i-1,j})\right| \leq \frac{1}{2^j}\cdot\mu(D_{2i-1,j})$$

for all $j\in\mathbb{N}$ and $i\in[N_j]$. Since $S_{2i-1,j}\triangle U=\bigcup_{k=1}^{2i-1}D_{k,j}$, we have the essential inclusion $D_{2i-1,j}\subseteq_\mu S_{2i-1,j}\triangle U$. Because of the way in which the locally inverted difference variation is defined, this means that

$$\begin{aligned}
&\left|\nabla F(S_{2i-1,j})(D_{2i-1,j})-\left(-\nabla F(U)(D_{2i-1,j})\right)\right| \\
&\quad\leq\left|\nabla F(S_{2i-1,j})+\nabla F(U)\right|(D_{2i-1,j}) \\
&\quad=\left(\nabla F(S_{2i-1,j})\underset{S_{2i-1,j}\triangle U}{\ominus}\nabla F(U)\right)(D_{2i-1,j}) \\
&\quad\leq L\cdot\mu(S_{2i-1,j}\triangle U)\cdot\mu(D_{2i-1,j}
\end{aligned}$$

and therefore

$$\begin{aligned}
&\left|F(S_{2i-1,j})-F(S_{2i-2,j})-\nabla F(U)(D_{2i-1,j})\right| \\
&\quad=\left|F(S_{2i-2,j})-F(S_{2i-1,j})-\left(-\nabla F(U)(D_{2i-1,j})\right)\right| \\
&\quad\leq\left(\frac{1}{2^j}+L\cdot\mu(S_{2i-1,j}\triangle U)\right)\cdot\mu(D_{2i-1,j})
\end{aligned}$$

for all $j\in\mathbb{N}$ and $i\in[N_j]$. By using the triangle inequality and the fact that $D_{2i,j}$ and $D_{2i-1,j}$ are essentially disjoint, these two estimates yield

$$\begin{aligned}
&\left|F(S_{2i,j})-F(S_{2i-2,j})-\nabla F(U)(S_{2i,j}\triangle S_{2i-2,j})\right| \\
&\quad\leq\left(\frac{1}{2^j}+L\cdot\underbrace{\mu(S_{2i-1,j})\triangle U)}_{=t_{i,j}}\right)\cdot\underbrace{\mu(S_{2i,j}\triangle S_{2i-2,j})}_{=s_{2i,j}-s_{2i-2,j}}.
\end{aligned}$$

We can then form the telescope sum

$$\begin{aligned}
&\left|F(V)-F(U)-\nabla F(U)(U\triangle V)\right| \\
&\quad=\left|\sum_{i=1}^{N_j}\left(F(S_{2i,j})-F(S_{2i-2,j})-\nabla F(U)(S_{2i,j}\triangle S_{2i-2,j})\right)\right| \\
&\quad\leq\sum_{i=1}^{N_j}\left|F(S_{2i,j})-F(S_{2i-2,j})-\nabla F(U)(S_{2i,j}\triangle S_{2i-2,j})\right| \\
&\quad\leq\sum_{i=1}^{N_j}\left(\frac{1}{2^j}+L\cdot t_{i,j}\right)\cdot(s_{2i,j}-s_{2i-2,j}) \\
&\quad=\frac{s_{2N_j,j}-s_{0,j}}{2^j}+L\cdot\sum_{i=1}^{N_j}t_{i,j}\cdot\underbrace{(s_{2i,j}-s_{2i-2,j})}_{\leq 2R_{i,j}}.
\end{aligned}$$

We can interpret the sum as a Riemann sum of the function $t\mapsto t$. We remember that we had chosen $R_{i,j}\leq\frac{1}{2^{j+1}}$ for all $j\in\mathbb{N}$ and $i\in[N_j]$. Therefore, the support grid of the Riemann sum becomes infinitesimally fine as $j\to\infty$. We also note

that $s_{0,j} = 0$ and $s_{2N_j,j} = \mu(U \triangle V)$ for all $j \in \mathbb{N}$. We therefore obtain

$$\begin{aligned}
&\big|F(V) - F(U) - \nabla F(U)(U \triangle V)\big| \\
&\quad\leq \lim_{j\to\infty}\left(\frac{s_{2N_j,j} - s_{0,j}}{2^j} + L \cdot \sum_{i=1}^{N_j} t_{i,j} \cdot (s_{2i,j} - s_{2i-2,j})\right) \\
&\quad= 0 + L \cdot \int_0^{\mu(U\triangle V)} t \,\mathrm{d}t \\
&\quad= \frac{L}{2} \cdot \big(\mu(U \triangle V)\big)^2.
\end{aligned}$$

Let $g(U)$ be the gradient density function of F in U. For cases in which $\mu(U \triangle V) \leq R$, we have

$$\begin{aligned}
F(V) &\geq F(U) + \nabla F(U)(U \triangle V) - \frac{L}{2}\big(\mu(U \triangle V)\big)^2 \\
&= F(U) + \int_{U\triangle V} g(U) \,\mathrm{d}\mu - \frac{L}{2}\big(\mu(U \triangle V)\big)^2 \\
&\geq F(U) + \int_X \min\{0,\ g(U)\} \,\mathrm{d}\mu - \frac{L}{2} \cdot R^2.
\end{aligned}$$

□

Let $F : \Sigma/_{\sim_\mu} \to \mathbb{R}$ be differentiable and let $g(U)$ be the gradient density function in $U \in \Sigma/_{\sim_\mu}$. We will subsequently use the quantity

$$\mathscr{C}_1(F, U) := \int_X \min\{0,\ g(U)\} \,\mathrm{d}\mu$$

as a measure for suboptimality for unconstrained problems. We use the index to distinguish it from other suboptimality measures we will later introduce for constrained problems.

Definition 2.4.9 (Gradient-Based Suboptimality Measure).
Let $(X,\ \Sigma,\ \mu)$ be a finite atomless measure space, and let $F : \Sigma/_{\sim_\mu} \to \mathbb{R}$ be differentiable. For each $U \in \Sigma/_{\sim_\mu}$, let $\nabla^- F(U)$ be the negative portion of the signed measure $\nabla F(U)$. We refer to

$$\mathscr{C}_1(F, U) := \nabla^- F(U)(X) \leq 0$$

as the *unconstrained suboptimality measure of F in U*. ◁

We stress that the unconstrained optimality measure is not a measure in the technical sense but rather a means by which we quantify the suboptimality of a differentiable function. We note that that the curvature criterion indicates that the unconstrained suboptimality measure is not always a lower bound on local improvements in functional value. In Section 2.5, we will introduce classes of functionals where this is the case.

2.4.2 Derivation: Banach Space

The argument presented in this section was first presented in a slightly simpler form in [HLS22].

We can derive set derivatives from the Fréchet derivatives of Fréchet differentiable functions in Banach spaces. In order to do this, we have to make several assumptions.

Assumption 2.4.10.
Let $(X,\ \Sigma,\ \mu)$, Y, $\nu\colon \Sigma \to Y$, and $f\colon Y \to \mathbb{R}$ satisfy the following assumptions:

(1) (X,Σ,μ) is a finite atomless measure space;

(2) Y is a Banach space;

(3) ν is a vector measure of bounded variation;

(4) there exists a constant $C \geq 0$ such that $|\nu|(D) \leq C \cdot \mu(D)\ \forall D \in \Sigma$;

(5) f is Fréchet differentiable on $\nu(\Sigma)$ in the sense that for each $U \in \Sigma$, there exists a linear map $\nabla f\big(\nu(U)\big)\colon Y \to \mathbb{R}$ and a constant $M \geq 0$ such that

$$\Big|\nabla f\big(\nu(U)\big)\big(\nu(V)-\nu(U)\big)\Big| \leq M \cdot \big\|\nu(V)-\nu(U)\big\|_Y \ \forall V \in \Sigma,$$
$$f\big(\nu(V)\big) - f\big(\nu(U)\big) - \nabla f\big(\nu(U)\big)\big(\nu(V)-\nu(U)\big) = o\Big(\big\|\nu(V)-\nu(U)\big\|_Y\Big) \ \forall V \in \Sigma. \triangleleft$$

Assumption 2.4.10 (5) is a relaxation of an overall Fréchet differentiability condition. It is trivially satisfied if f is Fréchet differentiable on Y. However, the assumption is more general because it only requires differentiability between vectors that are actually realized by the vector measure ν.

We note that Assumption 2.4.10 (4) implies that the vector measure ν has bounded variation and satisfies $\nu \ll \mu$. Therefore, ν assumes well-defined values for similarity classes with respect to μ.

Under Assumption 2.4.10, we can show that the map $F\colon \Sigma/_{\sim_\mu} \to \mathbb{R}$ with

$$F(U) \coloneqq f\big(\nu(U)\big) \quad \forall U \in \Sigma/_{\sim_\mu}$$

is differentiable and its derivative satisfies

$$\nabla F(U)(D) = \nabla f\big(\nu(U)\big)\big(\nu(D \setminus U) - \nu(D \cap U)\big) \quad \forall U \in \Sigma/_{\sim_\mu},\ D \in \Sigma.$$

The vector space derivative of f in $\nu(U)$ is a linear form. By concatenating this linear form with a sign-adjusted variant of the vector measure ν, we obtain a signed measure φ that will turn out to be a suitable derivative for F.

Proposition 2.4.11.
Let (X,Σ,μ) be a measure space, let Y be a Banach space, let $U \in \Sigma$, let $\nu\colon \Sigma \to Y$ be a vector measure that satisfies Assumption 2.4.10 (3), and let $A\colon Y \to \mathbb{R}$ be a linear operator that is bounded in the sense of Assumption 2.4.10 (5). Then the map $\varphi_U\colon \Sigma \to \mathbb{R}$ with

$$\varphi_U(D) \coloneqq A\big(\nu(D \setminus U) - \nu(D \cap U)\big) \quad \forall D \in \Sigma$$

is a finite signed measure with $\varphi_U \ll \mu$. $\triangleleft$

Proof. Because A is bounded in the sense of Assumption 2.4.10 (5), there exists a constant $M \geq 0$ such that

$$\Big|A\big(\nu(V)-\nu(U)\big)\Big| \leq M \cdot \big\|\nu(V)-\nu(U)\big\|_Y.$$

For each $D \in \Sigma$, we have

$$\begin{aligned}
\left|\varphi_U(D)\right| &\leq M \cdot \left\|\nu(D \setminus U) - \nu(D \cap U)\right\|_Y \\
&\leq M \cdot \left(\left\|\nu(D \setminus U)\right\|_Y + \left\|\nu(D \cap U)\right\|_Y\right) \\
&\leq M \cdot \left(|\nu|(D \setminus U) + |\nu|(D \cap U)\right) \\
&\leq M \cdot |\nu|(D) \\
&\leq M \cdot \|\nu\| \\
&< \infty.
\end{aligned}$$

If $\mu(D) = 0$, then we have

$$\begin{aligned}
\left|\varphi_U(D)\right| &= \left|A\left(\nu(D \setminus U) - \nu(D \cap U)\right)\right| \\
&\leq M \cdot \left(\left\|\nu(D \setminus U)\right\|_Y + \left\|\nu(D \cap U)\right\|_Y\right) \\
&\leq MC \cdot \mu(D) \\
&= 0.
\end{aligned}$$

This also implies $\varphi_U(\emptyset) = 0$. Once we have shown that φ_U is countably additive and therefore a signed measure, this means that $\varphi_U \ll \mu$.

Let $(D_i)_{i\in\mathbb{N}}$ be a sequence in Σ such that $D_i \cap D_j = \emptyset$ for all $i, j \in \mathbb{N}$ with $i \neq j$. Then we have

$$\nu\left(\left(\bigcup_{i=1}^{\infty} D_i\right) \setminus U\right) = \nu\left(\bigcup_{i=1}^{\infty} (D_i \setminus U)\right) = \sum_{i=1}^{\infty} \nu(D_i \setminus U).$$

Since ν is of bounded variation, we have

$$\sum_{i=1}^{\infty} \left\|\nu(D_i \setminus U)\right\|_Y \leq \sum_{i=1}^{\infty} |\nu|(D_i \setminus U) = |\nu|\left(\bigcup_{i=1}^{\infty} (D_i \setminus U)\right) \leq \|\nu\| < \infty.$$

Therefore, the sum is absolutely convergent. Similarly,

$$\nu\left(\left(\bigcup_{i=1}^{\infty} D_i\right) \cap U\right) = \sum_{i=1}^{\infty} \nu(D_i \cap U)$$

is absolutely convergent. We therefore find that

$$\begin{aligned}
\varphi_U\left(\bigcup_{i=1}^{\infty} D_i\right) &= A\left(\nu\left(\left(\bigcup_{i=1}^{\infty} D_i\right) \setminus U\right) - \nu\left(\left(\bigcup_{i=1}^{\infty} D_i\right) \cap U\right)\right) \\
&= A\left(\left(\sum_{i=1}^{\infty} \nu(D_i \setminus U)\right) - \left(\sum_{i=1}^{\infty} \nu(D_i \cap U)\right)\right) \\
&= A\left(\sum_{i=1}^{\infty} \left(\nu(D_i \setminus U) - \nu(D_i \cap U)\right)\right) \\
&= A\left(\lim_{n\to\infty} \sum_{i=1}^{n} \left(\nu(D_i \setminus U) - \nu(D_i \cap U)\right)\right) \\
&= \lim_{n\to\infty} A\left(\sum_{i=1}^{n} \left(\nu(D_i \setminus U) - \nu(D_i \cap U)\right)\right) \\
&= \lim_{n\to\infty} \sum_{i=1}^{n} A\left(\nu(D_i \setminus U) - \nu(D_i \cap U)\right)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{\infty} A\big(\nu(D_i \setminus U) - \nu(D_i \cap U)\big) \\
&= \sum_{i=1}^{\infty} \varphi_U(D_i).
\end{aligned}
$$

Here, we make use of the fact that

$$
\begin{aligned}
&\left| A\Big(\lim_{n\to\infty} \sum_{i=1}^{n} \big(\nu(D_i \setminus U) - \nu(D_i \cap U)\big)\Big) - A\Big(\sum_{i=1}^{m} \big(\nu(D_i \setminus U) - \nu(D_i \cap U)\big)\Big) \right| \\
&\quad = \left| A\Big(\lim_{n\to\infty} \sum_{i=m+1}^{n} \big(\nu(D_i \setminus U) - \nu(D_i \cap U)\big)\Big) \right| \\
&\quad \leq M \cdot \sum_{i=m+1}^{\infty} |\nu|(D_i) \\
&\quad \xrightarrow{m\to\infty} 0.
\end{aligned}
$$

Therefore, φ_U is countably additive. In conjunction with the fact that $\varphi_U(\emptyset) = 0$, this means that φ_U is a signed measure. As we had previously shown, we have $\varphi_U \ll \mu$. □

We can apply Proposition 2.4.11 to the derivative of f in $\nu(U)$ to obtain the derivative of F.

Proposition 2.4.12 (Local Derivatives for Banach Space Functions).
Let $n \in \mathbb{N}$, (X, Σ, μ), Y, $\nu : \Sigma \to Y$, and $f : Y \to \mathbb{R}$ satisfy Assumptions 2.4.10 (1) to 2.4.10 (4). Let $U \in \Sigma/_{\sim_\mu}$ be such that there exists a linear map $\nabla f\big(\nu(U)\big) : Y \to \mathbb{R}$ and a constant $M \geq 0$ such that

$$
\begin{aligned}
\left| \nabla f\big(\nu(U)\big)\big(\nu(V) - \nu(U)\big) \right| &\leq M \cdot \big\| \nu(V) - \nu(U) \big\|_Y \quad \forall V \in \Sigma, \\
f\big(\nu(V)\big) - f\big(\nu(U)\big) - \nabla f\big(\nu(U)\big)\big(\nu(V) - \nu(U)\big) &= o\Big(\big\| \nu(V) - \nu(U) \big\|_Y \Big) \quad \forall V \in \Sigma.
\end{aligned}
$$

Then the map $F : \Sigma/_{\sim_\mu} \to \mathbb{R}$ with

$$
F(V) := f\big(\nu(V)\big) \quad \forall V \in \Sigma/_{\sim_\mu}
$$

is differentiable in U and its derivative satisfies

$$
\nabla F(U)(D) = \nabla f\big(\nu(U)\big)\big(\nu(D \setminus U) - \nu(D \cap U)\big) \quad \forall D \in \Sigma. \qquad \triangleleft
$$

Proof. For all $D \in \Sigma/_{\sim_\mu}$, we have

$$
\begin{aligned}
F(U \triangle D) - F(U) &= f\big(\nu(U \triangle D)\big) - f\big(\nu(U)\big) \\
&= \nabla f\big(\nu(U)\big)\underbrace{\big(\nu(U \triangle D) - \nu(U)\big)}_{= \nu(D \setminus U) - \nu(D \cap U)} + o\Big(\big\| \nu(U \triangle D) - \nu(U) \big\|_Y \Big).
\end{aligned}
$$

Let $C \geq 0$ be the constant from Assumption 2.4.10 (4). We have

$$
\begin{aligned}
\big\| \nu(U \triangle D) - \nu(U) \big\|_Y &= \big\| \nu(D \setminus U) - \nu(D \cap U) \big\|_Y \\
&\leq \big\| \nu(D \setminus U) \big\|_Y + \big\| \nu(D \cap U) \big\|_Y \\
&\leq |\nu|(D) \\
&\leq C \cdot \mu(D).
\end{aligned}
$$

This allows us to simplify the small-O term, which becomes

$$F(U \triangle D) - F(U) = \nabla f(\nu(U))\big(\nu(D \setminus U) - \nu(D \cap U)\big) + o\big(\mu(D)\big)$$

Let $\varphi\colon \Sigma \to \mathbb{R}$ be given by

$$\varphi(D) := \nabla f(\nu(U))\big(\nu(D \setminus U) - \nu(D \cap U)\big) \quad \forall D \in \Sigma.$$

According to Proposition 2.4.11, φ is a finite signed measure with $\varphi \ll \mu$. Therefore, F is differentiable in U with $\nabla F(U) = \varphi$. □

Because the set derivative is directly calculated from a Fréchet derivative, continuity and Lipschitz continuity of the derivative follow from corresponding continuity of the Fréchet derivative.

Theorem 2.4.13 (Global Derivatives for Banach Space Functions). *Let $n \in \mathbb{N}$, (X, Σ, μ), Y, $\nu\colon \Sigma \to Y$, and $f\colon Y \to \mathbb{R}$ satisfy Assumptions 2.4.10 (1) to 2.4.10 (5). Then the set functional $F\colon \Sigma/_{\sim_\mu} \to \mathbb{R}$ with*

$$F(V) := f\big(\nu(V)\big) \quad \forall V \in \Sigma/_{\sim_\mu}$$

is differentiable. Let $\mathcal{N} := \{\nu(U) \,|\, U \in \Sigma\}$ and let

$$\Big\|\nabla f\big(\nu(V)\big) - \nabla f\big(\nu(U)\big)\Big\|_{\mathcal{N}} := \sup_{\substack{u,v \in \mathcal{N} \\ u \neq v}} \frac{\Big|\nabla f\big(\nu(V)\big)(v) - \nabla f\big(\nu(U)\big)(u)\Big|}{\|u - v\|_Y}$$

Then we have

$$\big(\nabla F(U) \operatorname*{\ominus}_{U \triangle V} \nabla F(V)\big)(D) \leq \Big\|\nabla f\big(\nu(V)\big) - \nabla f\big(\nu(U)\big)\Big\|_{\mathcal{N}} \cdot |\nu|(D)$$

for all $U, V, D \in \Sigma/_{\sim_\mu}$ and therefore

1. *F is continuously differentiable if for every $\varepsilon > 0$ and every $u \in \mathcal{N}$, there exists $\delta > 0$ such that*

$$\big\|\nabla f(v) - \nabla f(u)\big\|_{\mathcal{N}} \leq \varepsilon \quad \forall v \in \mathcal{N}\colon \|u - v\|_Y \leq \delta;$$

2. *F is uniformly continuously differentiable if for every $\varepsilon > 0$, there exists $\delta > 0$ such that*

$$\big\|\nabla f(v) - \nabla f(u)\big\|_{\mathcal{N}} \leq \varepsilon \quad \forall u, v \in \mathcal{N}\colon \|u - v\|_Y \leq \delta;$$

3. *F is Lipschitz continuously differentiable if there exists a constant $L \geq 0$ such that*

$$\big\|\nabla f(v) - \nabla f(u)\big\|_{\mathcal{N}} \leq L \cdot \|u - v\|_Y \quad \forall u, v \in \mathcal{N}.$$ ◁

PROOF. The differentiability of F follows from Proposition 2.4.12. To verify the locally inverted variation difference estimate, let $U, V \in \Sigma/_{\sim_\mu}$. For every $D \in \Sigma$, we have

$$\begin{aligned}\big(\nabla F(U) \operatorname*{\ominus}_{U \triangle V} \nabla F(V)\big)(D) = \quad & \big|\nabla F(U) - \nabla F(V)\big|\big(D \setminus (U \triangle V)\big) \\ & + \big|\nabla F(U) + \nabla F(V)\big|\big(D \cap (U \triangle V)\big)\end{aligned}$$

We consider $\left|\nabla F(U)(B) - \nabla F(V)(B)\right|$ and $\left|\nabla F(U)(B) + \nabla F(V)(B)\right|$ for all measurable subsets of $D \setminus (U \triangle V)$ and $D \cap (U \triangle V)$, respectively. Let $B \in \Sigma$ such that $B \subseteq D \setminus (U \triangle V)$. We have $B \cap (U \triangle V) = (B \cap U) \triangle (B \cap V) = \emptyset$ and therefore $B \cap V = B \cap U$. This also implies that $B \setminus V = B \setminus U$ and therefore

$$\begin{aligned}
\left|\nabla F(U)(B) - \nabla F(V)(B)\right| &= \Big| \quad \nabla f\big(\nu(U)\big)\big(\nu(B \setminus U) - \nu(B \cap U)\big) \\
&\quad - \nabla f\big(\nu(V)\big)\big(\nu(B \setminus V) - \nu(B \cap V)\big)\Big| \\
&= \Big|\Big(\nabla f\big(\nu(U)\big) - \nabla f\big(\nu(V)\big)\Big)\big(\nu(B \setminus U) - \nu(B \cap U)\big)\Big| \\
&\leq \Big\|\nabla f\big(\nu(U)\big) - \nabla f\big(\nu(V)\big)\Big\|_{\mathcal{N}} \cdot \big\|\nu(B \setminus U) - \nu(B \cap U)\big\|_Y.
\end{aligned}$$

We note that this makes use of the fact that $\nu(B \setminus U)$ and $\nu(B \cap U)$ are both in $\mathcal{N}$ and that $B \setminus U$ and $B \cap U$ are disjoint. For $B \in \Sigma$ with $B \subseteq D \cap (U \triangle V)$, we have $B \subseteq U \triangle V$. In this case, we have $B \cap V = B \setminus U$ and $B \cap U = B \setminus V$, which implies that

$$\begin{aligned}
\left|\nabla F(U)(B) + \nabla F(V)(B)\right| &= \Big| \quad \nabla f\big(\nu(U)\big)\big(\nu(B \setminus U) - \nu(B \cap U)\big) \\
&\quad + \nabla f\big(\nu(V)\big)\big(\nu(B \setminus V) - \nu(B \cap V)\big)\Big| \\
&= \Big| \quad \nabla f\big(\nu(U)\big)\big(\nu(B \setminus U) - \nu(B \cap U)\big) \\
&\quad - \nabla f\big(\nu(V)\big)\big(\nu(B \cap V) - \nu(B \setminus V)\big)\Big| \\
&= \Big| \quad \nabla f\big(\nu(U)\big)\big(\nu(B \setminus U) - \nu(B \cap U)\big) \\
&\quad - \nabla f\big(\nu(V)\big)\big(\nu(B \setminus U) - \nu(B \cap U)\big)\Big| \\
&\leq \Big\|\nabla f\big(\nu(U)\big) - \nabla f\big(\nu(V)\big)\Big\|_{\mathcal{N}} \cdot \big\|\nu(B \setminus U) - \nu(B \cap U)\big\|_Y.
\end{aligned}$$

Let $X_- \in \Sigma$ and its complement form a Hahn decomposition (see Theorem 2.1.5) of X with respect to $\nabla F(U) - \nabla F(V)$ such that X_- encompasses the non-positive part of $\nabla F(U) - \nabla F(V)$. We first consider the part of D that lies outside of $U \triangle V$ and is therefore not affected by the inversion. We make use of the fact that

$$\begin{aligned}
B \setminus (U \triangle V) \setminus U &= B \setminus (U \setminus V) \setminus (V \setminus U) \setminus U \\
&= B \setminus U \setminus (V \setminus U) \\
&= B \setminus (U \cup V),
\end{aligned}$$

$$\begin{aligned}
B \setminus (U \triangle V) \cap U &= B \setminus (U \triangle V) \cap U \\
&= B \setminus (U \setminus V) \setminus (V \setminus U) \cap U \\
&= B \cap (U \cap V) \setminus (V \setminus U) \\
&= B \cap (U \cap V)
\end{aligned}$$

holds for all sets B. The variation of $\nabla F(U) - \nabla F(V)$ on $D \setminus (U \triangle V)$ is given by

$$\begin{aligned}
&\big|\nabla F(U) - \nabla F(V)\big|\big(D \setminus (U \triangle V)\big) \\
&\quad = \big(\nabla F(U) - \nabla F(V)\big)\big((D \setminus X_-) \setminus (U \triangle V)\big) \\
&\quad - \big(\nabla F(U) - \nabla F(V)\big)\big((D \cap X_-) \setminus (U \triangle V)\big) \\
&\quad = \Big|\big(\nabla F(U) - \nabla F(V)\big)\big((D \setminus X_-) \setminus (U \triangle V)\big)\Big| \\
&\quad + \Big|\big(\nabla F(U) - \nabla F(V)\big)\big((D \cap X_-) \setminus (U \triangle V)\big)\Big| \\
&\quad \leq \Big\|\nabla f\big(\nu(U)\big) - \nabla f\big(\nu(V)\big)\Big\|_{\mathcal{N}} \cdot \Big\| \nu\big((D \setminus X_-) \setminus (U \cup V)\big) \\
&\qquad\qquad - \nu\big((D \setminus X_-) \cap (U \cap V)\big)\Big\|_Y \\
&\quad + \Big\|\nabla f\big(\nu(U)\big) - \nabla f\big(\nu(V)\big)\Big\|_{\mathcal{N}} \cdot \Big\| \nu\big((D \cap X_-) \setminus (U \cup V)\big) \\
&\qquad\qquad - \nu\big((D \cap X_-) \cap (U \cap V)\big)\Big\|_Y.
\end{aligned}$$

For the part of D that lies inside of $U \triangle V$, we can make a similar estimate. Here, we select X_+ such that it and its complement form a Hahn decomposition with respect to $\nabla F(U) + \nabla F(V)$ and X_+ encompasses the non-positive part of $\nabla F(U) + \nabla F(V)$. We make use of the fact that

$$\begin{aligned}
B \cap (U \triangle V) \setminus U &= B \cap (U \cup V) \setminus (U \cap V) \setminus U \\
&= B \cap (V \setminus U), \\
B \cap (U \triangle V) \cap U &= B \cap (U \cup V) \setminus (U \cap V) \cap U \\
&= B \cap U \setminus (U \cap V) \\
&= B \cap (U \setminus V)
\end{aligned}$$

holds for all sets B. The variation of $\nabla F(U) + \nabla F(V)$ on $D \cap (U \triangle V)$ is given by

$$\begin{aligned}
&\big|\nabla F(U) + \nabla F(V)\big|\big(D \cap (U \triangle V)\big) \\
&\quad = \big(\nabla F(U) + \nabla F(V)\big)\big((D \setminus X_+) \cap (U \triangle V)\big) \\
&\quad - \big(\nabla F(U) + \nabla F(V)\big)\big((D \cap X_+) \cap (U \triangle V)\big) \\
&\quad = \Big|\big(\nabla F(U) + \nabla F(V)\big)\big((D \setminus X_+) \cap (U \triangle V)\big)\Big| \\
&\quad + \Big|\big(\nabla F(U) + \nabla F(V)\big)\big((D \cap X_+) \cap (U \triangle V)\big)\Big| \\
&\quad \leq \Big\|\nabla f\big(\nu(U)\big) - \nabla f\big(\nu(V)\big)\Big\|_{\mathcal{N}} \cdot \Big\| \nu\big((D \setminus X_+) \cap (V \setminus U)\big) \\
&\qquad\qquad - \nu\big((D \setminus X_+) \cap (U \setminus V)\big)\Big\|_Y \\
&\quad + \Big\|\nabla f\big(\nu(U)\big) - \nabla f\big(\nu(V)\big)\Big\|_{\mathcal{N}} \cdot \Big\| \nu\big((D \cap X_+) \cap (V \setminus U)\big) \\
&\qquad\qquad - \nu\big((D \cap X_+) \cap (U \setminus V)\big)\Big\|_Y.
\end{aligned}$$

We now partition D into eight sets:

$$\begin{aligned}
D_1 &:= (D \setminus X_-) \setminus (U \cup V), & D_2 &:= (D \setminus X_-) \cap (U \cap V), \\
D_3 &:= (D \cap X_-) \setminus (U \cup V), & D_4 &:= (D \cap X_-) \cap (U \cap V), \\
D_5 &:= (D \setminus X_+) \cap (V \setminus U), & D_6 &:= (D \setminus X_+) \cap (U \setminus V), \\
D_7 &:= (D \cap X_+) \cap (V \setminus U), & D_8 &:= (D \cap X_+) \cap (U \setminus V).
\end{aligned}$$

It is evident that this is a partition of D because $U \setminus V$ and $V \setminus U$ partition $U \triangle V$ and because $(U \cup V)^{\complement}$ and $U \cap V$ partition $(U \triangle V)^{\complement}$. We then find that

$$\begin{aligned}
&\big(\nabla F(U) \mathbin{\ominus}_{U \triangle V} \nabla F(V)\big)(D) \\
&\quad = \big|\nabla F(U) - \nabla F(V)\big|\big(D \setminus (U \triangle V)\big) + \big|\nabla F(U) + \nabla F(V)\big|\big(D \cap (U \triangle V)\big) \\
&\quad \leq \Big\|\nabla f\big(\nu(U)\big) - \nabla f\big(\nu(V)\big)\Big\|_{\mathscr{N}} \cdot \Big(\big\|\nu(D_1) - \nu(D_2)\big\|_Y + \big\|\nu(D_3) - \nu(D_4)\big\|_Y \\
&\qquad\qquad + \big\|\nu(D_5) - \nu(D_6)\big\|_Y + \big\|\nu(D_7) - \nu(D_8)\big\|_Y \Big) \\
&\quad \leq \Big\|\nabla f\big(\nu(U)\big) - \nabla f\big(\nu(V)\big)\Big\|_{\mathscr{N}} \cdot \sum_{i=1}^{8} \big\|\nu(D_i)\big\|_Y \\
&\quad \leq \Big\|\nabla f\big(\nu(U)\big) - \nabla f\big(\nu(V)\big)\Big\|_{\mathscr{N}} \cdot |\nu|\Big(\bigcup_{i=1}^{8} D_i\Big) \\
&\quad = \Big\|\nabla f\big(\nu(U)\big) - \nabla f\big(\nu(V)\big)\Big\|_{\mathscr{N}} \cdot |\nu|(D),
\end{aligned}$$

which is the central estimate that we set out to prove. All that remains to be shown is the transferrability of the continuity results. We first note that, according to Assumption 2.4.10 (4), for all $U, V \in \Sigma$, we have

$$\big\|\nu(U) - \nu(V)\big\|_Y = \big\|\nu(U \setminus V) - \nu(V \setminus U)\big\|_Y \leq |\nu|(U \triangle V) \leq C \cdot \mu(U \triangle V).$$

Therefore, $\mu(U \triangle V) \leq \frac{\delta}{C}$ implies $\big\|\nu(U) - \nu(V)\big\|_Y \leq \delta$ for all $\delta > 0$. Furthermore, for every $L \geq 0$, we have $L \cdot \big\|\nu(U) - \nu(V)\big\|_Y \leq LC \cdot \mu(U \triangle V)$. We very briefly show the three estimates required.

Case 1 (Continuous Differentiability). Let $\varepsilon > 0$ and $U \in \Sigma/_{\sim_\mu}$. We choose a radius $\delta_0 > 0$ such that

$$\Big\|\nabla f(v) - \nabla f\big(\nu(U)\big)\Big\|_{\mathscr{N}} \leq \frac{\varepsilon}{C} \quad \forall v \in \mathscr{N} \colon \big\|\nu(U) - v\big\|_Y \leq \delta_0.$$

Then we set $\delta := \frac{\delta_0}{C} > 0$. Let $V \in \Sigma/_{\sim_\mu}$ with $\mu(U \triangle V) \leq \delta$. Because we have $\big\|\nu(U) - \nu(V)\big\|_Y \leq C \cdot \mu(U \triangle V) \leq \delta_0$, we can infer that

$$\begin{aligned}
\big(\nabla F(U) \mathbin{\ominus}_{U \triangle V} \nabla F(V)\big)(D) &\leq \Big\|\nabla f\big(\nu(U)\big) - \nabla f\big(\nu(V)\big)\Big\|_{\mathscr{N}} \cdot |\nu|(D) \\
&\leq \frac{\varepsilon}{C} \cdot C \cdot \mu(D) \\
&= \varepsilon \cdot \mu(D).
\end{aligned}$$

◁

Case 2 (Uniform Continuous Differentiability). Let $\varepsilon > 0$. We choose a radius $\delta_0 > 0$ such that

$$\big\|\nabla f(v) - \nabla f(u)\big\|_{\mathscr{N}} \leq \frac{\varepsilon}{C} \quad \forall u, v \in \mathscr{N} \colon \|u - v\|_Y \leq \delta_0.$$

Then we set $\delta := \frac{\delta_0}{C} > 0$. Let $U, V \in \Sigma/_{\sim_\mu}$ with $\mu(U \triangle V) \leq \delta$. Because we have $\big\|\nu(U) - \nu(V)\big\|_Y \leq C \cdot \mu(U \triangle V) \leq \delta_0$, we can infer that

$$\begin{aligned}
\big(\nabla F(U) \mathbin{\ominus}_{U \triangle V} \nabla F(V)\big)(D) &\leq \Big\|\nabla f\big(\nu(U)\big) - \nabla f\big(\nu(V)\big)\Big\|_{\mathscr{N}} \cdot |\nu|(D) \\
&\leq \frac{\varepsilon}{C} \cdot C \cdot \mu(D) \\
&= \varepsilon \cdot \mu(D).
\end{aligned}$$

◁

Case 3 (Lipschitz Continuous Differentiability). Let $\varepsilon > 0$. We have

$$\left\|\nabla f(v) - \nabla f(u)\right\|_{\mathcal{N}} \leq L \cdot \|u - v\|_Y \quad \forall u, v \in \mathcal{M}.$$

Then we set $L' \coloneqq LC^2 \geq 0$. Let $U, V \in \Sigma/_{\sim_\mu}$. We have $\left\|\nu(U) - \nu(V)\right\|_Y \leq C \cdot \mu(U \triangle V)$ and therefore

$$\begin{aligned}\left(\nabla F(U) \underset{U \triangle V}{\ominus} \nabla F(V)\right)(D) &\leq \left\|\nabla f\left(\nu(U)\right) - \nabla f\left(\nu(V)\right)\right\|_{\mathcal{N}} \cdot |\nu|(D) \\ &\leq LC \cdot \mu(U \triangle V) \cdot C \cdot \mu(D) \\ &= LC^2 \cdot \mu(U \triangle V) \cdot \mu(D) \\ &= L' \cdot \mu(U \triangle V) \cdot \mu(D).\end{aligned}$$

□

We end this section by showing that the straightforward approach of mapping a set A to its characteristic function χ_A satisfies Assumption 2.4.10. This is significant because it allows our methods to be applied to many conventional optimal control problems and allows us to import a lot of prior theoretical work into our framework.

Theorem 2.4.14 (Characteristic Function Vector Measure).
Let (X, Σ, μ) be a finite atomless measure space. The function space $L^1(\Sigma, \mu)$ is a Banach space and $\chi_\cdot \colon \Sigma \to L^1(\Sigma, \mu)$ with $U \mapsto \chi_U$ is an absolutely continuous vector measure such that

$$|\chi_\cdot|(U) = \|\chi_U\|_{L^1(\Sigma,\mu)} = \mu(U) \quad \forall U \in \Sigma. \qquad \triangleleft$$

PROOF. We first show that $\chi_\cdot$ is a vector measure. We evidently have $\chi_\emptyset \equiv 0$. Let $(A_i)_{i \in \mathbb{N}} \subseteq \Sigma$ such that $A_i \cap A_j = \emptyset$ for all $i, j \in \mathbb{N}$ with $i \neq j$. Since $\operatorname{supp} \chi_A = A$ for all $A \in \Sigma$, we have $x \in A_i \implies x \notin A_j$ for all $i, j \in \mathbb{N}$ with $i \neq j$. Therefore, we have

$$\begin{aligned}\chi_{\bigcup_{i=1}^\infty A_i}(x) &= \begin{cases} 1 & \text{if } \exists i \in \mathbb{N} \colon x \in A_i \\ 0 & \text{if } x \notin A_i \ \forall i \in \mathbb{N} \end{cases} \\ &= \sum_{i=1}^\infty \chi_{A_i}.\end{aligned}$$

Thus, $\chi_\cdot$ is a vector measure. We further have

$$\|\chi_U\|_{L^1(\Sigma,\mu)} = \int_U |\chi_U| \, \mathrm{d}\mu = \int_U 1 \, \mathrm{d}\mu = \mu(U) \qquad \forall U \in \Sigma.$$

Let $U \in \Sigma$. For every $N \in \mathbb{N}$ and every finite partition $(A_i)_{i \in [N]} \subseteq \Sigma$ of U, we have

$$\sum_{i=1}^N \|\chi_{A_i}\|_{L^1(\Sigma,\mu)} = \sum_{i=1}^N \mu(A_i) = \mu\Big(\bigcup_{i=1}^N A_i\Big) = \mu(U).$$

Thus, we have

$$|\chi_\cdot|(U) = \sup\Big\{\sum_{i=1}^N \|\chi_{A_i}\|_{L^1(\Sigma,\mu)} \,\Big|\, N \in \mathbb{N},\ (A_i)_{i \in [N]} \subseteq \Sigma \text{ partition of } U\Big\} = \mu(U)$$

for all $U \in \Sigma$. □

We note that the continuity requirement in Assumption 2.4.10 (4) can be transferred from one measure μ to some equivalent measures. This is important because it enables scaling. In Sections 2.4.3 and 2.4.4, we discuss functions whose evaluation involves the solution of differential equations. In such functions, small perturbations in control can propagate to cause large overall effects. Scaling can be used to mitigate the numerical issues that arise from this (see, e.g., [HLS22]).

Definition 2.4.15 (Scaling Density).
Let $(X,\ \Sigma,\ \mu)$ be a finite atomless measure space, let $f \in L^\infty(\Sigma,\mu)$. We refer to f as a *scaling density function for* μ if and only there exists a constant $C > 0$ such that $f(x) \geq C$ for almost all $x \in X$. ◁

Proposition 2.4.16 (Scaling Measures).
Let $(X,\ \Sigma,\ \mu)$ be a finite atomless measure space, let Y be a Banach space, and let $f \in L^\infty(\Sigma,\mu)$ be a scaling density function for μ. Then the measure $\phi\colon \Sigma \to \mathbb{R}_{\geq 0}$ with

$$\phi(U) \coloneqq \int_U f \,\mathrm{d}\mu \quad \forall U \in \Sigma$$

satisfies

$$\mu(U) \leq C \cdot \phi(U) \quad \forall U \in \Sigma$$

for an appropriately chosen constant $C > 0$. Therefore μ and ϕ are equivalent.

Let $\nu\colon \Sigma \to Y$ be a vector measure of bounded variation such that there exists a constant $C' \geq 0$ with

$$|\nu|(U) \leq C' \cdot \mu(U) \quad \forall U \in \Sigma,$$

then we have

$$|\nu|(U) \leq C \cdot C' \cdot \phi(U) \quad \forall U \in \Sigma.$$ ◁

Proof. Part 1 (ϕ is a measure). Because $f \geq M > 0$ almost everywhere, we have $\phi(U) \geq 0$ for all $U \in \Sigma$. Furthermore, $f \in L^1(\Sigma,\mu)$ implies $\phi(U) < \infty$ for all $U \in \Sigma$. We evidently have $\phi(\emptyset) = 0$. For $(U_i)_{i\in\mathbb{N}} \subseteq \Sigma$ with $U_i \cap U_j = \emptyset$ for $i \neq j$, we have

$$\sum_{i=1}^{N} \phi(U_i) = \sum_{i=1}^{N} \int_{U_i} f \,\mathrm{d}\mu = \int_{\bigcup_{i=1}^{N} U_i} f \,\mathrm{d}\mu = \phi\Big(\bigcup_{i=1}^{N} U_i\Big) \quad \forall N \in \mathbb{N}$$

because

$$A \cap B = \emptyset \implies \int_A f \,\mathrm{d}\mu + \int_B f \,\mathrm{d}\mu = \int_{A\cup B} f \,\mathrm{d}\mu \quad \forall A,B \in \Sigma.$$

Since $\phi(U_i) \geq 0$ for all $i \in \mathbb{N}$, the sum is monotonically increasing. Furthermore, it is bounded above by $\phi(X)$. Therefore, the series is absolutely convergent and we have

$$\phi\Big(\bigcup_{i=1}^{\infty} U_i\Big) = \sum_{i=1}^{\infty} \phi(U_i).$$

Part 2 ($\mu(U) \leq C \cdot \phi(U)\ \forall U \in \Sigma$). Let $M \geq 0$ be such that

$$f(x) \geq M \qquad \text{almost everywhere.}$$

Let $N \in \Sigma$ be a μ-nullset such that $f(x) \geq M$ for all $x \in N^{\complement}$. We define $C \coloneqq \frac{1}{M} > 0$. For all $U \in \Sigma$, we have

$$\begin{aligned}\phi(U) &= \phi(U \cap N) + \phi(U \setminus N) \\ &= \underbrace{\int_{U \cap N} f \,\mathrm{d}\mu}_{=0} + \underbrace{\int_{U \setminus N} f \,\mathrm{d}\mu}_{\geq M \cdot \mu(U \setminus N)} \\ &\geq M \cdot \mu(U \setminus N) \\ &= M \cdot \mu(U)\end{aligned}$$

By dividing both sides by $M > 0$, we obtain $\mu(U) \leq \frac{1}{M} \cdot \phi(U) = C \cdot \phi(U)$.

PART 3 ($\phi \ll \mu$ AND $\mu \ll \phi$). The absolute continuity of ϕ with respect to μ follows from the absolute continuity of the integral. Let $C > 0$ be the constant derived in the previous part. The absolute continuity $\mu \ll \phi$ follows because

$$\mu(N) \leq C \cdot \phi(N) = C \cdot 0 = 0 \qquad \forall N \in \Sigma \colon \phi(N) = 0.$$

PART 4 ($|\nu|(U) \leq C \cdot C' \cdot \phi(U)$). Let $\nu \colon \Sigma \to Y$ be a vector measure of bounded variation such that

$$|\nu|(U) \leq C' \cdot \mu(U) \qquad \forall U \in \Sigma$$

for a constant $C' \geq 0$. Then we have

$$|\nu|(U) \leq C' \cdot C \cdot \phi(U) \qquad \forall U \in \Sigma$$

where $C > 0$ is the constant derived in Part 2. □

2.4.3 Derivation: ODE Case

The differentiation method described in this section was first developed for a slightly simplified setting in [HLS22, Sec. 3.3.1].

In this section, we use results from Section 2.4.2 to show the differentiability of scalar functions whose evaluation involves the solution of ordinary differential equations. A fairly generic template for such functions is the Bolza-type cost functional which was first formulated in [Bol13]:

$$j(u) = \int_0^{t_f} l\big(x(t), u(t)\big) \,\mathrm{d}t + \Phi(x(t_f)).$$

Here $u \in L^1_\lambda\big([0,\ t_f],\ \mathbb{R}^{n_w}\big)$ is a real-valued control function whose component functions we ultimately expect to be the layers of a characteristic function in a layered measure space with $n_w \in \mathbb{N}$ layers. Theorem 2.1.20 guarantees that we can consider individual layer functions and the layered function interchangeably. The mapping $x \in W^{1,1}\big([0,\ t_f], \mathbb{R}^{n_x}\big)$ is a state function with $n_x \in \mathbb{N}$ components that is determined by solving an initial value problem of the form

$$\begin{aligned}&\dot{x}(t) = f\big(x(t), u(t)\big) \quad \text{for a.a. } t \in [0,\ t_f], \\ &x(0) = x_0\end{aligned} \tag{2.48}$$

for some fixed initial datum $x_0 \in \mathbb{R}^{n_x}$. Because the control function u will ultimately only assume the values in $\{0,\ 1\}^{n_w}$, we can use partial outer convexification to simplify the expressions for l and f. We assume that they take the form

$$l(x,u) = l_0(x) + \sum_{i=1}^{n_w} l_i(x) \cdot u_i,$$
$$f(x,u) = f_0(x) + \sum_{i=1}^{n_w} f_i(x) \cdot u_i.$$

We note that we only discuss the case of autonomous ODE systems, i.e., ODE systems that do not have an explicit time dependency. This is not a theoretical restriction. However, the assumptions we have to make are substantially simplified by limiting ourselves to this case. If we were to include non-autonomous ODEs in our discussion, we would have to make a number of additional measurability and integrability assumptions.

To prove the existence and uniqueness of our IVP solution, we invoke existence theory of Carathéodory type as described, for instance, in [Hal69, Sec. I.5]. We presume that the reader has prior knowledge of differentiability conditions stated therein. This theory yields absolutely continuous solutions. In $\mathbb{R}$, the set of absolutely continuous functions and $W^{1,1}$ are identical. We collect the assumptions for the ODE setting in Assumption 2.4.17.

Assumption 2.4.17 (Bolza-Type Cost Functional).
Let $n_x, n_w \in \mathbb{N}_0$; let $t_f > 0$; and let $x_0 \in \mathbb{R}^{n_x}$; and $D \subseteq \mathbb{R}^{n_x}$ satisfy:

(1) D is open and convex with $x_0 \in D$.

Let $f_i : D \to \mathbb{R}^{n_x}$ and $l_i : D \to \mathbb{R}^{n_x}$ for $i \in [n_w]_0$ satisfy:

(2) for all $i \in [n_w]_0$, f_i is continuously differentiable;

(3) for all $i \in [n_w]_0$, $\nabla_x f_i$ is locally Lipschitz continuous;

(4) $\exists L \geq 0 \colon \left\| f_i(x) - f_i(y) \right\| \leq L \cdot \|x - y\| \ \forall i \in [n_w]_0,\ x, y \in D$;

(5) there exists $\varepsilon > 0$ such that for all $u \in L^1_\lambda\big([0,\ t_f],\ [0,\ 1]^{n_w}\big)$, all $\tau \in (0,\ t_f]$, and all absolutely continuous functions $x \colon [0,\ \tau) \to D$ with $x(0) = x_0$ and $\dot{x}(t) = f_0\big(x(t)\big) + \sum_{i=1}^{n_w} f_i\big(x(t)\big) \cdot u_i(t)$ for almost all $t \in [0,\ \tau)$, $B_\varepsilon\big(x(t)\big) \subseteq D$ holds for all $t \in [0,\ \tau)$;

(6) for all $i \in [n_w]_0$, l_i is continuously differentiable;

(7) for all $i \in [n_w]_0$, $\nabla_x l_i$ is locally Lipschitz continuous.

Finally, let $\Phi \colon [0,\ t_f] \times D \to \mathbb{R}$ and $m \colon [t_0,\ t_f] \to \mathbb{R}^{n_w}$ satisfy:

(8) $\Phi(t,x)$ is continuously differentiable in x for fixed t;

(9) for each $i \in [n_w]$, $m_i \in L^\infty_\lambda\big([t_0,\ t_f]\big)$ is a scaling density function for λ. ◁

Throughout this section, we will use the shorthands

$$l(x,u) \coloneqq l_0(x) + \sum_{i=1}^{n_w} l_i(x) \cdot u_i,$$
$$f(x,u) \coloneqq f_0(x) + \sum_{i=1}^{n_w} f_i(x) \cdot u_i.$$

Under Assumption 2.4.17, we prove existence and uniqueness of the IVP solution for all relaxed controls $u \in L^1_\lambda\big([0,\ t_f],\ [0,\ 1]^{n_w}\big)$ as well as for relaxed controls in a small L^1 environment around such u. Proving the existence and uniqueness of solutions in a small environment around the relevant set of controls is essential for the applicability of the term "derivative." Generally speaking, derivatives describe the behavior of a function in a small environment around a point. Because all binary-valued control functions lie on the boundary of $L^1_\lambda\big([0,\ t_f],\ [0,\ 1]^{n_w}\big)$, some concerns may arise as to whether derivatives meaningfully exist around such functions. Proving the existence of solutions in a δ-environment around such controls removes this concern.

We note that the functions f_i are vector-valued and the functions $\nabla_x f_i$ are matrix-valued. We have not yet specified which norms are used in Assumption 2.4.17. In fact, we do not have to specify the precise norms used because all norms are equivalent on finite-dimensional $\mathbb{R}^n$. Unless otherwise specified, we use the Euclidean norm in $\mathbb{R}^{n_x}$ and the standard operator norm in $\mathbb{R}^{n_x \times n_x}$ that is induced by the Euclidean norm.

To simplify notation, we introduce three sets of control functions: binary-valued control functions, $[0,\ 1]$-valued control functions, and real-valued control functions that are nearly $[0,\ 1]$-valued in an L^1 sense.

Definition 2.4.18 (Control Function Classes for ODE Functions).
Let $n_w \in \mathbb{N}$, $t_f > 0$. We define

$$\begin{aligned}
&\mathscr{B}^{n_w}(t_f) \coloneqq L^1_\lambda\big([0,\ t_f],\ \{0,\ 1\}^{n_w}\big),\\
&\mathscr{B}_0^{n_w}(t_f) \coloneqq L^1_\lambda\big([0,\ t_f],\ [0,\ 1]^{n_w}\big),\\
&\mathscr{B}_\delta^{n_w}(t_f) \coloneqq \big\{u \in L^1_\lambda\big([0,\ t_f],\ \mathbb{R}^{n_w}\big) \,\big|\, \exists v \in \mathscr{B}_0^{n_w}\big([0,\ t_f]\big)\colon \|u - v\|_{L^1} \le \delta\big\}\ \forall \delta > 0. \qquad \lhd
\end{aligned}$$

We can easily show that

$$\mathscr{B}^{n_w}(t_f) \subseteq \mathscr{B}_0^{n_w}(t_f) \subseteq \mathscr{B}_{\delta_1}^{n_w}(t_f) \subseteq \mathscr{B}_{\delta_2}^{n_w}(t_f) \qquad \forall \delta_1, \delta_2 \in \mathbb{R}\colon 0 < \delta_1 \le \delta_2.$$

We can now show that the initial value problem (2.48) has a unique solution for all $\alpha \in \mathscr{B}_\delta^{n_w}(t_f)$ for suitably chosen $\delta > 0$.

Theorem 2.4.19 (Existence and Uniqueness of Solutions).
Let $n_x, n_w \in \mathbb{N}$; $t_f > 0$; $x_0 \in \mathbb{R}^{n_x}$; $D \subseteq \mathbb{R}^{n_x}$; and $f_i\colon [0,\ t_f] \times D \to \mathbb{R}^{n_x}$ for $i \in [n_w]_0$ satisfy Assumptions 2.4.17 (1), 2.4.17 (2), 2.4.17 (4) and 2.4.17 (5). Then there exist a constant $\delta > 0$ and a compact convex set $K \subseteq D$ with $x_0 \in K$ such that for every $u \in \mathscr{B}_\delta^{n_w}(t_f)$, there exists a unique absolutely continuous function $x_u\colon [0,\ t_f] \to K$ with

$$\begin{aligned}
&\dot{x}_u(t) = f(x_u(t), u(t)) \qquad \text{for a.a. } t \in [0,\ t_f],\\
&x_u(0) = x_0. \qquad \lhd
\end{aligned}$$

PROOF. Let $u \in L^1\big([0,\ t_f],\ \mathbb{R}^{n_w}\big)$ and let $v \in \mathscr{B}_0^{n_w}(t_f)$ such that $\|u - v\|_{L^1} \le \delta$ for a constant $\delta > 0$ that we have yet to specify. We note that the following arguments especially apply for $u = v$.

PART 1 (EXTENDING THE TIME INTERVAL). We extend all real-valued control functions $w \in L^1\big([0,\ t_f],\ \mathbb{R}^{n_w}\big)$ to $I \coloneqq (-1,\ t_f]$. This is to ensure that 0 does

not lie on the boundary of the time interval, which may cause issues with local existence theorems. We define $\overline{w}\colon I \to \mathbb{R}^{n_w}$ by

$$\overline{w}(t) \coloneqq \begin{cases} w(t) & \text{if } t \geq 0 \\ (0)_{i \in [n_w]} & \text{if } t < 0. \end{cases} \quad \forall w \in L^1\big([0,\ t_f],\ \mathbb{R}^{n_w}\big),\ t \in I.$$

This extension evidently satisfies $\overline{w} \in L^1_\lambda\big(I,\ \mathbb{R}^{n_w}\big)$.

Part 2 (Aggregated Right Hand Side). To simplify notation, for each real-valued control function $w \in L^1\big([0,\ t_f], \mathbb{R}^{n_w}\big)$, we define the aggregate right hand side function $f_w\colon I \times D \to \mathbb{R}^{n_x}$ with

$$f_w(t,x) \coloneqq f\big(x, w(t)\big).$$

In our discussion, we are specifically interested in f_u and f_v.

Because each f_i for $i \in [n_w]_0$ is constant in t for fixed x and because w is measurable in t, f_w is measurable in t for fixed x. Furthermore, because w is constant in x and because f is continuous in x, f_w is continuous in x. Because f_w is not constant over time, we now have to establish integrable bounds and Lipschitz constants for f_w. We first define $k_{1,w}\colon I \to \mathbb{R}_{\geq 0}$ with

$$k_{1,w}(t) \coloneqq \big\|f_0(x_0)\big\| + \sum_{i=1}^{n_w} \big\|f_i(x_0)\big\| \cdot \big|\overline{w}_i(t)\big| \quad \forall t \in I.$$

We then have

$$\big\|f_w(t,x_0)\big\| = \Big\|f\big(x_0, \overline{w}(t)\big)\Big\| \leq k_{1,w}(t) \quad \forall t \in I.$$

We also have

$$\int_I \big|k_{1,w}(t)\big|\,\mathrm{d}\lambda \leq (t_f+1) \cdot \big\|f_0(x_0)\big\| + \sum_{i=1}^{n_w} \big\|f_i(x_0)\big\| \cdot \|\overline{w}_i\|_{L^1} < \infty$$

and therefore $k_{1,w} \in L^1_\lambda(I)$. We define $k_{2,w}\colon I \to \mathbb{R}_{\geq 0}$ with

$$k_{2,w}(t) \coloneqq L \cdot \Big(1 + \big\|\overline{w}(t)\big\|_1\Big) \quad \forall t \in I.$$

We then have

$$\begin{aligned} \big\|f_w(t,x) - f_w(t,y)\big\| &\leq \big\|f_0(x) - f_0(y)\big\| + \sum_{i=1}^{n_w} \big\|f_i(x) - f_i(y)\big\| \cdot \big|\overline{w}_i(t)\big| \\ &= L \cdot \Big(1 + \sum_{i=1}^{n_w} \big|\overline{w}_i(t)\big|\Big) \cdot \|x - y\| \\ &= L \cdot \Big(1 + \big\|\overline{w}(t)\big\|_1\Big) \cdot \|x - y\| \\ &= k_{2,w}(t) \cdot \|x - y\| \end{aligned}$$

for all $t \in I$ and $x, y \in D$. We further have

$$\int_I \big|k_{2,w}(t)\big|\,\mathrm{d}\lambda \leq L \cdot \Big((t_f+1) + \int_I \big\|\overline{w}(t)\big\|_1\,\mathrm{d}\lambda\Big) = L \cdot \big(t_f + 1 + \|\overline{w}\|_{L^1}\big) < \infty,$$

which implies that $k_{2,w} \in L^1_\lambda(I)$.

PART 3 (LOCAL EXISTENCE). In the previous part of this proof, we showed that for every control function $w \in L^1\big([0,\ t_f], \mathbb{R}^{n_w}\big)$, the aggregate right hand side function $f_w : I \times D \to \mathbb{R}^{n_x}$ is measurable in t for fixed x, is continuous in x for fixed t, and satisfies

$$\begin{aligned} \big\| f_w(t,x_0) \big\| &\leq k_{1,w}(t) && \forall t \in I, \\ \big\| f_w(t,x) - f_w(t,y) \big\| &\leq k_{2,w}(t) \cdot \|x-y\| && \forall t \in I,\ x,y \in D \end{aligned}$$

for suitably chosen bound functions $k_{1,w}, k_{2,w} \in L^1_\lambda(I)$.

Therefore, for every $w \in L^1\big([0,\ t_f],\ \mathbb{R}^{n_w}\big)$ and every $(\tau,\ \xi) \in (-1,\ t_f) \times D$, there exist an open interval $J_{w,\tau,\xi} \subseteq (-1,\ t_f)$ with $\tau \in J_{w,\tau,\xi}$ and a unique absolutely continuous function $x_{w,\tau,\xi} : J_{y,\tau,\xi} \to D$ with

$$\begin{aligned} &\dot{x}_{w,\tau,\xi}(t) = f_y\big(t, x_{w,\tau,\xi}(t)\big) \quad \text{for a.a. } t \in J_{w,\tau,\xi}, \\ &x_{w,\tau,\xi}(\tau) = \xi. \end{aligned}$$

If $x_{w,\tau,\xi}(t)$ does not approach the boundary of $(-1,\ t_f) \times D$ as t approaches the boundary of $J_{w,\tau,\xi}$, then we can extend the solution beyond $J_{w,\tau,\xi}$. In this way, we extend $x_{w,\tau,\xi}$ to its maximal existence interval $I_{w,\tau,\xi} \subseteq (-1,\ t_f)$.

PART 4 (GLOBAL EXISTENCE). In this part of the proof we look specifically at $x_u := x_{u,0,x_0}$ and $x_v := x_{v,0,x_0}$ and their maximal existence intervals $I_u := I_{u,0,x_0}$ and $I_v := I_{v,0,x_0}$. We first look at the $[0,\ 1]$-valued control function v.

According to Assumption 2.4.17 (5), there exists a constant $\varepsilon > 0$ such that

$$\operatorname{dist}\big(x_v(t),\ \partial D\big) \geq \varepsilon \qquad \forall t \in [0,\ \sup I_v).$$

Because $\big(t,\ x_v(t)\big)$ approaches $\partial\big((-1,\ t_f) \times D\big) = \big(\{-1,\ t_f\} \times D\big) \cup \big([-1,\ t_f] \times \partial D\big)$ for $t \to \sup I_v$, it follows that $\sup I_v = t_f$. Therefore, we have $[0,\ t_f) \subseteq I_v$. This also ensures that $[0,\ t_f) \cap I_u \subseteq I_v$.

We can also establish a bound for the distance between $x_v(t)$ and x_0 for $t \in [0,\ t_f)$. To do so, we first find that for all $t \in [0,\ t_f)$, we have

$$\begin{aligned} \big\| x_v(t) - x_0 \big\| &\leq \underbrace{\big\| x_v(0) - x_0 \big\|}_{=0} + \int_0^t \Big\| f_v\big(s, x_v(s)\big) \Big\| \,\mathrm{d}s \\ &\leq \int_0^t \left(\Big\| f_0\big(x_v(s)\big) \Big\| + \sum_{i=1}^{n_w} \Big\| f_i\big(x_v(s)\big) \Big\| \cdot \underbrace{\big| v_i(s) \big|}_{\leq 1} \right) \mathrm{d}s \\ &\leq \int_0^t \Bigg(\underbrace{\Big\| f_0\big(x_v(s)\big) \Big\|}_{\leq \| f_0(x_0) \| + L \cdot \| x_v(s) - x_0 \|} + \sum_{i=1}^{n_w} \underbrace{\Big\| f_i\big(x_v(s)\big) \Big\|}_{\leq \| f_i(x_0) \| + L \cdot \| x_v(s) - x_0 \|} \Bigg) \mathrm{d}s \\ &\leq (1 + n_w) \cdot \big\| f(x_0) \big\|_\infty \cdot t + \int_0^t (1 + n_w) \cdot L \cdot \big\| x_v(s) - x_0 \big\| \,\mathrm{d}s \end{aligned}$$

where $L \geq 0$ is the Lipschitz constant from Assumption 2.4.17 (4). Because $t \mapsto (1 + n_w) \cdot \big\| f(x_0) \big\|_\infty \cdot t$ is non-decreasing, we can apply the simplified form of

Grönwall's lemma to obtain the estimate

$$\begin{aligned}\left\|x_v(t)-x_0\right\| &\le (1+n_w)\cdot\left\|f(x_0)\right\|_\infty\cdot t\cdot\exp\left(\int_0^t(1+n_w)\cdot L\,\mathrm{d}s\right)\\ &= (1+n_w)\cdot\left\|f(x_0)\right\|_\infty\cdot t\cdot e^{(1+n_w)\cdot L\cdot t}\\ &\le \underbrace{(1+n_w)\cdot\left\|f(x_0)\right\|_\infty\cdot t_f\cdot e^{(1+n_w)\cdot L\cdot t_f}}_{=:C_0}.\end{aligned}$$

for all $t\in[0,\ t_f)$. For all $t\ge 0$ with $t\in I_u$, we have

$$\begin{aligned}\left\|x_u(t)-x_v(t)\right\| &\le \underbrace{\left\|x_u(0)-x_v(0)\right\|}_{=0}+\int_0^t\left\|f_u\big(s,x_u(s)\big)-f_v\big(s,x_v(s)\big)\right\|\mathrm{d}s\\ &\le \int_0^t\left\|f_u\big(s,x_u(s)\big)-f_u\big(s,x_v(s)\big)\right\|+\left\|f_u\big(s,x_v(s)\big)-f_v\big(s,x_v(s)\big)\right\|\mathrm{d}s\\ &\le \int_0^t k_{2,u}(t)\cdot\left\|x_u(s)-x_v(s)\right\|+\sum_{i=1}^{n_w}\underbrace{\left\|f_i\big(x_v(s)\big)\right\|}_{\le\|f(x_0)\|_\infty+L\cdot\|x_v(s)-x_0\|}\cdot\left|u_i(s)-v_i(s)\right|\mathrm{d}s\\ &\le \int_0^t k_{2,u}(t)\cdot\left\|x_u(s)-x_v(s)\right\|+\underbrace{\left(\left\|f(x_0)\right\|_\infty+L\cdot C_0\right)}_{=:C_1}\cdot\left\|u(s)-v(s)\right\|_1\mathrm{d}s\\ &= C_1\cdot\int_0^t\left\|u(s)-v(s)\right\|_1\mathrm{d}s+\int_0^t k_{2,u}(t)\cdot\left\|x_u(t)-x_v(t)\right\|\mathrm{d}s.\end{aligned}$$

Here, we can once more apply the simplified form of Grönwall's lemma because $t\mapsto C_1\cdot\int_0^t\left\|u(s)-v(s)\right\|_1\mathrm{d}s$ is increasing. This yields the estimate

$$\begin{aligned}\left\|x_u(t)-x_v(t)\right\| &\le C_1\cdot\underbrace{\left(\int_0^t\left\|u(s)-v(s)\right\|_1\mathrm{d}s\right)}_{\le\|u-v\|_{L^1}\le\delta}\cdot\exp\left(\int_0^t k_{2,u}(s)\,\mathrm{d}s\right)\\ &\le C_1\cdot\delta\cdot\exp\left(\|k_{2,u}\|_{L^1}\right)\\ &\le C_1\cdot\delta\cdot e^{L\cdot(1+t_f+\|\overline{u}\|_{L^1})}\\ &= C_1\cdot\delta\cdot e^{L\cdot(1+t_f+\|u\|_{L^1})}\\ &\le C_1\cdot\delta\cdot e^{L\cdot(1+t_f+\|v\|_{L^1}+\|u-v\|_{L^1})}\\ &\le C_1\cdot\delta\cdot e^{L\cdot(1+t_f+n_w\cdot t_f+\delta)}\\ &= \underbrace{C_1e^{L+L\cdot(1+n_w)\cdot t_f}}_{=:C_2}\cdot\delta e^{\delta}\end{aligned}$$

for all $t\ge 0$ with $t\in I_u$. Since $\delta\mapsto\delta e^{\delta}$ is continuous and strictly increasing with $0\cdot e^0=0$ and $\delta e^{\delta}\xrightarrow{\delta\to\infty}\infty$, there exists $\delta>0$ such that

$$\delta e^{\delta}=\frac{\varepsilon}{2C_2}.$$

We choose this to be our as yet unspecified constant $\delta > 0$ and obtain the estimates

$$\begin{aligned}\left\|x_u(t) - x_v(t)\right\| &\leq \frac{\varepsilon}{2},\\ \left\|x_u(t) - x_0\right\| &\leq \|x_v(t) - x_0\| + \|x_u(t) - x_v(t)\|\\ &\leq \underbrace{C_0 + \frac{\varepsilon}{2}}_{=:C_3}\end{aligned}$$

for all $t \geq 0$ with $t \in I_u$. The first estimate implies that

$$\operatorname{dist}\big(x_u(t),\, \partial D\big) \geq \operatorname{dist}\big(x_v(t),\, \partial D\big) - \left\|x_u(t) - x_v(t)\right\| \geq \varepsilon - \frac{\varepsilon}{2} = \frac{\varepsilon}{2} > 0 \quad \forall t \in [0,\, \sup I_u).$$

This then implies that $\sup I_u = t_f$. Furthermore, the set

$$K := \left\{ x \in D \,\middle|\, \operatorname{dist}(x, \partial D) \geq \frac{\varepsilon}{2} \right\} \cap \overline{B}_{C_3}(x_0)$$

is closed because it is an intersection of two closed sets. It is also bounded because it is a subset of $\overline{B}_{C_3}(x_0)$. Therefore K is compact. By definition, we have $x_u(t) \in K$ for all $t \in [0,\, t_f)$. Because K is compact and x_u is continuous, $x_u(t)$ has a well-defined limit for $t \to t_f$ and we can continuously extend x_u to $[0,\, t_f]$. This does not affect absolute continuity.

PART 5 (UNIQUENESS). Let $x' \colon [0,\, t_f] \to D$ be any absolutely continuous function such that

$$\begin{aligned}\dot{x}'(t) &= f_0\big(t, x'(t)\big) + \sum_{i=1}^{n_w} f_i\big(t, x'(t)\big) \cdot u_i(t) \quad \text{for a.a. } t \in [0,\, t_f],\\ x'(0) &= x_0.\end{aligned}$$

We prove by contradiction that $x'(t) = x_u(t)$ for all $t \in [0,\, t_f]$. If we were to assume that there existed any $t \in [0,\, t_f]$ such that $x'(t) \neq x_u(t)$, then

$$t^* := \inf\big\{ t \in [0,\, t_f] \,\big|\, x'(t) \neq x_u(t) \big\}$$

would satisfy $t^* \geq 0$ because $x'(0) = x_u(0)$, and $t^* \leq t_f$ because there would exist $t \in [0,\, t_f]$ with $x'(t) \neq x_u(t)$.

If $t^* = 0$, then we would have $x'(t^*) = x_u(t^*)$ due to the initial condition. If $t^* > 0$, then we would have $x'(t) = x_u(t)$ for all $t \in [0,\, t^*)$. Continuity of x' and x_u would then imply $x'(t^*) = x_u(t^*)$. If $t^* = t_f$, then this would already imply that $x'(t) = x_u(t)$ for all $t \in [0,\, t_f]$. Therefore, if there were any $t \in [0,\, t_f]$ with $x'(t) \neq x_u(t^*)$, then we would have $t^* < t_f$.

If $t^* < t_f$, then both x' and x_u would be local Carathéodory-type solutions to the differential equation around $(t^*,\, x'(t^*)) = (t^*, x_u(t^*))$. The uniqueness of the local solution around that point then states that there exists $\eta > 0$ such that

$$x'(t) = x_u(t) \qquad \forall t \in [t^*,\, t^* + \eta).$$

This would contradict the definition of t^* which implies that for every $\eta > 0$, there would exist $t \in [t^*,\, t^* + \eta)$ with $x'(t) \neq x_u(t)$.

This contradiction shows that our initial assumption that there are $t \in [0,\, t_f]$ with $x'(t) \neq x_u(t)$ must be incorrect and therefore, we have $x'(t) = x_u(t)$ for all $t \in [0,\, t_f]$.

PART 6 (K IS CONVEX). Because the intersection of two convex sets is convex and because $\overline{B}_{C_3}(x_0)$ is convex due to the triangle inequality, it is sufficient to show that

$$D' \coloneqq \left\{ x \in D \;\middle|\; \operatorname{dist}(x,\, \partial D) \geq \frac{\varepsilon}{2} \right\}$$

is convex. Let $x, y \in D'$ and let $\rho \in (0,\, 1)$. If we were to assume that the convex combination $z \coloneqq \rho \cdot y + (1-\rho) \cdot x$ did not lie in D', then there would exist $z' \in \partial D$ with $\|z - z'\| < \frac{\varepsilon}{2}$.

Because $x, y \in D' \subseteq D$ and $\|z - z'\| < \frac{\varepsilon}{2}$, we would then have $x' \coloneqq x + (z' - z) \in D$ and $y' \coloneqq y + (z' - z) \in D$. These points would then satisfy

$$\begin{aligned} z' &= z + (z' - z) \\ &= \rho y + (1-\rho)x + \rho(z' - z) + (1-\rho)(z' - z) \\ &= \rho y' + (1-\rho)x'. \end{aligned}$$

Because $z' \in \partial D$ and because D is open, this would contradict the assumption that D is convex. Therefore, our initial assumption that $z \notin D'$ would have to have been wrong and D' must be convex. It then follows that K is also convex.□

To prove differentiability later on, we require a stability result that controls the pointwise difference in state with respect to the L^1 difference in control.

Lemma 2.4.20 (Stability).
Let n_x, n_w, t_f, x_0, D, f, l, and Φ satisfy Assumptions 2.4.17 (1), 2.4.17 (2), 2.4.17 (4) and 2.4.17 (5). Let $\delta > 0$ be constant and let $K \subseteq D$ be compact such that for every $u \in \mathscr{B}_\delta^{n_w}(t_f)$, there exists an absolutely continuous function $x_u : [0,\, t_f] \to K$ with

$$\begin{aligned} &\dot{x}_u(t) = f\big(x_u(t), u(t)\big) && \text{for a.a. } t \in [0,\, t_f], \\ &x_u(0) = 0. \end{aligned}$$

Then there exist constants $C_{1,\delta,K}, C_2 \geq 0$ such that

$$\big\|x_u(t) - x_v(t)\big\| \leq C_{1,\delta,K} \cdot e^{C_2 t} \cdot \int_0^t \big\|u(s) - v(s)\big\|_1 \,\mathrm{d}s \quad \forall u, v \in \mathscr{B}_\delta^{n_w}(t_f),\ t \in [0,\, t_f].$$

This implies the more relaxed estimate

$$\big\|x_u(t) - x_v(t)\big\| \leq C_{1,\delta,K} \cdot e^{C_2 t_f} \cdot \big\|u - v\big\|_{L^1} \quad \forall u, v \in \mathscr{B}_\delta^{n_w}(t_f),\ t \in [0,\, t_f]. \qquad \triangleleft$$

PROOF. We prove this result using Grönwall's lemma. Let $u, v \in \mathscr{B}_\delta^{n_w}(t_f)$. Let $L \geq 0$ be the constant from Assumption 2.4.17 (4). Because f_i is continuous for all $i \in [n_w]_0$ and K is compact, the quantity

$$M \coloneqq \max_{\substack{i \in [n_w]_0 \\ x \in K}} f_i(x) \in [0,\, \infty)$$

is constant. For each $t \in [0,\ t_f]$, we have

$$\begin{aligned}
\left\|x_u(t) - x_v(t)\right\| &\leq \underbrace{\left\|x_u(0) - x_v(0)\right\|}_{=0} + \int_0^t \left\| f\big(x_u(t), u(t)\big) - f\big(x_v(t), v(t)\big) \right\| \mathrm{d}s \\
&\leq \int_0^t \Bigg(\left\| f_0\big(x_u(s)\big) - f_0\big(x_v(s)\big) \right\| \\
&\qquad + \sum_{i=1}^{n_w} \left\| f_i\big(x_u(s)\big) u_i(s) - f_i\big(x_v(s)\big) v_i(s) \right\| \Bigg) \mathrm{d}s \\
&\leq \int_0^t \Bigg(L \cdot \left\|x_u(s) - x_v(s)\right\| + \sum_{i=1}^{n_w} \Bigg(\left\| f_i\big(x_u(s)\big) - f_i\big(x_v(s)\big) \right\| \cdot \left|u_i(s)\right| \\
&\qquad + \left\| f_i\big(x_v(s)\big) \right\| \cdot \left|u_i(s) - v_i(s)\right| \Bigg) \Bigg) \mathrm{d}s \\
&\leq M \cdot \int_0^t \left\|u(s) - v(s)\right\|_1 \mathrm{d}s + \int_0^t L \cdot \Big(1 + \left\|u(s)\right\|_1\Big) \cdot \left\|x_u(s) - x_v(s)\right\| \mathrm{d}s
\end{aligned}$$

Since $t \mapsto M \cdot \int_0^t \left\|u(s) - v(s)\right\|_1 \mathrm{d}s$ is montonically increasing, we can use the simplified version of Grönwall's lemma to obtain

$$\begin{aligned}
\left\|x_u(t) - x_v(t)\right\| &\leq M \cdot \int_0^t \left\|u(s) - v(s)\right\|_1 \mathrm{d}s \cdot \exp\Bigg(L \cdot \Bigg(t + \int_0^t \left\|u(s)\right\|_1 \mathrm{d}s \Bigg) \Bigg) \\
&\leq M \cdot \int_0^t \left\|u(s) - v(s)\right\|_1 \mathrm{d}s \cdot \exp\big(L \cdot t + L \cdot \underbrace{\|u\|_{L^1}}_{\leq n_w \cdot t + \delta} \big) \\
&\leq M e^{L\delta} \cdot e^{L \cdot (1 + n_w) \cdot t} \cdot \int_0^t \left\|u(s) - v(s)\right\|_1 \mathrm{d}s
\end{aligned}$$

Therefore, the claim holds with $C_{1,\delta,K} \coloneqq M e^{L\delta}$ and $C_2 \coloneqq L \cdot (1 + n_w)$. The looser estimate then follows from $t \leq t_f$ and $\int_0^t \left\|u(s) - v(s)\right\|_1 \mathrm{d}s \leq \|u - v\|_{L^1}$. □

In the next step, we show that the objective function value is well-defined and that the adjoint state equations have a unique solution. The adjoint state equations take the form

$$\begin{aligned}
\dot{\xi}(t) &= -\Big(\nabla_x l\big(x_u(t), u(t)\big)\Big)^T - \Big(\nabla_x f\big(x_u(t), u(t)\big)\Big)^T \xi(t) \quad \text{for a.a. } t \in [0,\ t_f], \\
\xi(t_f) &= \Big(\nabla_x \Phi\big(x_u(t_f)\big)\Big)^T.
\end{aligned}$$

The solution $\xi : [0,\ t_f] \to \mathbb{R}^{n_x}$ of this boundary value problem is generally known as the *adjoint state*. The adjoint state is usually referred to by the greek letter λ. We use ξ to avoid confusion with the Lebesgue measure.

Because the adjoint state depends on the control function u, we add an index ξ_u to clarify which control an adjoint state is associated with.

Proposition 2.4.21 (Existence and Uniqueness of the Adjoint State).
Let n_x, n_w, t_f, x_0, D, f, l, and Φ satisfy Assumptions 2.4.17 (1), 2.4.17 (2), 2.4.17 (4) to 2.4.17 (6) and 2.4.17 (8). Let $\delta > 0$ be constant and $K \subseteq D$ be convex and compact such that for every nearly $[0,\ 1]$-valued control function $u \in \mathscr{B}_\delta^{n_w}(t_f)$, there exists a unique absolutely continuous function $x_u : [0,\ t_f] \to K$ such that

$$\begin{aligned}
\dot{x}_u(t) &= f\big(x_u(t), u(t)\big) \quad \textit{for a.a. } t \in [0,\ t_f], \\
x_u(0) &= x_0.
\end{aligned}$$

Then the map $j\colon \mathscr{B}_\delta^{n_w}(t_f) \to \mathbb{R}$ *with*

$$j(u) \coloneqq \Phi\big(x_u(t_f)\big) + \int_0^{t_f} l\big(x_u(t), u(t)\big)\,\mathrm{d}t \quad \forall u \in \mathscr{B}_\delta^{n_w}(t_f)$$

is well-defined and for every $u \in \mathscr{B}_\delta^{n_w}(t_f)$*, there exists a unique absolutely continuous function* $\xi_u\colon [0,\ t_f] \to \mathbb{R}^{n_x}$ *such that*

$$\begin{aligned}\dot{\xi}_u(t) &= -\Big(\nabla_x l\big(x_u(t), u(t)\big)\Big)^T - \Big(\nabla_x f\big(x_u(t), u(t)\big)\Big)^T \xi_u(t) \quad \text{for a.a. } t \in [0,\ t_f],\\ \xi_u(t_f) &= \Big(\nabla_x \Phi\big(x_u(t_f)\big)\Big)^T.\end{aligned}$$

There exists a constant $M_{\xi,\delta,K} \geq 0$ *such that*

$$\big\|\xi_u(t)\big\| \leq M_{\xi,\delta,K} \quad \forall u \in \mathscr{B}_\delta^{n_w}(t_f),\ t \in [0,\ t_f]. \qquad \triangleleft$$

Proof. Part 1 (Well-Definedness of $j(u)$). Let $u \in \mathscr{B}_\delta^{n_w}(t_f)$. Because $x_u(t_f) \in K \subseteq D$, $\Phi\big(x_u(t_f)\big)$ is well-defined. It is therefore sufficient to show that the integral is well-defined to show that $j(u)$ is well-defined. The integrand $t \mapsto l\big(x_u(t), u(t)\big)$ is a sum of products between continuous functions of t and is therefore itself continuous. Because the interval $[0,\ t_f]$ is compact, this map is integrable over $[0,\ t_f]$. Therefore, $j(u)$ is well-defined.

Part 2 (Existence of ξ_u). The adjoint state ξ_u is the solution to the boundary value problem

$$\begin{aligned}\dot{\xi_u}(t) &= F\big(t, \xi_u(t)\big) \qquad \text{for a.a. } t \in [0,\ t_f],\\ \xi_u(t_f) &= \Big(\nabla_x \Phi\big(x_u(t_f)\big)\Big)^T\end{aligned}$$

with

$$F(t,\xi) \coloneqq -\Big(\nabla_x l\big(x_u(t), u(t)\big)\Big)^T - \Big(\nabla_x f\big(x_u(t), u(t)\big)\Big)^T \cdot \xi.$$

We note that for every fixed $\xi \in \mathbb{R}^{n_x}$, $t \mapsto F(t,\xi)$ is measurable. This is because for all $i \in [n_w]_0$, the maps $t \mapsto \nabla_x f_i\big(x_u(t)\big)$ and $t \mapsto \nabla_x l_i\big(x_u(t)\big)$ are compositions of continuous functions and are therefore themselves continuous. Since u_i is measurable in t and ξ is fixed, $t \mapsto F(t,\xi)$ is measurable in t. For fixed $t \in [0,\ t_f]$, $\xi \mapsto F(t,\xi)$ is linear and therefore continuous.

Next, we have to derive integrable bounds for F and an integrable Lipschitz constant. Since $x \mapsto \nabla_x f_i(x)$ and $x \mapsto \nabla_x l_i(x)$ are continuous for $i \in [n_w]_0$ and K is compact, there exist constants $C_{0,K}, C_{1,K} \geq 0$ such that

$$\begin{aligned}C_{0,K} &\geq \big\|\nabla_x l_i(x)\big\| \quad \forall i \in [n_w]_0,\ x \in K,\\ C_{1,K} &\geq \big\|\nabla_x f_i(x)\big\| \quad \forall i \in [n_w]_0,\ x \in K.\end{aligned}$$

To simplify notation, we define $\xi_0 \coloneqq \nabla_x \Phi\big(x_u(t)\big)$. For each $t \in [0,\ t_f]$, we have $x_u(t) \in K$ and therefore

$$\begin{aligned}\Big\|\nabla_x l\big(x_u(t), u(t)\big)\Big\| &\leq \Big\|\nabla_x l_0\big(x_u(t)\big)\Big\| + \sum_{i=1}^{n_w} \Big\|\nabla_x l_i\big(x_u(t)\big)\Big\| \cdot \big|u_i(t)\big|\\ &\leq C_{0,K} \cdot \Big(1 + \big\|u(t)\big\|_1\Big).\end{aligned}$$

Similarly, we find that

$$\left\| \left(\nabla_x f\left(x_u(t), u(t)\right) \right) \cdot \xi_0 \right\| \leq C_{1,K} \cdot \|\xi_0\| \cdot \left(1 + \left\|u(t)\right\|_1\right).$$

Thus, we have

$$\left\|F(t,\xi_0)\right\| \leq \underbrace{\left(C_{0,K} + C_{1,K} \cdot \|\xi_0\|\right) \cdot \left(1 + \left\|u(t)\right\|_1\right)}_{=:k_1(t)} \quad \forall t \in [0,\ t_f]$$

with $k_1 \in L^1([0,\ t_f])$ because

$$\int_0^{t_f} \left|k_1(t)\right| \mathrm{d}t = \left(C_{0,K} + C_{1,K} \cdot \|\xi_0\|\right) \cdot \left(t_f + \|u\|_{L^1}\right) < \infty.$$

Similarly, we have

$$\left\|F(t,\xi) - F(t,\zeta)\right\| \leq \underbrace{C_{1,K} \cdot \left(1 + \left\|u(t)\right\|_1\right)}_{=:k_2(t)} \cdot \|\xi - \zeta\| \quad \forall \xi, \zeta \in \mathbb{R}^{n_x},\ t \in [0,\ t_f]$$

with $k_2 \in L^1([0,\ t_f])$ because

$$\int_0^{t_f} \left|k_2(t)\right| \mathrm{d}t = C_{1,K} \cdot \left(t_f + \|u\|_{L^1}\right) < \infty.$$

Therefore, F satisfies the Carathéodory conditions on $[0,\ t_f]$ (see [Hal69]). If we extend F to $(-1,\ t_f + 1) \times \mathbb{R}^{n_x}$ by setting $F(t,\xi) := 0$ for all $t \notin [0,\ t_f]$, then we can similarly extend k_1 and k_2 by zero without breaking the Carathéodory conditions or the bound conditions.

Because the extended F satisfies the Carathéodory conditions, the integrable bound condition with bound k_1 and the Lipschitz condition with k_2, the ODE $\dot{\xi}(t) = F\left(t, \xi(t)\right)$ has unique local solutions around every initial datum $\left(t_0,\ \xi(t_0)\right) \in (-1,\ t_f + 1) \times \mathbb{R}^{n_x}$, i.e., for every $(t_0,\ \zeta_0) \in (-1,\ t_f + 1) \times \mathbb{R}^{n_x}$, there exist an open interval $I_{u,t_0,\zeta_0} \subseteq (-1,\ t_f + 1)$ with $t_0 \in I_{u,t_0,\zeta_0}$ and an absolutely continuous function $\xi_{u,t_0,\zeta_0} : I_{u,t_0,\zeta_0} \to \mathbb{R}^{n_x}$ such that

$$\begin{aligned} \dot{\xi}_{u,t_0,\zeta_0}(t) &= F\left(t, \xi_{u,t_0,\zeta_0}(t)\right) \quad \text{for a.a. } t \in I_{u,t_0,\zeta_0}, \\ \xi_{u,t_0,\zeta_0}(t_0) &= \zeta_0. \end{aligned}$$

Let ξ_u be the extension of ξ_{u,t_f,ξ_0} to its maximal existence interval which we refer to as $I_u \subseteq (-1,\ t_f + 1)$. We certainly have $t_f \in I_u$. Let $a := \inf I_u \in [-1,\ t_f]$. For $t \to a$ in I_u, we have either $t \to -1$ or $\left\|\xi_u(t)\right\| \to \infty$. For each $t \in I_u$ with $t < t_f$, we

have

$$\begin{aligned}
\left\|\xi_u(t)-\xi_0\right\| &\leq \int_0^{t_f-t} \left\|\dot{\xi}_u(t_f-s)\right\| \mathrm{d}s \\
&= \int_0^{t_f-t} \left\|F\big(t_f-s,\xi_u(t_f-s)\big)\right\| \mathrm{d}s \\
&\leq \int_0^{t_f-t} C_{0,K}\cdot\Big(1+\left\|u(t_f-s)\right\|_1\Big) \\
&\qquad +C_{1,K}\cdot\Big(1+\left\|u(t_f-s)\right\|_1\Big)\cdot\Big(\|\xi_0\|+\left\|\xi_u(t_f-s)-\xi_0\right\|\Big)\mathrm{d}s \\
&= \big(C_{0,K}+C_{1,K}\cdot\|\xi_0\|\big)\cdot\left(t_f-t+\int_0^{t_f-t}\left\|u(t_f-s)\right\|_1\mathrm{d}s\right) \\
&\quad +\int_0^{t_f-t} C_{1,K}\cdot\Big(1+\left\|u(t_f-s)\right\|_1\Big)\cdot\left\|\xi_u(t_f-s)-\xi_0\right\|\mathrm{d}s.
\end{aligned}$$

By applying Grönwall's lemma, we obtain the estimate

$$\begin{aligned}
\left\|\xi_u(t)-\xi_0\right\| \leq\ & \big(C_{0,K}+C_{1,K}\cdot\|\xi_0\|\big)\cdot\left(t_f-t+\int_0^{t_f-t}\left\|u(t_f-s)\right\|_1\mathrm{d}s\right) \\
&\cdot\exp\left(C_{1,K}\cdot\left(t_f-t+\int_0^{t_f-t}\left\|u(t_f-s)\right\|_1\mathrm{d}s\right)\right) \\
&= \big(C_{0,K}+C_{1,K}\cdot\|\xi_0\|\big)\cdot\Big(t_f-t+\left\|u|_{[t,\ t_f]}\right\|_{L^1}\Big)\cdot e^{C_{1,K}\cdot(t_f-t+\|u|_{[t,\ t_f]}\|_{L^1})} \\
&\leq \underbrace{\big(C_{0,K}+C_{1,K}\cdot\|\xi_0\|\big)\cdot\big(t_f+\|u\|_{L^1}\big)\cdot e^{C_{1,K}\cdot(t_f+\|u\|_{L^1})}}_{=:C_{2,u}} \\
&< \infty.
\end{aligned}$$

Thus, ξ_u is confined to a closed ball of radius $C_{2,u}$ around x_0, which implies that $a=\inf I_u<0$ and that therefore ξ_u can be extended to the entire interval $[0,\ t_f]$.

Because $K\subseteq D$ is compact and because $\nabla_x\Phi$ is continuous, there exists a constant $C_{3,K}\geq 0$ such that

$$\left\|\nabla_x\Phi(x)\right\|\leq C_{3,K}\quad\forall x\in K.$$

We can use the fact that $x_u(t_f)\in K$ to argue that $\|\xi_0\|\leq C_{3,K}$. According to the definition of $\mathscr{B}_\delta^{n_w}(t_f)$, we have

$$\|u\|_{L^1}\leq\underbrace{n_w\cdot t_f+\delta}_{=:C_{4,\delta}}\quad\forall u\in\mathscr{B}_\delta^{n_w}(t_f).$$

For all $t\in[0,\ t_f]$, we have

$$\begin{aligned}
\left\|\xi_u(t)\right\| &\leq \left\|\xi_0\right\|+\left\|\xi_u(t)-\xi_0\right\| \\
&\leq C_{3,K}+\big(C_{0,K}+C_{1,K}\cdot\|\xi_0\|\big)\cdot\big(t_f+\|u\|_{L^1}\cdot e^{C_{1,K}\cdot(t_f+\|u\|_{L^1})}\big) \\
&\leq \underbrace{C_{3,K}+(C_{0,K}+C_{1,K}\cdot C_{3,K})\cdot\big(t_f+C_{4,\delta}\cdot\cdot e^{C_{1,K}\cdot(t_f+C_{4,\delta})}\big)}_{=:M_{\xi,\delta,K}}-
\end{aligned}$$

Part 3 (Uniqueness of ξ_u). As we have argued in prior proofs, uniqueness follows by contradiction from the local uniqueness of Carathéodory solutions around every point $\big(t,\ \xi_u(t)\big)$, which follows from the Lipschitz condition on F.□

We can now argue that the map $j\colon \mathscr{B}_\delta^{n_w}(t_f) \to \mathbb{R}$ from Proposition 2.4.21 is Fréchet differentiable by exploiting the fact that for the Lagrangian relaxation

$$\mathscr{L}(x,u,\xi) \coloneqq \Phi\big(x(t_f)\big) + \int_0^{t_f} l\big(x(t),u(t)\big) + \xi^T(t)\Big(f\big(x(t),u(t)\big) - \dot{x}(t)\Big)\,\mathrm{d}t$$

satisfies

$$\begin{aligned}\mathscr{L}(x_u,u,\xi) &= \Phi\big(x_u(t_f)\big) + \int_0^{t_f} l\big(x_u(t),u(t)\big) + \xi^T(t)\cdot\Big(\underbrace{f\big(x_u(t),u(t)\big) - \dot{x}_u(t)}_{=0}\Big)\,\mathrm{d}t\\ &= \Phi\big(x_u(t_f)\big) + \int_0^{t_f} l\big(x_u(t),u(t)\big)\,\mathrm{d}t\\ &= j(u)\end{aligned}$$

for all $u \in \mathscr{B}_\delta^{n_w}(t_f)$. This is independent of the particular choice of ξ. By applying partial integration to the last summand of the integrand in $L(x,u,\xi)$, we further obtain

$$\begin{aligned}\mathscr{L}(x,u,\xi) = \quad & \Phi\big(x(t_f)\big) + \xi^T(0)x(0) - \xi^T(t_f)x(t_f)\\ & + \int_0^{t_f}\Big(l\big(x(t),u(t)\big) + \xi^T(t)f\big(x(t),u(t)\big) + \dot{\xi}^T(t)x(t)\Big)\,\mathrm{d}t.\end{aligned}$$

We can use the special form of the adjoint state ξ_u to further simplify this expression.

Proposition 2.4.22 (Fréchet differentiability of ODE-based functions). *Let n_x, n_w, t_f, x_0, D, f, l, and Φ satisfy Assumptions 2.4.17 (1) to 2.4.17 (8). Let $\delta > 0$ be constant and let $K \subset D$ be compact and convex such that for each $u \in \mathscr{B}_\delta^{n_w}(t_f)$, there exists a unique absolutely continuous function $x_u\colon [0,\ t_f] \to K$ with*

$$\begin{aligned}&\dot{x}_u(t) = f\big(x_u(t),u(t)\big) \quad \text{for a.a. } t \in [0,\ t_f],\\ &x_u(0) = x_0.\end{aligned}$$

Then the functional $j\colon \mathscr{B}_\delta^{n_w}(t_f) \to \mathbb{R}$ with

$$j_u(t_f) \coloneqq \Phi\big(x_u(t_f)\big) + \int_0^{t_f} l\big(x_u(t),u(t)\big)\,\mathrm{d}t \quad \forall u \in \mathscr{B}_\delta^{n_w}(t_f)$$

is Lipschitz-continuously Fréchet differentiable on $\mathscr{B}_\delta^{n_w}(t_f)$. Its Fréchet derivative satisfies

$$Dj(u)d = \int_0^{t_f}\Big(\nabla_u l\big(x_u(t),u(t)\big) + \xi_u^T(t)\nabla_u f\big(x_u(t),u(t)\big)\Big)\cdot d(t)\,\mathrm{d}t$$

for all $d \in L^1\big([0,\ t_f],\mathbb{R}^{n_w}\big)$. ◁

PROOF. According to Lemma 2.4.20, there also exists a constant $L_{0,\delta,K} \geq 0$ such that

$$\big\|x_u(t) - x_v(t)\big\| \leq L_{0,\delta,K}\cdot\|u-v\|_{L^1} \quad \forall t \in [0,\ t_f],\ u,v \in \mathscr{B}_\delta^{n_w}(t_f).$$

Furthermore, since $K \subseteq D$ is compact and $\nabla_x f_i$ and $\nabla_x l_i$ are continuous for all $i \in [n_w]_0$, there exist constants $L_{1,K} \geq 0$ and $L_{2,K} \geq 0$ such that

$$\begin{aligned} \left\|\nabla_x l_i(x)\right\| &\leq L_{1,K} \quad \forall i \in [n_w]_0,\ x \in K, \\ \left\|\nabla_x f_i(x)\right\| &\leq L_{2,K} \quad \forall i \in [n_w]_0,\ x \in K. \end{aligned}$$

Assumptions 2.4.17 (3) and 2.4.17 (7) state that there are constants $L_{3,K} \geq 0$ and $L_{4,K} \geq 0$ such that

$$\begin{aligned} \left\|\nabla_x l_i(x) - \nabla_x l_i(y)\right\| &\leq L_{3,K} \cdot \|x-y\| \quad \forall i \in [n_w]_0,\ x,y \in K, \\ \left\|\nabla_x f_i(x) - \nabla_x f_i(y)\right\| &\leq L_{4,K} \cdot \|x-y\| \quad \forall i \in [n_w]_0,\ x,y \in K. \end{aligned}$$

For every $u \in \mathscr{B}_\delta^{n_w}(t_f)$, there exists $u' \in \mathscr{B}_0^{n_w}(t_f)$. Therefore, we have

$$\|u\|_{L^1} \leq \|u'\|_{L^1} + \|u-u'\|_{L^1} \leq n_w \cdot t_f + \delta.$$

Finally, according to Proposition 2.4.21, there exists a constant $M_{\xi,\delta,K}$ such that

$$\left\|\xi_u(t)\right\| \leq M_{\xi,\delta,K} \quad \forall u \in \mathscr{B}_\delta^{n_w}(t_f),\ t \in [0,\ t_f].$$

Part 1 (Initial Reformulation). Let $u, v \in B_\delta^{n_w}(t_f)$ be control functions. Let x_u, x_v be the absolutely continuous IVP solutions corresponding to u and v, respectively, and let ξ_v be the adjoint state function in v according to Proposition 2.4.21. The hypothesized expression for the Fréchet derivative yields the linearized change

$$Dj(v)(u-v) = \int_0^{t_f} \Big(\nabla_u l\big(x_v(t), v(t)\big) + \xi_v^T(t) \nabla_u f\big(x_v(t), v(t)\big)\Big)\big(u(t) - v(t)\big)\,\mathrm{d}t.$$

We have

$$\begin{aligned} & j(u) - j(v) - Dj(v)(u-v) \\ &\quad = \mathscr{L}(x_u, u, \xi_v) - \mathscr{L}(x_v, v, \xi_v) \\ &\qquad - \int_0^{t_f} \Big(\nabla_u l\big(x_v(t), v(t)\big) + \xi_v^T(t) \nabla_u f\big(x_v(t), v(t)\big)\Big)\big(u(t) - v(t)\big)\,\mathrm{d}t \\ &\quad = \Phi\big(x_u(t_f)\big) - \Phi\big(x_v(t_f)\big) + \big(\xi_v(0) - \xi_v(0)\big)^T x_0 - \xi_v^T(t_f)\big(x_u(t_f) - x_v(t_f)\big) \\ &\qquad + \int_0^{t_f} \Big(l\big(x_u(t), u(t)\big) + \xi_v^T(t) f\big(x_u(t), u(t)\big) + \dot{\xi}_v^T(t) x_u(t)\Big)\,\mathrm{d}t \\ &\qquad - \int_0^{t_f} \Big(l\big(x_v(t), v(t)\big) + \xi_v^T(t) f\big(x_v(t), v(t)\big) + \dot{\xi}_v^T(t) x_v(t)\Big)\,\mathrm{d}t \\ &\qquad - \int_0^{t_f} \Big(\nabla_u l\big(x_v(t), v(t)\big) + \xi_v^T(t) \nabla_u f\big(x_v(t), v(t)\big)\Big)\big(u(t) - v(t)\big)\,\mathrm{d}t \end{aligned}$$

$$
\begin{aligned}
&= \Phi\big(x_u(t_f)\big) - \Phi\big(x_v(t_f)\big) - \nabla_x \Phi\big(x_v(t_f)\big)\big(x_u(t_f) - x_v(t_f)\big) \\
&\quad + \int_0^{t_f} \Big(l\big(x_u(t), u(t)\big) - l\big(x_v(t), v(t)\big) \Big) \,\mathrm{d}t \\
&\quad + \int_0^{t_f} \xi_v^T(t) \Big(f\big(x_u(t), u(t)\big) - f\big(x_v(t), v(t)\big) \Big) \,\mathrm{d}t \\
&\quad - \int_0^{t_f} \nabla_x l\big(x_v(t), v(t)\big)\big(x_u(t) - x_v(t)\big) \,\mathrm{d}t \\
&\quad - \int_0^{t_f} \xi_v^T(t) \nabla_x f\big(x_v(t), v(t)\big)\big(x_u(t) - x_v(t)\big) \,\mathrm{d}t \\
&\quad - \int_0^{t_f} \Big(\nabla_u l\big(x_v(t), v(t)\big) + \xi_v^T(t) \nabla_u f\big(x_v(t), v(t)\big) \Big)\big(u(t) - v(t)\big) \,\mathrm{d}t \\
&= \Phi\big(x_u(t_f)\big) - \Phi\big(x_v(t_f)\big) - \nabla_x \Phi\big(x_v(t_f)\big)\big(x_u(t_f) - x_v(t_f)\big) \\
&\quad + \int_0^{t_f} \Big(\; l\big(x_u(t), u(t)\big) - l(x_v(t), v(t)) \\
&\qquad\qquad - \nabla_x l\big(x_v(t), v(t)\big)\big(x_u(t) - x_v(t)\big) \\
&\qquad\qquad - \nabla_u l\big(x_v(t), v(t)\big)\big(u(t) - v(t)\big) \;\Big) \,\mathrm{d}t \\
&\quad + \int_0^{t_f} \xi_v(t) \Big(\; f\big(x_u(t), u(t)\big) - f\big(x_v(t), v(t)\big) \\
&\qquad\qquad\quad - \nabla_x f\big(x_v(t), v(t)\big)\big(x_u(t) - x_v(t)\big) \\
&\qquad\qquad\quad - \nabla_u f\big(x_v(t), v(t)\big)\big(u(t) - v(t)\big) \;\Big) \,\mathrm{d}t.
\end{aligned}
$$

We now derive estimates for each of the three summands of the error term.

PART 2 (MAYER TERM ESTIMATE). For the first estimate, we use Taylor's theorem. Because Φ is differentiable, there exists a continuous residual function $R_\Phi : D \to \mathbb{R}$ with $R_\Phi(x) \to 0$ for $x \to x_v(t_f)$ and

$$
\Phi(x) - \Phi\big(x_v(t_f)\big) - \nabla_x \Phi\big(x_v(t_f)\big)\big(x - x_v(t_f)\big) = R_\Phi(x) \cdot \big\| x - x_v(t_f) \big\| \quad \forall x \in D.
$$

Thus, we have

$$
\begin{aligned}
&\Big| \Phi\big(x_u(t_f)\big) - \Phi\big(x_v(t_f)\big) - \nabla_x \Phi\big(x_v(t_f)\big)\big(x_u(t_f) - x_v(t_f)\big) \Big| \\
&\qquad = \Big| R_\Phi\big(x_u(t_f)\big) \Big| \cdot \big\| x_u(t_f) - x_v(t_f) \big\| \\
&\qquad \le \Big| R_\Phi\big(x_u(t_f)\big) \Big| \cdot L_{0,\delta,K} \cdot \| u - v \|_{L^1}.
\end{aligned}
$$

PART 3 (LAGRANGE TERM ESTIMATE). The second term takes the form

$$
\begin{aligned}
&\int_0^{t_f} \Big(\; l\big(x_u(t), u(t)\big) - l\big(x_v(t), v(t)\big) - \nabla_x l\big(x_v(t), v(t)\big)\big(x_u(t) - x_v(t)\big) \\
&\qquad - \nabla_u l\big(x_v(t), v(t)\big)\big(u(t) - v(t)\big) \Big) \,\mathrm{d}t.
\end{aligned}
$$

For every $t \in [0,\ t_f]$, we have

$$
\begin{aligned}
&l\big(x_u(t), u(t)\big) - l\big(x_v(t), v(t)\big) \\
&\qquad = l_0\big(x_u(t)\big) - l_0\big(x_v(t)\big) + \sum_{i=1}^{n_w} \Big(l_i\big(x_u(t)\big) u_i(t) - l_i\big(x_v(t)\big) v_i(t) \Big)
\end{aligned}
$$

$$= l_0\big(x_u(t)\big) - l_0\big(x_v(t)\big) + \sum_{i=1}^{n_w} \Big(l_i\big(x_u(t)\big) - l_i\big(x_v(t)\big)\Big) \cdot u_i(t) + \sum_{i=1}^{n_w} l_i\big(x_v(t)\big) \cdot \big(u_i(t) - v_i(t)\big).$$

For the derivative $\nabla_u l$, we have

$$\nabla_u l\big(x_v(t), v(t)\big)\big(u(t) - v(t)\big) = \sum_{i=1}^{n_w} l_i\big(x_v(t)\big)\big(u_i(t) - v_i(t)\big) \quad \forall t \in [0,\ t_f].$$

We can use the fact that $x_u(t) \in K$ and $x_v(t) \in K$ for all $t \in [0,\ t_f]$ and that K is convex. According to the fundamental theorem of calculus, we have

$$\begin{aligned}
&\Big| l_i\big(x_u(t)\big) - l_i\big(x_v(t)\big) - \nabla_x l_i\big(x_v(t)\big)\big(x_u(t) - x_v(t)\big)\Big| \\
&\quad = \left| \int_0^1 \Big(\nabla_x l_i\big((1-s)\cdot x_v(t) + s \cdot x_u(t)\big) - \nabla_x l_i\big(x_v(t)\big)\Big)\big(x_u(t) - x_v(t)\big)\, \mathrm{d}s \right| \\
&\quad \le \int_0^1 \Big\| \nabla_x l_i\big((1-s)\cdot x_v(t) + s \cdot x_u(t)\big) - \nabla_x l_i\big(x_v(t)\big)\Big\| \cdot \big\| x_u(t) - x_v(t)\big\|\, \mathrm{d}s \\
&\quad \le L_{3,K} \cdot \int_0^1 s \cdot \big\| x_u(t) - x_v(t)\big\|^2\, \mathrm{d}s \\
&\quad \le \frac{L_{3,K} \cdot L_{0,\delta,K}^2}{2} \cdot \|u - v\|_{L^1}^2
\end{aligned}$$

for all $i \in [n_w]_0$ and $t \in [0,\ t_f]$. In summary, we obtain the estimate

$$\begin{aligned}
&\left| \int_0^{t_f} \Big(\ l\big(x_u(t), u(t)\big) - l\big(x_v(t), v(t)\big) - \nabla_x l\big(x_v(t), v(t)\big)\big(x_u(t) - x_v(t)\big) \right. \\
&\qquad \left. - \nabla_u l\big(x_v(t), v(t)\big)\big(u(t) - v(t)\big)\Big)\, \mathrm{d}t \right| \\
&\le \int_0^{t_f} \Big| \ \ l_0\big(x_u(t)\big) - l_0\big(x_v(t)\big) - \nabla_x l_0\big(x_v(t)\big) \cdot \big(x_u(t) - x_v(t)\big) \\
&\qquad + \sum_{i=1}^{n_w} \Big(\Big(l_i\big(x_u(t)\big) - l_i\big(x_v(t)\big)\Big) \cdot u_i(t) - \nabla_x l_i\big(x_v(t)\big)\big(x_u(t) - x_v(t)\big) \cdot v_i(t)\Big) \Big|\, \mathrm{d}t \\
&\le \int_0^{t_f} \Big(\ \ \frac{L_{3,K} \cdot L_{0,\delta,K}^2}{2} \cdot \|u - v\|_{L^1}^2 \cdot \big(1 + \|u(t)\|_1\big) \\
&\qquad + \sum_{i=1}^{n_w} \Big\| \nabla_x l_i\big(x_v(t)\big)\Big\| \cdot \big\| x_u(t) - x_v(t)\big\| \cdot \big| u_i(t) - v_i(t)\big| \Big)\, \mathrm{d}t \\
&\le \int_0^{t_f} \Big(\ \ \frac{L_{3,K} \cdot L_{0,\delta,K}^2}{2} \cdot \|u - v\|_{L^1}^2 \cdot \big(1 + \|u(t)\|_1\big) \\
&\qquad + L_{1,K} \cdot L_{0,\delta,K} \cdot \|u - v\|_{L^1} \cdot \big\| u(t) - v(t)\big\|_1 \Big)\, \mathrm{d}t \\
&\le \left(\frac{L_{3,K} \cdot L_{0,\delta,K}^2}{2} \cdot \big(t_f + \|u\|_{L^1}\big) + L_{1,K} \cdot L_{0,\delta,K} \right) \cdot \|u - v\|_{L^1}^2 \\
&\le \underbrace{\left(\frac{L_{3,K} \cdot L_{0,\delta,K}^2}{2} \cdot \big((1 + n_w) \cdot t_f + \delta\big) + L_{1,K} \cdot L_{0,\delta,K} \right)}_{=: C_{l,\delta,K}} \cdot \|u - v\|_{L^1}^2.
\end{aligned}$$

PART 4 (MULTIPLIER TERM ESTIMATE). The estimate for the third term is derived very similarly to that for the second term. Here, we have

$$\begin{aligned}
&\Big\| f_i\big(x_u(t)\big) - f_i\big(x_v(t)\big) - \nabla_x f_i\big(x_v(t)\big)\big(x_u(t) - x_v(t)\big)\Big\| \\
&\quad = \left\| \int_0^1 \Big(\nabla_x f_i\big((1-s)\cdot x_v(t) + s\cdot x_u(t)\big) - \nabla_x f_i\big(x_v(t)\big)\Big)\cdot\big(x_u(t) - x_v(t)\big)\,\mathrm{d}s \right\| \\
&\quad \leq \int_0^1 \Big\|\nabla_x f_i\big((1-s)\cdot x_v(t) + s\cdot x_u(t)\big) - \nabla_x f_i\big(x_v(t)\big)\Big\| \cdot \big\|x_u(t) - x_v(t)\big\|\,\mathrm{d}s \\
&\quad \leq L_{4,K}\cdot \int_0^1 s\cdot \big\|x_u(t) - x_v(t)\big\|^2\,\mathrm{d}s \\
&\quad \leq \frac{L_{4,K}\cdot L_{0,\delta,K}^2}{2}\cdot \|u-v\|_{L^1}^2
\end{aligned}$$

for all $t \in [0,\ t_f]$. We therefore obtain the final estimate

$$\begin{aligned}
&\left| \int_0^{t_f} \xi_v(t)\cdot\Big(\ f\big(x_u(t),u(t)\big) - f\big(x_v(t),v(t)\big) - \nabla_x f\big(x_v(t),v(t)\big)\cdot\big(x_u(t)-x_v(t)\big) \right. \\
&\qquad\qquad \left. -\nabla_u f\big(x_v(t),v(t)\big)\big(u(t)-v(t)\big)\Big)\,\mathrm{d}t \right| \\
&\leq \int_0^{t_f} \big\|\xi_v(t)\big\| \cdot \Big\| \ f_0\big(x_u(t)\big) - f_0\big(x_v(t)\big) - \nabla_x f_0\big(x_v(t)\big)\cdot\big(x_u(t)-x_v(t)\big) \\
&\qquad\qquad + \sum_{i=1}^{n_w}\Big(\ \Big(f_i\big(x_u(t)\big) - f_i\big(x_v(t)\big)\Big)\cdot u_i(t) \\
&\qquad\qquad\qquad -\nabla_x f_i\big(x_v(t)\big)\big(x_u(t)-x_v(t)\big)\cdot v_i(t)\Big)\Big\|\,\mathrm{d}t \\
&\leq M_{\xi,\delta,K}\cdot \int_0^{t_f}\Big(\ \frac{L_{4,K}+L_{0,\delta,K}^2}{2}\cdot\Big(1+\big\|u(t)\big\|_1\Big)\cdot\|u-v\|_{L^1}^2 \\
&\qquad\qquad + \sum_{i=1}^{n_w}\Big\|\nabla_x f_i\big(x_v(t)\big)\Big\|\cdot\big\|x_u(t)-x_v(t)\big\|\cdot\big|u_i(t)-v_i(t)\big|\Big)\,\mathrm{d}t \\
&\leq M_{\xi,\delta,K}\cdot \int_0^{t_f}\Big(\ \frac{L_{4,K}+L_{0,\delta,K}^2}{2}\cdot\Big(1+\big\|u(t)\big\|_1\Big)\cdot\|u-v\|_{L^1}^2 \\
&\qquad\qquad + L_{2,K}\cdot L_{0,\delta,K}\cdot\|u-v\|_{L^1}\cdot\big\|u(t)-v(t)\big\|_1\Big)\,\mathrm{d}t \\
&= M_{\xi,\delta,K}\cdot\left(\frac{L_{4,K}+L_{0,\delta,K}^2}{2}\cdot\big(t_f+\|u\|_{L^1}\big)+L_{2,K}\cdot L_{0,\delta,K}\right)\cdot\|u-v\|_{L^1}^2 \\
&= \underbrace{M_{\xi,\delta,K}\cdot\left(\frac{L_{4,K}+L_{0,\delta,K}^2}{2}\cdot\big((1+n_w)\cdot t_f+\delta\big)+L_{2,K}\cdot L_{0,\delta,K}\right)}_{=:C_{f,\delta,K}}\cdot\|u-v\|_{L^1}^2.
\end{aligned}$$

PART 5 ($Dj(v)$ IS THE FRÉCHET DERIVATIVE OF j IN v). We can now combine the three estimates with our initial reformulation to obtain

$$\begin{aligned}
&\big|j(u)-j(v)-Dj(v)(u-v)\big| \\
&\quad \leq \Big|R_\Phi\big(x_u(t_f)\big)\Big|\cdot L_{0,\delta,K}\cdot\|u-v\|_{L^1} + (C_{l,\delta,K}+C_{f,\delta,K})\cdot\|u-v\|_{L^1}^2
\end{aligned}$$

$$= \Bigg(\underbrace{\Big| R_\Phi\big(x_u(t_f)\big) \Big| \cdot L_{0,\delta,K} + (C_{l,\delta,K} + C_{f,\delta,K}) \cdot \|u-v\|_{L^1}}_{\xrightarrow{u \to v} 0} \Bigg) \cdot \|u-v\|_{L^1}.$$

Thus, we have

$$j(u) = j(v) + Dj(v)(u-v) + o\big(\|u-v\|_{L^1}\big)$$

for all $u \in \mathscr{B}_\delta^{n_w}(t_f)$. In order to show that j is Fréchet differentiable around v, we still have to show that $d \mapsto Dj(v)d$ is a bounded linear operator. Linearity follows from the linearity of the integral. Boundedness follows because for all $d \in L^1\big([0,\ t_f], \mathbb{R}^{n_w}\big)$, we have

$$\begin{aligned} \big|Dj(v)d\big| &\leq \int_0^{t_f} \sum_{i=1}^{n_w} \big| l_i(x_v(t)) + \xi_v^T(t) f_i(x_v(t)) \big| \cdot \big| d_i(t) \big| \,\mathrm{d}t \\ &\leq (L_{1,K} + M_{\xi,\delta,K} \cdot L_{2,K}) \cdot \int_0^{t_f} \big\| d(t) \big\|_1 \,\mathrm{d}t \\ &\leq (L_{1,K} + M_{\xi,\delta,K} \cdot L_{2,K}) \cdot \|d\|_{L^1}. \end{aligned}$$

Therefore, $d \mapsto Dj(v)d$ is a bounded linear operator for all $v \in \mathscr{B}_\delta^{n_w}(t_f)$.

Part 6 ($u \mapsto Dj(u)$ is Lipschitz continuous). Let $u, v \in \mathscr{B}_\delta^{n_w}(t_f)$. For all $i \in [n_w]_0$ and $t \in [0,\ t_f]$, the fundamental theorem of analysis shows that

$$\begin{aligned} \Big| l_i\big(x_u(t)\big) - l_i\big(x_v(t)\big) \Big| &\leq \int_0^1 \Big\| \nabla_x l_i \big(\underbrace{s \cdot x_u(t) + (1-s) \cdot x_v(t)}_{\in K} \big) \Big\| \cdot \big\| x_u(t) - x_v(t) \big\| \,\mathrm{d}s \\ &\leq \frac{L_{1,K} \cdot L_{0,\delta,K}}{2} \cdot \|u-v\|_{L^1}, \end{aligned}$$

$$\begin{aligned} \Big\| f_i\big(x_u(t)\big) - f_i\big(x_v(t)\big) \Big\| &\leq \int_0^1 \Big\| \nabla_x f_i \big(s \cdot x_u(t) + (1-s) \cdot x_v(t) \big) \Big\| \cdot \big\| x_u(t) - x_v(t) \big\| \,\mathrm{d}s \\ &\leq \frac{L_{2,K} \cdot L_{0,\delta,K}}{2} \cdot \|u-v\|_{L^1}. \end{aligned}$$

For all $d \in L^1\big([0,\ t_f], \mathbb{R}^{n_w}\big)$, we therefore have

$$\begin{aligned} &\big| Dj(u)d - Dj(v)d \big| \\ &\quad \leq \int_0^{t_f} \sum_{i=1}^{n_w} \Big(\Big| l_i\big(x_u(t)\big) - l_i\big(x_v(t)\big) \Big| + \big\| \xi_v(t) \big\| \cdot \Big\| f_i\big(x_u(t)\big) - f_i\big(x_v(t)\big) \Big\| \Big) \cdot |d_i| \,\mathrm{d}t \\ &\quad \leq (L_{1,K} + M_{\xi,\delta,K} \cdot L_{2,K}) \cdot L_{0,\delta,K} \cdot \|u-v\|_{L^1} \cdot \|d\|_{L^1}. \end{aligned}$$

Therefore, the operator norm of $Dj(u) - Dj(v)$ satisfies

$$\big\| Dj(u) - Dj(v) \big\| \leq (L_{1,K} + M_{\xi,\delta,K} \cdot L_{2,K}) \cdot L_{0,\delta,K} \cdot \|u-v\|_{L^1} \quad \forall u, v \in \mathscr{B}_\delta^{n_w}(t_f).$$

This means that the Fréchet derivative of j is Lipschitz continuous on $\mathscr{B}_\delta^{n_w}(t_f)$.□

Having proven that Bolza-type functionals of nearly [0, 1]-valued control functions are Fréchet differentiable, we can use the general Banach space theory derived in Section 2.4.2. We use the canonical vector measure described in Theorem 2.4.14 which maps a measurable set to its characteristic function.

Theorem 2.4.23 (Set Differentiability of ODE-based Functions).
Let n_x, n_w, t_f, x_0, D, f, l, Φ, and m satisfy Assumptions 2.4.17 (1) to 2.4.17 (9). For each $i \in [n_w]$, let $\big([0,\ t_f],\ \mathscr{B}([0,\ t_f]),\ \mu_i\big)$ be the measure space of Lebesgue measurable subsets of $[0,\ t_f]$ equipped with the measure $\mu_i : \mathscr{B}([0,\ t_f]) \to \mathbb{R}_{\geq 0}$ with

$$\mu_i(A) \coloneqq \int_A m_i \,\mathrm{d}\lambda < \infty \quad \forall A \in \mathscr{B}([0,\ t_f]).$$

Let $(X,\ \Sigma,\ \mu)$ be the layered measure space whose layers are the original measure spaces $\big([0,\ t_f],\ \mathscr{B}([0,\ t_f]),\ \mu_i\big)$ for $i \in [n_w]$. For each $U \in \Sigma/_{\sim_\mu}$, let $x_U : [0,\ t_f] \to D$ and $\xi_U : [0,\ t_f] \to \mathbb{R}^{n_x}$ be the unique solutions to the boundary value problems

$$\begin{aligned}
&\dot{x}_U(t) = f\big(x_U(t),\ \chi_U(t)\big) && \text{for a.a. } t \in [0,\ t_f],\\
&x_U(0) = x_0,\\
&\dot{\xi}_U(t) = -\Big(\nabla_x l\big(x_U(t),\ \chi_U(t)\big)\Big)^T - \Big(\nabla_x f\big(x_U(t),\ \chi_{U_i}(t)\big)\Big)^T \xi_U(t) && \text{for a.a. } t \in [0,\ t_f],\\
&\xi_U(t_f) = \Big(\nabla_x \Phi\big(x_U(t_f)\big)\Big)^T,
\end{aligned}$$

where $\chi_U \coloneqq \big(\chi_{U_i}\big)_{i \in [n_w]}$ for all $U \in \Sigma/_{\sim_\mu}$ and U_i is the i-th layer of U. Then $J : \Sigma/_{\sim_\mu} \to \mathbb{R}$ with

$$J(U) \coloneqq \Phi\big(x_U(t_f)\big) + \int_0^{t_f} l\big(x_U(t),\ \chi_U(t)\big)\big)\,\mathrm{d}t \quad \forall U \in \Sigma/_{\sim_\mu}$$

is Lipschitz-continuously differentiable and satisfies

$$\nabla J(U)(W) = \sum_{i=1}^{n_w} \int_{W_i} \frac{1 - 2\chi_{U_i}(t)}{m_i(t)} \cdot \Big(l_i\big(x_U(t)\big) + \xi_U^T(t) f_i\big(x_U(t)\big)\Big)\,\mathrm{d}\mu_i \quad \forall U, W \in \Sigma/_{\sim_\mu}.$$

◁

PROOF. PART 1 (EXISTENCE AND UNIQUENESS OF x_U, ξ_U). According to Theorem 2.4.19, there exist a constant $\delta > 0$ and a compact set $K \subseteq D$ such that for every nearly $[0,\ 1]$-valued control function $u \in \mathscr{B}_\delta^{n_w}(t_f)$, there exists a unique absolutely continuous function $x_u : [0,\ t_f] \to \mathbb{R}^{n_x}$ with

$$\begin{aligned}
&\dot{x}_u(t) = f\big(x_u(t), u(t)\big) \quad \text{for a.a. } t \in [0,\ t_f],\\
&x_u(0) = x_0.
\end{aligned}$$

According to Proposition 2.4.21, there then also exists a unique absolutely continuous function $\xi_u : [0,\ t_f] \to \mathbb{R}^{n_x}$ with

$$\begin{aligned}
&\dot{\xi}_u(t) = -\Big(\nabla_x l\big(x_u(t),\ u(t)\big)\Big)^T - \Big(\nabla_x f\big(x_u(t),\ u(t)\big)\Big)^T \xi_u(t) \quad \text{for a.a. } t \in [0,\ t_f],\\
&\xi_u(t_f) = \Big(\nabla_x \Phi\big(x_u(t_f)\big)\Big)^T.
\end{aligned}$$

This specifically applies for binary-valued controls $u = \chi_U \in \mathscr{B}^{n_w}(t_f)$ and for a δ environment around each χ_U.

PART 2 (DIFFERENTIABILITY OF J). As shown in Proposition 2.4.22, the functional $j : \mathscr{B}_\delta^{n_w}(t_f) \to \mathbb{R}$ with

$$j(u) \coloneqq \Phi\big(x_u(t_f)\big) + \int_0^{t_f} l\big(x_u(t),\ u(t)\big)\,\mathrm{d}t \quad \forall u \in \mathscr{B}_\delta^{n_w}(t_f)$$

is Lipschitz-continuously Fréchet differentiable and its derivative satisfies

$$Dj(u)d = \int_0^{t_f} \Big(\nabla_u l\big(x_u(t), u(t)\big) + \xi_u^T(t) \nabla_u f\big(x_u(t), u(t)\big)\Big) d(t)\,\mathrm{d}t$$

for all $u \in \mathscr{B}_\delta^{n_w}(t_f)$ and $d \in L_\lambda^1\big([0,\ t_f],\ \mathbb{R}^{n_w}\big)$.

We now verify Assumption 2.4.10. To do so, we have to bridge the gap between the vector measures $\nu_i \colon \mathscr{B}\big([0,\ t_f]\big) \to L_\lambda^1\big([0,\ t_f]\big)$ with $\nu_i(U_i) = \chi_{U_i}$ that operate on individual layers of the layered space, and the composite vector measure $\nu \colon \Sigma \to L_\lambda^1\big([0,\ t_f],\ \mathbb{R}^{n_w}\big)$ with $\nu(U) = \big(\nu_i(U_i)\big)_{i \in [n_w]}$.

For an individual layer with index $i \in [n_w]$, Theorem 2.4.14 and Proposition 2.4.16 prove that ν_i is a vector measure of bounded variation with $|\nu_i|(V) \leq C_i \cdot \mu_i$. The fact that ν is a vector measure follows from the elementwise measure properties of the individual ν_i. For every $U \in \Sigma$, we have

$$\begin{aligned}
\big\|\nu(U)\big\| &= \Big\| \Big(\big\|\nu_i(U_i)\big\|_{L_\lambda^1([0,\ t_f])} \Big)_{i \in [n_w]} \Big\|_1 \\
&= \sum_{i=1}^{n_w} \big\|\nu_i(U_i)\big\|_{L_\lambda^1([0,\ t_f])} \\
&\leq \sum_{i=1}^{n_w} C_i \cdot \mu_i(U_i) \\
&\leq \underbrace{\big(\max_{i \in [n_w]} C_i\big)}_{=:C} \cdot \sum_{i=1}^{n_w} \mu_i(U_i) \\
&= C \cdot \mu(U).
\end{aligned}$$

Here, the fact that all norms are equivalent on $\mathbb{R}^{n_w}$ makes the particular choice of the outer norm irrelevant. For every partition $(U^{(i)})_{i \in \mathbb{N}} \in \Sigma^{\mathbb{N}}$ of U, we also obtain the estimate

$$\begin{aligned}
\sum_{i=1}^{\infty} \big\|\nu(U^{(i)})\big\| &\leq \sum_{i=1}^{\infty} C \cdot \mu(U^{(i)}) \\
&= C \cdot \mu\Big(\bigcup_{i=1}^{\infty} U^{(i)}\Big) \\
&= C \cdot \mu(U),
\end{aligned}$$

which implies that

$$|\nu|(U) \leq C \cdot \mu(U) \quad \forall U \in \Sigma.$$

Because $\mu(X) < \infty$, this implies that ν is of bounded variation. In conjunction with Theorem 2.1.17 and the fact that $Y := L^1\big([0,\ t_f],\ \mathbb{R}^{n_w}\big)$ is a Banach space, this demonstrates Assumptions 2.4.10 (1) to 2.4.10 (4). Assumption 2.4.10 (5) follows from Proposition 2.4.22.

We can now apply Theorem 2.4.13 to show that J is Lipschitz-continuously differentiable because $J(U) = j\big(\nu(U)\big)$ for all $U^{\Sigma/\sim_\mu}$. According to Proposition 2.4.12, the derivative of J in $U \in {}^{\Sigma/\sim_\mu}$ takes the form

$$\begin{aligned}
&\nabla F(U)(D) \\
&\quad = Dj\big(\nu(U)\big)\big(\nu(D \setminus U) - \nu(D \cap U)\big)
\end{aligned}$$

$$
\begin{aligned}
&= \int_0^{t_f} \Big(\nabla_u l\big(x_U(t), \chi_U(t)\big) + \big(\xi_U(t)\big)^\mathsf{T} \nabla_u f\big(x_U(t), \chi_U(t)\big) \Big) \cdot \big(\nu(D \setminus U) - \nu(D \cap U) \big) \, \mathrm{d}t \\
&= \sum_{i=1}^{n_w} \left(\int_{D_i \setminus U_i} \big(l_i(x_U) + \xi_U^\mathsf{T} f_i(x_U) \big) \, \mathrm{d}\lambda - \int_{D_i \cap U_i} \big(l_i(x_U) + \xi_U^\mathsf{T} f_i(x_U) \big) \, \mathrm{d}\lambda \right) \\
&= \sum_{i=1}^{n_w} \int_{D_i} (1 - 2\chi_{U_i}) \cdot \big(l_i(x_U) + \xi_U^\mathsf{T} f_i(x_U) \big) \, \mathrm{d}\lambda \\
&= \sum_{i=1}^{n_w} \int_{D_i} \frac{1 - 2\chi_{U_i}}{m_i} \cdot \big(l_i(x_U) + \xi_U^\mathsf{T} f_i(x_U) \big) \, \mathrm{d}\mu_i
\end{aligned}
$$

for all $D \in \Sigma/_{\sim_\mu}$. We note that we can safely divide by m here because Definition 2.4.15 demands that m be bounded away from zero. We also make use of the fact that

$$
1 - 2\chi_{U_i}(t) = \begin{cases} 1 & \text{if } t \notin U_i, \\ -1 & \text{if } t \in U_i \end{cases}
$$

which gives us a very convenient way to encode the change in sign between the two summand integrals. □

Of course, the explicit form of $\nabla F(U)$ written as a sum of integrals in individual layers of the layered space that we have derived in Theorem 2.4.23 makes it very convenient to determine the gradient density function in this case. According to Theorem 2.1.20, the gradient density function is simply the layered function made up of integrands of the integral and therefore takes the form

$$
g(U)(i,t) = \frac{1 - 2\chi_{U_i}(t)}{m_i(t)} \cdot \Big(l_i\big(x_U(t)\big) + \big(\xi_U(t)\big)^\mathsf{T} f_i\big(x_U(t)\big) \Big) \quad \forall t \in [0,\ t_f]
$$

for all $i \in [n_w]$ and $U \in \Sigma/_{\sim_\mu}$.

Using this explicit form of the gradient density functions, we can see an interesting parallel between the necessary optimality criterion we state in Proposition 2.4.7 and Pontryagin's minimum principle. The necessary optimality criterion is satisfied if and only if

$$
\int_0^{t_f} \min\big\{ g(U)(i,t),\ 0 \big\} \, \mathrm{d}\mu_i(t) = 0 \quad \forall i \in [n_w],
$$

which is equivalent to $g(U)(i,t) \geq 0$ for almost all $t \in [0,\ t_f]$. If we consider the simplified case that $m_i(t) = 1$ for all i and t, then $g(U)(i,t) \geq 0$ is equivalent to

$$
\begin{aligned}
l_i\big(x_U(t)\big) + \big(\xi_U(t)\big)^\mathsf{T} f_i\big(x_U(t)\big) &\geq 0 \quad \text{for a.a. } t \in [0,\ t_f] \setminus U_i, \\
l_i\big(x_U(t)\big) + \big(\xi_U(t)\big)^\mathsf{T} f_i\big(x_U(t)\big) &\leq 0 \quad \text{for a.a. } t \in U_i.
\end{aligned}
$$

If we take into account that $l_i\big(x_U(t)\big) + \big(\xi_U(t)\big)^\mathsf{T} f_i(x_U(t))$ is the derivative of the Hamiltonian function with respect to the i-th component of the control function which is at its upper bound 1 for all points in U_i and at its lower bound 0 for all points outside of U_i, we can see that this is equivalent to the statement that the Hamiltonian function, which is differentiable in this setting, cannot be improved locally by changing the control function. In conjunction with the choice of ξ_U, this means that the point $(x_U,\ \chi_U,\ \xi_U)$ would satisfy Pontryagin's minimum principle. A detailed discussion of Pontryagin's minimum priciple is beyond the scope of

this thesis. Such discussions can be found in most basic texts on the foundations of optimal control theory (see, e.g., [Ber74; MZ21]).

Intuitively, directly solving for a stationary point with this type of gradient density can be thought of as partitioning the time horizon into separate regions and picking the control configuration that minimizes the Hamiltonian for each region separately, though this intuition is somewhat complicated by nonlinearities in the set functional. Similar approaches for mixed-integer optimal control have previously been proposed under the name "competing Hamiltonians" [BL85; Boc+17].

2.4.4 Derivation: PDE Case

The differentiation method described in this section was first developed for a slightly simplified setting in [HLS22, Sec. 3.3.2].

In this section, we discuss in a very generalized way how to obtain derivatives in a PDE setting. PDEs require much more problem-specific treatment than ODEs. Therefore, we will not be able to state an explicit term for the gradient density function. However, there are certain recurring steps in PDE solution that allow for some general statements.

PDEs are often solved in a weak form. This means that rather than solving for a solution to the original PDE, we solve for a function that behaves like a solution with respect to a selection of linear *test functionals*. Any strong solution to the original PDE is then also a weak solution. If a weak solution can be found and shown to be unique, a strong solution, if it exists, must be equal to it by whatever concept of equality applies in the search space.

It is important to stress the "concept of equality" aspect of this statement because weak solutions are usually determined within spaces of integrable functions and Sobolev spaces. Equality in these spaces usually allows for differences on nullsets. There are, however, sometimes embedding results that show that weak solutions can be made into strong solutions by adjusting them on nullsets.

Working with weak formulations in Banach function spaces is convenient because it effectively turns the PDE into an equation system in a Banach space. The function spaces in question are then generally approximated using subspaces of finite dimension. This process usually involves subdividing the problem domain into a mesh. To make more concrete statements, we formulate the set of assumptions for this setting.

We once more introduce two categories of control functions.

Definition 2.4.24 (Control Function Classes).
Let $d \in \mathbb{N}$, let $n_w \in \mathbb{N}$, and let $\Omega \subseteq \mathbb{R}^d$ be Lebesgue measurable with $\lambda(\Omega) < \infty$. We refer to

$$\mathcal{W}^{n_w}(\Omega) \coloneqq L^\infty_\lambda\big(\Omega,\, [0,\, 1]^{n_w}\big),$$
$$\mathcal{W}^{n_w}_\delta(\Omega) \coloneqq \big\{w \in L^\infty_\lambda(\Omega,\, \mathbb{R}^{n_w}) \,\big|\, \exists v \in \mathcal{W}^{n_w}(\Omega)\colon \|w - v\|_{L^\infty} < \delta\big\} \quad \forall \delta > 0.$$

as the set of *relaxed control functions* and its *δ-neighborhood*, respectively. ◁

Because $\lambda(\Omega) < \infty$, these control functions are always integrable. The fact that they are essentially bounded is significant because for some PDE-constrained optimization problems, differentiability can only be proven within L^∞ neighborhoods to maintain essential separation of coefficients from values at which

the PDE solution ceases to exist or be unique. We discuss one such problem in Section 4.2.

We ultimately derive these control functions from similarity classes by using the indicator function mapping as a vector measure. For $n_w > 1$, the indicator function is an integrable function in a layered measure space where each layer acts as one component of the control function.

The problem with determining Fréchet derivatives with L^∞ as codomain is that the L^∞ norm of the difference between two indicator functions does not gradually decrease to zero as the two indicator functions grow "closer." The difference between the two indicator functions must be measured using the L^1 norm to accurately reflect the distance between the two similarity classes that are being represented. To accommodate this, we have to incorporate a special condition into our preconditions that ensures that the derivative that is originally determined for infinitesimal L^∞ perturbations also functions as a derivative with respect to L^1 perturbations as long as the start and end points of the perturbation are both in $\mathcal{W}^{n_w}(\Omega)$.

The advantage of this two-stage approach is that we can use the stronger guarantees of an L^∞ perturbation to prove the existence of the derivative and can then use known properties of the PDE solution, such as regularity results and maximum principles, to retrospectively broaden error estimates to apply to L^1 perturbations.

Assumption 2.4.25.
Let $d \in \mathbb{N}$, $n_w \in \mathbb{N}$, Ω, Y, Z, G, $\delta > 0$, $f : G \times \mathcal{W}_\delta^{n_w}(\Omega) \to Z$, $j : G \times \mathcal{W}_\delta^{n_w}(\Omega) \to \mathbb{R}$ satisfy the following assumptions:

(1) $\Omega \subseteq \mathbb{R}^d$ is Lebesgue measurable with $\lambda(\Omega) < \infty$;

(2) Y and Z are Banach spaces;

(3) $G \subseteq Y$ is open;

(4) f is continuously F-differentiable on $G \times \mathcal{W}_\delta^{n_w}(\Omega)$;

(5) j is continuously F-differentiable on $G \times \mathcal{W}_\delta^{n_w}(\Omega)$;

(6) for $(x,\ w) \in G \times \mathcal{W}_\delta^{n_w}(\Omega)$, $D_x f(x,w) \in \mathscr{L}(Y,\ Z)$ has a bounded inverse;

(7) for each $w \in \mathcal{W}_\delta^{n_w}(\Omega)$, there is exactly one $x_w \in G$ with $f(x_w, w) = 0$;

(8) for $v \in \mathcal{W}^{n_w}(\Omega)$ with $v(x) \in \{0,\ 1\}^{n_w}$ almost everywhere and $x_v \in Y$ with $f(x_v, v) = 0$, there exists a constant $L \geq 0$ such that the linear form T_v with

$$T_v := -D_x j(x_v, v) \circ \big(D_x f(x_v, v)\big)^{-1} \circ D_w f(x_v, v)(w - v) + D_w j(x_v, v)$$

satisfies

$$\big|T(w - v)\big| \leq L \cdot \|w - v\|_{L^1}$$

as well as

$$j(x_w, w) - j(x_v, v) - T_v(w - v) = o\big(\|w - v\|_{L^1}\big)$$

for all $w \in \mathcal{W}^{n_w}(\Omega)$ with $w(x) \in \{0,\ 1\}^{n_w}$ almost everywhere and $x_w \in Y$ with $f(x_w, w) = 0$.

Let $m : \Omega \to \mathbb{R}^{n_w}$ satisfy the following assumption:

(9) for each $i \in [n_w]$, $m_i \in L^\infty_\lambda(\Omega)$ is a scaling density function of λ.

Furthermore, let $(n_i)_{i\in\mathbb{N}_0} \subseteq \mathbb{N}$ and $(T_{i,j})_{i\in\mathbb{N}_0,\, j\in[n_i]} \subseteq \mathscr{B}(\Omega)$ satisfy the following assumptions:

(10) $\lambda(T_{i,j}) > 0 \;\forall i \in \mathbb{N}_0,\; j \in [n_i]$;

(11) for each $i \in \mathbb{N}_0$, $(T_{i,j})_{j\in[n_i]}$ is a partition of Ω;

(12) $\max_{j\in[n_i]} \lambda(T_{i,j}) \xrightarrow{i\to\infty} 0$;

(13) $\forall i > 0,\; j \in [n_i]\; \exists j' \in [n_{i-1}]\colon T_{i,j} \subseteq T_{i-1,\,j'}$;

(14) there exists $C > 0$ such that for all $i \in \mathbb{N}_0$ and $j \in [n_i]$, there exists a ball $B_{i,j}$ such that $T_{i,j} \subseteq B_{i,j}$ and $\lambda(T_{i,j}) \geq C \cdot \lambda(B_{i,j})$. ◁

Assumption 2.4.25 (8) appears at first glance to be a very strong assumption. However, we will see that the existence of T_v as an F-derivative of the mapping $w \mapsto j(x_w, w)$ from $L^\infty_\lambda(\Omega, \mathbb{R}^{n_w})$ to $\mathbb{R}$ follows from Assumptions 2.4.25 (1) to 2.4.25 (7). Therefore, Assumption 2.4.25 (8) is simply an assumption that both the derivative and the residual are bounded with respect to the L^1 as well as the L^∞ norm. We note that Assumption 2.4.25 (8) can be inferred from boundedness assumptions on the derivatives of f and j with respect to the control function w by replicating the proof of the Implicit Function Theorem (as stated, e.g., in [Pat18, Thm. 3.13]) with restricted controls. However, the assumption as we have stated it here gives us additional flexibility. We could, for instance, use the fact that actual PDE solutions x_w lie within a more restricted solution space $Y' \subset Y$ for binary-valued control functions w to prove L^1 boundedness.

Let subsequently $(X,\, \Sigma,\, \mu)$ be the layered measure space obtained by layering $\mathscr{B}(\Omega)$ with the Lebesgue measure for each control component:

$$\begin{aligned} X &\coloneqq \bigcup_{i=1}^{n_w} \big(\{i\} \times \Omega\big), \\ \Sigma &\coloneqq \big\{U \subseteq X \,\big|\, U_i \coloneqq \{x \in \Omega \,|\, (i,\, x) \in U\} \in \mathscr{B}(\Omega)\;\forall i \in [n_w]\big\}, \\ \mu(U) &\coloneqq \sum_{i=1}^{n_w} \lambda(U_i) \qquad\qquad \forall U \in \Sigma. \end{aligned}$$

At first glance, this may appear problematic because the spaces $L^p(\Sigma, \mu)$ and $L^p(\mathscr{B}(\Omega), \lambda, \mathbb{R}^{n_w})$ are not the same. The map $T\colon L^p(\Sigma, \mu) \to L^p(\mathscr{B}(\Omega), \lambda, \mathbb{R}^{n_w})$ with

$$T(f) \coloneqq x \mapsto \big(f(i,x)\big)_{i\in[n]} \quad \forall f \in L^p(\Sigma, \mu)$$

is straightforwardly an isomorphism with

$$T^{-1}(f) = (i,x) \mapsto f_i(x).$$

This isomorphism preserves integrals in the sense that

$$\begin{aligned} \int_X f(i,x)\,\mathrm{d}\mu(i,x) &= \sum_{i=1}^{n_w} \int_\Omega f(i,x)\,\mathrm{d}\lambda(x) \\ &= \int_\Omega \sum_{i=1}^{n} \big(T(f)\big)_i(x)\,\mathrm{d}\lambda(x) \end{aligned}$$

and L^p norms in the sense that

$$\begin{aligned}
\|f\|_{L^p(\Sigma,\mu)} &= \left(\int_X \left|f(i,x)\right|^p \mathrm{d}\mu(i,x)\right)^{\frac{1}{p}} \\
&= \left(\int_X \left|f(i,x)\right|^p \mathrm{d}\mu(i,x)\right)^{\frac{1}{p}} \\
&= \left(\int_\Omega \sum_{i=1}^{n_w} \left(T\left(|f|^p\right)\right)_i (x)\,\mathrm{d}\lambda(x)\right)^{\frac{1}{p}} \\
&= \left(\int_\Omega \sum_{i=1}^{n_w} \left|\left(T(f)\right)_i (x)\right|^p \mathrm{d}\lambda(x)\right)^{\frac{1}{p}} \\
&= \left(\int_\Omega \left\|T(f)(x)\right\|_p^p \,\mathrm{d}\lambda(x)\right)^{\frac{1}{p}} \\
&= \left\|T(f)\right\|_{L^p(\mathscr{B}(\Omega),\,\lambda,\,\mathbb{R}^{n_w})}
\end{aligned}$$

for $1 \leq p < \infty$ and

$$\begin{aligned}
\|f\|_{L^p(\Sigma,\mu)} &= \operatorname*{ess\,sup}_{(i,x)\in X} \left|f(i,x)\right| \\
&= \inf\left\{a \in \mathbb{R} \;\middle|\; \mu\left(|f|^{-1}\left((a,\,\infty)\right)\right) = 0\right\} \\
&= \inf\left\{a \in \mathbb{R} \;\middle|\; \sum_{i=1}^{n_w} \lambda\left(|f|^{-1}\left((a,\,\infty)\right) \cap \left(\{i\} \times \Omega\right)\right) = 0\right\} \\
&= \inf\left\{a \in \mathbb{R} \;\middle|\; \lambda\left(|f|^{-1}\left((a,\,\infty)\right) \cap \left(\{i\} \times \Omega\right)\right) = 0 \;\forall i \in [n_w]\right\} \\
&= \inf\left\{a \in \mathbb{R} \;\middle|\; \lambda\left(\left|\left(T(f)\right)_i\right|^{-1}\left((a,\,\infty)\right)\right) = 0 \;\forall i \in [n_w]\right\} \\
&= \inf\left\{a \in \mathbb{R} \;\middle|\; \lambda\left(\bigcup_{i=0}^{n_w} \left|\left(T(f)\right)_i\right|^{-1}\left((a,\,\infty)\right)\right) = 0\right\} \\
&= \inf\left\{a \in \mathbb{R} \;\middle|\; \lambda\left(\left\|T(f)\right\|_\infty^{-1}\left((a,\,\infty)\right)\right) = 0\right\} \\
&= \operatorname*{ess\,sup}_{x\in\Omega} \left\|T(f)(x)\right\|_\infty \\
&= \left\|T(f)\right\|_{L^p(\mathscr{B}(\Omega),\,\lambda,\,\mathbb{R}^{n_w})}
\end{aligned}$$

for $p = \infty$. Thus, $L^p(\Sigma,\mu)$ isometrically embeds into $L^p\left(\mathscr{B}(\Omega),\lambda,\mathbb{R}^{n_w}\right)$ for all exponents p with $1 \leq p \leq \infty$. We reiterate that the distinction between $\mathscr{B}(\Omega)$ and $\mathscr{L}(\Omega)$ is largely irrelevant for our purposes because the Borel- and Lebesgue-σ-algebras only differ in nullsets, which have no effect on differentiable objective functionals.

We consider the functional $J : \Sigma/_{\sim_\mu} \to \mathbb{R}$ with

$$J(U) := j(x_U, \chi_U) \qquad \forall U \in \Sigma/_{\sim_\mu}$$

where $x_U \in Y$ is the unique solution to the equation

$$f(x_U, \chi_U) = 0.$$

We can break Assumption 2.4.25 up into two parts. Assumptions 2.4.25 (1) to 2.4.25 (8) ensure the differentiability of $w \mapsto x_w$ and imply the set differentiability of the functional J. Assumption 2.4.25 (9) allows us to scale the variables. Assumptions 2.4.25 (10) to 2.4.25 (14) are not necessary to prove the set differentiability of J. Instead, they describe a sequence of increasingly refined meshes which we can use to approximate the gradient density function.

Proposition 2.4.26.
Let d, n_w, Ω, Y, Z, G, δ, f, j, and m satisfy Assumptions 2.4.25 (1) to 2.4.25 (8), then $J\colon \Sigma/_{\sim_\mu} \to \mathbb{R}$ with $J(U) \coloneqq j(x_U, \chi_U)$ is a differentiable set functional and the derivative of J has the form

$$\begin{aligned}\nabla J(U)(D) = &-D_x j(x_U, \chi_U)\big(D_x f(x_U, \chi_U)\big)^{-1} D_w f(x_U, \chi_U)(\chi_{D\setminus U} - \chi_{D\cap U}) \\ &+ D_w j(x_U, \chi_U)(\chi_{D\setminus U} - \chi_{D\cap U})\end{aligned}$$

for all $D \in \Sigma/_{\sim_\mu}$. ◁

Proof. Part 1 (F-differentiability of $w \mapsto x_w$). In the first step, we demonstrate that $w \mapsto x_w$ as a mapping from $\mathcal{W}_\delta^{n_w}(\Omega) \subset L^\infty(\mathcal{B}(\Omega),\ \lambda,\ \mathbb{R}^{n_w})$ to $G \subset X$ is Fréchet differentiable. This is relatively easy to do using the Implicit Function Theorem (see, e.g., [Hin+09, Thm. 1.41] or [Pat18, Thm. 3.13])[1]. By Assumption 2.4.25 (7), x_w exists and is a unique element of G for all $w \in \mathcal{W}_\delta^{n_w}(\Omega)$. Thus

$$\mathcal{U} \coloneqq G \times \mathcal{W}_\delta^{n_w}(\Omega)$$

is a neighborhood of $(x_w,\ w)$ for all w. By Assumptions 2.4.25 (4) and 2.4.25 (6), f is continuously F-differentiable around $(x_w,\ w)$ and $D_y F(x_w, w)$ exists and is a bounded isomorphism. According to the Implicit Function Theorem, this establishes the uniqueness and continuous differentiability of $w \mapsto x_w$ in a smaller neighborhood around $(x_w,\ w)$ and

$$D_w(w \mapsto x_w)(w) = -\big(D_x f(x_w, w)\big)^{-1} \circ D_w f(x_w, w).$$

Global uniqueness and continuous differentiability then follows from the fact that $\mathcal{W}_\delta^{n_w}(\Omega)$ is pathwise connected, that every path through it can be covered with a finite number of overlapping neighborhoods obtained from the Implicit Function Theorem, and that x_w is the unique solution of $f(x, w) = 0$ because of Assumption 2.4.25 (7). We note that without the uniqueness assumption, the Implicit Function Theorem would allow for multiple sets of solutions as long as they are strictly separated from one another.

Part 2 (Composition with j). Because j is continuously F-differentiable, we can apply the chain rule to show that $w \mapsto j(x_w, w)$ is continuously F-differentiable with

$$\begin{aligned}&D\big(w \mapsto j(x_w, w)\big)(d_w) \\ &\quad = Dj(x_w, w)\big(D(w \mapsto x_w)(d_w),\ d_w\big) \\ &\quad = -D_x j(x_w, w)\big(D_x f(x_w, w)\big)^{-1} D_w f(x_w, w) d_w + D_w j(x_w, w) d_w\end{aligned}$$

[1] Both of the cited sources make a sign error in the expression for the derivative. However, [Pat18] provides a proof in which the correct sign is used.

for L^∞ input perturbations d_w. To transfer this result to L^1 perturbations between binary-valued inputs, we use Assumption 2.4.25 (8) to show that the derivative mapping

$$T_v \coloneqq -D_x j(x_w, w)\big(D_x f(x_w, w)\big)^{-1} D_w f(x_w, w) + D_w j(x_w, w)$$

is bounded with respect to the L^1 norm for inputs that are differences between binary-valued functions and that the truncation error can be vanishes relative to the L^1 norm of the perturbation.

PART 3 (DIFFERENTIABILITY OF J). We want to invoke the results from Section 2.4.2. Therefore, we need to verify Assumption 2.4.10. (X, Σ, μ) is a layered measure space composed of finite atomless measure spaces. Therefore, it is itself a finite atomless measure space. The codomain Y is $L^1(\Sigma, \mu)$, which is a Banach space. The vector measure ν is the indicator function mapping, which satisfies Assumptions 2.4.10 (3) and 2.4.10 (4) according to Theorem 2.4.14 and

$$\big\|\nu(V) - \nu(U)\big\|_Y = \mu(U \triangle V)$$

for all $U, V \in \Sigma$. The indicator function mapping maps all sets to binary-valued integrable functions. As we have already shown, the mapping $\big(w \mapsto j(x_w, w)\big)$ is F-differentiable between binary-valued functions in the sense of Assumption 2.4.10 (5). Therefore, we can apply Proposition 2.4.12 and Theorem 2.4.13 to demonstrate the local and global differentiability of J. □

We note that Theorem 2.4.13 also demonstrates the continuity of the derivative. Given Assumption 2.4.25 (9), we can combine the components of the scaling density functions m into a layered function which then satisfies Definition 2.4.15 and allows us to invoke Proposition 2.4.16 to scale gradient densities.

Knowing that the derivative exists is not sufficient if we cannot approximate its density function. Assumptions 2.4.25 (10) to 2.4.25 (14) ensure that we can approximate the density function by using the gradient of the functional with respect to discretized control vectors on a hierarchy of meshes.

Theorem 2.4.27 (Approximation of the Derivative).
Let $d \in \mathbb{N}$, $n_w \in \mathbb{N}$, Ω, $\delta > 0$, Y, Z, G, f, j, and m satisfy Assumptions 2.4.25 (1) to 2.4.25 (7) and 2.4.25 (9). Let (X, Σ, μ) be the layered measure space composed of layers (Ω, Σ, μ_i) where μ_i is the measure whose density function with respect to λ is m_i. Then the functional $J : \Sigma/_{\sim_\mu} \to \mathbb{R}$ with

$$J(U) \coloneqq j(x_{\chi_U}, \chi_U) \quad \forall U \in \Sigma/_{\sim_\mu}$$

is continuously differentiable and we have

$$\begin{aligned}\nabla J(U)(D) = &-\Big(D_x j(x_{\chi_U}, \chi_U) \circ \big(D_x f(x_{\chi_U}, \chi_U)\big)^{-1} \circ D_w f(x_{\chi_U}, \chi_U)\Big)\big(\chi_{D\setminus U} - \chi_{D\cap U}\big) \\ &+ D_w j(x_{\chi_U}, \chi_U)\big(\chi_{D\setminus U} - \chi_{D\cap U}\big)\end{aligned}$$

for all $U, D \in \Sigma/_{\sim_\mu}$. Let $(n_i)_{i\in\mathbb{N}_0} \subseteq \mathbb{N}$ and $(T_{i,j})_{i\in\mathbb{N}_0,\, j\in[n_i]} \subseteq \mathscr{B}(\Omega)$ satisfy Assumptions 2.4.25 (10) to 2.4.25 (14). For each $x \in \Omega$, there exists a unique sequence $\big(j_i(x)\big)_{i\in\mathbb{N}_0} \subseteq \mathbb{N}$ such that $j_i(x) \in [n_i]$ and $x \in T_{i,j_i(x)}$ for all $i \in \mathbb{N}_0$. We then find that the k-th gradient density function of F satisfies

$$g(k, x) = \lim_{i\to\infty} \frac{\nabla J(U)(\{k\} \times T_{i,j_i(x)})}{\mu(\{k\} \times T_{i,j_i(x)})} \quad \text{for a.a. } x \in \Omega$$

for all $k \in [n_w]$. ◁

PROOF. The differentiability and form of the derivative follow from Proposition 2.4.26 Let g be the density function of $\nabla J(U)$. At first we only consider one $k \in [n_w]$ and one $x \in \Omega$.

Our first concern is the existence of the sequence $\big(j_i(x)\big)_{i\in\mathbb{N}_0}$. According to Assumption 2.4.25 (11), $(T_{i,j})_{j\in[n_i]}$ is a partition of Ω for all $i \in \mathbb{N}_0$. Therefore, for every $i \in \mathbb{N}_0$, there exists a unique $j_i(x) \in [n_i]$ such that $x \in T_{i,j_i(x)}$.

We use Lebesgue's differentiation theorem as stated in [BC09, Thm. 8.4.6]. There, the stated condition on the sequence of sets used to approximate a point x is "measure-metrizable convergence of sets to a point" [BC09, Def. 8.4.1]. As noted there, this is a particularly weak form of convergence because it does not even require that x is a member of any of the sets used to approximate it. In particular, this means that x does not need to lie in the interior of the mesh cells used to approximate it and we do not even have to exclude the nullset of all cell boundaries from our considerations.

The first criterion for measure-metrizable convergence is that for all $i \in \mathbb{N}_0$, there exists a radius $r_i(x)$ such that $T_{i,j_i(x)} \subseteq B_{r_i(x)}(x)$. This is true because of Assumption 2.4.25 (14). Let $C > 0$, $(\bar{x}_{i,j})_{i\in\mathbb{N}_0,\, j\in[n_i]} \subseteq \mathbb{R}^d$, and $(r_{i,j})_{i\in\mathbb{N}_0,\, j\in[n_i]} \subseteq \mathbb{R}$ be such that for all $i \in \mathbb{N}_0$ and $j \in [n_i]$, we have $T_{i,j} \subseteq B_{i,j} \coloneqq B_{r_{i,j}}(\bar{x}_{i,j})$ and $\lambda(T_{i,j}) \geq C \cdot \lambda(B_{i,j})$.

For each $i \in \mathbb{N}_0$ and $x \in \Omega$, we have $x \in T_{i,j_i(x)} \subseteq B_{i,j_i(x)}$. Therefore, for all $y \in T_{i,j_i(x)}$, we have

$$\|x - y\| \leq \|x - \bar{x}_{i,j_i(x)}\| + \|y - \bar{x}_{i,j_i(x)}\| \leq 2r_{i,j_i(x)}.$$

It then follows that $T_{i,j_i(x)} \subseteq B_{2r_{i,j_i(x)}}(x)$. We write $r_i(x) \coloneqq 2r_{i,j_i(x)}$. We note that the inclusion implies $\mu_k(T_{i,j_i(x)}) \leq \mu_k\big(B_{r_i(x)}(x)\big)$.

We now have to derive an upper bound on $r_{i,j_i(x)}$. We can derive this using the d-dimensional volume of a Euclidean ball which is given by

$$\lambda(B_{i,j}) = \frac{\pi^{d/2}}{\Gamma(d/2+1)} r_{i,j}^d \qquad \forall i \in \mathbb{N}_0,\ j \in [n_w]$$

where Γ is Euler's gamma function. This equation implies

$$\begin{aligned} r_i(x) &= 2r_{i,j_i(x)} \\ &= 2\sqrt[d]{\frac{\Gamma(d/2+1)}{\pi^{d/2}} \lambda(B_{i,j_i(x)})} \\ &\leq 2\sqrt[d]{\frac{\Gamma(d/2+1)}{C\pi^{d/2}} \lambda(T_{i,j_i(x)})}. \end{aligned}$$

In conjunction with Assumption 2.4.25 (12), this means that

$$\lim_{i\to\infty} r_i(x) = 0$$

for all $x \in \Omega$, which is the second criterion for measure-metrizable convergence.

The third and final criterion for measure-metrizable convergence follows directly from Assumption 2.4.25 (14). Let $C' > 0$ be such that $\lambda(U) \leq C' \cdot \mu_k(U)$ for all $U \in \mathscr{B}(\Omega)$ and $k \in [n_w]$. Such C' exist according to Proposition 2.4.16. For all

$i \in \mathbb{N}_0$ and $x \in \Omega$, we have

$$\begin{aligned}
\frac{C}{2^d C' \|m_k\|_{L^\infty}} \mu_k\big(B_{r_i(x)}(x)\big) &\leq \frac{C}{2^d C'} \cdot \lambda\big(B_{r_i(x)}(x)\big) \\
&= \frac{C}{2^d C'} \cdot \frac{\pi^{d/2}}{\Gamma(d/2+1)} r_i^d(x) \\
&= \frac{C}{C'} \cdot \frac{\pi^{d/2}}{\Gamma(d/2+1)} r_{i,j_i(x)}^d \\
&= \frac{C}{C'} \cdot \lambda(B_{i,j_i(x)}) \\
&\leq \frac{\lambda(T_{i,j_i(x)})}{C'} \\
&\leq \mu_k(T_{i,j_i(x)}).
\end{aligned}$$

Therefore, we have $\mu_k(T_{i,j_i(x)}) \geq \frac{C}{2^d C' \|m_k\|_{L^\infty}} \mu_k\big(B_{r_i(x)}(x)\big)$. We have $\frac{C}{2^d C' \|m_k\|_{L^\infty}} > 0$. We also have $\mu_k(T_{i,j_i(x)}) \leq \mu_k(B_{r_i(x)}(x))$, which ensures that $\frac{C}{2^d C' \|m_k\|_{L^\infty}} \leq 1$.

We have shown that $(T_{i,j_i(x)})_{i \in \mathbb{N}_0}$ converges measure-metrizably to x for all $x \in \Omega$. Since μ_k is a weighted version of the Lebesgue measure, μ_k is also a regular Borel measure and we can apply Lebesgue's differentiation theorem, which states that

$$\begin{aligned}
g(k,x) &= \lim_{i \to \infty} \frac{1}{\mu_k(T_{i,j_i(x)})} \underbrace{\int_{T_{i,j_i(x)}} g(k,x) \, \mathrm{d}\mu_k(x)}_{= \nabla J(U)(\{k\} \times T_{i,j_i(x)})} \\
&= \lim_{i \to \infty} \frac{\nabla J(U)(\{k\} \times T_{i,j_i(x)})}{\mu(\{k\} \times T_{i,j_i(x)})}
\end{aligned}$$

for almost all $x \in \Omega$ and all $k \in [n_w]$. □

2.5 CONVEX SET FUNCTIONS

Convexity is of great relevance in mathematical optimization as a distinguishing feature of "simple" optimization problems. In convex optimization problems, an algorithm that gradually improves an existing feasible solution until further improvement is no longer possible yields an optimal or nearly optimal solution. In non-convex problems, the existence of local optima can make such an algorithm yield suboptimal results.

The reason why we have to specifically address convexity is because it is usually defined using the *secant inequality*. This criterion states that a function $f : D \to \mathbb{R}$ mapping a convex set D to $\mathbb{R}$ is said to be convex if and only if

$$f\big(ty + (1-t)x\big) \leq tf(y) + (1-t)f(x) \quad \forall x, y \in Y,\ t \in [0,\ 1].$$

This definition depends on scalar multiplication, which is not available in similarity spaces. Therefore, we must define convexity differently in our problem setting.

The most straightforward way to define convexity would be to use geodesics instead of convex combinations. We discuss this in Section 2.5.1. There, we show that this way of defining convexity yields unexpected and undesirable results.

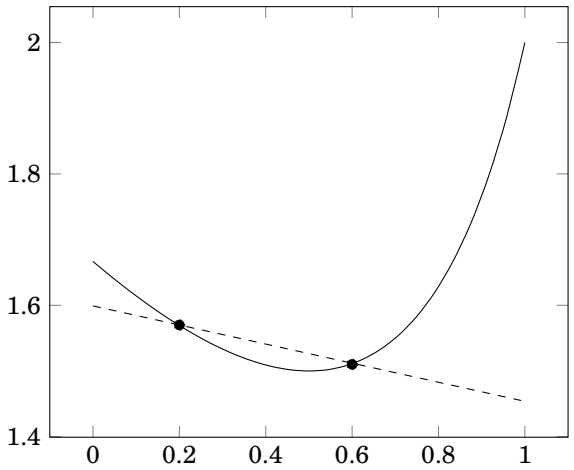


Figure 2.8: Illustration of the secant inequality. The graph of the function $f(x) = 1 - x + \frac{1}{\frac{3}{2} - x}$ underestimates its secant between $x = 0.2$ and $y = 0.6$.

Notably, even trivial set functionals such as signed measures can be non-convex under such a definition.

Because we are primarily interested in differentiable set functionals, we instead define convexity using an adapted version of the *tangent inequality*. We investigate this way of defining convexity in Section 2.5.2. We show that this definition yields a much more flexible and more universally applicable concept of convexity that has similar implications for optimization and can be inferred from vector space convexity in set functionals that are derived from vector space functionals in the manner described in Section 2.4.2.

In Section 2.5.3, we address pseudoconvexity, which is a generalization of convexity that is sometimes used in mathematical optimization. We show that pseudoconvexity, being based on derivatives, is easily transferrable to differentiable set functionals. However, we argue that its usefulness in our setting is very limited.

2.5.1 Secant Inequality

The first definition of convexity that we discuss here uses the secant inequality. For functions $f : D \to \mathbb{R}$ defined on a convex subset $D \subseteq V$ of a vector space V, the secant inequality takes the form

$$f\big(\eta \cdot y + (1-\eta) \cdot x\big) \leq \eta \cdot f(y) + (1-\eta) \cdot f(x) \quad \forall x, y \in D,\ \eta \in [0,\ 1].$$

The left hand side of this inequality is the evaluation of the function f at an intermediate point between x and y. This intermediate point is found by convex combination of the endpoints x and y. The right hand side is the corresponding convex combination of the function values $f(x)$ and $f(y)$ which corresponds to the value of a secant supported by the points x and y. Figure 2.8 illustrates the meaning of the secant inequality for two fixed support points.

Because convex combinations require scalar multiplication, which is not available in similarity spaces, we cannot apply the usual secant inequality. However,

we can replace the convex combination with an intermediate point on a geodesic connecting both endpoints.

While this may appear straightforward, there is no "conventional" geodesic connecting two endpoints. By using the sparse interpolation theorem (see Theorem 2.3.22), we can create geodesics connecting $U, V \in \Sigma/_{\sim_\mu}$ via any intermediate point $W \in \Sigma/_{\sim_\mu}$ that satisfies $\mu(U \triangle W) + \mu(V \triangle W) = \mu(U \triangle V)$. The secant inequality therefore becomes

$$F(W) = \frac{\mu(U \triangle W)}{\mu(U \triangle V)} \cdot F(V) + \frac{\mu(V \triangle W)}{\mu(U \triangle V)} \cdot F(U) \tag{2.49}$$

for all $U, V, W \in \Sigma/_{\sim_\mu}$ with $\mu(U \triangle W) + \mu(V \triangle W) = \mu(U \triangle V)$. This is equivalent to demanding that the secant inequality holds along every geodesic.

Equation (2.49) is overly restrictive. To show this, we consider the simplest type of differentiable set functional: an absolutely continuous signed measure. In our analogy to vector spaces, this is the analogue of a linear functional. We would therefore expect them to be convex. However, in the sense of Equation (2.49), they are generally not.

Proposition 2.5.1 (Triviality of Secant Convex Signed Measures). *Let (X, Σ, μ) be a finite atomless measure space with $\mu(X) > 0$, let $\varphi \colon \Sigma \to \mathbb{R}$ be a signed measure with $\varphi \ll \mu$, and let $f \in L^1(\Sigma, \mu)$ be the density function of φ. If there exists a constant $\eta \in \mathbb{R}$ such that*

$$\mu(\{f < \eta\}) > 0,$$
$$\mu(\{f > \eta\}) > 0.$$

Then φ does not satisfy Equation (2.49)*.*

Conversely, if φ is secant convex in the sense of Equation (2.49)*, then there exists $\eta^* \in \mathbb{R}$ such that $f \equiv \eta^*$ almost everywhere and we have $\varphi = \eta^* \cdot \mu$.* ◁

PROOF. Let $\eta \in \mathbb{R}$ be such that $\mu\big(\{f < \eta\}\big) > 0$ and $\mu\big(\{f > \eta\}\big) > 0$. We define

$$U := [\varnothing]_{\sim_\mu},$$
$$V := \big[\{f \neq \eta\}\big]_{\sim_\mu},$$
$$W := \big[\{f > \eta\}\big]_{\sim_\mu}.$$

$U \triangle W = W$ and $V \triangle W = \big[\{f < \eta\}\big]_{\sim_\mu}$ are essentially disjoint. Therefore, we have

$$\begin{aligned}\mu(U \triangle V) &= \mu(V) \\ &= \mu(V \triangle W \triangle W) \\ &= \mu\big((V \triangle W) \cup (U \triangle W)\big) \\ &= \mu(U \triangle W) + \mu(V \triangle W).\end{aligned}$$

However, for the values of the signed measure φ, we obtain

$$\begin{aligned}\varphi(W) &= \int_{\{f > \eta\}} f \, \mathrm{d}\mu \\ &> \eta \cdot \underbrace{\mu\big(\{f > \eta\}\big)}_{>0}\end{aligned}$$

$$
\begin{aligned}
&= \eta \cdot \mu(U \triangle W) \\
&= \underbrace{\varphi(U)}_{=0} + \eta \cdot \mu(U \triangle W),
\end{aligned}
$$

$$
\begin{aligned}
\varphi(V) &= \int_V f \,\mathrm{d}\mu \\
&= \int_{\{f>\eta\}} f \,\mathrm{d}\mu + \int_{\{f<\eta\}} f \,\mathrm{d}\mu \\
&= \varphi(W) + \int_{\{f<\eta\}} f \,\mathrm{d}\mu \\
&< \varphi(W) + \eta \cdot \mu\big(\{f<\eta\}\big) \\
&= \varphi(W) + \eta \cdot \mu(V \triangle W).
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
\varphi(W) &= \frac{\overbrace{\mu(U \triangle W) + \mu(V \triangle W)}^{=\mu(U \triangle V) > 0}}{\mu(U \triangle V)} \cdot \varphi(W) \\
&= \frac{\mu(U \triangle W)}{\mu(U \triangle V)} \cdot \underbrace{\varphi(W)}_{>\varphi(V) - \eta \cdot \mu(V \triangle W)} + \frac{\mu(V \triangle W)}{\mu(U \triangle V)} \cdot \underbrace{\varphi(W)}_{>\varphi(U) + \eta \cdot \mu(U \triangle W)} \\
&> \frac{\mu(U \triangle W)}{\mu(U \triangle V)} \cdot \varphi(V) + \frac{\mu(V \triangle W)}{\mu(U \triangle V)} \cdot \varphi(U) \\
&\qquad + \eta \cdot \frac{\mu(V \triangle W) \cdot \mu(U \triangle W) - \mu(U \triangle W) \cdot \mu(V \triangle W)}{\mu(U \triangle V)} \\
&= \frac{\mu(U \triangle W)}{\mu(U \triangle V)} \cdot \varphi(V) + \frac{\mu(V \triangle W)}{\mu(U \triangle V)} \cdot \varphi(U).
\end{aligned}
$$

This means that φ does not satisfy Equation (2.49).

Conversely, if φ satisfies Equation (2.49), there exists no $\eta \in \mathbb{R}$ such that both $\{f > \eta\}$ and $\{f < \eta\}$ are sets of nonzero measure. The measure of the sublevel set $\{f < \eta\}$ is increasing in η. If $\mu\big(\{f < \eta\}\big) = 0$ for all $\eta \in \mathbb{R}$, then we would have $f \equiv \infty$ almost everywhere. Because $\mu(X) > 0$, this would contradict $f \in L^1(\Sigma, \mu)$. Therefore, there must be $\eta \in \mathbb{R}$ such that $\mu\big(\{f < \eta\}\big) > 0$. Let

$$
\eta^* \coloneqq \sup\big\{\eta \in \mathbb{R} \,\big|\, \mu(\{f < \eta\}) = 0\big\} \in \mathbb{R} \cup \{-\infty\}.
$$

We first have to exclude the possibility that $\eta^* = -\infty$. If this were the case, then we would have $\mu\big(\{f < \eta\}\big) > 0$ for all $\eta \in \mathbb{R}$. This would then imply that $\mu\big(\{f > \eta\}\big) = 0$ for all $\eta \in \mathbb{R}$, which would mean that $f \equiv -\infty$ almost everywhere. This would contradict the integrability of f. Therefore, η^* must be a real number.

Because $\{f < \eta^*\}$ can be written as a countable union of $\{f < \eta\}$ with $\eta < \eta^*$ and $\{f \leq \eta^*\}$ can be written as a countable intersection of $\{f < \eta\}$ with $\eta > \eta^*$, we have

$$
\mu\big(\{f < \eta^*\}\big) = 0 \quad \text{and} \quad \mu\big(\{f \leq \eta^*\}\big) > 0.
$$

This means that $\mu\big(\{f = \eta^*\}\big) > 0$. However, because $\{f = \eta^*\} \subseteq \{f < \eta\}$ for all $\eta > \eta^*$, this also implies that $\mu\big(\{f > \eta\}\big) = 0$ for all $\eta > \eta^*$. We can then rewrite $\{f > \eta^*\}$ as a countable union of $\{f > \eta\}$ with $\eta > \eta^*$ and obtain

$$
\mu\big(\{f > \eta^*\}\big) = 0.
$$

This shows that $\mu(\{f > \eta^*\}) = 0$ and $\mu(\{f < \eta^*\}) = 0$. Together, this means that $f \equiv \eta^*$ almost everywhere. We then have

$$\varphi(A) = \int_A f \, \mathrm{d}\mu = \eta^* \cdot \mu(A) \quad \forall A \in \Sigma. \qquad \square$$

Proposition 2.5.1 shows that only a very small subset of very simple set functionals satisfies the naïve secant inequality. We have used signed measures as an example. However, we could apply the same type of argument to any differentiable set functional by choosing sufficiently small steps. We choose signed measures because the density function makes it much easier to construct steps that violate the secant inequality by using level sets.

We take this as an indication that any attempt of extracting a useful concept of convexity from the secant inequality likely requires much more intricate considerations. One possible variation would be to define a set functional as convex if it is convex only along a specific geodesic as opposed to being convex along all geodesics.

Definition 2.5.2 (Weak Secant Convexity).
Let $(X,\ \Sigma,\ \mu)$ be a finite atomless measure space, let $\mathscr{D} \subseteq \Sigma/_{\sim_\mu}$, and let $F\colon \mathscr{D} \to \mathbb{R}$. We refer to F as *weakly secant convex* if for all $U, V \in \Sigma/_{\sim_\mu}$ there exists a geodesic $\gamma\colon [0,\ 1] \to \mathscr{D}$ such that $\gamma(0) = U$, $\gamma(1) = V$, and

$$F\big(\gamma(t)\big) \leq t \cdot F(V) + (1-t) \cdot F(U) \quad \forall t \in [0,\ 1]. \qquad \triangleleft$$

Definition 2.5.2 is a more plausibly useful definition of convexity. For signed measures with density functions, constant mean geodesics can be used to easily prove weak secant convexity. However, for more complex set functionals, the construction of a suitable geodesic is a significant obstacle. Thankfully, the tangent inequality provides a much more convenient and useful definition of convexity. We develop this type of convexity in the following section.

2.5.2 Tangent Inequality

For differentiable functions, the tangent inequality provides a second, usually equivalent way of determining convexity. The tangent inequality is satisfied if tangents to the function graph at any point in the domain underestimate the function throughout the domain. Figure 2.9 on the facing page illustrates this for the tangent at one specific point.

For differentiable set functionals, the tangent inequality can be applied with very little modification.

Definition 2.5.3 (Tangent Convexity).
Let $(X,\ \Sigma,\ \mu)$ be a finite atomless measure space. Let $F\colon \Sigma/_{\sim_\mu} \to \mathbb{R}$ be a differentiable set functional. We refer to F as *tangent convex* if and only if

$$F(V) - F(U) \geq \nabla F(U)(U \triangle V) \quad \forall U, V \in \Sigma/_{\sim_\mu}. \tag{2.50}$$

We refer to F as *strictly tangent convex* if and only if

$$F(V) - F(U) > \nabla F(U)(U \triangle V) \quad \forall U, V \in \Sigma/_{\sim_\mu}\colon U \neq V. \tag{2.51}$$

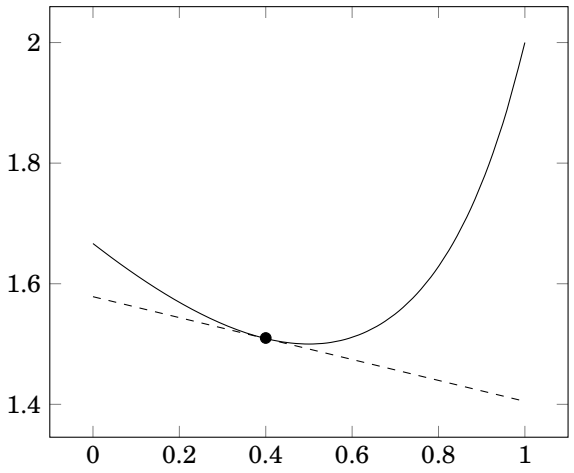


Figure 2.9: Illustration of the tangent inequality. The graph of the function $f(x) = 1 - x + \frac{1}{\frac{3}{2}-x}$ on the domain $[0,\ 1]$ is above its tangent around $x = 0.4$.

We refer to F as *strongly tangent convex* if and only if there exists a constant $C > 0$ such that

$$F(V) - F(U) \geq \nabla F(U)(U \triangle V) + C \cdot \big(\mu(U \triangle V)\big)^2 \quad \forall U, V \in \Sigma/_{\sim_\mu}. \tag{2.52}$$

◁

An interesting byproduct of tangent convexity is that it implies a stronger suboptimality estimator than the one that we had proposed in Proposition 2.4.8. For tangent convex functions, we can completely omit the curvature term in that estimator.

Proposition 2.5.4 (Suboptimality Estimation for Convex Functionals). *Let $(X,\ \Sigma,\ \mu)$ be a finite atomless measure space, and let $F : \Sigma/_{\sim_\mu} \to \mathbb{R}$ be a tangent convex differentiable set functional. Let $U \in \Sigma/_{\sim_\mu}$ and let $g_U : X \to \mathbb{R}$ be the density function of $\nabla F(U)$. Then we have*

$$F(V) \geq F(U) + \int_{U \triangle V} \min\{g_U,\ 0\}\, \mathrm{d}\mu \quad \forall V \in \Sigma/_{\sim_\mu}. \tag{2.53}$$

◁

Proof. Let $V \in \Sigma/_{\sim_\mu}$. Because F is tangent convex, we have

$$\begin{aligned} F(V) - F(U) &\geq \nabla F(U)(U \triangle V) \\ &= \int_{U \triangle V} g_U \, \mathrm{d}\mu \\ &\geq \int_{U \triangle V} \min\{g_U,\ 0\}\, \mathrm{d}\mu. \end{aligned}$$

□

This makes the unconstrained suboptimality measure $\mathscr{C}_1(F, U)$ that we had introduced in Definition 2.4.9 much more meaningful. Rather than simply knowing that $\mathscr{C}_1(F, U) = 0$ in a hypothetical optimal point which may not exist for many binary optimal control problems, convexity allows us to use $F(U) + \mathscr{C}_1(F, U)$ as a global lower bound on the value of F. Therefore, $\big|\mathscr{C}_1(F, U)\big|$ can be used

to estimate the global optimality gap and $\mathscr{C}_1(F,U_i) \xrightarrow{i\to\infty} 0$ indicates that the solution sequence $(U_i)_{i\to\mathbb{N}}$ becomes nearly globally optimal over time.

Unless we can prove that there are actual tangent convex set functionals, the mere definition of the term is meaningless. For proof, we turn to Section 2.4 in which we described our main way of finding derivatives of set functionals: deriving them from Fréchet derivatives. If a set functional is derived from a functional in a Banach space, then both its value and its derivative mirror that of the Banach space functional. Therefore, if the underlying Banach space functional satisfies the traditional tangent inequality, then we would expect the set functional to be tangent convex. This is indeed the case.

Theorem 2.5.5 (Tangent Convexity: Banach Space Functionals).
Let $(X,\ \Sigma,\ \mu)$, Y, $\nu\colon \Sigma \to Y$, and $f\colon Y \to \mathbb{R}$ satisfy Assumption 2.4.10. Let further f be Fréchet differentiable and convex. Then the set functional $F\colon \Sigma/_{\sim_\mu} \to \mathbb{R}$ with

$$F(U) \coloneqq f\big(\nu(U)\big) \quad \forall U \in \Sigma_i/_{\sim_{\mu_i}}$$

is tangent convex. ◁

PROOF. According to Theorem 2.4.13, F is differentiable. For $U,V \in \Sigma/_{\sim_\mu}$, we have

$$\begin{aligned} F(V) - F(U) &= f\big(\nu(V)\big) - f\big(\nu(U)\big) \\ &\geq \nabla f\big(\nu(U)\big)\big(\nu(V) - \nu(U)\big) \\ &= \nabla f\big(\nu(U)\big)\big(\nu(V \setminus U) - \nu(U \setminus V)\big). \end{aligned}$$

We note that $D \coloneqq V \mathbin{\triangle} U = (V \setminus U) \cup (U \setminus V)$. Therefore, we have $D \setminus U = V \setminus U$ and $D \cap U = U \setminus V$, which implies that

$$\begin{aligned} F(V) - F(U) &\geq \nabla f\big(\nu(U)\big)\big(\nu(V \setminus U) - \nu(U \setminus V)\big) \\ &= \nabla f\big(\nu(U)\big)\big(\nu(D \setminus U) - \nu(D \cap U)\big) \\ &= \nabla F(U)(D) \\ &= \nabla F(U)(U \mathbin{\triangle} V). \end{aligned}$$

Here, we make use of the expression for $\nabla F(U)(D)$ that we had derived in Proposition 2.4.12 on page 165. □

Theorem 2.5.5 implies that if any of the set functionals shown to be differentiable in Sections 2.4.2 to 2.4.4 are convex in a vector space sense, then they are also tangent convex in our context. This means that tangent convexity is a very broadly applicable concept. As the proof suggests, strict and strong convexity can be similarly transferred.

In vector space optimization, strict convexity usually implies the uniqueness of the optimum. We cannot expect to transfer this result to problems where the optimum does not necessarily exist. However, we expect that in a strictly or strongly convex setting, a sequence of solutions whose unconstrained suboptimality measure $\mathscr{C}_1(F,U)$ goes to zero would be funnelled into an increasingly small region of solutions. In essence, the solution sequence would still reach a consensus structure even though the perfect consensus may not be achievable. This form of consensus without convergence is best expressed by the solution sequence being a Cauchy sequence.

We leave the question open of whether this is true for strictly tangent convex set functionals. However, it is relatively easy to show that this holds for strongly convex set functionals.

Proposition 2.5.6 (Strong Convexity and Cauchy Sequences).
Let (X, Σ, μ) be a finite atomless measure space, and let $F\colon \Sigma/_{\sim_\mu} \to \mathbb{R}$ be a strongly tangent convex differentiable set functional. Let $(U_j)_{j\in\mathbb{N}} \in (\Sigma/_{\sim_\mu})^{\mathbb{N}}$ be a sequence such that $\big(F(U_j)\big)_{j\in\mathbb{N}}$ is decreasing and

$$\mathscr{C}_1(F, U_j) \xrightarrow{j\to\infty} 0.$$

Then $(U_j)_{j\in\mathbb{N}}$ is a Cauchy sequence. ◁

Proof. Part 1 ($(F(U_j))_{j\in\mathbb{N}}$ is Cauchy). Since F is strongly tangent convex, F is also tangent convex. Because F is tangent convex, according to Proposition 2.5.4, we have

$$F(U_k) \geq F(U_j) + \nabla F(U_j)(U_j \triangle U_k) \geq F(U_j) + \mathscr{C}_1(F, U_j) > -\infty \quad \forall j,k \in \mathbb{N}.$$

Let $\varepsilon > 0$ and let $j_0 \in \mathbb{N}$ be such that

$$\mathscr{C}_1(F, U_j) \geq -\varepsilon \quad \forall j \geq j_0.$$

We note that $\mathscr{C}_1(F, U_j) \leq 0$ always holds due to the way in which $\mathscr{C}_1$ is defined. Let $j,k \in \mathbb{N}$ with $j \geq j_0$, and $k \geq j_0$. Without loss of generality, let $F(U_j) \geq F(U_k)$. Then we have

$$\begin{aligned} \big|F(U_j) - F(U_k)\big| &= F(U_j) - F(U_k) \\ &\leq \underbrace{F(U_j) - F(U_j)}_{=0} - \underbrace{\mathscr{C}_1(F, U_j)}_{\geq -\varepsilon} \\ &\leq \varepsilon. \end{aligned}$$

The existence of such a j_0 for every $\varepsilon > 0$ implies that $\big(F(U_j)\big)_{j\in\mathbb{N}}$ is a Cauchy sequence. Alternatively, we could use the estimate from Proposition 2.5.4 to establish that the sequence is bounded below. Because it is also monotonically decreasing, this implies that the sequence is convergent and therefore also a Cauchy sequence.

Part 2 ($(U_j)_{j\in\mathbb{N}}$ is Cauchy). Now, we make use of the fact that F is strongly tangent convex. Let $C > 0$ be such that

$$F(V) \geq F(U) + \nabla F(U)(U \triangle V) + C \cdot \big(\mu(U \triangle V)\big)^2 \quad \forall U, V \in \Sigma/_{\sim_\mu}.$$

Such a constant C exists according to the definition of strong tangent convexity. For all $U, V \in \Sigma/_{\sim_\mu}$, we have

$$F(V) - F(U) - \nabla F(U)(U \triangle V) \geq C \cdot \big(\mu(U \triangle V)\big)^2 \geq 0,$$

and therefore

$$\mu(U \triangle V) \leq \sqrt{\frac{F(V) - F(U) - \nabla F(U)(U \triangle V)}{C}}.$$

Due to the way in which the suboptimality measure $\mathscr{C}_1$ is defined, we have $\nabla F(U)(U \triangle V) \geq \mathscr{C}_1(F,U)$. Because $\big(F(U_j)\big)_{j\in\mathbb{N}}$ is a Cauchy sequence and $\mathscr{C}_1(F,U_j) \to 0$ for $j \to 0$, we can choose $j_0 \in \mathbb{N}$ such that

$$\begin{aligned} \mathscr{C}_1(F,U_j) &\geq -\frac{C\cdot\varepsilon^2}{2} \quad \forall j \geq j_0, \\ \big|F(U_j) - F(U_k)\big| &\leq \frac{C\cdot\varepsilon^2}{2} \quad \forall j \geq j_0,\ k \geq j_0. \end{aligned}$$

Let $j,k \in \mathbb{N}$ with $j \geq j_0$ and $k \geq j_0$. Then, we have

$$\begin{aligned} \mu(U_j \triangle U_k) &\leq \sqrt{\frac{F(U_k) - F(U_j) - \nabla F(U_j)(U_j \triangle U_k)}{C}} \\ &\leq \sqrt{\frac{\big|F(U_k) - F(U_j)\big| - \mathscr{C}_1(F,U_j)}{C}} \\ &\leq \sqrt{\frac{\frac{C\cdot\varepsilon^2}{2} + \frac{C\cdot\varepsilon^2}{2}}{C}} \\ &= \sqrt{\frac{C\cdot\varepsilon^2}{C}} \\ &= \varepsilon. \end{aligned}$$

The fact that such a j_0 exists for every $\varepsilon > 0$ shows that $(U_j)_{j\in\mathbb{N}}$ is a Cauchy sequence. □

We now have two viable concepts of convexity: weak secant convexity and tangent convexity. In conventional optimization, these concepts are equivalent for differentiable functions. Proving a similar result in our setting requires the construction of special convexity-realizing geodesics. Outside of the special case of signed measures with density functions, where convexity is realized by constant mean geodesics, we are unable to provide a method to construct such geodesics. Therefore, the question of whether these concepts are equivalent or whether one implies the other is an open question. We pose the following as an open conjecture without proof.

Hypothesis 2.5.7 (Weak Secant Convexity).
Let $(X,\ \Sigma,\ \mu)$ be a finite atomless measure space, and let $F\colon \Sigma/_{\sim_\mu} \to \mathbb{R}$ be a differentiable set functional that is tangent convex. Then F is also weakly secant convex. ◁

If Hypothesis 2.5.7 is true, then weak secant convexity is a generalization of tangent convexity. Much of the theoretical work presented in Section 2.3 was originally developed in an attempt to construct suitable geodesics for tangent convex differentiable set functionals through iterative rearrangement of geodesics. However, we are unable to prove that such a process would converge in any meaningful sense at this time.

2.5.3 Pseudoconvexity

There is a well-known generalization of convexity that we want to briefly address here: pseudoconvexity. Because it is based on derivatives, the concept of pseudoconvexity is equally easy to transfer to our setting.

Definition 2.5.8 (Pseudoconvexity).
Let $(X,\ \Sigma,\ \mu)$ be a finite atomless measure space, and let $F\colon \Sigma/_{\sim_\mu} \to \mathbb{R}$ be a differentiable set functional. We refer to F as *pseudoconvex* if and only if

$$F(V) < F(U) \implies \nabla F(U)(U \triangle V) < 0 \quad \forall U, V \in \Sigma/_{\sim_\mu}. \qquad \triangleleft$$

The concept of pseudoconvexity allows for a transfer of the necessary optimality criterion, i.e., every similarity class U^* minimizing a pseudoconvex differentiable set functional F must satisfy $\mathscr{C}_1(F, U^*) = 0$. However, pseudoconvexity does not imply any gradient-based limit on the optimality gap. Because actual optima frequently do not exist for optimization problems in similarity spaces, this means that pseudoconvexity is likely of little use for optimization in similarity spaces. Therefore, we do not investigate its potential use further.

Chapter 3

Algorithms

In this chapter, we develop practical optimization algorithms for several types of optimization problems. Because our primary goal is to solve problems with differential equation constraints, we expect function evaluations to be subject to substantial discretization errors. Therefore, we will design these algorithms to be resilient to such errors.

In Section 3.1, we develop a framework for gradient-based unconstrained optimization. We develop a method of approximating a steepest descent step and embed it within a trust-region framework to obtain an optimization loop that provably produces a sequence of solutions whose instationarity converges to zero. According to Proposition 2.4.8, this implies convergence of the objective value to a locally optimal value for Lipschitz-continuously differentiable set functionals. For convex set functionals, it implies convergence to the globally optimal objective value according to Proposition 2.5.4.

In Section 3.2, we turn our attention to solving optimization problems with two different kinds of constraints: logical and scalar-valued. The former encodes logical relationships between set memberships and could be enforced through a modification of the steepest descent step finding routine. The latter encodes a conventional nonlinear constraint of, e.g., the form $F(x) \leq 0$ where F is a differentiable set functional. For these types of constraints, we can adapt the quadratic penalty method to show that it is possible to transfer existing algorithms for constrained nonlinear optimization to our setting. We do not put forward a concrete algorithm for a steepest descent method with logical constraints. However, we briefly discuss how a modified step-finding method could be implemented.

3.1 UNCONSTRAINED OPTIMIZATION

A variant of the unconstrained trust region loop without error control was first developed in [HLS22].

A differentiable unconstrained set-valued optimization problem has the form

$$\inf J(U) \quad \text{s.t.} \quad U \in \Sigma/_{\sim_\mu} \tag{3.1}$$

where (X, Σ, μ) is a finite atomless measure space, and $J : \Sigma/_{\sim_\mu} \to \mathbb{R}$ is continuously differentiable in the sense of Definition 2.4.4. We intentionally state this as

an infimization problem because we generally do not expect the problem to have an optimum.

We judge the suboptimality of a solution $U \in \Sigma/_{\sim_\mu}$ by the numeric criterion $\mathscr{C}_1(J,U)$ that we have defined in Definition 2.4.9. True optimality is often unachievable in set-valued optimization problems because similarity spaces are generally not compact. Therefore, solution sequences generally do not have accumulation points. We instead rely on suboptimality estimators such as those developed in Propositions 2.4.8 and 2.5.4 to prove an asymptotic optimality gap of zero.

3.1.1 Evaluation and Step-Finding Framework

The general framework that we will follow is that of an iterative trust-region descent method. Given a suboptimal solution $U_i \in \Sigma/_{\sim_\mu}$ with $i \in \mathbb{N}$, we derive a step $D_i \in \Sigma/_{\sim_\mu}$ such that $J(U_i \mathbin{\triangle} D_i) < J(U_i)$ and set $U_{i+1} \coloneqq U_i \mathbin{\triangle} D_i$. To ensure termination and asymptotic stationarity, we must ascertain that the step realizes at least a certain fraction of the projected possible descent, i.e., there must be a fraction $\omega > 0$ such that

$$J(U_{i+1}) \leq J(U_i) + \omega \cdot \mathscr{C}_1(F, U_i) \quad \forall i.$$

The challenge is to achieve this sufficient descent criterion despite the presence of evaluation errors. Since set-valued variables in atomless measure spaces of non-zero measure are infinite-dimensional, most practical solvers must work with finite-dimensional approximations, which means that evaluation methods may be subject to substantial discretization errors.

This is particularly problematic because of the way in which we exploit atomlessness. Because our intention is to increase the spatial resolution of binary choices until making the "wrong" choice in a few of them becomes nearly irrelevant to the overall optimization problem, we have to compare solutions with different control meshes and cannot simply ignore discretization errors as is frequently done in first-discretize-then-optimize approaches that work with the same mesh throughout.

In this section, we develop a general framework and terminology to refer to different parts of the overall optimization algorithm. We also make prescriptions about the way in which the algorithm should interact with error controlled evaluation routines.

Definition 3.1.1 (Error-Controlled Evaluation Methods).
Let $(X,\ \Sigma,\ \mu)$ be a tuple of finite atomless measure spaces, and let $\mathscr{F}(\Sigma/_{\sim_\mu}, \mathbb{R})$ be the set of all differentiable set functionals $F\colon \Sigma/_{\sim_\mu} \to \mathbb{R}$. For a given set functional $F \in \mathscr{F}(\Sigma/_{\sim_\mu}, \mathbb{R})$ and $U \in \Sigma/_{\sim_\mu}$, let $g_F(U)$ be the density function of the signed measure $\nabla F(U)$.

An *error-controlled evaluation method* for a set functional $F \in \mathscr{F}(\Sigma/_{\sim_\mu}, \mathbb{R})$ is a mapping $\widetilde{F}\colon \Sigma/_{\sim_\mu} \times \left(\mathbb{R}_{>0} \cup \{\infty\}\right) \to \mathbb{R} \times \mathbb{R}_{\geq 0}$ such that

$$\left|F(U) - f\right| \leq \varepsilon \leq \overline{\varepsilon} \quad \forall U \in \Sigma/_{\sim_\mu},\ \overline{\varepsilon} > 0,\ (f,\ \varepsilon) = \widetilde{F}(U, \overline{\varepsilon}). \tag{3.2}$$

An *L^1-controlled gradient evaluation method* for a differentiable set functional $F \in \mathscr{F}(\Sigma/_{\sim_\mu}, \mathbb{R})$ is a mapping $\widetilde{g}\colon \Sigma/_{\sim_\mu} \times \left(\mathbb{R}_{>0} \cup \{\infty\}\right) \to L^1(\Sigma, \mu) \times \mathbb{R}_{\geq 0}$ such that

$$\left\|g_F(U) - f\right\|_{L^1(\Sigma,\mu)} \leq \varepsilon \leq \overline{\varepsilon} \quad \forall U \in \Sigma/_{\sim_\mu},\ \overline{\varepsilon} > 0,\ (f,\ \varepsilon) = \widetilde{g}(U, \overline{\varepsilon}). \tag{3.3}$$

An L^∞*-controlled gradient evaluation method* for a differentiable set functional $F \in \mathscr{F}(\Sigma/_{\sim_\mu}, \mathbb{R})$ is a function $\widetilde{g}\colon \Sigma/_{\sim_\mu} \times \left(\mathbb{R}_{>0} \cup \{\infty\}\right) \to L^1(\Sigma,\mu) \times \mathbb{R}_{\geq 0}$ such that

$$\left\| g_F(U) - f \right\|_{L^\infty(\Sigma,\mu)} \leq \varepsilon \leq \overline{\varepsilon} \quad \forall U \in \Sigma/_{\sim_\mu},\, \overline{\varepsilon} > 0, (f,\, \varepsilon) = \widetilde{g}(U, \overline{\varepsilon}). \tag{3.4}$$

◁

It is clear that an L^∞-controlled gradient evaluation method is generally only achievable for benignly differentiable set functionals. Nonetheless, we allow for non-benign derivatives as long as the error is essentially bounded, because stronger assumptions are not required for unconstrained optimization.

Aside from discretization errors, we also need to account for errors made in the step-finding procedure itself. In our framework, steps are derived from approximations of the gradient density functions of the objective functional. While we expect the result to be subject to some error on account of working with approximations rather than the actual gradient density functions, the returned step may also not be accurate for the approximate gradient density function. For instance, we will later derive the steepest descent step using a bisection method that generates a truncation error.

Definition 3.1.2 (Controlled Unconstrained Step-Finding).
Let $(X,\ \Sigma,\ \mu)$ be a finite atomless measure space. We refer to a mapping $\mathscr{S}\colon L^1(\Sigma,\mu) \times \mathbb{R}_{>0} \times \mathbb{R}_{>0} \to \Sigma/_{\sim_\mu} \times \mathbb{R}_{\geq 0}$ as a *controlled unconstrained step-finding routine with quality* θ if and only if $\theta > 0$, and we have

$$\mu(D) \leq \Delta, \tag{3.5}$$

$$\delta \leq \varepsilon, \tag{3.6}$$

$$\int_D g \,\mathrm{d}\mu \leq \theta \cdot \min\left\{1,\ \frac{\Delta}{\mu(\{g<0\})}\right\} \cdot \left(\int_{\{g<0\}} g \,\mathrm{d}\mu\right) + \delta \tag{3.7}$$

for all $g \in L^1(\Sigma,\mu)$, $\Delta > 0$, $\varepsilon > 0$, and $(D,\ \delta) = \mathscr{S}(g, \Delta, \varepsilon)$. For the purposes of this definition, we consider $\frac{\Delta}{0} = \infty$ such that the factor becomes 1 if $\mu(\{g<0\}) = 0$, though the step-finding routine should never be invoked in this case. ◁

Because it is impossible to consistently find sets with better-than-average integral, there cannot be a controlled unconstrained step-finding routine with quality $\theta > 1$.

Proposition 3.1.3.
Let $(X,\ \Sigma,\ \mu)$ *be a finite atomless measure space. There is no controlled unconstrained step-finding routine* $\mathscr{S}\colon L^1(\Sigma,\mu) \times \mathbb{R}_{>0} \times \mathbb{R}_{>0} \to \Sigma/_{\sim_\mu} \times \mathbb{R}_{>0}$ *with quality* $\theta > 1$. ◁

Proof. We prove the claim by contradiction. If we assumed that there is a controlled unconstrained step-finding routine $\mathscr{S}\colon L^1(\Sigma,\mu) \times \mathbb{R}_{>0} \times \mathbb{R}_{>0} \to \Sigma/_{\sim_\mu} \times \mathbb{R}_{>0}$ with quality $\theta > 1$, then we could invoke $\mathscr{S}$ with the following arguments:

$$\begin{aligned} g &\equiv -1 && \text{almost everywhere,} \\ \Delta &:= \mu(X), \\ \varepsilon &:= \frac{(\theta-1)\cdot\Delta}{2}. \end{aligned}$$

Let $(D,\ \delta) := \mathscr{S}(g, \Delta, \varepsilon)$. According to Definition 3.1.2, we would have

$$\begin{aligned}
-\Delta &\leq -\mu(D) \\
&= \int_D g \,\mathrm{d}\mu \\
&\leq \underbrace{\theta \cdot \min\left\{1,\ \frac{\mu(X)}{\mu(\underbrace{\{g<0\}}_{=X})}\right\}}_{=1} \cdot \underbrace{\left(\int_{\{g<0\}} g \,\mathrm{d}\mu\right)}_{=-\mu(X)} + \underbrace{\delta}_{\leq\varepsilon} \\
&\leq -\theta \cdot \underbrace{\mu(X)}_{=\Delta} + \frac{(\theta-1)\cdot\Delta}{2} \\
&= \frac{\overbrace{-(1+\theta)}^{<-2}\cdot\Delta}{2} \\
&< -\Delta.
\end{aligned}$$

This contradiction shows that $\mathscr{S}$ cannot exist. □

We later develop a controlled unconstrained step-finding routine with $\theta = 1$. With appropriately error-controlled evaluation routines and a controlled step-finding routine, we can formulate a generic framework for the controlled descent algorithm.

3.1.2 Error Control and Bound Tuning

We assume that error is generated at three distinct points in our optimization procedure: objective evaluation, gradient evaluation, and step finding. We cannot provide a general account of how to perform the error-controlled evaluations of objective and gradient. This is highly problem-specific and should be left to problem domain experts. We restrict ourselves to the generic framework assumptions stated in the previous section. We may, however, assume that tight error bounds can cause evaluations to have significant resource requirements. Therefore, we have to strike a balance between the three error sources that requires none of the three error control procedures to shoulder too much of the overall burden of error control. In this section, we discuss how to tune our various error bounds such that

- ε-stationarity detection correctly indicates nearly stationary solutions;
- nearly stationary solutions are detected as ε-stationary;
- steps are only accepted if they sufficiently decrease the objective;
- sufficiently good steps are accepted.

In other words, we have to control for both false positives and false negatives in both the ε-stationarity test and the step acceptance test. It is important to bear in mind that we have to control for both types of errors. Generally, preventing one does not prevent the other.

This means that all decision thresholds, such as acceptance or termination thresholds will be subject to error and we generally only guarantee their satisfaction up to a certain error margin. This is unavoidable in settings where significant evaluation errors can occur.

Error control is a computationally expensive process. We generally want to keep our error margins as loose as possible for as long as possible. Instead of immediately setting error bounds that would guarantee sufficient accuracy, we evaluate each quantity following the general evaluation loop stated in Algorithm 1.

Algorithm 1 Error-Controlled Evaluation Loop

Require:
- Non-empty "parameter" and "value" sets P and V;
- "Evaluator" procedure $\phi\colon P \times \big(\mathbb{R}_{>0} \cup \{\infty\}\big) \to V \times \mathbb{R}_{\geq 0}$ such that all output tuples $(x,\ e)$ of $\phi(p,\beta)$ for $p \in P$ and $\beta \in \mathbb{R}_{>0} \cup \{\infty\}$ satisfy $e \leq \beta$;
- "Parameter oracle" mapping $\omega\colon P \times V \to P \times \big(\mathbb{R}_{>0} \cup \{\infty\}\big)$ such that there is a constant $\overline{e} > 0$ with

$$e \leq \overline{e} \implies \big(\omega_1(p,x) = p \wedge e \leq \omega_2(p,x)\big)$$

  for all output tuples $(x,\ e)$ of $\phi(p,\beta)$ for all $p \in P$ and $\beta \in \mathbb{R}_{>0} \cup \{\infty\}$;
- Initial parameter $p_0 \in P$ and error bound $\beta_0 \in \mathbb{R}_{>0} \cup \{\infty\}$;
- Decay rate $\xi \in (0,\ 1)$.

Ensure: Yields a tuple $(p,\ x,\ e) \in P \times V \times \mathbb{R}_{\geq 0}$ such that $(x,\ e)$ is an output tuple of $\phi(p,\beta^*)$ for some $\beta^* \in \mathbb{R}_{>0} \cup \{\infty\}$ and we have $p = \omega_1(p,x)$, and $e \leq \omega_2(p,x)$.

1: **procedure** CONTROLLEDEVAL(ϕ, ω, p_0, β_0, ξ)
2: $\quad i \leftarrow 0$
3: $\quad (x_0,\ e_0) \leftarrow \phi(p_0,\beta_0)$
4: $\quad (p_1,\ \beta_1) \leftarrow \omega(p_0,x_0)$
5: $\quad$ **while** $p_i \neq p_{i+1} \vee e_i > \beta_{i+1}$ **do**
6: $\quad\quad i \leftarrow i+1$
7: $\quad\quad (x_i,e_i) \leftarrow \phi\big(p_i,\ \xi \cdot \min\{e_{i-1},\ \beta_i\}\big)$
8: $\quad\quad (p_{i+1},\ \beta_{i+1}) \leftarrow \omega(p_i,\ x_i)$
9: $\quad$ **end while**
10: $\quad$ **return** $(p_i,\ x_i,\ e_i)$
11: **end procedure**

The idea behind Algorithm 1 is straightforward. We terminate when the error bound e_i satisfies $e_i \leq \omega_2(p_i,x_i)$. If $e_i > \omega_2(p_i,x_i)$, then setting the error bound on the subsequent evaluation to a value not exceeding $\xi \cdot \omega_2(p_i,x_i) < \xi \cdot e_i$ is guaranteed to lower the error e_{i+1} relative to e_i by a factor of at least $\xi < 1$. The fixed decay rate ξ guarantees an exponential decrease of β_i to zero, which implies that $e_i \xrightarrow{i\to\infty} 0$. The termination of the loop is guaranteed by the existence of $\overline{e} > 0$ such that the termination criterion is always satisfied for $e_i \leq \overline{e}$.

We allow for additional variable parameters hidden in the opaque parameter object $p \in P$. This is in preparation for the penalty method that we develop in Section 3.2.1.3. There, it is necessary to adapt the penalty parameter during gradient evaluation to ensure that stationarity implies feasibility. When such

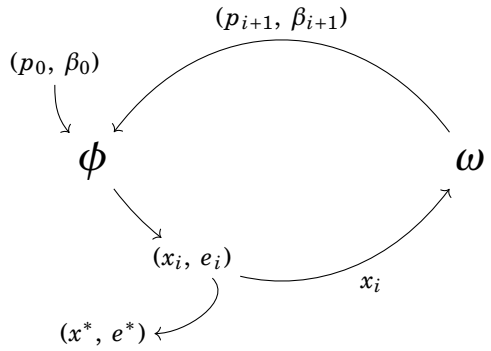


Figure 3.1: Illustration of Algorithm 1 on the preceding page. An initial bound β_0 is fed into the evaluator ϕ to obtain an approximate value x_0, which is then fed into the parameter oracle ω to recursively tighten the bound β_i and adjust the parameters p_i for the evaluator in order to obtain more accurate values x_i until the error estimate e_i is smaller than $\omega_2(p_i, x_i)$ and no further parameter adjustments are needed.

a parameter variation is not necessary, P can be chosen to be the set of empty tuples, which is a non-empty set with exactly one element. In this case, we omit p from our notation by convention.

Both the value and parameter objects can be implicit tuples in the sense that we write them as a sequence of several arguments instead of as a tuple. Distinguishing parameters and values can be challenging in this simplified notation. To distinguish the two, it is easiest to examine the signature of the evaluator ϕ. If ϕ has more than two inputs, then the parameter is the tuple consisting of all but the last input, the last being the error bound. Similarly, if ϕ has more than two outputs, the tuple consisting of all outputs except for the last one is the value tuple. The last output is always the error estimate.

Theorem 3.1.4 (Correctness of Algorithm 1).
Let P and V be non-empty sets. Let

- $\phi\colon P \times \big(\mathbb{R}_{>0} \cup \{\infty\}\big) \to V \times \mathbb{R}_{\geq 0}$ *be a procedure such that for all parameter tuples $(p,\ \beta) \in P \times \mathbb{R}_{>0} \cup \{\infty\}$ and all output tuples $(x,\ e) \in V \times \mathbb{R}_{\geq 0}$ of $\phi(p, \beta)$, we have $e \leq \beta$;*
- $\omega\colon P \times V \to P \times \big(\mathbb{R}_{>0} \cup \{\infty\}\big)$ *be such that there exists $\overline{e} > 0$ such that for all $(p,\ \beta) \in P \times \big(\mathbb{R}_{>0} \cup \{\infty\}\big)$ and all output tuples $(x,\ e)$ of $\phi(p, \beta)$, we have*

$$e \leq \overline{e} \implies \big(p = \omega_1(p, x) \wedge e \leq \omega_2(p, x)\big)$$

- $\beta_0 \in \mathbb{R}_{>0} \cup \{\infty\}$, $p_0 \in P$, *and* $\xi \in (0,\ 1)$.

Then Algorithm 1 terminates in a finite number of steps and returns a tuple $(p,\ x,\ e)$ such that $(x,\ e)$ is an output tuple of $\phi(p, \beta)$ for some bound $\beta \in \mathbb{R}_{>0} \cup \{\infty\}$ such that $p = \omega_1(p, x)$ and $e \leq \omega_2(p, x)$. ◁

PROOF. We first note that for all $i \in \mathbb{N}_0$ prior to termination, $(x_i,\ e_i)$ is an output tuple of $\phi(p_i, \beta_i)$. This implies that $(x_i,\ e_i) \in V \times \mathbb{R}_{\geq 0}$ with $e_i \leq \beta_i$ for all such

$i \in \mathbb{N}_0$. Because the loop terminates if and only if $p_i = \omega_1(p_i, x_i)$ and $e_i \leq \omega_2(x_i)$, we only need to show that the loop terminates.

If $p_0 = \omega_1(p_0, x_0)$ and $e_0 \leq \omega_2(p_0, x_0)$, then the loop terminates immediately. If $p_0 \neq \omega_1(p_0, x_0)$ or $e_0 > \omega_2(p_0, x_0)$, then we show inductively that $e_i \leq \xi^i e_0$ for all $i \in \mathbb{N}_0$ prior to loop termination. For $i = 0$, this is trivially true.

Let $i \in \mathbb{N}$ be such that $e_{i-1} \leq \xi^{i-1} e_0$. If $p_i \neq \omega_1(p_i, x_i)$ or $e_i > \omega_2(p_i, x_i)$, then (x_{i+1}, e_{i+1}) is the output of

$$\phi\Big(\omega_1(p_i, x_i),\ \xi \cdot \min\big\{e_i,\ \omega_2(p_i, x_i)\big\}\Big).$$

Because this is an output tuple of ϕ, it is guaranteed that

$$e_{i+1} \leq \xi \cdot \min\big\{e_i,\ \omega_2(p_i, x_i)\big\} \leq \xi e_i \leq \xi^{i+1} e_0.$$

This demonstrates inductively that $e_i \leq \xi^i e_0$ for all $i \in \mathbb{N}_0$ prior to termination. Because $0 \leq e_0 < \infty$ and $\xi \in (0,\ 1)$, $i \mapsto \xi^i e_0$ converges to zero for $i \to \infty$ and there exists $i_0 \in \mathbb{N}_0$ such that $\xi^{i_0} e_0 \leq \overline{e}$. If the loop does not terminate prior to $i = i_0$, then we have $e_{i_0} \leq \overline{e}$ and therefore $p_{i_0} = \omega_1(p_{i_0}, x_{i_0})$ and $e_{i_0} \leq \omega_2(p_{i_0}, x_{i_0})$. □

To reduce the number of arguments that we have to pass between procedures, we will always choose a decay rate of $\xi = \frac{1}{2}$. There is a tradeoff between frequency of evaluation and tightness of error bounds to make when choosing this parameter. We will not devote any time to its fine-tuning. However, it is important to note that this decay rate can technically be seen as an additional tuning parameter.

Remark 3.1.5 (Mappings and Procedures).
It is a simplification to work with the evaluator as if it was a mapping from a bound to an output. In practice, the precise output of a complex evaluator can depend on additional internal state variables or even random variables. Therefore, the output of the evaluator ϕ may differ on subsequent invocations with the same bound argument. We refer to ϕ as a "procedure" to emphasize that it need not be a mapping in the technical sense.

The correctness of Algorithm 1 is not impacted by this because we only invoke the evaluator once per argument tuple and use stored output for subsequent uses. It is important to bear this in mind during implementation. ◁

Algorithm 1 requires an error-controlled evaluator, a bound oracle and an initial guess for the parameters and error bound. The remainder of this section is dedicated to designing these oracles and guesses to control the error of the two notable decision quantities of the unconstrained trust region method.

Let subsequently $(X,\ \Sigma,\ \mu)$ be a finite atomless measure space, let $U, V \in \Sigma/_{\sim_\mu}$, let $D := U \triangle V$ be the step between U and V, let f_U and f_V be approximations of $F(U)$ and $F(V)$, respectively, and let $g_D := \int_D \widetilde{g}\,\mathrm{d}\mu$ be the approximate projected change of F from U to V, obtained by integrating the approximate density function $\widetilde{g}$ of $\nabla F(U)$ over the step D.

To simplify notation of errors, we write $e_{f,U} := f_U - F(U)$, $e_{f,V} := f_V - F(V)$, $e_g := \widetilde{g} - g$ where g is the actual gradient density function of F in U, and $e_{g,D} := \int_D e_g\,\mathrm{d}\mu$.

We assume that the step D has been found using a controlled unconstrained step-finding routine of quality $\theta \in (0,\ 1]$. Therefore, we have

$$g_D \leq \theta \cdot \min\left\{1,\ \frac{\Delta}{\mu(\{\widetilde{g} < 0\})}\right\} \cdot \int_{\{\widetilde{g}<0\}} \widetilde{g}\,\mathrm{d}\mu + \delta$$

for a dynamically configurable step error margin $\delta > 0$ where $\Delta > 0$ is the trust region radius and D satisfies $\mu(D) \leq \Delta$. Our goal is to determine loose error bounds ω_τ, ω_g, and ω_f such that

$$\int_{\{g<0\}\cup\{\widetilde{g}<0\}} |e_g| \, \mathrm{d}\mu \leq \omega_\tau, \tag{3.8}$$

$$|e_{g,D}| \leq \omega_g, \tag{3.9}$$

$$|e_{f,U}| \leq \omega_f, \tag{3.10}$$

$$|e_{f,V}| \leq \omega_f. \tag{3.11}$$

ensures the approximate correctness of both stationarity and step acceptance tests.

3.1.2.1 INSTATIONARITY AND STATIONARITY TESTING

The two quantities that control the trust region loop are *instationarity* and *step quality*. The gradient of the objective enters into both. Notably, it enters into the termination criterion, which is $-\mathscr{C}_1(F,U) \leq \varepsilon$. Let subsequently

$$\tau(f) := -\int_X \min\{0, f\} \, \mathrm{d}\mu \quad \forall f \in L^1(\Sigma, \mu).$$

It is evident that $\tau(g) = -\mathscr{C}_1(F,U)$. As a shorthand, we write $\tau := \tau(g)$ for the actual instationarity of F at U and $\widetilde{\tau} := \tau(\widetilde{g})$ for the approximation that we can actually calculate. The difference between the two is bounded by

$$\begin{aligned}
|\tau - \widetilde{\tau}| &= \left| \int_X \big(\min\{0,\ g\} - \min\{0,\ \widetilde{g}\} \big) \, \mathrm{d}\mu \right| \\
&\leq \int_X \underbrace{\big| \min\{0,\ g\} - \min\{0,\ \widetilde{g}\} \big|}_{=0 \text{ on } \{g\geq 0\}\cap\{\widetilde{g}\geq 0\}} \, \mathrm{d}\mu \\
&\leq \int_{\{g<0\}\cup\{\widetilde{g}<0\}} \big| g - \widetilde{g} \big| \, \mathrm{d}\mu \\
&= \int_{\{g<0\}\cup\{\widetilde{g}<0\}} |e_g| \, \mathrm{d}\mu \\
&\overset{(3.8)}{\leq} \omega_\tau.
\end{aligned}$$

We introduce an error tuning parameter $\xi_\tau \in (0,\ 1)$ that controls the permissible relative instationarity error and set the error bound ω_τ to

$$\omega_\tau(\widetilde{g}) := \xi_\tau \cdot \max\big\{\varepsilon,\ \tau(\widetilde{g}) - \varepsilon\big\} \geq \xi_\tau \cdot \varepsilon > 0. \tag{3.12}$$

This term is bounded from below and can be used as a bound oracle in Algorithm 1. We note that

$$\int_{\{g<0\}\cup\{\widetilde{g}<0\}} |e_g| \, \mathrm{d}\mu \leq \omega_\tau(\widetilde{g})$$

implies

$$\begin{aligned}
\tau(\widetilde{g}) \leq \varepsilon \implies \underbrace{-\mathscr{C}_1(F,U)}_{=\tau(g)} &\leq \varepsilon + \underbrace{\omega_\tau(\widetilde{g})}_{=\xi_\tau\cdot\varepsilon} && = (1+\xi_\tau)\cdot\varepsilon;\\
\tau(\widetilde{g}) > \varepsilon \implies -\mathscr{C}_1(F,U) &\geq \tau(\widetilde{g}) - \omega_\tau(\widetilde{g}) = \begin{cases} \tau(\widetilde{g}) - \xi_\tau\cdot\varepsilon & \text{if } \tau(\widetilde{g}) \leq 2\varepsilon \\ (1-\xi_\tau)\cdot\tau(\widetilde{g}) + \xi_\tau\cdot\varepsilon & \text{if } \tau(\widetilde{g}) > 2\varepsilon \end{cases}\\
&> \begin{cases} (1-\xi_\tau)\cdot\varepsilon & \text{if } \tau(\widetilde{g}) \leq 2\varepsilon \\ \varepsilon & \text{if } \tau(\widetilde{g}) > 2\varepsilon \end{cases}\\
&\geq (1-\xi_\tau)\cdot\varepsilon.
\end{aligned}$$

This means that the ε-stationarity test will not indicate stationarity unless the true instationarity is at most $(1+\xi_\tau)\cdot\varepsilon$ and will always indicate stationarity if it is less than or equal to $(1-\xi_\tau)\cdot\varepsilon$. This ensures that the ε-stationarity test still yields meaningful information.

We terminate the optimization loop as soon as $\tau(\widetilde{g}) \leq \varepsilon$. This means that for every evaluation taking place after the ε-stationarity test, we may assume strictly positive lower bounds $\tau(\widetilde{g}) > \varepsilon$ and $\tau(g) > (1-\xi_\tau)\cdot\varepsilon$. In subsequent bound oracles, we will therefore replace $\tau(\widetilde{g})$ with $\max\{\tau(\widetilde{g}),\ \varepsilon\}$ to accomodate cases where $\tau(\widetilde{g})$ is projected to be a value below ε but the bound does not become relevant unless $\tau(\widetilde{g}) > \varepsilon$. This helps to ensure that subsequent bound oracles are bounded away from zero.

3.1.2.2 STEP QUALITY AND ACCEPTANCE

Step quality calculation and acceptance testing only takes place if the ε-stationarity test fails. Therefore, we may assume $\widetilde{\tau} > \varepsilon$ and $\mathscr{C}_1(F,U) > (1-\xi_\tau)\cdot\varepsilon$. We also note that $\widetilde{\tau} > 0$ implies $\mu\big(\{\widetilde{g} < 0\}\big) > 0$. Because the step is found using a controlled unconstrained step-finding routine, we have

$$g_D \leq -\theta\cdot\min\left\{1,\ \frac{\Delta}{\mu(\{\widetilde{g} < 0\})}\right\}\cdot\widetilde{\tau} + \delta.$$

To simplify matters, we set a fixed relative error margin for step determination, i.e., we introduce an error tuning parameter $\xi_\delta \in (0,\ 1)$ and demand that

$$\delta \leq \xi_\delta\cdot\theta\cdot\min\left\{1,\ \frac{\Delta}{\mu(\{\widetilde{g} < 0\})}\right\}\cdot\widetilde{\tau} \tag{3.13}$$

which gives us

$$\begin{aligned}
g_D &\leq -(1-\xi_\delta)\cdot\theta\cdot\min\left\{1,\ \frac{\Delta}{\mu(\{\widetilde{g} < 0\})}\right\}\cdot\widetilde{\tau}\\
&< -(1-\xi_\delta)\cdot\theta\cdot\min\left\{1,\ \frac{\Delta}{\mu(\{\widetilde{g} < 0\})}\right\}\cdot\varepsilon\\
&< 0.
\end{aligned}$$

We define

$$\overline{\tau}_D(\widetilde{g}) := (1-\xi_\delta)\cdot\theta\cdot\min\left\{1,\ \frac{\Delta}{\mu(\{\widetilde{g} < 0\})}\right\}\cdot\max\{\tau(\widetilde{g}),\ \varepsilon\} > 0$$

as a lower bound on the absolute value of g_D for use in bound oracles. Because $\Delta > 0$, we can interpret the fraction as an infinity if $\mu(\{\widetilde{g} < 0\}) = 0$ due to numerical errors. This does not impact the minimum between the fraction and 1. Furthermore, we have prior guarantees that $\mu(\{\widetilde{g} < 0\}) > 0$.

Remark 3.1.6 (Practical Evaluation Order).
At first glance, it may appears counterintuitive to define a lower bound on $|g_D|$ that depends on $\widetilde{g}$. If we have access to $\widetilde{g}$, then we could simply find D and calculate $|g_D|$. This is particularly true considering that we primarily plan to use $\overline{\tau}_D(\widetilde{g})$ in bound oracles for functional evaluation, which are technically only used after $\widetilde{g}$ has already been determined.

We use an estimate of $|g_D|$ instead of the exact value due to a quirk in the evaluation order. For the optimization algorithm, we evaluate the gradient first, then determine the step, and then evaluate functional values at start and end of the step. However, many complex functionals, particularly those that involve the solutions of differential equations, require that we evaluate the functional value first and then derive the gradient from data gathered during functional evaluations.

This means that functional evaluation is a prerequisite for gradient evaluation. Therefore, a meaningful functional value bound should be set during gradient evaluation to maximize the chance of obtaining a reusable function value during gradient evaluation. Using a lower bound for $|g_D|$ instead of $|g_D|$ allows us to always calculate value and gradient error bounds together, regardless of which stage of evaluation we are currently in. ◁

For the step acceptance test, we have to control the difference between the true step quality

$$\rho := \frac{F(V) - F(U)}{\int_D g \,\mathrm{d}\mu}$$

and the approximate step quality

$$\widetilde{\rho} := \frac{f_V - f_U}{g_D}.$$

However, estimates of the error between the two are much easier to derive if we also introduce a *semi-approximate step quality* as an intermediate quantity between the two

$$\overline{\rho} := \frac{F(V) - F(U)}{g_D}.$$

The semi-approximate step quality uses exact functional values and the approximate projected descent. The advantage of working with the semi-approximate step quality is that the error estimation can be done in two steps. The difference between true and semi-approximate step quality is multiplicative and controlled by the gradient error, while the difference between semi-approximate and approximate step quality is additive and controlled by the functional value error.

The absolute difference between semi-approximate and approximate step quality is

$$\begin{aligned}
|\overline{\rho} - \widetilde{\rho}| &\leq \frac{1}{|g_D|} \cdot \left|F(V) - f_v - (F(U) - f_U)\right| \\
&\leq \frac{1}{|g_D|} \cdot \left(\left|F(V) - f_v\right| + \left|F(U) - f_U\right|\right) \\
&\overset{(3.10)}{\leq} \frac{1}{|g_D|} \cdot \left(\left|F(V) - f_V\right| + \omega_f\right) \\
&\overset{(3.11)}{\leq} \frac{2\omega_f}{|g_D|} \\
&\leq \frac{2\omega_f}{\overline{\tau}_D(\widetilde{g})}.
\end{aligned}$$

We can calculate $\widetilde{\rho}$ from approximate quantitites. By controlling the error bound ω_f, we can establish an interval within which $\overline{\rho}$ must lie. The change from an exact value to an interval requires that we define three decision thresholds instead of the usual two for our trust region algorithm:

- $\sigma_0 \in (0,\ 1)$ is the *acceptance threshold*, i.e., we accept steps if the lower bound for $\overline{\rho}$ is above σ_0;
- $\sigma_1 \in (\sigma_0,\ 1)$ is the *rejection threshold*, i.e., we reject unaccepted steps if the upper bound for $\overline{\rho}$ is below σ_1;
- $\sigma_2 > \sigma_0$ is the *trust region expansion threshold*, i.e., we expand the trust region if a step is accepted and satisfies $\widetilde{\rho} \geq \sigma_2$.

By separating the acceptance and rejection thresholds, we ensure that once the upper and lower bounds on $\overline{\rho}$ are less than $\frac{\sigma_1 - \sigma_0}{2}$ apart, an acceptance or rejection decision can always be made. We can immediately ensure this by enforcing $\frac{2\omega_f}{|g_D|} \leq \frac{\sigma_1 - \sigma_0}{2}$, which is equivalent to $\omega_f \leq |g_D| \cdot \frac{\sigma_1 - \sigma_0}{4}$. However, this error bound is likely too strict. An acceptance decision can be made if $\overline{\rho}$ is sufficiently separated from *either* decision threshold. Therefore, we can relax this error bound by only demanding that

$$\frac{2\omega_f}{|g_D|} \leq \frac{2\omega_f}{\overline{\tau}_D(\widetilde{g})} \leq \max\{\widetilde{\rho} - \sigma_0,\ \sigma_1 - \widetilde{\rho}\}$$

where $\widetilde{\rho}$ is a guess for the approximate step quality. This gives us an $\widetilde{\rho}$- and $\widetilde{g}$-dependent functional error bound of

$$\omega_f(\widetilde{\rho}, \widetilde{g}) := \overline{\tau}_D(\widetilde{g}) \cdot \frac{\max\{\widetilde{\rho} - \sigma_0,\ \sigma_1 - \widetilde{\rho}\}}{2}. \tag{3.14}$$

This is our bound oracle for functional evaluation. We note that $\omega_f(\widetilde{\rho}, \widetilde{g})$ assumes its minimal value over $\widetilde{\rho}$ at $\widetilde{\rho} = \frac{\sigma_0 + \sigma_1}{2}$ where $\omega_f(\widetilde{\rho}, \widetilde{g}) = \overline{\tau}_D(\widetilde{g}) \cdot \frac{\sigma_1 - \sigma_0}{4} > 0$. Because $\overline{\tau}_D(\widetilde{g})$ is bounded from below, so is the bound oracle.

Next, we have to consider the difference between true step quality and semi-approximate step quality. By definition, we have

$$\overline{\rho} = \frac{\int_D g \, \mathrm{d}\mu}{g_D} \cdot \rho.$$

The ratio is well-defined because g_D is bounded away from zero whenever step quality becomes a concern. Using the previously introduced error symbol $e_{g,D}$, we have

$$\left| g_D - \int_D g \, \mathrm{d}\mu \right| = |e_{g,D}| \leq \omega_{g,D}.$$

To avoid division by zero in the true step quality, we enforce that

$$\omega_{g,D} \leq \xi_g \cdot \overline{\tau}_D(\widetilde{g}) \leq \xi_g \cdot |g_D| \tag{3.15}$$

for some tuning parameter $\xi_g \in (0,\ 1)$. Once more, the use of a lower bound for $|g_D|$ allows us to calculate bounds for $|e_{g,D}|$ without having to know D. We define the $\widetilde{g}$-dependent error bound

$$\omega_{g,D}(\widetilde{g}) \coloneqq \xi_g \cdot \overline{\tau}_D(\widetilde{g}) \geq \xi_g \cdot \overline{\tau}_D(0) > 0.$$

Enforcing this error bound ensures that

$$g_D < 0 \implies \int_D g \, \mathrm{d}\mu \leq g_D + e_{g,D} \leq (1 - \xi_g) \cdot g_D < 0.$$

This also guarantees that $\overline{\rho}$ and ρ have the same sign because they have the same numerator and denominators of the same sign.

With regard to the absolute value of ρ, we have

$$\begin{aligned} |\rho| &= \frac{|g_D|}{\left|\int_D g \, \mathrm{d}\mu\right|} \cdot |\overline{\rho}| \\ &\in \left[\frac{|g_D|}{|g_D| + |e_{g,D}|} \cdot |\overline{\rho}|,\ \frac{|g_D|}{|g_D| - |e_{g,D}|} \cdot |\overline{\rho}| \right] \\ &\subseteq \left[\frac{1}{1+\xi_g} \cdot |\overline{\rho}|,\ \frac{1}{1-\xi_g} \cdot |\overline{\rho}| \right]. \end{aligned}$$

Every accepted step satisfies $\overline{\rho} \geq \sigma_0$, which implies

$$\rho \geq \frac{1}{1+\xi_g} \cdot \overline{\rho} \geq \frac{\sigma_0}{1+\xi_g}.$$

Therefore, every step that is accepted based on the semi-approximate acceptance criterion also has a true step quality that satisfies an acceptance threshold that is lowered by a factor of $\frac{1}{1+\xi_g}$. Every rejected step, on the other hand, satisfies $\overline{\rho} \leq \sigma_1$ and therefore

$$\rho \leq \frac{1}{1-\xi_g} \cdot \overline{\rho} \leq \frac{\sigma_1}{1-\xi_g}.$$

Here, we have to demand that $\frac{\sigma_1}{1-\xi_g} < 1$. Otherwise, it is not guaranteed that $\rho \to 1$ always yields an acceptable step. If we do not ensure this, then an infinite sequence of step rejections is possible. To avoid this, we have to choose $\xi_g \in (0,\ 1-\sigma_1) \subset (0,\ 1)$. This ensures that sufficiently good steps are always acceptable according to the semi-approximate acceptance criterion.

The exact way to incorporate the upper bound on $\omega_{g,D}$ with the instationarity bound oracle depends on the type of gradient error control and will be discussed later.

3.1.2.3 SUMMARY

In summary, we have introduced three error tuning parameters:

- $\xi_\tau \in (0,\ 1)$ is a relative error margin for instationarity;
- $\xi_\delta \in (0,\ 1)$ is a relative error margin for step finding;
- $\xi_g \in (0,\ 1-\sigma_1)$ is a relative error margin for the projected descent.

We have reformulated the step acceptance test in terms of three decision thresholds:

- the acceptance threshold $\sigma_0 \in (0,\ 1)$;
- the rejection threshold $\sigma_1 \in (\sigma_0,\ 1)$;
- the trust region expansion threshold $\sigma_2 > \sigma_0$.

To ensure that our algorithmic decision-making is sound, we enforce the following error bounds:

$$\int_{\{g<0\}\cup\{\widetilde{g}<0\}} |e_g|\,\mathrm{d}\mu \leq \omega_\tau(\widetilde{g}), \tag{3.12}$$

$$|\delta| \leq \xi_\delta \cdot \theta \cdot \min\left\{1,\ \frac{\Delta}{\mu(\{\widetilde{g}<0\})}\right\} \cdot \tau(\widetilde{g}), \tag{3.13}$$

$$|e_{f,U/V}| \leq \omega_f(\widetilde{\rho}, \widetilde{g}), \tag{3.14}$$

$$|e_{g,D}| \leq \omega_{g,D}(\widetilde{g}), \tag{3.15}$$

where δ is the step finding error margin, $\widetilde{g}$ is the approximate step quality, $\tau(\widetilde{g})$ is the apparent instationarity, and $\widetilde{\rho}$ is the apparent step quality.

Equations (3.12), (3.14) and (3.15) are cases where the error bound itself depends on the outcome of the calculation whose error is being bounded. Therefore, these three bounds each require an evaluation feedback loop similar to Algorithm 1 on page 215.

Equations (3.12) and (3.15) are a special case because they are both bounds on the gradient evaluation error and are therefore not strictly independent.

When all of these error-bounds are applicable and satisfied, we argue the validity of the trust region loop as follows:

1. Equation (3.12) ensures the validity of the instationarity test, which either causes loop termination or guarantees sufficient potential for descent;
2. Equation (3.13) guarantees that the step causes sufficient apparent descent;
3. Equation (3.15) ensures that the apparent descent corresponds to sufficient true descent and links the true step quality to the semi-approximate step quality;
4. Equation (3.14) ensures the correctness of the step acceptance test by linking the semi-approximate and approximate step quality.

3.1.2.4 EVALUATORS AND BOUND ORACLES

In order for us to use the bound oracles developed in this section in Algorithm 1, they must be paired with appropriate evaluators. In addition, we must couple the bound oracles from Equations (3.12) and (3.15) into a single bound oracle. The precise bound oracle is also different based on whether the gradient evaluator is L^1- or L^∞-controlled.

If the gradient evaluation routine is L^1-controlled, then we obtain an estimate of the form

$$\int_X |e_g| \, \mathrm{d}\mu \leq \omega_{g,L^1}(\widetilde{g})$$

where $\omega_{g,L^1}(\widetilde{g})$ is the currently unknown bound oracle for L^1-controlled evaluation.

Of course, we have

$$\begin{aligned}
\int_{\{g<0\}\cup\{\widetilde{g}<0\}} |e_{g,i}| \, \mathrm{d}\mu &\leq \int_X |e_g| \, \mathrm{d}\mu \\
&\leq \omega_{g,L^1}(\widetilde{g}), \\
|e_{g,D}| &= \left| \int_D e_g \, \mathrm{d}\mu \right| \\
&\leq \int_D |e_g| \, \mathrm{d}\mu \\
&\leq \int_X |e_g| \, \mathrm{d}\mu \\
&\leq \omega_{g,L^1}(\widetilde{g}),
\end{aligned}$$

which means that it is sufficient to choose

$$\omega_{g,L^1}(\widetilde{g}) := \min\{\omega_\tau(\widetilde{g}),\ \omega_{g,D}(\widetilde{g})\}.$$

and pair the bound oracle with with an evaluator that yields the approximate gradient density function alongside an estimate of the L^1 error.

If the gradient evaluation routine is L^∞-controlled, then we obtain an estimate of the form

$$\|e_{g,i}\|_{L^\infty(\Sigma,\mu)} \leq \omega_{g,L^\infty}(\widetilde{g}).$$

We then have

$$\begin{aligned}
\int_{\{g<0\}\cup\{\widetilde{g}<0\}} |e_g| \, \mathrm{d}\mu &\leq \omega_{g,L^\infty}(\widetilde{g}) \cdot \mu\big(\{g<0\}\cup\{\widetilde{g}<0\}\big) \\
&\leq \omega_{g,L^\infty}(\widetilde{g}) \cdot \mu\Big(\big\{\widetilde{g} < \omega_{g,L^\infty}(\widetilde{g})\big\}\Big)
\end{aligned}$$

and

$$\int_D |e_g| \, \mathrm{d}\mu \leq \omega_{g,L^\infty}(\widetilde{g}) \cdot \mu(D).$$

Therefore, for L^∞-controlled gradient evaluation methods, we use a bound oracle that uses the error estimate η in addition to the approximate gradient density:

$$\omega_{g,L^\infty}(\widetilde{g},\eta) := \min\left\{ \frac{\omega_\tau(\widetilde{g})}{\mu\big(\{\widetilde{g}<\eta\}\big)},\ \frac{\omega_{g,D}(\widetilde{g})}{\Delta} \right\}.$$

Algorithm 2 Gradient evaluator for set functional $F\colon \Sigma/_{\sim_\mu} \to \mathbb{R}$

Require:
- $q \in \{1, \infty\}$, $U \in \Sigma/_{\sim_\mu}$, $\varepsilon > 0$, $\xi_\tau \in (0,\ 1)$, $\xi_g \in (0,\ 1)$, $\xi_\delta \in (0,\ 1)$, $\Delta > 0$ with $\Delta \le \mu(X)$, $\theta \in (0,\ 1]$;
- $\beta_0 \in \mathbb{R}_{>0} \cup \{\infty\}$, $\xi \in (0,\ 1)$;
- $\widetilde{g}$ is an L^q-controlled gradient evaluation method for F.

Ensure: Yields $(g^*, e^*) \in L^1(\Sigma, \mu) \times \mathbb{R}_{\ge 0}$ with $\|g - g^*\|_{L^q} \le e^*$,

$$\left| \int_X \min\{0, g\}\,\mathrm{d}\mu - \int_X \min\{0, g^*\}\,\mathrm{d}\mu \right| \le \xi_\tau \cdot \max\left\{\varepsilon,\ -\int_X \min\{0, g^*\}\,\mathrm{d}\mu - \varepsilon\right\},$$

and

$$\left| \int_D g\,\mathrm{d}\mu - \int_D g^*\,\mathrm{d}\mu \right| \le -\xi_g \cdot \int_D g^*\,\mathrm{d}\mu$$

for all $D \in \Sigma/_{\sim_\mu}$ with $\mu(D) \le \Delta$ and

$$-\int_D g^*\,\mathrm{d}\mu \ge (1 - \xi_\delta) \cdot \theta \cdot \frac{\Delta}{\max\{\Delta,\ \mu(\{g^* < 0\})\}} \cdot \max\left\{-\int_X \min\{0, g^*\}\,\mathrm{d}\mu,\ \varepsilon\right\},$$

where $g \in L^1(\Sigma, \mu)$ is the density function of $\nabla F(U)$.

function $\omega_{q,\varepsilon,\Delta}(g,\ e)$ ▷ Parameter oracle
  $\widetilde{\tau} \leftarrow -\int_X \min\{0,\ g\}\,\mathrm{d}\mu$
  $\omega_\tau \leftarrow \xi_\tau \cdot \max\{\varepsilon,\ \widetilde{\tau} - \varepsilon\}$
  $\overline{\tau}_D \leftarrow (1 - \xi_\delta) \cdot \theta \cdot \frac{\Delta}{\max\{\Delta,\ \mu(\{g<0\})\}} \cdot \max\{\widetilde{\tau},\ \varepsilon\}$
  $\omega_D \leftarrow \xi_g \cdot \overline{\tau}_D$
  if $q = 1$ **then** ▷ Bounds for L^1-controlled gradients
    return $\min\{\omega_\tau,\ \omega_D\}$
  else ▷ Bounds for L^∞-controlled gradients
    return $\min\left\{\frac{\omega_\tau}{\mu(\{g<e\})},\ \frac{\omega_D}{\Delta}\right\}$
  end if
end function

procedure $\phi_U(\beta)$ ▷ Inner evaluator
  $(g,\ e) \leftarrow \widetilde{g}(U; \beta)$
  return $(g,\ e,\ e)$
end procedure

procedure EvalGrad$_q(\widetilde{g},\ U,\ \varepsilon,\ \Delta;\ \xi,\ \beta_0)$ ▷ Gradient evaluator
  $(g,\ e,\ e) \leftarrow$ ControlledEval$(\phi_U,\ \omega_{q,\varepsilon,\Delta},\ \beta_0,\ \xi)$
  return $(g,\ e)$
end procedure

The evaluator then simply needs to return a tuple of approximate gradient density and a copy of the L^∞ error bound as its value. The combined evaluator for the gradient is stated in Algorithm 2 on the preceding page.

Theorem 3.1.7 (Correctness of Algorithm 2).
Let $U \in \Sigma/_{\sim_\mu}$, let $F\colon \Sigma/_{\sim_\mu} \to \mathbb{R}$ be differentiable in U, let $\widetilde{g}$ be an L^q-controlled gradient evaluation method for F for $q \in \{1, \infty\}$. Let $\varepsilon > 0$, $\xi_\tau \in (0,\ 1)$, $\xi_g \in (0,\ 1)$, $\xi_\delta \in (0,\ 1)$, $\Delta > 0$ with $\Delta \leq \mu(X)$, $\theta \in (0,\ 1]$, $\beta_0 \in \mathbb{R}_{>0} \cup \{\infty\}$, and $\xi \in (0,\ 1)$.

Then the procedure $\textsc{EvalGrad}_q(\widetilde{g}, U, \varepsilon, \Delta; \xi, \beta_0)$ *as stated in Algorithm 2 terminates in finite time and yields an output tuple $(g^*,\ e^*) \in L^1(\Sigma,\mu) \times \mathbb{R}_{\geq 0}$ such that, if $g \in L^1(\Sigma,\mu)$ is the density function of $\nabla F(U)$, we have*

$$\|g - g^*\|_{L^q} \leq e^*,$$

$$\left|\int_X \min\{0,g\}\,\mathrm{d}\mu - \int_X \min\{0,g^*\}\,\mathrm{d}\mu\right| \leq \xi_\tau \cdot \max\left\{\varepsilon,\ -\int_X \min\{0,g^*\}\,\mathrm{d}\mu - \varepsilon\right\},$$

and

$$\left|\int_D g\,\mathrm{d}\mu - \int_D g^*\,\mathrm{d}\mu\right| \leq -\xi_g \cdot \int_D g^*\,\mathrm{d}\mu$$

for all $D \in \Sigma/_{\sim_\mu}$ with $\mu(D) \leq \Delta$ and

$$-\int_D g^*\,\mathrm{d}\mu \geq (1-\xi_\delta)\cdot\theta\cdot\frac{\Delta}{\max\{\Delta,\ \mu(\{g^* < 0\})\}}\cdot\max\left\{-\int_X \min\{0,g^*\}\,\mathrm{d}\mu,\ \varepsilon\right\}. \qquad \triangleleft$$

PROOF. Our argument is mostly based on the correctness of the CONTROLLEDEVAL procedure as demonstrated in Theorem 3.1.4. Accordingly, the proof can be split into three steps. First, we show that ϕ satisfies the requirements for an evaluation procedure in Algorithm 1. This is straightforward because $\widetilde{g}$ is an L^q-controlled evaluation method for the gradient density of F. According to Definition 3.1.1, this means that every output tuple $(g',\ e,\ e) \in L^1(\Sigma,\mu) \times \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0}$ of $\phi_U(\beta)$ satisfies

$$\|g - g'\|_{L^q} \leq e \leq \beta$$

where $g \in L^1(\Sigma,\mu)$ is the true density function of $\nabla F(U)$.

Next, we have to prove that $\omega_{q,\varepsilon,\Delta}$ is a suitable parameter oracle for Algorithm 1. This is also not difficult. First, we note that

$$\omega_\tau = \xi_\tau \cdot \max\{\varepsilon,\ \widetilde{\tau} - \varepsilon\} \geq \xi_\tau \cdot \varepsilon > 0.$$

We also have the lower bound

$$\begin{aligned}
\omega_D &= \xi_g \cdot (1-\xi_\delta)\cdot\theta\cdot\underbrace{\frac{\Delta}{\max\{\Delta,\ \mu(\{g<0\})\}}}_{\leq \mu(X)}\cdot\underbrace{\max\{\widetilde{\tau},\ \varepsilon\}}_{\geq\varepsilon} \\
&\geq \xi_g \cdot (1-\xi_\delta)\cdot\theta\cdot\frac{\Delta}{\mu(X)}\cdot\varepsilon \\
&> 0.
\end{aligned}$$

For all $(g,\ e) \in L^1(\Sigma,\mu) \times \mathbb{R}_{\geq 0}$, we therefore have

$$\begin{aligned}
\omega_{1,\varepsilon,\Delta}(g,e) &= \min\{\omega_\tau,\ \omega_D\} \\
&\geq \min\left\{\varepsilon,\ \frac{\xi_g \cdot (1-\xi_\delta)\cdot\theta\cdot\Delta}{\mu(X)}\cdot\varepsilon\right\} \\
\omega_{\infty,\varepsilon,\Delta}(g,e) &= \min\left\{\frac{\omega_\tau}{\mu(\{g<e\})},\ \frac{\omega_D}{\Delta}\right\} \\
&\geq \min\left\{\frac{\varepsilon}{\mu(X)},\ \frac{\xi_g\cdot(1-\xi_\delta)\cdot\theta}{\mu(X)}\cdot\varepsilon\right\}.
\end{aligned}$$

In either case, $\omega_{q,\varepsilon,\Delta}$ is bounded below by a constant $\overline{e}_q > 0$ that depends only on q. We then have

$$e \leq \overline{e}_q \implies e \leq \omega_{q,\varepsilon,\Delta}(g,e)$$

for all output tuples (g,e,e) of ϕ, regardless of the error bound β. Because there are no additional varying parameters, this means that $\omega_{q,\varepsilon,\Delta}$ is suitable as a bound oracle for Algorithm 1.

The overall procedure EVALGRAD$_q$ invokes Algorithm 1 with ϕ_U as evaluator, $\omega_{q,\varepsilon,\Delta}$ as bound oracle, ξ as decay rate, and β_0 as initial bound. This terminates in finite time according to Theorem 3.1.4 and yields an output tuple $(g^*,\ e^*,\ e^*) \in L^1(\Sigma,\mu) \times \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0}$ of ϕ such that

$$e^* \leq \omega_{q,\varepsilon,\Delta}(g^*,e^*).$$

The last two entries of any output tuple of ϕ are equal. Let subsequently $g \in L^1(\Sigma,\mu)$ be the true density function of $\nabla F(U)$. The inequality

$$\|g - g^*\|_{L^q} \leq e^*$$

follows directly from the fact that $\widetilde{g}$ is an L^q-controlled gradient density evaluation method for F. For the remaining inequalities, we have to make different arguments based on the value of q.

Case 1 ($q = 1$). If $q = 1$, then the output of $\omega_{q,\varepsilon,\Delta}$ satisfies

$$\omega_{q,\varepsilon,\Delta}(g^*,e^*) \leq \omega_\tau = \xi_\tau \cdot \max\left\{\varepsilon,\ -\int_X \min\{0,\ g^*\}\,\mathrm{d}\mu - \varepsilon\right\}$$

and therefore, we have

$$\begin{aligned}
&\left|\int_X \min\{0,\ g\}\,\mathrm{d}\mu - \int_X \min\{0,\ g^*\}\,\mathrm{d}\mu\right| \\
&\quad\leq \int_X \left|\min\{0,\ g\} - \min\{0,\ g^*\}\right|\mathrm{d}\mu \\
&\quad\leq \int_X |g - g^*|\,\mathrm{d}\mu \\
&\quad= \|g - g^*\|_{L^1} \\
&\quad\leq e^* \\
&\quad\leq \omega_{1,\varepsilon,\Delta}(g^*,e^*) \\
&\quad\leq \xi_\tau \cdot \max\left\{\varepsilon,\ -\int_X \min\{0,\ g^*\}\,\mathrm{d}\mu - \varepsilon\right\}.
\end{aligned}$$

We also have

$$\omega_{1,\varepsilon,\Delta}(g^*,e^*) \le \omega_D = \xi_g \cdot (1-\xi_\delta) \cdot \frac{\Delta}{\max\{\Delta,\ \mu(\{g^*<0\})\}} \cdot \max\left\{-\int_X \min\{0,\ g^*\}\,\mathrm{d}\mu,\ \varepsilon\right\},$$

which implies that for every $D \in \Sigma/_{\sim_\mu}$, we have

$$\begin{aligned}\left|\int_D g\,\mathrm{d}\mu - \int_D g^*\,\mathrm{d}\mu\right| &\le \|g-g^*\|_{L^1} \\ &\le \omega_{1,\varepsilon,\Delta}(g^*,e^*) \\ &\le \omega_D.\end{aligned}$$

Therefore, if

$$-\int_D g^*\,\mathrm{d}\mu \ge (1-\xi_\delta) \cdot \frac{\Delta}{\max\{\Delta,\ \mu(\{g^*<0\})\}} \cdot \max\left\{-\int_X \min\{0,\ g^*\}\,\mathrm{d}\mu,\ \varepsilon\right\},$$

then

$$\left|\int_D g\,\mathrm{d}\mu - \int_D g^*\,\mathrm{d}\mu\right| \le \omega_D \le -\xi_g \cdot \int_D g^*\,\mathrm{d}\mu.$$

◁

Case 2 ($q=\infty$). If $q=\infty$, then $\omega_{q,\varepsilon,\Delta}(g^*,e^*)$ satisfies

$$\omega_{\infty,\varepsilon,\Delta}(g^*,e^*) \le \frac{\omega_\tau}{\mu\big(\{g^*<e^*\}\big)}$$

and we have

$$\|g-g^*\|_{L^\infty} \le e^* \le \omega_{\infty,\varepsilon,\Delta}(g^*,e^*).$$

Because e^* bounds the absolute difference between g and g^* almost everywhere, we have

$$\{g<0\} \cup \{g^*<0\} \subseteq_\mu \{g^*<e^*\}.$$

We therefore find that

$$\begin{aligned}&\left|\int_X \min\{0,\ g\}\,\mathrm{d}\mu - \int_X \min\{0,\ g^*\}\,\mathrm{d}\mu\right| \\ &\quad\le \int_X \underbrace{\left|\min\{0,\ g\} - \min\{0,\ g^*\}\right|}_{=0 \text{ outside of } \{g<0\}\cup\{g^*<0\}}\,\mathrm{d}\mu \\ &\quad\le \int_{\{g<0\}\cup\{g^*<0\}} |g-g^*|\,\mathrm{d}\mu \\ &\quad\le \|g-g^*\|_{L^\infty} \cdot \mu\big(\{g<0\}\cup\{g^*<0\}\big) \\ &\quad\le \|g-g^*\|_{L^\infty} \cdot \mu\big(\{g^*<e^*\}\big) \\ &\quad\le \omega_\tau \\ &\quad= \xi_\tau \cdot \max\left\{\varepsilon,\ -\int_X \min\{0,\ g^*\}\,\mathrm{d}\mu - \varepsilon\right\}.\end{aligned}$$

We also have

$$\omega_{\infty,\varepsilon,\Delta}(g^*,e^*) \le \frac{\omega_D}{\Delta}.$$

For $D \in \Sigma/_{\sim_\mu}$ with $\mu(D) \leq \Delta$, we have

$$\begin{aligned}\left|\int_D g \,\mathrm{d}\mu - \int_D g^* \,\mathrm{d}\mu\right| &\leq \|g - g^*\|_{L^\infty} \cdot \mu(D) \\ &\leq \omega_{\infty,\varepsilon,\Delta}(g^*, e^*) \cdot \Delta \\ &\leq \omega_D.\end{aligned}$$

Therefore, if

$$-\int_D g^* \,\mathrm{d}\mu \geq (1-\xi_\delta) \cdot \frac{\Delta}{\max\{\Delta,\ \mu(\{g^* < 0\})\}} \cdot \max\left\{-\int_X \min\{0,\ g^*\} \,\mathrm{d}\mu,\ \varepsilon\right\},$$

then

$$\left|\int_D g \,\mathrm{d}\mu - \int_D g^* \,\mathrm{d}\mu\right| \leq \omega_D \leq -\xi_g \cdot \int_D g^* \,\mathrm{d}\mu. \qquad \square$$

Rather than writing a functional evaluation loop that is executed twice, we make functional evaluation an implicit part of an evaluator for the step quality. This evaluator evaluates the objective at both start and end point of the step, calculates the approximate step quality, and returns the step quality alongside an aggregate error estimate. The resulting evaluator for the step quality is stated in Algorithm 3 on the next page.

Theorem 3.1.8 (Correctness of Algorithm 3).
Let $U \in \Sigma/_{\sim_\mu}$, $D \in \Sigma/_{\sim_\mu}$, $\varepsilon > 0$, $\xi_\delta \in (0,\ 1)$, $\sigma_0 \in (0,\ 1)$, $\sigma_1 \in (\sigma_0,\ 1)$, $\theta \in (0,\ 1]$, $\Delta > 0$, $\beta_0 \in \mathbb{R}_{>0} \cup \{\infty\}$, $\xi \in (0,\ 1)$, and $g \in L^1(\Sigma, \mu)$ such that $\mu(D) \leq \Delta$ and $\int_D g \,\mathrm{d}\mu < 0$. Let $\widetilde{F} \colon \Sigma/_{\sim_\mu} \times (\mathbb{R}_{>0} \cup \{\infty\}) \to \mathbb{R} \times \mathbb{R}_{\geq 0}$ be a controlled evaluator for $F \colon \Sigma/_{\sim_\mu} \to \mathbb{R}$. Then Algorithm 3 terminates in finite time and returns a tuple $(\rho,\ e_\rho) \in \mathbb{R} \times \mathbb{R}_{\geq 0}$ such that

$$\left|\frac{F(U \mathbin{\triangle} D) - F(U)}{\int_D g \,\mathrm{d}\mu} - \rho\right| \leq e_\rho \leq \max\{\rho - \sigma_0,\ \sigma_1 - \rho\}. \qquad \triangleleft$$

Proof. Again, we rely mostly on the correctness of Algorithm 1, which we use without additional varying parameters p. First, we show that ω is a suitable bound oracle. This is a relatively simple matter, because we have a guarantee that $\sigma_0 < \sigma_1$. Therefore, we have

$$\omega(\rho) = \max\{\rho - \sigma_0,\ \sigma_1 - \rho\} \geq \frac{\sigma_1 - \sigma_0}{2} > 0,$$

which demonstrates that ω is bounded below and therefore a suitable parameter oracle.

Next, we show that $\phi_{g,U,D}$ is a suitable inner evaluator. Because $\widetilde{F}$ is a controlled evaluator for F, the output tuples $(x_1,\ e_1)$ and $(x_2,\ e_2)$ from $\widetilde{F}(U,\ -\frac{\beta}{2} \cdot \int_D g \,\mathrm{d}\mu)$ and $\widetilde{F}(U \mathbin{\triangle} D,\ -\frac{\beta}{2} \cdot \int_D g \,\mathrm{d}\mu)$ satisfy

$$\begin{aligned}\left|F(U) - x_1\right| &\leq e_1 \leq -\frac{\beta}{2} \cdot \int_D g \,\mathrm{d}\mu, \\ \left|F(U \mathbin{\triangle} D) - x_2\right| &\leq e_2 \leq -\frac{\beta}{2} \cdot \int_D g \,\mathrm{d}\mu.\end{aligned}$$

Algorithm 3 Step quality evaluator for a set functional $F\colon \Sigma/_{\sim_\mu} \to \mathbb{R}$

Require:
- $U \in \Sigma/_{\sim_\mu}$, $D \in \Sigma/_{\sim_\mu}$, $\varepsilon > 0$, $\xi_\delta \in (0,\ 1)$, $\sigma_0 \in (0,\ 1)$, $\sigma_1 \in (\sigma_0,\ 1)$, $\theta \in (0,\ 1]$, $\Delta > 0$ with $\Delta \geq \mu(D)$;
- $\beta_0 \in \mathbb{R}_{>0} \cup \{\infty\}$, $\xi \in (0,\ 1)$;
- $g \in L^1(\Sigma, \mu)$ such that
$$\int_D g \,\mathrm{d}\mu < 0;$$
- $\widetilde{F}\colon \Sigma/_{\sim_\mu} \times (\mathbb{R} \cup \{\infty\}) \to \mathbb{R} \times \mathbb{R}_{\geq 0}$ is a controlled evaluator for $F\colon \Sigma/_{\sim_\mu} \to \mathbb{R}$.

Ensure: EVALRHO yields $(\rho,\ e_\rho) \in \mathbb{R} \times \mathbb{R}_{\geq 0}$ such that
$$\left| \frac{F(U \mathbin{\triangle} D) - F(U)}{\int_D g \,\mathrm{d}\mu} - \rho \right| \leq e_\rho \leq \max\{\rho - \sigma_0,\ \sigma_1 - \rho\}.$$

1: **function** $\omega(\rho)$ ▷ Parameter oracle
2: **return** $\max\{\rho - \sigma_0,\ \sigma_1 - \rho\}$
3: **end function**

4: **procedure** $\phi_{g,U,D}(\beta)$ ▷ Inner evaluator
5: $(x_1,\ e_1) \leftarrow \widetilde{F}\left(U; -\frac{\beta}{2} \cdot \int_D g \,\mathrm{d}\mu\right)$
6: $(x_2,\ e_2) \leftarrow \widetilde{F}\left(U \mathbin{\triangle} D; -\frac{\beta}{2} \cdot \int_D g \,\mathrm{d}\mu\right)$
7: **return** $\left(\frac{x_2 - x_1}{\int_D g \,\mathrm{d}\mu},\ \frac{e_1 + e_2}{-\int_D g \,\mathrm{d}\mu}\right)$
8: **end procedure**

9: **procedure** EVALRHO($g, U, D; \beta_0, \xi$) ▷ Quality evaluator
10: **return** CONTROLLEDEVAL($\phi_{g,U,D}, \omega, \beta_0, \xi$)
11: **end procedure**

Therefore, we have
$$e_1 + e_2 \leq -\left(\frac{\beta}{2} + \frac{\beta}{2}\right) \cdot \int_D g \,\mathrm{d}\mu = -\beta \cdot \int_D g \,\mathrm{d}\mu,$$
which implies that
$$\frac{e_1 + e_2}{-\int_D g \,\mathrm{d}\mu} \leq \beta.$$
This demonstrates that ϕ is suitable as an evaluation procedure.

In addition, every output tuple $(\rho,\ e_\rho)$ of $\phi_{g,U,D}$ satisfies
$$\begin{aligned}
\left| \frac{F(U \mathbin{\triangle} D) - F(U)}{\int_D g \,\mathrm{d}\mu} - \rho \right| &= \left| \frac{F(U \mathbin{\triangle} D) - F(U)}{\int_D g \,\mathrm{d}\mu} - \frac{x_2 - x_1}{\int_D g \,\mathrm{d}\mu} \right| \\
&\leq \frac{\left|F(U \mathbin{\triangle} D) - x_2\right| + \left|F(U) - x_1\right|}{\left|\int_D g \,\mathrm{d}\mu\right|} \\
&= \frac{e_2 + e_1}{-\int_D g \,\mathrm{d}\mu} \\
&= e_\rho.
\end{aligned}$$

Finally, we turn to the main procedure EVALRHO. It obtains an output tuple $(\rho,\ e_\rho)$ from Algorithm 1 with evaluator $\phi_{g,U,D}$, parameter oracle ω, initial bound β_0, and decay rate ξ. This tuple is an output tuple from $\phi_{g,U,D}$ with

$$e_\rho \leq \omega(\rho) = \max\{\rho - \sigma_0,\ \rho - \sigma_1\}.$$

Therefore, we have

$$\left|\frac{F(U \triangle D) - F(U)}{\int_D g \,\mathrm{d}\mu} - \rho\right| \leq e_\rho \leq \max\{\rho - \sigma_0,\ \rho - \sigma_1\}. \qquad \square$$

3.1.3 Trust Region Loop

We can now define the main trust region loop. This loop is essentially identical to other trust-region loops except for the use of the error controlled evaluation loop that we had stated in Algorithm 1. To simplify the algorithm statement, we remove some of the preconditions for the algorithm to Assumption 3.1.9.

Assumption 3.1.9 (Preconditions for Algorithm 4).
Let

(1) $(X,\ \Sigma,\ \mu)$ be finite atomless measure spaces;

(2) $F\colon \Sigma/_{\sim_\mu} \to \mathbb{R}$ be uniformly continuously differentiable and bounded below;

(3) $\widetilde{F}\colon \Sigma/_{\sim_\mu} \times \left(\mathbb{R}_{>0} \cup \{\infty\}\right) \to \mathbb{R} \times \mathbb{R}_{\geq 0}$ be an error-controlled evaluation method for the functional F;

(4) $\widetilde{g}\colon \Sigma/_{\sim_\mu} \times \left(\mathbb{R}_{>0} \cup \{\infty\}\right) \to L^1(\Sigma,\mu) \times \mathbb{R}_{\geq 0}$ be an L^q-controlled gradient evaluation routine for F with $q \in \{1, \infty\}$;

(5) $\mathscr{S}\colon L^1(\Sigma,\mu) \times \mathbb{R}_{>0} \times \mathbb{R}_{>0} \to \Sigma/_{\sim_\mu} \times \mathbb{R}_{\geq 0}$ be a controlled unconstrained step-finding routing with quality $\theta > 0$;

(6) $\sigma_0,\ \sigma_1,\ \sigma_2 \in \mathbb{R}$ such that $0 < \sigma_0 < \sigma_2$ and $\sigma_0 < \sigma_1 < 1$;

(7) $\xi_\tau \in (0,\ 1)$, $\xi_\delta \in (0,\ 1)$, and $\xi_g \in (0,\ 1 - \sigma_1)$;

(8) $U_0 \in \Sigma/_{\sim_\mu}$, $\Delta_0 \in \left(0,\ \mu(X)\right]$. $\triangleleft$

We note that these assumptions differ from those in [HLS22]. This is partially because of the inclusion of error control. However, we have also added the assumption that the objective is bounded below. In [HLS22], this is inferred from the finiteness of the measure space by using the curvature suboptimality estimator (Proposition 2.4.8). This required a Lipschitz continuously differentiable objective. By allowing for alternate ways of proving boundedness, we can generalize the correctness result to any uniformly continuously differentiable objective functional.

Under Assumption 3.1.9, we can formulate the main trust region loop for unconstrained problems, which is stated in Algorithm 4 on the next page. To simplify notation, we define

$$\tau(g) := -\int_X \min\{0, g\} \,\mathrm{d}\mu \quad \forall g \in L^1(\Sigma,\mu),$$

which is in line with the definition of τ in Section 3.1.2. We stress that the parameter q in Algorithm 2 is determined by the type of error control provided

Algorithm 4 Controlled trust-region descent framework

Require: Let $(X,\ \Sigma,\ \mu)$, F, $\widetilde{F}$, $\widetilde{g}$, q, $\mathscr{S}$, σ_0, σ_1, σ_2, ε, ξ_τ, ξ_δ, ξ_g, U_0, Δ_0, ω_g, ϕ_g, ω_f, and ϕ_f satisfy Assumption 3.1.9 on the preceding page. Let $K \in \mathbb{N}_0 \cup \{\infty\}$.

Ensure: $k \in \mathbb{N}_0$ with $k \leq K$ and $U^* \in \Sigma/_{\sim_\mu}$ such that $\mathscr{C}_1(F, U^*) \geq -(1+\xi_\tau)\cdot\varepsilon$ or $k \geq K$.

$(g_0, e_{g,0}) \leftarrow \textsc{EvalGrad}_q\left(\widetilde{g}, U_0, \varepsilon, \Delta_0; \infty, \frac{1}{2}\right)$ ▷ Algorithm 2
$(j,\ k) \leftarrow (0,\ 0)$
while $k < K$ and $\tau(g_j) > \varepsilon$ **do**
    $D_j \leftarrow \mathscr{S}_1\left(g_j, \Delta_j, \xi_\delta \cdot \theta \cdot \min\left\{1, \frac{\Delta_j}{\mu(\{g_j < 0\})}\right\} \cdot \tau(g_j)\right)$
    $(\widetilde{\rho}_j,\ e_{\rho,j}) \leftarrow \textsc{EvalRho}\left(g_j, U_j, D_j; \infty, \frac{1}{2}\right)$ ▷ Algorithm 3
    if $\widetilde{\rho}_j - e_{\rho,j} \geq \sigma_0$ **then** ▷ Accept step
        $(U_{j+1}, k) \leftarrow (U_j \triangle D_j, k+1)$
        $\Delta_{j+1} \leftarrow \min\{2\Delta_j, \mu(X)\}$ **if** $\widetilde{\rho}_j \geq \sigma_2$ **else** Δ_j
    else ▷ Reject step
        $(U_{j+1}, \Delta_{j+1}) \leftarrow \left(U_j, \frac{\Delta_j}{2}\right)$
    end if
    $(g_{j+1}, e_{g,j+1}) \leftarrow \textsc{EvalGrad}_q\left(\widetilde{g}, U_{j+1}, \varepsilon, \Delta_{j+1}; \infty, \frac{1}{2}\right)$
    $j \leftarrow j+1$
end while
$U^* \leftarrow U_j$

by $\widetilde{g}$ and is not explicitly passed as an argument. We now prove that Algorithm 4 terminates and yields a correct result. This is primarily ensured by Theorem 3.1.4 and the error control results from Section 3.1.2.

Theorem 3.1.10 (Correctness of Algorithm 4).
Let (X, Σ, μ), F, $\widetilde{F}$, $\widetilde{g}$, q, $\mathscr{S}$, σ_0, σ_1, σ_2, ε, ξ_τ, ξ_δ, ξ_g, U_0, and Δ_0 satisfy Assumption 3.1.9. Then for every $K \in \mathbb{N}_0 \cup \{\infty\}$, Algorithm 4 terminates in finite time with either $k = K$ or $\mathscr{C}_1(F, U^*) \geq -(1+\xi_\tau)\cdot\varepsilon$. There exists $K_0 \in \mathbb{N}_0$ such that for $K \geq K_0$, the algorithm terminates with $\mathscr{C}_1(F, U^*) \geq -(1+\xi_\tau)\cdot\varepsilon$. ◁

PROOF. We note that $\mu(X) > 0$ is implied by the fact that otherwise, there would be no valid choice for $\Delta_0 \in \left(0,\ \mu(X)\right]$.

PART 1 (INSTATIONARITY ERROR). For each $j \in \mathbb{N}_0$ prior to termination, let $g_j^* \in L^1(\Sigma, \mu)$ be the exact density function of $\nabla F(U_j)$ with respect to μ.

The approximate gradient is always evaluated by using the procedure $\textsc{EvalGrad}_q$ from Algorithm 2 with the current solution U_j and Δ_j. According to Theorem 3.1.7, this guarantees that

$$\left|\tau(g_j^*) - \tau(g_j)\right| \leq \xi_\tau \cdot \max\{\varepsilon, \tau(g_j) - \varepsilon\}.$$

As we had discussed in Section 3.1.2, this means that

$$\begin{aligned} \tau(g_j) > \varepsilon &\implies -\mathscr{C}_1(F, U) > (1-\xi_\tau)\cdot\varepsilon, \\ \tau(g_j) \leq \varepsilon &\implies -\mathscr{C}_1(F, U) \leq (1+\xi_\tau)\cdot\varepsilon. \end{aligned}$$

Because the loop terminates if $\tau(g_j) \leq \varepsilon$, we can assume for all subsequent arguments about step quality calculation that $\tau(g_j) > \varepsilon$. According to Theorem 3.1.7, the gradient evaluation procedure then guarantees that

$$\left| \int_D g_j^* \, \mathrm{d}\mu - \int_D g_j \, \mathrm{d}\mu \right| \leq \xi_g \cdot \left(- \int_D g_j \, \mathrm{d}\mu \right) \tag{3.16}$$

holds for all steps $D \in \Sigma/_{\sim_\mu}$ with $\mu(D) \leq \Delta_j$ and

$$- \int_D g_j \, \mathrm{d}\mu \geq (1 - \xi_\delta) \cdot \theta \cdot \frac{\Delta_j}{\max\{\Delta_j, \mu(\{g_j < 0\})\}} \cdot \tau(g_j).$$

It also implies that the lower bound $\overline{\tau}_D$ on the projected descent that we had introduced in Section 3.1.2.2 satisfies

$$\overline{\tau}_D(g_j) = (1 - \xi_\delta) \cdot \theta \cdot \min\left\{1, \frac{\Delta_j}{\mu(\{g_j < 0\})}\right\} \cdot \tau(g_j).$$

Part 2 (Step Quality Error). In Line 4 of Algorithm 4, we determine the step $D_j \in \Sigma/_{\sim_\mu}$. Because $\mathscr{S}$ is a controlled step-finding method with quality θ, the projected descent satisfies

$$\begin{aligned} \int_{D_j} g_j \, \mathrm{d}\mu &\leq \theta \cdot \min\left\{1, \frac{\Delta_j}{\mu(\{g_j < 0\})}\right\} \cdot \left(-\tau(g_j)\right) + \delta \\ &= (1 - \xi_\delta) \cdot \theta \cdot \min\left\{1, \frac{\Delta_j}{\mu(\{g_j < 0\})}\right\} \cdot \left(-\tau(g_j)\right) \\ &= -\overline{\tau}_D(g_j) \end{aligned}$$

This bounds the apparent descent away from zero and ensures that the approximate step quality is well-defined. It also allows us to apply Algorithm 3 to evaluate the approximate step quality and Equation (3.16) to bound the gradient error on D_j. The latter means that we have

$$\left| \int_{D_j} g_j^* \, \mathrm{d}\mu - \int_{D_j} g_j \, \mathrm{d}\mu \right| \leq \xi_g \cdot \left(- \int_{D_j} g_j \, \mathrm{d}\mu \right).$$

We evaluate the approximate step quality $\widetilde{\rho}_j$ by using the procedure EvalRho as defined in Algorithm 3. According to Theorem 3.1.8, this guarantees that

$$\left| \frac{F(U_j \triangle D_j) - F(U)}{\int_D g \, \mathrm{d}\mu} - \widetilde{\rho}_j \right| \leq \max\{\widetilde{\rho}_j - \sigma_0, \sigma_1 - \widetilde{\rho}_j\}.$$

As we had discussed in Section 3.1.2.2, these guarantees imply that

$$\begin{aligned}
\widetilde{\rho}_j - e_{\rho_j} \geq \sigma_0 \implies \frac{F(U_j \triangle D_j) - F(U_j)}{\nabla F(U_j)(D_j)} &\geq \frac{1}{1+\xi_g} \cdot \frac{F(U_j \triangle D_j) - F(U_j)}{\int_{D_j} g_j \,\mathrm{d}\mu} \\
&\geq \frac{\widetilde{\rho}_j - e_{\rho,j}}{1+\xi_g} \\
&\geq \frac{\sigma_0}{1+\xi_g},
\end{aligned}$$

$$\begin{aligned}
\widetilde{\rho}_j - e_{\rho,j} < \sigma_0 &\implies \widetilde{\rho}_j + e_{\rho,j} \leq \sigma_1 \\
&\implies \frac{F(U_j \triangle D_j) - F(U_j)}{\nabla F(U_j)(D_j)} \leq \frac{1}{1-\xi_g} \cdot \frac{F(U_j \triangle D_j) - F(U_j)}{\int_{D_j} g_j \,\mathrm{d}\mu} \\
&\leq \frac{\widetilde{\rho}_j + e_{\rho,j}}{1-\xi_g} \\
&\leq \frac{\sigma_1}{1-\xi_g}.
\end{aligned}$$

For the latter of the two implications, we make use of the fact that $\widetilde{\rho}_j - e_{\rho,j} < \sigma_0$ implies $e_{\rho,j} > \widetilde{\rho}_j - \sigma_0$. However, because $e_{\rho,j} \leq \max\{\widetilde{\rho}_j - \sigma_0, \sigma_1 - \widetilde{\rho}_j\}$, we must then have $e_{\rho,j} \leq \sigma_1 - \widetilde{\rho}_j$, which implies that $\widetilde{\rho}_j + e_{\rho,j} \leq \sigma_1$. In both cases, the relation between the true and semi-approximate step quality follows from the gradient evaluation and step-finding error guarantees in the manner that we had derived in Section 3.1.2.

PART 3 (TRUST REGION BOUND). In Parts 1 and 2 we have established that the use of Algorithm 1 in conjunction with the error bound set for $\mathscr{S}$ and the main loop termination condition ensures Equations (3.12) to (3.15) (see Page 223) and all implications thereof, as discussed in those parts of the proof and Section 3.1.2. We subsequently consider these established.

We first note that Δ_j is evidently bounded below by $\Delta_j > 0$ and above by $\Delta_j \leq \mu(X)$. This can be shown inductively. In each iteration, the radius is either halved, which preserves both relations, or doubled and clamped to the upper bound $\mu(X)$, which also preserves both relations. To establish a tighter lower bound bound, we only have to look at the circumstances in which the trust region radius is halved, because this is the only circumstance in which the radius decreases.

The radius is halved if and only if the iterate fails the stationarity test and $\widetilde{\rho}_j - e_{\rho,j} < \sigma_0$. As we had discussed in Section 3.1.2 and Part 2, this implies $\rho_j \leq \frac{\sigma_1}{1-\xi_g}$. Conversely, if

$$\rho_j > \frac{\sigma_1}{1-\xi_g},$$

then the step is always accepted and can therefore never be reduced any further. We have to demonstrate that there is a lower bound on the step size below which every step is above this quality threshold.

Because F is uniformly continuously differentiable, according to Definition 2.4.4, for every $\varepsilon' > 0$, there exists $\delta > 0$ such that

$$\big(\nabla F(U) \underset{U \triangle V}{\ominus} \nabla F(V)\big)(D) \leq \varepsilon' \cdot \mu(D) \quad \forall U, V, D \in \Sigma/_{\sim_\mu} \colon \mu(U \triangle V) \leq \delta.$$

As a first step, we want to demonstrate that if $\Delta_j \leq \delta$, then

$$\left|F(U_j \triangle D_j) - F(U_j) - \nabla F(U_j)(D_j)\right| \leq \varepsilon' \cdot \mu(D_j) \leq \varepsilon' \cdot \Delta_j.$$

Because (X, Σ, μ) is finite and atomless, there exists a geodesic $\gamma\colon I \to \Sigma/_{\sim_\mu}$ that connects U_j with $U_j \triangle D_j$. Without loss of generality, let $I = [0,\ \mu(D_j)]$ with $\gamma(0) = U_j$ and $\gamma\big(\mu(D_j)\big) = U_j \triangle D_j$. Evidently, $C_\gamma = 1$ is a geodesic constant of γ.

Let $\varepsilon'' > 0$. According to the Taylor criterion (Definition 2.4.1), for each $t \in I$, there exists $R(t, \varepsilon'') > 0$ such that

$$\Big|F(V) - F\big(\gamma(t)\big) - \nabla F\big(\gamma(t)\big)\big(V \triangle \gamma(t)\big)\Big| \leq \varepsilon'' \cdot \mu\big(V \triangle \gamma(t)\big)$$

holds for all $V \in \Sigma/_{\sim_\mu}$ with $\mu\big(V \triangle \gamma(t)\big) \leq R(t, \varepsilon'')$.

Because $C_\gamma = 1$ is a geodesic constant of γ, we have $\mu\big(\gamma(s) \triangle \gamma(t)\big) = |s - t|$ for all $s, t \in I$, which means that

$$|s - t| \leq R(t, \varepsilon'') \implies \mu\big(\gamma(s) \triangle \gamma(t)\big) \leq R(t, \varepsilon'') \quad \forall s, t \in I.$$

Because I is a compact subset of $\mathbb{R}$, we can apply the Heine-Borel theorem to find a finite tuple of support points $(t_i)_{i \in [n_t]} \subseteq I$ with $n_t \in \mathbb{N}$ such that

$$I \subseteq \bigcup_{i=1}^{n_t} B_{R(t_i, \varepsilon'')}(t_i).$$

As we have discussed in more detail in the proof of Proposition 2.4.8, we may assume without loss of generality that the tuple $(t_i)_{i \in [n_t]}$ is strictly monotonically increasing and that

$$\begin{aligned} i > j &\implies t_i + R(t_i, \varepsilon'') > t_j + R(t_j, \varepsilon'') \quad \forall i, j \in [n_t], \\ i < j &\implies t_i - R(t_i, \varepsilon'') < t_j - R(t_j, \varepsilon'') \quad \forall i, j \in [n_t], \end{aligned}$$

which ensures that

$$\begin{aligned} 0 &\in B_{R(t_1, \varepsilon'')}(t_1), \\ \mu(D_j) &\in B_{R(t_{n_t}, \varepsilon'')}(t_{n_t}), \\ B_{R(t_i, \varepsilon')}(t_i) \cap B_{R(t_{i+1}, \varepsilon'')}(t_{i+1}) &\neq \emptyset \quad \forall i \in [n_t - 1]. \end{aligned}$$

We can now define a second support grid with support points $(s_i)_{i \in [n_t]_0} \subseteq I$ with

$$\begin{aligned} s_0 &\coloneqq 0, \\ s_{n_t} &\coloneqq \mu(D_j), \\ s_i &\in (t_i,\ t_{i+1}) \cap B_{R(t_i, \varepsilon'')}(t_i) \cap B_{R(t_{i+1}, \varepsilon'')}(t_{i+1}) \quad \forall i \in [n_t - 1]. \end{aligned}$$

This choice ensures that $\{s_{i-1}, s_i\} \subseteq B_{R(t_i, \varepsilon'')}(t_i)$ and therefore that $\nabla F\big(\gamma(t_i)\big)$ can be used to approximate the difference between $F\big(\gamma(s_{i-1})\big)$ and $F\big(\gamma(s_i)\big)$. We have

$$\begin{aligned} &F(U_j \triangle D_j) - F(U_j) - \nabla F(U_j)(D_j) \\ &\quad = F\big(\gamma(s_{n_t})\big) - F\big(\gamma(s_0)\big) - \nabla F(U_j)\Big(\bigcup_{i=1}^{n_t} \big(\gamma(s_i) \triangle \gamma(s_{i-1})\big)\Big) \\ &\quad = \sum_{i=1}^{n_t} \Big(F\big(\gamma(s_i)\big) - F\big(\gamma(s_{i-1})\big) - \nabla F(U_j)\big(\gamma(s_i) \triangle \gamma(s_{i-1})\big)\Big). \end{aligned}$$

We now consider an individual index $i \in [n_t]$. We have

$$F\big(\gamma(s_i)\big) - F\big(\gamma(s_{i-1})\big) = F\big(\gamma(s_i)\big) - F\big(\gamma(t_i)\big) + F\big(\gamma(t_i)\big) - F\big(\gamma(s_{i-1})\big)$$

with

$$\begin{aligned}
\Big|F\big(\gamma(s_i)\big) - F\big(\gamma(t_i)\big) - \nabla F\big(\gamma(t_i)\big)\big(\gamma(s_i) \triangle \gamma(t_i)\big)\Big| &\leq \varepsilon'' \cdot \mu\big(\gamma(s_i) \triangle \gamma(t_i)\big) \\
&= \varepsilon'' \cdot |s_i - t_i| \\
&= \varepsilon'' \cdot (s_i - t_i), \\
\Big|F\big(\gamma(t_i)\big) - F\big(\gamma(s_{i-1})\big) + \nabla F\big(\gamma(t_i)\big)\big(\gamma(s_{i-1}) \triangle \gamma(t_i)\big)\Big| &\leq \varepsilon'' \cdot \mu\big(\gamma(s_{i-1}) \triangle \gamma(s_i)\big) \\
&= \varepsilon'' \cdot |s_{i-1} - t_i| \\
&= \varepsilon'' \cdot (t_i - s_{i-1}),
\end{aligned}$$

because

$$t_i - R(t_i, \varepsilon'') < s_{i-1} \leq t_i \leq s_i < t_i + R(t_i, \varepsilon'') \quad \forall i \in [n_t]$$

holds by construction. We note that equality between t_i and s_{i-1} or s_i is usually explicitly prohibited, but can occur for $t_i = 0$ or $t_i = \mu(D_j)$. We find that

$$\begin{aligned}
&\Big|F\big(\gamma(s_i)\big) - F\big(\gamma(s_{i-1})\big) - \nabla F\big(\gamma(t_i)\big)\big(\gamma(s_i) \triangle \gamma(t_i)\big) + \nabla F\big(\gamma(t_i)\big)\big(\gamma(s_{i-1}) \triangle \gamma(t_i)\big)\Big| \\
&\quad \leq \Big|F\big(\gamma(s_i)\big) - F\big(\gamma(t_i)\big) - \nabla F\big(\gamma(t_i)\big)\big(\gamma(s_i) \triangle \gamma(t_i)\big)\Big| \\
&\qquad + \Big|F\big(\gamma(t_i)\big) - F\big(\gamma(s_{i-1})\big) + \nabla F\big(\gamma(t_i)\big)\big(\gamma(s_{i-1}) \triangle \gamma(t_i)\big)\Big| \\
&\quad \leq \varepsilon'' \cdot (s_i - t_i + t_i - s_{i-1}) \\
&\quad = \varepsilon'' \cdot (s_i - s_{i-1}).
\end{aligned}$$

We can now aggregate these estimates to obtain an estimate for the deviation of F from its linearization around U_j on each segment of the geodesic γ.

$$\begin{aligned}
&\Big|F\big(\gamma(s_i)\big) - F\big(\gamma(s_{i-1})\big) - \nabla F(U_j)\big(\gamma(s_i) \triangle \gamma(s_{i-1})\big)\Big| \\
&\overset{U_j = \gamma(0)}{\leq} \Big|F\big(\gamma(s_i)\big) - F\big(\gamma(s_{i-1})\big) - \nabla F\big(\gamma(t_i)\big)\big(\gamma(s_i) \triangle \gamma(t_i)\big) \\
&\qquad + \nabla F\big(\gamma(t_i)\big)\big(\gamma(s_{i-1}) \triangle \gamma(t_i)\big)\Big| \\
&\qquad + \Big|\nabla F\big(\gamma(t_i)\big)\big(\gamma(s_i) \triangle \gamma(t_i)\big) - \nabla F\big(\gamma(t_i)\big)\big(\gamma(s_{i-1}) \triangle \gamma(t_i)\big) \\
&\qquad - \nabla F\big(\gamma(0)\big)\big(\gamma(s_i) \triangle \gamma(s_{i-1})\big)\Big| \\
&\leq \varepsilon'' \cdot (s_i - s_{i-1}) + \Big|\Big(\nabla F\big(\gamma(t_i)\big) - \nabla F\big(\gamma(0)\big)\Big)\big(\gamma(s_i) \triangle \gamma(t_i)\big) \\
&\qquad - \Big(\nabla F\big(\gamma(t_i)\big) + \nabla F\big(\gamma(0)\big)\Big)\big(\gamma(t_i) \triangle \gamma(s_{i-1})\big)\Big| \\
&\leq \varepsilon'' \cdot (s_i - s_{i-1}) + \Big|\nabla F\big(\gamma(t_i)\big) - \nabla F(U_j)\Big|\big(\gamma(s_i) \triangle \gamma(t_i)\big) \\
&\quad + \Big|\nabla F\big(\gamma(t_i)\big) + \nabla F(U_j)\Big|\big(\gamma(t_i) \triangle \gamma(s_{i-1})\big) \\
&= \varepsilon'' \cdot (s_i - s_{i-1}) + \Big|\nabla F\big(\gamma(t_i)\big) - \nabla F(U_j)\Big|\Big(\big(\gamma(s_i) \triangle \gamma(s_{i-1})\big) \setminus \big(\gamma(t_i) \triangle U_j\big)\Big) \\
&\quad + \Big|\nabla F\big(\gamma(t_i)\big) + \nabla F(U_j)\Big|\Big(\big(\gamma(s_i) \triangle \gamma(s_{i-1})\big) \cap \big(\gamma(t_i) \triangle U_j\big)\Big)
\end{aligned}$$

$$\overset{\text{Def. 2.4.3}}{=} \varepsilon'' \cdot (s_i - s_{i-1}) + \Big(\nabla F\big(\gamma(t_i)\big) \underset{\gamma(t_i) \triangle U_j}{\ominus} \nabla F(U_j)\Big)\big(\gamma(s_i) \triangle \gamma(s_{i-1})\big).$$

By combining these estimates for all segments, we obtain the overall estimate

$$\begin{aligned}
&\big|F(U_j \triangle D_j) - F(U_j) - \nabla F(U_j)(D_j)\big| \\
&\quad \le \sum_{i=1}^{n_t} \Big|F\big(\gamma(s_i)\big) - F\big(\gamma(s_{i-1})\big) - \nabla F(U_j)\big(\gamma(s_i) \triangle \gamma(s_{i-1})\big)\Big| \\
&\quad \le \sum_{i=1}^{n_t} \big(\varepsilon'' \cdot (s_i - s_{i-1})\big) + \sum_{i=1}^{n_t} \Big(\nabla F\big(\gamma(t_i)\big) \underset{\gamma(t_i) \triangle U_j}{\ominus} \nabla F(U_j)\Big)\big(\gamma(s_i) \triangle \gamma(s_{i-1})\big) \\
&\quad = \varepsilon'' \cdot \mu(D_j) + \sum_{i=1}^{n_t} \Big(\nabla F\big(\gamma(t_i)\big) \underset{\gamma(t_i) \triangle U_j}{\ominus} \nabla F(U_j)\Big)\big(\gamma(s_i) \triangle \gamma(s_{i-1})\big).
\end{aligned}$$

If $\Delta_j \le \delta$, then we have $\mu(D_j) \le \Delta_j \le \delta$, which implies that

$$\mu\big(\gamma(t_i) \triangle U_j\big) = t_i \le \mu(D_j) \le \delta \quad \forall i \in [n_t].$$

This means that we have

$$\Big(\nabla F\big(\gamma(t_i)\big) \underset{\gamma(t_i) \triangle U_j}{\ominus} \nabla F(U_j)\Big)\big(\gamma(s_i) \triangle \gamma(s_{i-1})\big) \le \varepsilon' \cdot \mu\big(\gamma(s_i) \triangle \gamma(s_{i-1})\big) = \varepsilon' \cdot (s_i - s_{i-1})$$

holds for all $i \in [n_t]$, which allows us to further simplify the overall estimate to

$$\begin{aligned}
&\big|F(U_j \triangle D_j) - F(U_j) - \nabla F(U_j)(D_j)\big| \\
&\quad \le \varepsilon'' \cdot \mu(D_j) + \sum_{i=1}^{n_t} \Big(\nabla F\big(\gamma(t_i)\big) \underset{\gamma(t_i) \triangle U_j}{\ominus} \nabla F(U_j)\Big)\big(\gamma(s_i) \triangle \gamma(s_{i-1})\big) \\
&\quad \le \varepsilon'' \cdot \mu(D_j) + \sum_{i=1}^{n_t} \varepsilon' \cdot (s_i - s_{i-1}) \\
&\quad = (\varepsilon'' + \varepsilon') \cdot \mu(D_j).
\end{aligned}$$

Because this holds for every $\varepsilon'' > 0$, we arrive at the final estimate that, if $\delta > 0$ is chosen such that

$$\big(\nabla F(U) \underset{U \triangle V}{\ominus} \nabla F(V)\big)(D) \le \varepsilon' \cdot \mu(D) \quad \forall U, V, D \in \Sigma/_{\sim_\mu} \colon \mu(U \triangle V) \le \delta,$$

and if $\Delta_j \le \delta$, then we also have

$$\big|F(U_j \triangle D_j) - F(U_j) - \nabla F(U_j)(D_j)\big| \le \varepsilon' \cdot \mu(D_j) \le \varepsilon' \cdot \Delta_j.$$

As we had shown in Parts 1 and 2, we have

$$\begin{aligned}
\nabla F(U_j)(D_j) &= \int_{D_j} g_j^* \, \mathrm{d}\mu \\
&\le (1 - \xi_g) \cdot \int_{D_j} g_j \, \mathrm{d}\mu \\
&\le (1 - \xi_g) \cdot (1 - \xi_\delta) \cdot \theta \cdot \min\left\{1, \frac{\Delta_j}{\mu(\{g_j < 0\})}\right\} \cdot \big(-\tau(g_j)\big) \\
&\le -\underbrace{\left((1 - \xi_g) \cdot (1 - \xi_\delta) \cdot \theta \cdot \min\left\{1, \frac{\Delta_j}{\mu(\{g_j < 0\})}\right\} \cdot \varepsilon\right)}_{>0} \\
&\overset{\Delta_j \le \mu(X)}{\le} -\left((1 - \xi_g) \cdot (1 - \xi_\delta) \cdot \theta \cdot \frac{\Delta_j}{\mu(X)} \cdot \varepsilon\right),
\end{aligned}$$

where $\varepsilon > 0$ is the overall stationarity tolerance. Let $\varepsilon_1 \coloneqq \frac{(1-\xi_g)\cdot(1-\xi_\delta)\cdot\theta\cdot\varepsilon}{\mu(X)} > 0$. We subsequently assume that $\varepsilon' \leq \varepsilon_1$. Then the previous estimate becomes $\nabla F(U_j)(D_j) \leq -\varepsilon_1 \cdot \Delta_j$ and we have

$$\begin{aligned} F(U_j \triangle D_j) - F(U_j) &\leq \nabla F(U_j)(D_j) + \varepsilon' \cdot \Delta_j \\ &\leq -\varepsilon_1 \cdot \Delta_j + \varepsilon' \cdot \Delta_j \\ &\leq 0. \end{aligned}$$

Knowing that $F(U_j \triangle D_j) - F(U_j)$ and $\nabla F(U_j)(D_j)$ are non-positive, we can then make the following estimate for the true step quality:

$$\begin{aligned} \rho_j &= \frac{F(U_j \triangle D_j) - F(U_j)}{\nabla F(U_j)(D_j)} \\ &\geq \frac{\nabla F(U_j)(D_j) + \varepsilon' \cdot \Delta_j}{\nabla F(U_j)(D_j)} \\ &= 1 - \frac{\varepsilon' \cdot \Delta_j}{-\nabla F(U_j)(D_j)} \\ &\geq 1 - \frac{\varepsilon' \cdot \Delta_j}{\varepsilon_1 \cdot \Delta_j} \\ &\overset{\Delta_j > 0}{=} 1 - \frac{\varepsilon'}{\varepsilon_1} \\ &= \frac{\varepsilon_1 - \varepsilon'}{\varepsilon_1}. \end{aligned}$$

If we further choose

$$\varepsilon' \leq \varepsilon_2 \coloneqq \left(1 - \frac{1}{2} \cdot \left(1 + \frac{\sigma_1}{1-\xi_g}\right)\right) \cdot \varepsilon_1,$$

which is possible because $\xi_g < 1 - \sigma_1$ and therefore $\frac{\sigma_1}{1-\xi_g} \in (\sigma_1, 1)$. If $\frac{\sigma_1}{1-\xi_g}$ is strictly less than 1, then the arithmetic mean between 1 and $\frac{\sigma_1}{1-\xi_g}$ is also strictly between $\frac{\sigma_1}{1-\xi_g}$ and 1, which implies both $0 < \varepsilon_2 \leq \varepsilon_1$ and

$$\rho_j \geq 1 - \frac{\varepsilon'}{\varepsilon_1} \geq 1 - \frac{\varepsilon_2}{\varepsilon_1} = 1 - 1 + \underbrace{\frac{1}{2} \cdot \left(1 + \frac{\sigma_1}{1-\xi_g}\right)}_{> \frac{\sigma_1}{1-\xi_g}} > \frac{\sigma_1}{1-\xi_g}$$

for all $\varepsilon' \leq \varepsilon_2$. Let $\varepsilon' \coloneqq \min\{\varepsilon_1, \varepsilon_2\}$ and let $\Delta_{\text{acc}} > 0$ be such that

$$\left(\nabla F(U) \underset{U \triangle V}{\ominus} \nabla F(V)\right)(D) \leq \varepsilon' \cdot \mu(D) \quad \forall U, V, D \in \Sigma/_{\sim_\mu} \colon \mu(U \triangle V) \leq \Delta_{\text{acc}}.$$

Then for every $j \in \mathbb{N}_0$ prior to termination such that $\tau(g_j) > \varepsilon$ and $\Delta_j \leq \Delta_{\text{acc}}$, we have $\rho_j > \frac{\sigma_1}{1-\xi_g}$, which implies that $\widetilde{\rho}_j - e_{\rho,j} \geq \sigma_0$, which means that the step D_j is accepted. As a consequence, because Δ_j is only ever reduced if the step is rejected and because it is only ever reduced by the fixed factor $\frac{1}{2}$, we can show by contradiction that

$$\Delta_j \geq \underbrace{\min\left\{\Delta_0, \frac{\Delta_{\text{acc}}}{2}\right\}}_{=:\Delta_{\min}} \quad \forall j \in \mathbb{N}_0 \text{ prior to termination.} \tag{3.17}$$

Assume that there existed $j \in \mathbb{N}_0$ with $\Delta_j < \Delta_{\min}$. Without loss of generality, let j be the minimal index with that property. Because $\Delta_j < \Delta_0$, we would have $j > 0$. Because $\Delta_i \geq \Delta_{\min}$ for all $i < j$, a trust region reduction would have to have taken place immediately prior to the j-th iteration, which means that $\Delta_{j-1} = 2\Delta_j$. However, because $\Delta_j < \frac{\Delta_{\text{acc}}}{2}$, this would imply $\Delta_{j-1} < \Delta_{\text{acc}}$, which would contradict the fact that the step was rejected. This contradiction proves that the tighter lower bound for Δ_j must hold.

PART 4 (MAIN INDUCTION ARGUMENT). In this section, we show inductively that the following invariants hold throughout all iterations of the main loop:

- Equation (3.17) holds for all j;
- $F(U_{j+1}) \leq F(U_j)$ for all $j \in \mathbb{N}_0$ prior to termination;
- There exists $\overline{F} > 0$ such that step acceptance implies $F(U_{j+1}) \leq F(U_j) - \overline{F}$.

Equation (3.17) is significant to prove the minimal descent for accepted steps. It also establishes that there is never an infinite unbroken sequence of rejected steps. The minimal descent, along with the boundedness of the objective then ensures the termination of the algorithm.

We prove these invariants by induction over j. For $j = 0$, we self-evidently have $\Delta_j = \Delta_0 \geq \Delta_{\min}$ by definition of $\Delta_{\min}$. We then consider the transition from the j-th to the $(j+1)$-th iterate.

For subsequent iterations, we have to bear in mind that a transition to the next iteration only takes place if the termination criterion of the main loop is not met. Therefore, the induction step may assume that the loop condition is satisfied.

Specifically, this means that we have $\widetilde{\tau}_j = \tau(g_j) > \varepsilon$. In Part 1, we have demonstrated that this implies $\tau(g_j^*) > (1 - \xi_\tau) \cdot \varepsilon$. Due to the choice of step tolerance in Algorithm 4 of Algorithm 4, we have

$$\int_{D_j} g_j \, \mathrm{d}\mu \leq -(1-\xi_\delta) \cdot \theta \cdot \min\left\{1, \frac{\Delta_j}{\mu(\{g_j < 0\})}\right\} \cdot \varepsilon < 0$$

where the fraction on the right hand side is always well defined because $\widetilde{\tau}_j < 0$. If $\Delta_j \geq \Delta_{\min}$, then this implies that

$$\begin{aligned}\int_{D_j} g_j \, \mathrm{d}\mu &\leq -(1-\xi_\delta) \cdot \theta \cdot \min\left\{1, \frac{\Delta_j}{\mu(\{g_j < 0\})}\right\} \cdot \varepsilon \\ &\leq -(1-\xi_\delta) \cdot \theta \cdot \min\left\{1, \frac{\Delta_{\min}}{\mu(X)}\right\} \cdot \varepsilon.\end{aligned}$$

In addition, we had established that Equation (3.15) holds, which means that

$$\begin{aligned}\int_{D_j} g_j^* \, \mathrm{d}\mu &\leq \underbrace{\int_{D_j} g_j \, \mathrm{d}\mu}_{<0} + \xi_g \cdot \left| \int_{D_j} g_j \, \mathrm{d}\mu \right| \\ &\leq (1-\xi_g) \cdot \int_{D_j} g_j \, \mathrm{d}\mu \\ &\leq -(1-\xi_g) \cdot (1-\xi_\delta) \cdot \theta \cdot \min\left\{1, \frac{\Delta_{\min}}{\mu(X)}\right\} \cdot \varepsilon.\end{aligned}$$

This gives us a negative upper bound on the true projected descent that only depends on fixed algorithmic parameters. In conjunction with a lower bound on the true step quality, this yields a lower bound on the true decrease of the objective function for an accepted step. If the step is accepted, then, as we have established in Part 2, we have $\rho_j \geq \frac{\sigma_0}{1+\xi_g}$. This means that

$$\begin{aligned} F(U_{j+1}) - F(U_j) &\leq \rho_j \cdot \int_{D_j} g_j^* \,\mathrm{d}\mu \\ &\leq \underbrace{-\sigma_0 \cdot \frac{1-\xi_g}{1+\xi_g} \cdot (1-\xi_\delta) \cdot \theta \cdot \min\left\{1, \frac{\Delta_{\min}}{\mu(X)}\right\}}_{=:\overline{F}>0} \cdot \varepsilon. \end{aligned}$$

If the step is accepted, then we have $F(U_{j+1}) \leq F(U_j) - \overline{F} < F(U_j)$ and $\Delta_{j+1} \in \{\Delta_j, \min\{2\Delta_j, \overline{M}\}\}$, which implies $\Delta_{j+1} \geq \Delta_j \geq \overline{\Delta}$.

Conversely, if the step is rejected, then $F(U_{j+1}) = F(U_j) \leq F(U_j)$ and we have $\Delta_{j+1} = \frac{\Delta_j}{2}$. We had previously shown in Part 3 that there exists $\Delta_{\mathrm{acc}} > 0$ such that the step is always accepted if $\Delta_j \leq \Delta_{\mathrm{acc}}$. Therefore, rejection of the step implies $\Delta_j > \Delta_{\mathrm{acc}}$, which establishes the crucial invariant Equation (3.17):

$$\Delta_{j+1} = \frac{\Delta_j}{2} > \frac{\Delta_{\mathrm{acc}}}{2} \geq \Delta_{\min}.$$

This inductively proves that Equation (3.17) holds. As the trust region radius is always halved on rejection and is bounded from below by a strictly positive bound, there can never be an infinite string of rejected steps. Conversely, as long as the loop does not terminate, there is an unending chain of accepted steps with only a finite number of rejections between them.

Because we have established that every accepted step reduces the true objective by at least $\overline{F} > 0$ and because the objective is bounded from below, there can at most be

$$K_0 := \left\lfloor \frac{F(U_0) - F^-}{\overline{F}} \right\rfloor$$

accepted steps where $F^- \in \mathbb{R}$ is such that $F(U) \geq F^-$ for all $U \in \Sigma/_{\sim_\mu}$.

Because the number of accepted steps is bounded by K_0 and there cannot be an infinite sequence of rejected steps, the main loop must terminate, even if the accepted step limit K is infinite. If the number of accepted steps k satisfies $k < K$ after the loop terminates, which is always the case if $K \geq K_0$, then the loop terminates because $\widetilde{\tau}_j \leq \varepsilon$, which implies $\mathscr{C}_1(F, U^*) = -\tau(g_j^*) \geq -(1+\xi_\tau) \cdot \varepsilon$. □

3.1.4 Trust-Region Steepest Descent

The controlled trust-region descent framework is an algorithmic framework with three degrees of freedom. For each problem, a domain expert has to specify objective and gradient evaluation methods. These are specific to the underlying function and can differ greatly based on the numerical methods we use to calculate objective and gradient. Therefore, we will not discuss generalized evaluation methods here.

The third degree of freedom is the step-finding routine. Here, we can provide a generalized algorithm in an application-agnostic manner. According to Definition 3.1.2, a controlled unconstrained step-finding routine of quality $\theta \in (0,\ 1]$ finds a product similarity class $D \in \Sigma/_{\sim_\mu}$ such that $\mu(D) \leq \Delta$ and

$$\int_D g \,\mathrm{d}\mu \leq \theta \cdot \min\left\{1,\ \frac{\Delta}{\mu(\{g < 0\})}\right\} \cdot \int_{\{g<0\}} g \,\mathrm{d}\mu + \delta$$

for given $\Delta > 0$ and $\delta > 0$. Here, $g \in L^1(\Sigma, \mu)$ is an approximate gradient density function, $\Delta > 0$ is a trust region radius, and δ is an error bound based on the instationarity of the current iterate. If we formulate our algorithm using only a limited subset of operations that can be applied to all L^1-functions, then the encoding of g and the method by which it is calculated are irrelevant.

Here, we rely on the large amount of preparatory theoretical work that we have done in Chapter 2. There, we have broken down complex operations on integrable functions down into combinations of their level and sublevel sets. In a very similar manner, rather than assuming a specific encoding of g, we will only assume that there is a fast and accurate way to generate these sets. We will assume that the methods by which these sets are determined are free from error, i.e., that they yield exact representations of the sets up to true nullsets.

In Section 2.3.5, we had discussed how to translate level set functions into geodesics. Specifically, we had shown the existence of *minimal mean geodesics* in Theorem 2.3.54. A minimal mean geodesic $\gamma\colon I \to \Sigma/_{\sim_\mu}$ has the property that

$$\int_{\gamma(t)} g \,\mathrm{d}\mu \leq \int_A g \,\mathrm{d}\mu \quad \forall t \in I,\ A \in \Sigma/_{\sim_\mu}\colon \mu(A) \leq \mu\big(\gamma(t)\big).$$

This means that minimal mean geodesics achieve the minimal mean value of the functions that they are generated from, which very closely resembles the criterion for a controlled unconstrained step-finding routine. We can create such a routine of quality $\theta = 1$ by approximating the theoretical construction of a minimal mean geodesic.

Selecting a step set by picking a set which minimizes the mean gradient value means that we achieve the best possible ratio between projected descent and step size. Hence, we will refer to this step as both the *minimal mean step* and the *steepest descent step*. We state our approximation algorithm in Algorithm 5 on the following page.

As we had stated in Theorem 2.3.54, minimal mean geodesics are generated by taking the supremum of all levels for which the non-strict sublevel sets of the level set functions are smaller than the desired size and adding an appropriately sized piece of the exact level sets corresponding to that supremum to the strict sublevel sets. The main issue that we have to address in Algorithm 5 is that the supremum generally cannot be determined with perfect accuracy. Instead, we have to approximate it. We perform this approximation with a bisection algorithm.

Theorem 3.1.11 (Correctness of Algorithm 5).
Let $(X,\ \Sigma,\ \mu)$ be a finite atomless measure space. For $g \in L^1(\Sigma, \mu)$, $\Delta > 0$, and $\overline{\delta} > 0$, the procedure AMMStep *stated in Algorithm 5 returns $(D,\ \delta) \in \Sigma/_{\sim_\mu} \times [0,\ \overline{\delta}]$ such that $\mu(D) \leq \Delta$ and*

$$\int_D g \,\mathrm{d}\mu \leq \int_A g \,\mathrm{d}\mu + \delta \quad \forall A \in \Sigma/_{\sim_\mu}\colon \mu(A) \leq \Delta. \tag{3.18}$$

Algorithm 5 Approximate Minimal Mean Step

Require: $(X,\ \Sigma,\ \mu)$ finite atomless measure space, $g \in L^1(\Sigma,\mu)$, $\Delta > 0$, $\overline{\delta} > 0$
Ensure: Returns $(D,\ \delta) \in \Sigma/_{\sim_\mu} \times (0,\ \overline{\delta}]$ such that $\mu(D) \leq \Delta$ and such that for all $B \in \Sigma/_{\sim_\mu}$ with $\mu(B) \leq \Delta$, we have $\int_D g \,\mathrm{d}\mu \leq \int_B g \,\mathrm{d}\mu + \delta$.

1: **procedure** AMMSTEP(g, Δ, $\overline{\delta}$)
2: **if** $\mu\big(\{g < 0\}\big) \leq \Delta$ **then** ▷ Short-circuit if full step is admissible
3: **return** $\Big(\big[\{g < 0\}\big]_{\sim_\mu},\ 0\Big)$
4: **end if**

▷ Approximate $\eta^* := \sup\{\eta \in \mathbb{R} \,\big|\, \mu\big(\{g \leq \eta\}\big) < \Delta\}$ to precision $\frac{\overline{\delta}}{2\Delta}$
5: $(\eta_0^-,\ \eta_0^+,\ j) \leftarrow (-\overline{\delta},\ 0,\ 0)$
6: **while** $\mu\big(\{g \leq \eta_j^-\}\big) > \Delta$ **do** ▷ Find initial bounds
7: $(\eta_{j+1}^-,\ \eta_{j+1}^+,\ j) \leftarrow (2\eta_j^-,\ \eta_j^-,\ j+1)$
8: **end while**

9: **while** $\eta_j^+ - \eta_j^- > \frac{\overline{\delta}}{2\Delta}$ **do** ▷ Main bisection loop
10: $\eta_j^\sim \leftarrow \frac{1}{2}(\eta_j^- + \eta_j^+)$ ▷ Calculate midpoint
11: **if** $\mu\big(\{g \leq \eta_j^\sim\}\big) \leq \Delta$ **then**
12: $(\eta_{j+1}^-,\ \eta_{j+1}^+,\ j) \leftarrow (\eta_j^\sim,\ \eta_j^+,\ j+1)$ ▷ Adjust lower bound
13: **else**
14: $(\eta_{j+1}^-,\ \eta_{j+1}^+,\ j) \leftarrow (\eta_j^-,\ \eta_j^\sim,\ j+1)$ ▷ Adjust upper bound
15: **end if**
16: **end while**

▷ Derive the step set
17: $D^\pm \leftarrow \big[\{g \leq \eta_j^\pm\}\big]_{\sim_\mu}$ ▷ Lower and upper bounds for the step set
18: Select $D^\sim \subseteq_\mu D^+ \setminus D^-$ with $\Delta - \frac{\overline{\delta} - \Delta\cdot(\eta_j^+ - \eta_j^-)}{|\eta_j^-|} \leq \mu(D^\sim \cup D^-) \leq \Delta$
19: **return** $\Big(D^- \cup D^\sim,\ (\eta_j^+ - \eta_j^-)\cdot\Delta + |\eta_j^-|\cdot\big(\Delta - \mu(D^- \cup D^\sim)\big)\Big)$
20: **end procedure**

By Definition 3.1.2, this means that AMMSTEP *is a controlled unconstrained step finding method of quality* $\theta = 1$. ◁

PROOF. PART 1 (SHORT-CIRCUIT CONDITION). Before we turn our attention to the main bisection algorithm, we discuss the short-circuit conditional in Algorithms 5 to 5. This conditional deals with cases in which the "full step" does not exceed the trust region radius. The "full step" is the step that consists of the similarity class corresponding to the set D that comprises all points at which g is strictly negative, i.e.,

$$D \coloneqq \big[\{g < 0\}\big]_{\sim_\mu}.$$

The conditional is only triggered if

$$\mu(D) = \mu\big(\{g < 0\}\big) \leq \Delta,$$

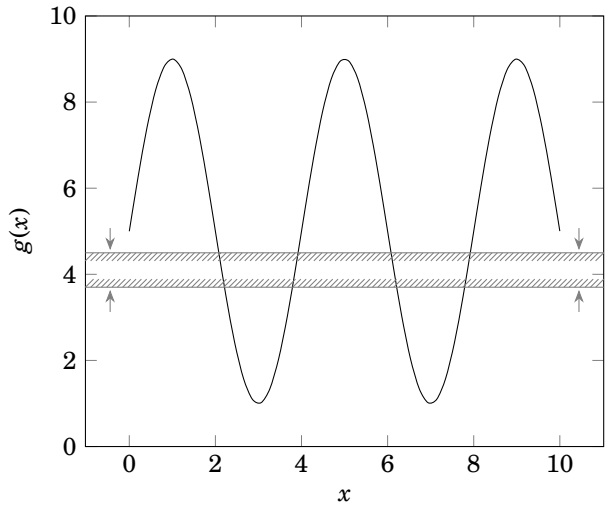


Figure 3.2: Illustrations of the approximation of the supremal level for the minimal mean geodesic. The horizontal lines represent the upper and lower levels η^+ and η^-. These lines are then brought together until the difference between the values is negligible.

which means that $\mu(D) \leq \Delta$ is satisfied. The returned error margin δ is equal to 0. For all $A \in \Sigma/_{\sim_\mu}$, we have

$$\int_A g \,\mathrm{d}\mu + \underbrace{\delta}_{=0} = \int_{A\cap D} g \,\mathrm{d}\mu + \int_{A\setminus D} \underbrace{g}_{\geq 0} \,\mathrm{d}\mu \geq \int_D \underbrace{g}_{<0} \,\mathrm{d}\mu.$$

Part 2 (Bisection). If the short-circuit transitional is not triggered, then we have

$$\mu\big(\{g < 0\}\big) > \Delta > 0.$$

In this case, we approximate the supremal level

$$\eta^* = \sup\Big\{\eta \in \mathbb{R} \,\Big|\, \mu\big(\{g \leq \eta\}\big) < \Delta\Big\}$$

with η_j^- and η_j^+ such that $\eta_j^- \leq \eta^* \leq \eta_j^+$. Because $\eta \mapsto \mu\big(\{g \leq \eta\}\big)$ is monotonically increasing and η^* is a supremum, we have

$$\begin{aligned}
\mu\big(\{g \leq \eta\}\big) \leq \Delta &\implies \eta \leq \eta^* \quad \forall \eta \in \mathbb{R}, \\
\mu\big(\{g \leq \eta\}\big) > \Delta &\implies \eta \geq \eta^* \quad \forall \eta \in \mathbb{R}.
\end{aligned}$$

We can therefore bisect based on the measure $\mu\big(\{g \leq \eta\}\big)$ rather than direct comparison to the unknown value η^*. We first show that the setup loop in Algorithms 5 to 5 terminates in finite time and establishes the relations

$$\begin{aligned}
\mu\big(\{g \leq \eta_j^-\}\big) &\leq \Delta, \\
\mu\big(\{g \leq \eta_j^+\}\big) &> \Delta.
\end{aligned}$$

We first consider the values of $\eta_0^\pm$. Because this setup code is only invoked when the short-circuit condition is not satisfied, we know that

$$\mu\big(\{g \leq \eta_0^+\}\big) = \mu\big(\{g \leq 0\}\big) \geq \mu\big(\{g < 0\}\big) > \Delta.$$

We also know that $\eta_0^- = -\overline{\delta} < \eta_0^+ = 0$. We now consider subsequent iterations of the setup loop. Let $j \in \mathbb{N}_0$ be such that $\eta_j^- < \eta_j^+ \leq 0$, $\mu(\{g \leq \eta_j^-\}) > \Delta$. We evidently have

$$\eta_{j+1}^- = 2 \underbrace{\eta_j^-}_{<0} < \underbrace{\eta_j^-}_{=\eta_{j+1}^+} < 0$$

and

$$\mu(\{g \leq \eta_{j+1}^+\}) = \mu(\{g \leq \eta_j^-\}) > \Delta.$$

We now perform a proof by contradiction. If the setup loop did not terminate, then we would have $\eta_{j+1}^- = -2^j\overline{\delta} \to -\infty$ for $j \to \infty$. Because

$$\mu(\{g \leq \eta_j^-\}) > \Delta$$

for all $j \in \mathbb{N}_0$, we would then have

$$\mu(\{g = -\infty\}) = \mu\left(\bigcap_{j=0}^{\infty} \{g \leq \eta_j^-\}\right) = \lim_{j\to\infty} \mu(\{g \leq \eta_j^-\}) \geq \Delta > 0.$$

However, this would contradict $g \in L^1(\Sigma, \mu)$. Therefore, the setup loop must terminate. Let $j' \in \mathbb{N}_0$ be the value of j upon termination of the setup loop. Then we have

$$\mu(\{g \leq \eta_{j'}^-\}) \leq \Delta,$$
$$\mu(\{g \leq \eta_{j'}^+\}) > \Delta.$$

The former follows from the termination criterion of the setup loop and the latter follows from the loop invariant that we have already proven.

We now use these inequalities as a starting point for an inductive argument for the main bisection loop in Algorithms 5 to 5. In this argument, we prove that $\eta_j^- < \eta_j^+ \leq 0$ for all $j \geq j'$ and that

$$\mu(\{g \leq \eta_j^-\}) \leq \Delta \qquad \forall j \geq j',$$
$$\mu(\{g \leq \eta_j^+\}) > \Delta \qquad \forall j \geq j',$$
$$\eta_j^+ - \eta_j^- = \frac{\eta_{j'}^+ - \eta_{j'}^-}{2^{j-j'}} \qquad \forall j \geq j'.$$

For $j = j'$ we have already shown the first two inequalities. The third equation is evidently true for $j = j'$. Let $j \geq j'$ be such that $\eta_j^- < \eta_j^+ \leq 0$ with $\eta_j^+ - \eta_j^- > \frac{\overline{\delta}}{\Delta}$ and

$$\mu(\{g \leq \eta_j^-\}) \leq \Delta,$$
$$\mu(\{g \leq \eta_j^+\}) > \Delta,$$
$$\eta_j^+ - \eta_j^- = \frac{\eta_{j'}^+ - \eta_{j'}^-}{2^{j-j'}}.$$

We distinguish between two cases.

Case 1 ($\mu(\{g \leq \eta_j^\sim\}) \leq \Delta$). In this case, we replace the lower bound. We then have

$$\eta_{j+1}^- = \eta_j^\sim = \frac{1}{2}(\eta_j^- + \eta_j^+) < \underbrace{\eta_j^+}_{=\eta_{j+1}^+} \leq 0,$$

$$\begin{aligned}
\mu\big(\{g \le \eta_{j+1}^-\}\big) &= \mu\big(\{g \le \tilde{\eta_j}\}\big) \le \Delta,\\
\mu\big(\{g \le \eta_{j+1}^+\}\big) &= \mu\big(\{g \le \eta_j^+\}\big) > \Delta,\\
\eta_{j+1}^+ - \eta_{j+1}^- &= \eta_j^+ - \frac{1}{2}(\eta_j^+ + \eta_j^-)\\
&= \frac{\eta_j^+ - \eta_j^-}{2}\\
&= \frac{\eta_{j'}^+ - \eta_{j'}^-}{2^{j+1-j'}}.
\end{aligned}$$ ◁

Case 2 ($\mu(\{g \ge \tilde{\eta_j}\}) > \Delta$). In this case, we replace the upper bound and obtain

$$\begin{aligned}
\eta_{j+1}^- = \eta_j^- &< \underbrace{\frac{1}{2}(\eta_j^- + \eta_j^+)}_{=\tilde{\eta_j} = \eta_{j+1}^+} < \eta_j^+ \le 0,\\
\mu\big(\{g \le \eta_{j+1}^-\}\big) &= \mu\big(\{g \le \eta_j^-\}\big) \le \Delta,\\
\mu\big(\{g \le \eta_{j+1}^+\}\big) &= \mu\big(\{g \le \tilde{\eta_j}\}\big) > \Delta,\\
\eta_{j+1}^+ - \eta_{j+1}^- &= \frac{1}{2}(\eta_j^+ + \eta_j^-) - \eta_j^-\\
&= \frac{\eta_j^+ - \eta_j^-}{2}\\
&= \frac{\eta_{j'}^+ - \eta_{j'}^-}{2^{j+1-j'}}.
\end{aligned}$$ ◁

This proves the induction assumption for $j+1$. Because $\frac{\overline{\delta}}{2\Delta} > 0$, there exists a minimal index $j'' \ge j'$ such that $\eta_{j''}^+ - \eta_{j''}^- \le \frac{\overline{\delta}}{2\Delta}$. When j'' is reached, the bisection loop terminates.

Part 3 (Finalization). Finally we turn our attention to the finalization of the step in Algorithms 5 to 5. At this point, we have $j = j''$. Because $\eta_j^+ \ge \eta_j^-$, $D^+ \in \Sigma/_{\sim_\mu}$ and $D^- \in \Sigma/_{\sim_\mu}$ satisfy $D^- \subseteq_\mu D^+$. In addition, we have

$$\begin{aligned}
\mu(D^+) &= \mu\big(\{g \le \eta_{j''}^+\}\big) > \Delta,\\
\mu(D^-) &= \mu\big(\{g \le \eta_{j''}^-\}\big) \le \Delta
\end{aligned}$$

because we have shown these relations to remain invariant during the bisection loop. As a result, $D^+ \setminus D^-$ is a similarity class that is essentially disjoint from D^- and satisfies

$$\mu(D^+ \setminus D^-) = \mu(D^+) - \mu(D^-) \ge \Delta - \mu(D^-).$$

Because the measure space (X, Σ, μ) is atomless, $D^+ \setminus D^-$ has an essential subset $D^\sim \subseteq_\mu D^+ \setminus D^-$ such that

$$\mu(D^\sim) \in \left[\Delta - \mu(D^-) - \frac{\overline{\delta} - \Delta \cdot (\eta_j^+ - \eta_j^-)}{|\eta_j^-|},\ \Delta - \mu(D^-)\right].$$

We note that

$$\overline{\delta} - \Delta \cdot (\eta_j^+ - \eta_j^-) \geq \overline{\delta} - \Delta \cdot \frac{\overline{\delta}}{2\Delta} = \frac{\overline{\delta}}{2}.$$

Therefore, the range of admissible measures for $D^\sim$ is always of strictly positive width.

Let $\delta \coloneqq \Delta \cdot (\eta_j^+ - \eta_j^-) + |\eta_j^-| \cdot \big(\Delta - \mu(D^-)\big)$. Bearing in mind that η_j^+ is non-positive and that η_j^- is strictly negative with $|\eta_j^-| \geq |\eta_j^+|$ and that $\mu(D^\sim) \leq \Delta - \mu(D^-)$, we have $\delta \geq 0$. Let $D \coloneqq D^- \cup D^\sim$.

Let $A \in \Sigma/_{\sim_\mu}$ with $\mu(A) \leq \Delta$. We have

$$\int_D g \,\mathrm{d}\mu - \int_A g \,\mathrm{d}\mu = \int_{D \setminus A} g \,\mathrm{d}\mu - \int_{A \setminus D} g \,\mathrm{d}\mu.$$

This is significant because we have $D \setminus A \subseteq_\mu D \subseteq_\mu D^+$ and therefore $g(x) \leq \eta_j^+$ almost everywhere in $D \setminus A$. On the other hand, we have $D^- \subseteq_\mu D$ and therefore $g(x) > \eta_j^-$ almost everywhere in $A \setminus D$. Let

$$M^- \coloneqq \min\big\{\mu(D \setminus A), \mu(A \setminus D)\big\} \leq \Delta.$$

Then we have

$$\begin{aligned}
\int_D g \,\mathrm{d}\mu - \int_A g \,\mathrm{d}\mu &= \int_{D \setminus A} \underbrace{g}_{\leq \eta_j^+} \mathrm{d}\mu - \int_{A \setminus D} \underbrace{g}_{> \eta_j^-} \mathrm{d}\mu \\
&\leq \big(\eta_j^+ - \eta_j^-\big) \cdot M^- + \underbrace{\eta_j^+}_{\leq 0} \cdot \underbrace{\big(\mu(D \setminus A) - M^-\big)}_{\geq 0} - \eta_j^- \cdot \big(\mu(A \setminus D) - M^-\big) \\
&\leq \big(\eta_j^+ - \eta_j^-\big) \cdot M^- - \eta_j^- \cdot \big(\mu(A \setminus D) - M^-\big) \\
&= \big(\eta_j^+ - \eta_j^-\big) \cdot M^- + |\eta_j^-| \cdot \max\big\{0, \mu(A \setminus D) - \mu(D \setminus A)\big\}.
\end{aligned}$$

We now make use of the fact that $D \setminus A = D \setminus (D \cap A)$ and $A \setminus D = A \setminus (D \cap A)$, which implies that

$$\begin{aligned}
\mu(A \setminus D) - \mu(D \setminus A) &= \mu(A) - \mu(D \cap A) - \mu(D) + \mu(D \cap A) \\
&= \mu(A) - \mu(D) \\
&\leq \Delta - \mu(D) \\
&= \Delta - \mu(D^-) - \mu(D^\sim) \\
&\leq \frac{\overline{\delta} - \Delta \cdot (\eta_j^+ - \eta_j^-)}{|\eta_j^-|}.
\end{aligned}$$

In conjunction with the fact that $M^- \leq \Delta$, this yields

$$\begin{aligned}
\int_D g \,\mathrm{d}\mu - \int_A g \,\mathrm{d}\mu &\leq \underbrace{\big(\eta_j^+ - \eta_j^-\big) \cdot \Delta + |\eta_j^-| \cdot \big(\Delta - \mu(D)\big)}_{=\delta} \\
&\leq \big(\eta_j^+ - \eta_j^-\big) \cdot \Delta + |\eta_j^-| \cdot \frac{\overline{\delta} - \Delta \cdot (\eta_j^+ - \eta_j^-)}{|\eta_j^-|} \\
&= \overline{\delta}.
\end{aligned}$$

PART 4 (STEP-FINDING ROUTINE). We first consider the simple case where $\Delta \geq \mu(\{g < 0\})$. In this case, we can set $A := [\{g < 0\}]_{\sim_\mu}$ and obtain

$$\begin{aligned}\int_D g \,\mathrm{d}\mu &\leq \int_A g \,\mathrm{d}\mu + \delta \\ &= \underbrace{\min\left\{1, \frac{\Delta}{\mu(\{g < 0\})}\right\}}_{=1} \cdot \int_{\{g<0\}} g \,\mathrm{d}\mu + \delta.\end{aligned}$$

Next, we address the case where $\Delta < \mu(\{g < 0\})$. In the previous parts of this proof, we had briefly referred to the supremum

$$\eta^* = \sup\left\{\eta \in \mathbb{R} \,\middle|\, \mu(\{g \leq \eta\}) < \Delta\right\},$$

which is similar to the η^* that we use to construct the minimal mean geodesic. Because $\mu(\{g < 0\}) > \Delta$, we know that $\eta^* \leq 0$. We define

$$\begin{aligned}A^- &:= [\{g < \eta^*\}]_{\sim_\mu} \in \Sigma/_{\sim_\mu}, \\ A^+ &:= [\{g \leq \eta^*\}]_{\sim_\mu} \in \Sigma/_{\sim_\mu}, \\ A^= &:= A^+ \setminus A^-.\end{aligned}$$

Because A^- can be written as a countable union of monotonically increasing similarity classes of non-strict sublevel sets corresponding to levels below η^*, we have

$$\mu(A^-) \leq \Delta.$$

Similarly, because A^+ can be written as a countable intersection of monotonically decreasing similarity classes of non-strict sublevel sets corresponding to levels above η^*, we have

$$\mu(A^+) \geq \Delta.$$

This means that $\mu(A^=) \geq \Delta - \mu(A^-)$. Because (X, Σ, μ) is atomless, we can select $A^\sim \subseteq_\mu A^=$ with

$$\left\|\mu(A^\sim)\right\| = \Delta - \left\|\mu(A^-)\right\|.$$

Let $A := A^- \cup A^\sim$. Because this is an essentially disjoint union, we have $\mu(A) = \mu(A^-) + \mu(A^\sim) = \Delta$ and therefore

$$\int_D g \,\mathrm{d}\mu \leq \int_A g \,\mathrm{d}\mu + \delta.$$

Because $\Delta < \mu(\{g < 0\})$, we can be certain that even if $\eta^* = 0$, we have

$A \cap \big[\{g = 0\}\big]_{\sim_\mu} = [\emptyset]_{\sim_\mu}$ and therefore $A \subseteq_\mu \{g < 0\}$. We can exploit this because

$$\begin{aligned}
\int_{\{g<0\}} g \,\mathrm{d}\mu_i &= \int_A g \,\mathrm{d}\mu + \int_{[\{g<0\}]_{\sim_\mu} \setminus A} \underbrace{g}_{\geq \eta^*} \,\mathrm{d}\mu \\
&\geq \int_A g \,\mathrm{d}\mu + \eta^* \cdot \Big(\mu\big(\{g<0\}\big) - \mu(A)\Big) \\
&= \mu(A) \cdot \underbrace{\frac{1}{\mu(A)} \cdot \int_A \overbrace{g}^{\leq \eta^*} \,\mathrm{d}\mu}_{\leq \eta^*} + \eta^* \cdot \underbrace{\Big(\mu\big(\{g<0\}\big) - \mu(A)\Big)}_{\geq 0} \\
&\geq \frac{\mu(A) + \mu\big(\{g<0\}\big) - \mu(A)}{\mu(A)} \cdot \int_A g \,\mathrm{d}\mu \\
&\geq \frac{\mu\big(\{g<0\}\big)}{\mu(A)} \cdot \int_A g \,\mathrm{d}\mu.
\end{aligned}$$

By multiplying both sides with $\frac{\mu(A)}{\mu\left(\{g<0\}\right)} = \frac{\Delta}{\mu\left(\{g<0\}\right)}$, we obtain

$$\int_A g \,\mathrm{d}\mu \leq \frac{\Delta}{\mu\big(\{g<0\}\big)} \cdot \int_{\{g<0\}} g \,\mathrm{d}\mu$$

and therefore

$$\begin{aligned}
\int_D g \,\mathrm{d}\mu &\leq \int_A g \,\mathrm{d}\mu + \delta \\
&\leq \frac{\Delta}{\mu\big(\{g<0\}\big)} \cdot \int_{\{g<0\}} g \,\mathrm{d}\mu + \delta \\
&\leq \min\left\{1, \frac{\Delta}{\mu\big(\{g<0\}\big)}\right\} \cdot \int_{\{g<0\}} g \,\mathrm{d}\mu + \delta,
\end{aligned}$$

which is the desired overall estimate. In summary, both cases yield the estimate

$$\int_D g \,\mathrm{d}\mu \leq \min\left\{1, \frac{\Delta}{\mu\big(\{g<0\}\big)}\right\} \cdot \int_{\{g<0\}} g \,\mathrm{d}\mu + \delta,$$

which is the defining estimate for a controlled unconstrained step-finding method of quality $\theta = 1$. □

The steepest descent step achieves a step quality of $\theta = 1$. While this is the best possible step quality, this does not mean that it is the only step finding method that achieves this quality. By definition, approximating the behavior of a constant mean geodesic (see Definition 2.3.56 on page 133) could also achieve $\theta = 1$. It is also conceivable that methods with $\theta < 1$ could yield more desirable steps for some applications. The steepest descent step is special in that it always approximates the best possible projected descent for a step below the given size threshold.

The step quality θ cannot reflect this distinctive feature of the steepest descent step because it is part of a *worst case* estimate. In the absolute worst case, the steepest descent step achieves the average descent obtained dividing the integral of g over the set $\{g < 0\}$ by the set's measure. However, this worst

case is only realized if either g has constant value almost everywhere in $\{g < 0\}$, or if $\Delta \geq \mu\big(\{g < 0\}\big)$. In Proposition 3.1.3, we have used these exact edge cases to prove the theoretical upper limit to θ.

However, if the gradient density function does not have constant value almost everywhere in $\{g < 0\}$ and $\Delta < \mu\big(\{g < 0\}\big)$, then the steepest descent step preferentially selects points where the gradient density function is below the average. Therefore, it will achieve projected descents that are better than the average, i.e., achieve step qualities better than $\theta = 1$. The theoretical step quality $\theta = 1$ is only realized in the worst case.

It is important to note that Algorithm 5 is still not a "complete" algorithm. It does not specifiy how the similarity class $D^{\sim}$ is determined. It only ensures that the conditions which $D^{\sim}$ must satisfy are satisfiable. This omission is deliberate. In order to specify precisely how $D^{\sim}$ should be selected, we would have to make assumptions about the way in which sets are encoded. However, this is dictated by the numerical method that we use and is better left to domain experts to determine on a problem-by-problem basis.

For example, the PDE-constrained problem that we discuss in Section 4.2 uses finite element methods (FEM) and encodes sets as a selection of a finite number of mesh cells. In this case, it is convenient to select $D^{\sim}$ to be a union of mesh cells if possible and refine those mesh cells as necessary.

However, the ODE-constrained problem that we discuss in Section 4.1 uses an adaptive Runge-Kutta method which does not require prior definition of a control or integration mesh. Therefore, it is more convenient to encode a set as a vector of switching points that are placed arbitrarily on the continuous time axis. This requires a different selection method for $D^{\sim}$. By leaving the method of determining $D^{\sim}$ unspecified, we do not artificially constrain our optimization algorithm to only one type of problem.

3.1.4.1 EQUIVALENCE TO LINE SEARCH

Because the steepest descent step is calculated by approximating the construction method of a minimal mean geodesic, the controlled descent framework could also be stated as a line search algorithm or, more appropriately, a "geodesic search algorithm." In such an algorithm, the search direction would be a canonical geodesic (e.g., a minimal mean geodesic of g) and we would adjust the parameter of the geodesic to find a step length that yields the necessary descent guarantees.

As long as our step-finding routine approximates a geodesic and we adjust the trust-region radius and the step length in the exact same way, this is equivalent in outcome. In our discussion of constrained optimization in Section 3.2, we will encounter situations where not every step-finding routine is an approximation of a geodesic. However, for unconstrained optimization, we can make an informed conjecture that these two approaches are theoretically equivalent.

The greater concern is that, while trust region and line search methods are equivalent *in outcome*, they are not equally simple *in implementation*. In set-valued optimization, a trust region approach is much easier to implement and provides much greater flexibility. There are reasons for this and we will briefly go through one of the primary issues with the line search approach to explain why we do not dedicate more time to it.

Approximation and Stability

Consider a minimal mean geodesic of a function g. Even if the same tie-breaker geodesic is used, very minor changes to g can potentially lead to large changes in the geodesic. Consider $X = [0,\ 1]$, $\Sigma = \mathscr{B}(X)$, $\mu = \lambda$ and

$$g(x) := \begin{cases} -1 + \frac{\varepsilon}{2} & \text{if } x < \frac{1}{2}, \\ -1 - \frac{\varepsilon}{2} & \text{if } x \geq \frac{1}{2} \end{cases} \quad \forall x \in X$$

for some very small $\varepsilon \in (0,\ 1)$. We can see that this would be a valid density function of a signed measure. A minimal mean geodesic γ of g would first act on points in $\left[\frac{1}{2},\ 1\right]$ and then points in $\left[0,\ \frac{1}{2}\right)$. However, if we adjust the value by only ε in each point, we obtain

$$g'(x) := \begin{cases} -1 - \frac{\varepsilon}{2} & \text{if } x < \frac{1}{2}, \\ -1 + \frac{\varepsilon}{2} & \text{if } x \geq \frac{1}{2} \end{cases} \quad \forall x \in X.$$

A minimal mean geodesic of g' would display the exact opposite behavior. It would first act on points in the lower half of X and then proceed to points in the upper half. Regardless of the fact that both the pointwise and the L^1 change between the two functions is only ε which can be an arbitrarily small number, the resulting change in the minimal mean geodesic is substantial. This is problematic because the gradient density function is generally only available as an approximation and may need to be re-evaluated as the step-size decreases. Essentially, we would need to recalculate the geodesic every time this happens, eliminating the potential performance advantage that line search methods could derive from reuse of the pre-calculated search direction.

3.1.4.2 REMARKS ON CONSTANT MEAN STEPS

It is conceivable that we could derive step-finding routines from other geodesics. Constant mean geodesics appear to be good candidates for this because they produce exactly the average projected descent and would therefore also achieve $\theta = 1$.

However, constant mean geodesics are more difficult to approximate than minimal mean geodesics. They require that we balance two cutoff levels, one below the average and one above, such that their deviations from the mean roughly cancel each other out and they jointly achieve the desired trust-region radius. This causes algorithmic complications that require a much greater computational effort to resolve than does the steepest descent step. Therefore, we will not pursue this further here.

It is unclear whether deliberately choosing a suboptimal step could yield any advantages. It is possible that, because a step with average projected descent has a lower *projected* improvement, it would produce less step rejections and therefore maintain a larger trust region radius. Because evaluation error bounds depend on the trust region radius, this would then mean that the error bounds for gradient and function evaluation would not need to be as tight. However, this would at best be of practical rather than theoretical benefit.

3.2 CONSTRAINED OPTIMIZATION IN MEASURE SPACES

In the preceding section, we have exclusively dealt with unconstrained problems. We do not include differential equation constraints because we assume that we

are working with reduced problems where the solving of differential equations is hidden within opaque function evaluations.

Since we are specifically interested in problems with differential equation constraints, we should be mindful that, in addition to whatever solution constraints users may want to impose, most differential equation models have a limited region of validity outside of which they break down. It is therefore arguably more common than not to want to impose constraints other than reducible equality constraints on such problems.

In this section, we look at two different types of constraints: logical constraints and scalar-valued inequality constraints. We show that scalar-valued inequality constraints are similar to conventional NLP constraints and can be dealt with in a similar way. We transfer a naïve quadratic penalty method to our setting as a proof of concept. Our discussion of logical constraints is mostly speculative in nature, but may provide a good starting point for future research.

3.2.1 Scalar-Valued Inequality Constraints

We refer to constraints as "scalar-valued inequality constraints" if they have the form $\Phi(U) \leq b$ where Φ is a set functional that maps a similarity class U to a real number (i.e., a scalar), and $b \in \mathbb{R}$. We assume generally that Φ is benignly continuously differentiable (see Definition 2.4.5 on page 154) because this allows us to use to the chain rule (see Theorem 2.4.6 on page 154). With the chain rule, we can reconstruct many of the algorithms that are commonly used in constrained nonlinear optimization. For the sake of brevity, we only implement a very simple quadratic penalty in the spirit of [NW06, sec. 17.1]. However, other types of algorithms could also conceivably be transferred.

We do not address problems with equality constraints. Section 3.2.1.4 details why we do not attempt to develop a theoretical basis for optimization under scalar-valued equality constraints.

Before we begin defining the algorithm, we need to address some theoretical concerns. Constrained optimization is often considered in terms of Lagrangian relaxations and we need to transfer this concept to our setting before we can transfer further concepts from conventional NLP theory. We follow the line of argumentation from [UU12, Sec. 16.3] closely, though we need to re-prove most results because of the difference in setting.

For the remainder of this section, we consider an optimization problem of the form

$$\begin{aligned} &\inf F(U) \\ &\text{s.t. } G_j(U) \leq 0 \quad \forall j \in [n] \\ &\qquad U \in \Sigma_{\sim}, \end{aligned} \tag{3.19}$$

where $\Sigma_{\sim} \coloneqq \Sigma/_{\sim_\mu}$ for a finite atomless measure space (X, Σ, μ), $n \in \mathbb{N}_0$, $F \colon \Sigma_{\sim} \to \mathbb{R}$ is a benignly differentiable set functional, and $G_j \colon \Sigma_{\sim} \to \mathbb{R}$ are benignly differentiable set functionals for every $j \in [n]$.

In this section, we frequently make arguments considering the behavior of differentiable set functionals along geodesics. As we have done in previous sections, we have to account for the fact that the gradient locally inverts on the step set whenever we make a step. To this end, we had previously introduced the locally inverted difference variation (see Definition 2.4.3 on page 153). As a reminder, the R-locally inverted difference variation between two signed measures φ and ν

is a non-negative measure that is given by

$$(\varphi \underset{R}{\ominus} \nu)(D) := |\varphi - \nu|(D \setminus R) + |\varphi + \nu|(D \cap R) \quad \forall D \in \Sigma_\sim.$$

The locally inverted difference variation is a specialized construct that is designed exclusively to compare derivatives of the same function at different points. Whether it is useful in any other context is questionable.

Definition 3.2.1 (Feasible and Optimal Points).
A point $U \in \Sigma_\sim$ is called *feasible* for Problem (3.19) if

$$G_j(U) \leq 0 \quad \forall j \in [n].$$

The set of all feasible points of the problem is

$$\mathscr{F}_G := \{U \in \Sigma_\sim \,|\, G_j(U) \leq 0 \ \forall j \in [n]\}.$$

U is called *locally optimal* or a *local optimum* if there exists a neighborhood $\mathscr{N} \subseteq \Sigma_\sim$ of U such that

$$F(V) \geq F(U) \qquad \forall V \in \mathscr{N} \cap \mathscr{F}_G.$$

U is called *globally optimal* or a *global optimum* if

$$F(V) \geq F(U) \qquad \forall V \in \mathscr{F}_G. \qquad \triangleleft$$

Optimality is generally not tested directly by using Definition 3.2.1. Instead, we formulate more accessible *optimality conditions* that can serve as a substitute. As with unconstrained optimization, we are primarily interested in necessary optimality conditions because they can be formulated without second derivatives. In constrained optimization, necessary first-order optimality conditions usually come in the form of so-called Karush-Kuhn-Tucker (KKT) conditions. To formulate KKT conditions for our setting, we first have to discuss constraint qualifications.

3.2.1.1 KKT CONDITIONS

The KKT conditions are based on the concept of tangent and normal cones. In real Hilbert spaces such as $\mathbb{R}^n$, the tangent cone to a set $X \subseteq \mathbb{R}^n$ in $x \in X$ represents the set of all directions that either point into the inside of X or are at least tangential to its boundary. The normal cone is the set of all directions $y \in \mathbb{R}^n$ such that $\langle d,\ y\rangle \leq 0$ for all d in the tangent cone. Figure 3.3 on the next page illustrates tangent and normal cone for a simple example in $\mathbb{R}^2$.

Tangent and normal cones can also be constructed in non-Hilbert spaces, though they no longer necessarily reside in the same vector space. Instead, the normal cone resides in the dual space of the original search space.

The difficulty with transferring this concept is that similarity spaces are not vector spaces. There are two notable issues that arise when we try to transfer tangent and normal cone to optimization problems in similarity spaces:

1. There are no cones in similarity spaces because there is no scaling;
2. Similarity spaces are not self-dual, so the normal cone likely lies in a different space.

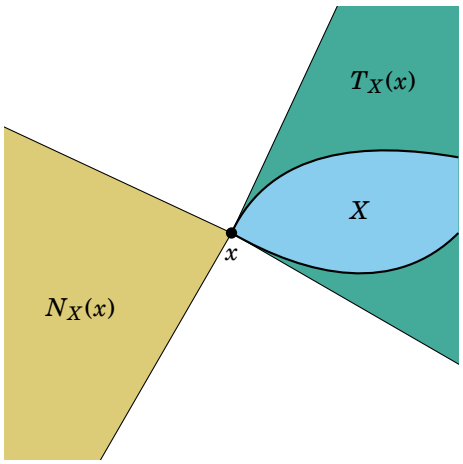


Figure 3.3: Illustration of tangent and normal cone in $\mathbb{R}^2$. The tangent cone of the set X in $x \in X$ is designated $T_X(x)$ while the normal cone is designated $N_X(x)$.

We can work around the first issue by interpreting cones as "sets of directions." We have previously noted that canonical geodesics can serve as a replacement for vectors as representations of "directions." To emphasize the analogy, we will subsequently refer to a suitable set of geodesics as a *pseudo-cone*.

Definition 3.2.2 (Directions, Prefixes, and Pseudo-Cones).
We refer to a minimizing canonical geodesic $\gamma\colon [0,\ L) \to \Sigma_\sim$ with $L > 0$ as a *direction* in $\Sigma_\sim$.

Let $\gamma\colon [0,\ a) \to \Sigma_\sim$ and $\delta\colon [0,\ b) \to \Sigma_\sim$ be directions with $0 < a \leq b$. We call γ a *prefix* of δ if $\gamma = \delta|_{[0,\ a)}$.

A set C of directions in $\Sigma_\sim$ is called a *pseudo-cone* if the following logical equivalence holds:

$$\delta \in C \iff \big(\gamma \in C\ \forall \gamma\colon \gamma \text{ prefix of } \delta\big).$$

We refer to this logical equivalence as the *prefix criterion*. We refer to

$$\mathrm{Dir}(\Sigma_\sim) := \{\gamma \,|\, \gamma \text{ is a direction in } \Sigma_\sim\}$$

as the *universal pseudo-cone of* $\Sigma_\sim$. ◁

Directions must be minimizing and have non-empty half-open parameter intervals. This excludes the edge cases where the parameter interval is empty or equal to $\{0\}$ or the geodesic is constant. Any geodesic with a geodesic constant $C > 0$ can be made into a minimizing geodesic by applying a linear parameter transformation. A half-open parameter interval also leaves open the possibility of $L = \infty$. Although infinite measure spaces are not discussed in this thesis, it is still notable that this theoretical construct does not require finiteness.

The prefix criterion appears difficult to verify at first. However, all pseudo-cones that we discuss here are defined based on the behavior for parameter values $t \to 0$. If δ is a direction and γ is a prefix of δ, then we always have $\delta(t) = \gamma(t)$ for t close to zero. The limit behavior is therefore always the same for δ and γ. We

skip individual verifications of the prefix criterion in favor of this general rule for the sake of brevity.

The dual space of a real vector space X is the space of all bounded linear forms on X. An analogue for a similarity space $\Sigma_\sim$ would be the set of all maps $\varphi\colon \Sigma_\sim \to \mathbb{R}$ that satisfy $\varphi(\emptyset) = 0$, are σ-additive, and satisfy $\big|\varphi(U)\big| \leq L \cdot \mu(U)$ for some $L \geq 0$. The first two properties form an analogue to linearity. The third property is analogous to boundedness. It guarantees that $\varphi \ll \mu$ and therefore that φ has a measurable density function f. It also guarantees that f is essentially bounded.

Thus, our equivalent is the space of all signed measures that are absolutely continuous with respect to μ and have density functions in $L^\infty(\Sigma, \mu)$. We will subsequently refer to these measures as *benign* measures in accordance with our prior definition of benign differentiability (see Definition 2.4.5).

A restriction to benignly differentiable set functionals also gives access to the chain rule (Theorem 2.4.6). We will use this to prove the differentiability of penalty terms.

Definition 3.2.3 (Benign Measures).
Let $(X,\ \Sigma,\ \mu)$ be a measure space. We refer to a signed measure $\varphi\colon \Sigma \to \mathbb{R} \cup \{\pm\infty\}$ with $\varphi \ll \mu$ for which there exists a density function $g \in L^\infty(\Sigma, \mu)$ such that

$$\varphi(A) = \int_A g \,\mathrm{d}\mu \qquad \forall A \in \Sigma$$

as a *benign measure*. ◁

The set of benign measures is a vector space because it is isomorphic to $L^\infty(\Sigma, \mu)$. We can easily verify that any benign measure whose underlying measure space is finite is itself finite-valued. Therefore, the space of benign measures is a subspace of the Banach space of finite signed measures.

Definition 3.2.4 (Spaces of Benign Measures).
Let (X, Σ, μ) be a finite atomless measure space, and let $\Sigma_\sim := {}^{\Sigma}\!/_{\sim_\mu}$. Let

$$M(\Sigma_\sim) := \big\{\varphi\colon \Sigma \to \mathbb{R} \,\big|\, \varphi \text{ is a benign measure in } (X,\ \Sigma,\ \mu)\big\}$$

be the vector space of benign measures in $(X,\ \Sigma,\ \mu)$, which is well-defined in relation to the associated similarity space $\Sigma_\sim$ because the fact that $\varphi \ll \mu$ for all $\varphi \in M(\Sigma_\sim)$ implies that φ assigns the same value to each member of a μ-similarity class.

We further introduce

$$M^+(\Sigma_\sim) := \big\{\varphi \in M(\Sigma_\sim) \,\big|\, \varphi(S) \geq 0 \ \forall S \in \Sigma_\sim\big\},$$
$$M^-(\Sigma_\sim) := \big\{\varphi \in M(\Sigma_\sim) \,\big|\, \varphi(S) \leq 0 \ \forall S \in \Sigma_\sim\big\}$$

to refer to the cones of all finite non-negative and non-positive benign measures on $(X,\ \Sigma,\ \mu)$, respectively. ◁

Lemma 3.2.5.
$M(\Sigma_\sim)$ is a real vector space and $M^\pm(\Sigma_\sim)$ are convex cones. For finite signed measures $\varphi^\pm\colon \Sigma \to \mathbb{R}$ for which there exist sets $P, N \in \Sigma$ with $\varphi^+(P) > 0$ and $\varphi^-(N) < 0$, respectively, we have $\operatorname{dist}\big(\varphi^+, M^-(\Sigma_\sim)\big) > 0$ *and* $\operatorname{dist}\big(\varphi^-, M^+(\Sigma_\sim)\big) > 0$. ◁

PROOF. Let $\varphi,\ \theta \in M(\Sigma_\sim)$ and let $r \in \mathbb{R}$. Then

$$\omega := \varphi + r \cdot \theta$$

is a signed measure. We further have

$$\begin{aligned}
\omega(A) &= \underbrace{\varphi(A)}_{\in\mathbb{R}} + r \cdot \underbrace{\theta(A)}_{\in\mathbb{R}} \in \mathbb{R} && \forall A \in \Sigma, \\
\omega(N) &= \underbrace{\varphi(N)}_{=0} + r \cdot \underbrace{\theta(N)}_{=0} = 0 && \forall N \in \Sigma\colon \mu(N) = 0.
\end{aligned}$$

Thus, ω is finite and we have $\omega \ll \mu$. Let $g, f \in L^\infty(\Sigma,\mu)$ be density functions of φ and θ, respectively. Then $w := g + r \cdot f \in L^\infty(\Sigma,\mu)$ is a density function of ω. Therefore, we have $\omega \in M(\Sigma_\sim)$.

This proves that $M(\Sigma_\sim)$ is a real vector space. To show that $M^+(\Sigma_\sim)$ is a cone, let $\varphi \in M^+(\Sigma_\sim)$ and let $r \geq 0$. Then we have

$$r \cdot \underbrace{\varphi(A)}_{\geq 0} \geq 0 \qquad \forall A \in \Sigma$$

and therefore $r\varphi \in M^+(\Sigma_\sim)$.

To show that $M^+(\Sigma_\sim)$ is convex, let $\varphi,\ \tau \in M^+(\Sigma_\sim)$, and let $r \in [0,\ 1]$. Then we have

$$\big(r \cdot \theta + (1-r) \cdot \varphi\big)(A) = r \cdot \underbrace{\theta(A)}_{\geq 0} + (1-r) \cdot \underbrace{\varphi(A)}_{\geq 0} \geq 0$$

and therefore $r \cdot \theta + (1-r) \cdot \varphi \in M^+(\Sigma_\sim)$.

Let $\varphi^- : \Sigma \to \mathbb{R}$ be a finite signed measure such that there exists a set $N \in \Sigma$ with $\varphi^-(N) < 0$. We note that φ^- need neither be benign nor absolutely continuous with respect to μ. For every measure $\psi \in M^+(\Omega)$, we have

$$\begin{aligned}
\|\psi - \varphi^-\| &= \sup\Big\{ \sum_{i=1}^{\infty} \big|\psi(A_i) - \varphi(A_i)\big| \,\Big|\, (A_i)_{i\in\mathbb{N}} \text{ pairwise disjoint partition of } X \Big\} \\
&\geq \big| \underbrace{\psi(N)}_{\geq 0} - \underbrace{\varphi^-(N)}_{<0} \big| + \underbrace{\big|\psi(N^\complement) - \varphi^-(N^\complement)\big|}_{\geq 0} \\
&\geq \big|\varphi^-(N)\big| \\
&> 0.
\end{aligned}$$

Therefore, we have $\operatorname{dist}\big(\varphi^-, M^+(\Sigma_\sim)\big) \geq \big|\varphi^-(N)\big| > 0$.

Finally, $M^-(\Sigma_\sim)$ is a convex cone because $M^-(\Sigma_\sim) = -M^+(\Sigma_\sim)$. The distance estimate follows by applying the same reasoning with $\varphi^- := -\varphi^+$. □

We note that the arguments proving Lemma 3.2.5 are not dependent on the restriction to benign measures. They hold equally for non-benign signed measures. We choose to restrict the space to benign measures because this simplifies notation later on, when we assume that all measures involved are benign. In the more general space, the distance condition simply means that the cones $M^\pm(\Sigma_\sim)$ are closed. For our restricted definition, it means that the closure of $M^\pm(\Sigma_\sim)$ within the space of finite signed measures consists entirely of non-negative or non-positive measures, respectively.

Even though pseudo-cones themselves are not cones, their "polar cone," which would reside in the "dual space" $M(\Sigma_\sim)$, is actually a proper cone in a normed vector space. To demonstrate this, we first have to properly define the polar cone of a pseudo-cone.

Definition 3.2.6 (Polar Cone of a Pseudo-Cone).
Let $C \subseteq \mathrm{Dir}(\Sigma_\sim)$ be a pseudo-cone in $\Sigma_\sim$. We refer to the set

$$C^\circ := \left\{ \varphi \in M(\Sigma_\sim) \,\middle|\, \limsup_{\substack{t \to 0 \\ t>0}} \frac{(\varphi \circ \gamma)(t)}{t} \leq 0 \ \forall \gamma \in C \right\}$$

as the *polar cone* of C. ◁

With the limes superior, we can apply bounds to the "derivative" of $\varphi \circ \gamma$ near $t = 0$ without having to assume that the limit exists. The existence of the limit would be a very strong requirement that would only be applicable to a small, carefully selected subset of geodesics.

Lemma 3.2.7.
Let $C \subseteq \mathrm{Dir}(\Sigma_\sim)$ be a pseudo-cone in $\Sigma_\sim$. Then C° is a convex cone in $M(\Sigma_\sim)$. ◁

PROOF. PART 1 (C° IS A CONE). Let $\varphi \in C^\circ$, let $r \geq 0$, and let $\omega := r\varphi \in M(\Sigma_\sim)$. For every $\gamma \in C$ and every sequence $(t_k)_{k\in\mathbb{N}}$ in $\mathrm{dom}(\gamma) \setminus \{0\}$ with $t_k \to 0$ for $k \to \infty$, we have

$$\begin{aligned}
\limsup_{k\to\infty} \frac{(\omega \circ \gamma)(t_k)}{t_k} &= \inf_{k_0 \in \mathbb{N}} \sup_{k \geq k_0} \frac{\omega\big(\gamma(t_k)\big)}{t_k} \\
&= \inf_{k_0 \in \mathbb{N}} \sup_{k \geq k_0} \frac{r \cdot \varphi\big(\gamma(t_k)\big)}{t_k} \\
&= r \cdot \inf_{k_0 \in \mathbb{N}} \sup_{k \geq k_0} \frac{\varphi\big(\gamma(t_k)\big)}{t_k} \\
&= r \cdot \limsup_{k\to\infty} \frac{\varphi\big(\gamma(t_k)\big)}{t_k} \\
&\leq r \cdot 0 \\
&= 0
\end{aligned}$$

where $r \geq 0$ allows us to exchange the scalar multiplication with the supremum and the infimum. Thus, we have $\omega \in C^\circ$.

PART 2 (C° IS CONVEX). Let $\varphi,\ \psi \in C^\circ$, and let $r \in [0,\ 1]$. Once more, we make use of the fact that multiplication with $r \geq 0$ and $1 - r \geq 0$ can be exchanged with forming the supremum and infimum, and that the supremum and infimum of a Minkowski sum are equal to the sum of suprema and infima, respectively.

Let $\omega := r\varphi + (1-r)\psi$. For every $\gamma \in C$ and every sequence $(t_k)_{k\in\mathbb{N}}$ in $\mathrm{dom}(\gamma)\setminus\{0\}$, we have

$$\begin{aligned}
\limsup_{k\to\infty} \frac{(\omega \circ \gamma)(t_k)}{t_k} &= \inf_{k_0 \in \mathbb{N}} \sup_{k \geq k_0} \frac{r\varphi\big(\gamma(t_k)\big) + (1-r)\psi\big(\gamma(t_k)\big)}{t_k} \\
&= \inf_{k_0 \in \mathbb{N}} \sup_{k \geq k_0} \left(r \cdot \frac{\varphi\big(\gamma(t_k)\big)}{t_k} + (1-r) \cdot \frac{\psi\big(\gamma(t_k)\big)}{t_k} \right)
\end{aligned}$$

$$
\begin{aligned}
&= r \cdot \left(\limsup_{k\to\infty} \frac{\varphi\big(\gamma(t_k)\big)}{t_k} \right) + (1-r) \cdot \left(\limsup_{k\to\infty} \frac{\psi\big(\gamma(t_k)\big)}{t_k} \right) \\
&\leq r \cdot 0 + (1-r) \cdot 0 \\
&= 0.
\end{aligned}
$$

Therefore, we have $\omega \in C^\circ$. □

We note that C° is generally not closed in the topology induced by the L^1 norm of the density function. This becomes clear when we consider that for a signed measure φ with density function f, we have

$$
\frac{\varphi\big(\gamma(t)\big)}{t} = \frac{\varphi\big(\gamma(t)\big)}{\mu\big(\gamma(t)\big)} = \frac{1}{\mu\big(\gamma(t)\big)} \cdot \int_{\gamma(t)} f \,\mathrm{d}\mu \qquad \forall t > 0.
$$

Thus, the quantity of interest is really the limes superior of the *mean value* of f over the infinitesimal similarity classes produced by γ for $t \to 0$. In L^1, even for very small $\varepsilon > 0$, an ε-perturbation can cause arbitrary changes in the average function value on very small similarity classes. We can therefore not reasonably expect C° to be closed according to L^1 topology. In L^∞ topology, however, C° may very well be closed, though we do not attempt to prove this here. The convex cone $M^-(\Sigma_\sim)$ of tuples of non-positive benign measures has a special role because it is the polar cone of the universal pseudo-cone.

Lemma 3.2.8.
The convex cone $M^-(\Sigma_\sim)$ is the polar cone of $\mathrm{Dir}(\Sigma_\sim)$. ◁

Proof. Part 1 ("⊆"). Let $\gamma \in \mathrm{Dir}(\Sigma_\sim)$ be a direction, and let $\varphi \in M^-(\Sigma_\sim)$. We have

$$
\varphi\big(\gamma(t)\big) \leq 0 \quad \forall t \in \mathrm{dom}(\gamma).
$$

Therefore, we have

$$
\limsup_{\substack{t\to 0 \\ t>0}} \frac{\varphi\big(\gamma(t)\big)}{t} = \inf_{t>0} \sup_{0<s\leq t} \frac{\overbrace{\varphi\big(\gamma(s)\big)}^{\leq 0}}{\underbrace{s}_{>0}} \leq 0.
$$

As this holds for all γ, φ is in the polar cone of $\mathrm{Dir}(\Sigma_\sim)$ and we have $M^-(\Sigma_\sim) \subseteq \big(\mathrm{Dir}(\Sigma_\sim)\big)^\circ$.

Part 2 ("⊇"). Conversely, let $\varphi \in M(\Sigma_\sim) \setminus M^-(\Sigma_\sim)$. Let $f \in L^\infty(\Sigma, \mu)$ be the density function of φ and let $D \coloneqq \{f > 0\}$. If $\mu(D) = 0$, then $\varphi \ll \mu$ would imply

$$
\varphi(U) = \underbrace{\varphi(U \cap D)}_{=0} + \int_{U\setminus D} \underbrace{f}_{\leq 0} \,\mathrm{d}\mu \leq 0 \quad \forall U \in \Sigma.
$$

Because φ is specifically chosen to not be non-positive, this cannot be the case. This indirectly demonstrates that we must have $\mu(D) > 0$. Let

$$
K \coloneqq \frac{1}{\mu(D)} \cdot \int_D f \,\mathrm{d}\mu = \frac{\varphi(D)}{\mu(D)} > 0
$$

be the mean value of φ over D. Let $E := \{f \geq K/2\} \subseteq D$. If $\mu(E) = 0$, then we would have

$$\frac{1}{\mu(D)} \cdot \int_D \underbrace{f}_{< \frac{K}{2} \text{ a.e.}} \mathrm{d}\mu \leq \frac{K}{2} < K.$$

This would contradict the definition of K. Therefore, we must have $\mu(E) > 0$.

According to Theorem 2.3.71, there exist an interval $I \subseteq \mathbb{R}$, a geodesic $\gamma\colon I \to \Sigma_\sim$, and $a,\ b \in I$ such that $\gamma(a) = \emptyset$ and $\gamma(b) = E$. Because $E \neq \emptyset$, we have $a \neq b$ and I contains at least two distinct points, which implies that γ has a unique geodesic constant. Furthermore, $\mu(E) > 0$ implies that the geodesic constant is not zero.

By affine linear parameter transformation and restriction, we may assume without loss of generality that $a = 0$, $b = \mu(E) > 0$, and $I = \big[0,\ \mu(E)\big]$. This makes γ a minimizing canonical geodesic. By excluding $\mu(E)$ from I, γ becomes a direction. For all $t \in \big[0,\ \mu(E)\big)$, we have $\gamma(t) \subseteq E$ because γ is canonical. This means that we have

$$\limsup_{\substack{t \to 0 \\ t > 0}} \frac{\varphi\big(\gamma(t)\big)}{t} = \limsup_{\substack{t \to 0 \\ t > 0}} \frac{1}{\mu\big(\gamma(t)\big)} \cdot \int_{\gamma(t)} \underbrace{f}_{\geq \frac{K}{2}} \mathrm{d}\mu \geq \frac{K}{2} > 0.$$

Thus, there exists $\gamma \in \mathrm{Dir}(\Sigma_\sim)$ such that

$$\limsup_{\substack{t \to 0 \\ t > 0}} \frac{\varphi(\gamma(t))}{t} > 0$$

and we have $\varphi \notin \big(\mathrm{Dir}(\Sigma_\sim)\big)^\circ$. In conclusion, we have

$$\big(M^-(\Sigma_\sim)\big)^\complement \subseteq \Big(\big(\mathrm{Dir}(\Sigma_\sim)\big)^\circ\Big)^\complement.$$

By taking the complement of both sides, we obtain the the subset relation

$$M^-(\Sigma_\sim) \supseteq \big(\mathrm{Dir}(\Sigma_\sim)\big)^\circ.$$

In conjunction with the converse inclusion that we have shown in the previous part of the proof, we have therefore demonstrated that

$$\big(\mathrm{Dir}(\Sigma_\sim)\big)^\circ = M^-(\Sigma_\sim). \qquad \square$$

Next, we define the tangent and normal "cones" of the feasible set.

Definition 3.2.9 (Tangent Pseudo-Cone and Normal Cone).
Let $\mathscr{X} \subseteq \Sigma_\sim$, and let $X \in \mathscr{X}$. The *tangent pseudo-cone* of $\mathscr{X}$ in X is given by

$$T_{\mathscr{X}}(X) := \left\{ \gamma \in \mathrm{Dir}(\Sigma_\sim) \,\middle|\, \lim_{\substack{t \to 0 \\ t > 0}} \frac{\mathrm{dist}\big(X \triangle \gamma(t), \mathscr{X}\big)}{t} = 0 \right\}$$

where

$$\mathrm{dist}(Y, \mathscr{X}) = \inf_{Z \in \mathscr{X}} \mu(Y \triangle Z)$$

The polar cone of $T_{\mathscr{X}}(X)$ is called the *normal cone* of $\mathscr{X}$ in X and is written as

$$N_{\mathscr{X}}(X) := \big(T_{\mathscr{X}}(X)\big)^\circ \subseteq M(\Sigma_\sim). \qquad \triangleleft$$

Because the definition of $T_{\mathscr{X}}(X)$ is solely based on the behavior of directions for $t \to 0$, the prefix criterion holds and $T_{\mathscr{X}}(X)$ is a pseudo-cone. With these definitions, it is already possible to formulate a necessary optimality criterion for optimization problems with scalar inequality constraints.

Proposition 3.2.10 (Necessary Optimality Under Scalar Inequalities). *Let $U \in \mathscr{F}_G$ be a locally optimal solution of Problem* (3.19)*. Then we have*

$$-\nabla F(U) \in N_{\mathscr{F}_G}(U). \qquad \triangleleft$$

Proof. Part 1 (Prerequisites). We prove the claim indirectly. Let U be such that $-\nabla F(U) \notin N_{\mathscr{F}_G}(U)$. Then there exists a direction $\gamma \in T_{\mathscr{F}_G}(U)$ such that

$$\limsup_{\substack{t \to 0 \\ t>0}} \frac{-\big(\nabla F(U) \circ \gamma\big)(t)}{t} > 0.$$

By Definition 3.2.2, γ has geodesic constant $C_\gamma = 1$ and $L > 0$.

This specifically means that we can choose a constant $M > 0$ and a sequence $(t_k)_{k\in\mathbb{N}}$ in $\operatorname{dom}(\gamma) \setminus \{0\}$ such that $t_k \to 0$ for $k \to \infty$ and $-\nabla F(U)\big(\gamma(t_k)\big) > M \cdot t_k$ for all $k \in \mathbb{N}$. For every $k \in \mathbb{N}$, let $V_k := U \triangle \gamma(t_k)$.

The objective functional F is benignly differentiable. Therefore, $\nabla F(U)$ has a density function $g_U \in L^\infty(\Sigma, \mu)$. Let

$$K := \|g_U\|_{L^\infty(\Sigma,\mu)}.$$

Our objective is to prove that for every $R > 0$, there exists a point $W_R \in \mathscr{F}_G$ such that $\mu(U \triangle W_R) < R$ and $F(W_R) < F(U)$. This then implies that U is not a local optimum.

Part 2 (Feasible Sequence W_k). While $V_k \to U$ for $k \to \infty$, we have no guarantee that $V_k \in \mathscr{F}_G$ for any k. This is where the defining property of the tangent pseudo-cone comes in. Because $\gamma \in T_{\mathscr{F}_G}(U)$, there exists $k_0 \in \mathbb{N}$ such that

$$\frac{\operatorname{dist}(V_k, \mathscr{F}_G)}{t_k} < \frac{M}{2K} \quad \forall k \geq k_0.$$

Due to the way in which distance is defined, this implies that for all $k \geq k_0$, there exists $W_k \in \mathscr{F}_G$ such that

$$\mu(V_k \triangle W_k) < \frac{M}{2K} \cdot t_k \quad \forall k \geq k_0.$$

Let subsequently

$$s_k := \mu(U \triangle W_k) \geq 0 \quad \forall k \geq k_0.$$

For each $k \geq k_0$, we have

$$s_k \leq \underbrace{\mu(U \triangle V_k)}_{=t_k} + \underbrace{\mu(V_k \triangle W_k)}_{<\frac{M}{2K}\cdot t_k} < \frac{2K+M}{2K} \cdot t_k \xrightarrow{k\to\infty} 0.$$

Let further

$$d_k := \frac{F(W_k) - F(U) - \nabla F(U)(W_k \triangle U)}{s_k} \quad \forall k \geq k_0.$$

According to the Taylor criterion (see Definition 2.4.1), we have $d_k \to 0$ for $k \to \infty$ because $s_k \to 0$ for $k \to \infty$.

PART 3 (DEVIATION OF LINEARIZATION). For brevity, let $D_1^k := U \triangle V_k$ and $D_2^k := U \triangle W_k$ for all $k \geq k_0$. We find that

$$\begin{aligned}\nabla F(U)(D_2^k) &= \nabla F(U)(D_1^k) + \nabla F(U)(D_2^k) - \nabla F(U)(D_1^k)\\ &= \nabla F(U)(D_1^k) + \int_{D_2^k} g_U \,\mathrm{d}\mu - \int_{D_1^k} g_U \,\mathrm{d}\mu\\ &= \nabla F(U)(D_1^k) + \int_{D_2^k \setminus D_1^k} g_U \,\mathrm{d}\mu - \int_{D_1^k \setminus D_2^k} g_U \,\mathrm{d}\mu.\end{aligned}$$

From this, we conclude that

$$\begin{aligned}\left|\nabla F(U)(D_2^k) - \nabla F(U)(D_1^k)\right| &\leq \left|\int_{D_2^k \setminus D_1^k} g_U \,\mathrm{d}\mu\right| + \left|\int_{D_1^k \setminus D_2^k} g_U \,\mathrm{d}\mu\right|\\ &\leq \int_{D_2^k \setminus D_1^k} |g_U| \,\mathrm{d}\mu + \int_{D_1^k \setminus D_2^k} |g_U| \,\mathrm{d}\mu\\ &= \int_{D_2^k \triangle D_1^k} |g_U| \,\mathrm{d}\mu\\ &\leq K \cdot \mu(D_2^k \triangle D_1^k).\end{aligned}$$

We note that

$$D_2^k \triangle D_1^k = U \triangle W_k \triangle U \triangle V_k = V_k \triangle W_k \quad \forall k \geq k_0$$

and therefore

$$K \cdot \mu(D_2^k \triangle D_1^k) = K \cdot \mu(V_k \triangle W_k) < \frac{KM}{2K} \cdot t_k = \frac{M}{2} \cdot t_k \quad \forall k \geq k_0,$$

which yields the overall estimate

$$\left|\nabla F(U)(D_2^k) - \nabla F(U)(D_1^k)\right| < \frac{M}{2} \cdot t_k \quad \forall k \geq k_0.$$

PART 4 (AGGREGATE ESTIMATE). For any given $R > 0$, there exists $k_1 \geq k_0$ such that

$$|d_k| \leq \frac{KM}{2K+M} \qquad \forall k \geq k_1.$$

In aggegate, for $k \geq k_1$, we have

$$\begin{aligned}F(W_k) &= F(U) + \nabla F(U)(U \triangle W_k) + d_k s_k\\ &= F(U) + \nabla F(U)(U \triangle V_k) + \big(\nabla F(U)(U \triangle W_k) - \nabla F(U)(U \triangle V_k)\big) + d_k s_k\\ &\leq F(U) + \nabla F(U)(U \triangle V_k) + \left|\nabla F(U)(U \triangle W_k) - \nabla F(U)(U \triangle V_k)\right| + |d_k| s_k\\ &< F(U) - M \cdot t_k + \frac{M}{2} \cdot t_k + |d_k| s_k\\ &\leq F(U) - M \cdot t_k + \frac{M}{2} \cdot t_k + \frac{KM}{2K+M} \cdot s_k\\ &\leq F(U) - M \cdot t_k + \frac{M}{2} \cdot t_k + \frac{KM}{2K+M} \cdot \frac{2K+M}{2K} \cdot t_k\\ &= F(U) - M \cdot t_k + \frac{M}{2} \cdot t_k + \frac{M}{2} \cdot t_k\\ &= F(U).\end{aligned}$$

Thus, we have $F(W_k) < F(U)$ for all $k \geq k_1$. This also implies $W_k \neq U$. However, because $s_k \to 0$, there exists $k_2 \geq k_1$ such that $s_k \leq R$ for all $k \geq k_2$. Let $W_R \coloneqq W_{k_2}$. Then we have

$$\mu(U \triangle W_R) = s_{k_2} \leq R,$$
$$F(W_R) < F(U).$$

As such a W_R exists for all $R > 0$, U is not a local minimum.

We have thus proven that $-\nabla F(U) \notin N_{\mathscr{F}_G}(U)$ implies that U is not a local minimum. This indirectly proves that

$$-\nabla F(U) \in N_{\mathscr{F}_G}(U)$$

is a necessary condition for U being a local minimum of F within $\mathscr{F}_G$. □

This gives us a technical necessary optimality criterion. We need to refine this in two significant aspects:

1. Proposition 3.2.10 only makes a statement about points that are local optima. As we can never depend on the existence of such points, we must find suboptimality estimators that make statements about points that are nearly feasible and nearly optimal.
2. Proposition 3.2.10 can only be applied directly if we can describe the normal cone in a way that is computationally tractible, which is why we require constraint qualifications.

We first focus on the second point. Ideally, we would like to express the tangent pseudo-cone and the normal cone in terms of the derivatives of the set functionals G_j. This would be a close analogue to how this is done in more conventional optimization.

Definition 3.2.11 (Linearized Tangent Pseudo-Cone).
Let $U \in \mathscr{F}_G$. We refer to

$$\widetilde{T}_G(U) \coloneqq \left\{ \gamma \in \mathrm{Dir}(\Sigma_\sim) \,\middle|\, \limsup_{\substack{t \to 0 \\ t>0}} \frac{\nabla G_j(U)\big(\gamma(t)\big)}{t} \leq 0 \ \forall j \in [n] \right\}$$

as the *linearized tangent pseudo-cone* of $\mathscr{F}_G$ in U. ◁

Definition 3.2.12 (Linearized Normal Cone).
Let $U \in \mathscr{F}_G$. We refer to

$$\widetilde{N}_G(U) \coloneqq M^-(\Sigma_\sim) + \left\{ \sum_{j=1}^{n} q_j \cdot \nabla G_j(U) \,\middle|\, q \in \mathbb{R}^n_{\geq 0},\ q_j = 0 \ \forall j \in [n] \colon G_j(U) < 0 \right\}$$

as the *linearized normal cone* of $\mathscr{F}_G$ in U. ◁

In Definition 3.2.12, we can already clearly see the familiar structure of the KKT conditions with Lagrange multipliers q. The role of constraint qualifications as preconditions to the KKT conditions is to ensure that the normal cone is equal to its linearized counterpart. The most basic of these constraint qualifications is the Guinard constraint qualification.

Definition 3.2.13 (Guinard Constraint Qualification (GCQ)).
Problem (3.19) is said to satisfy a *Guinard constraint qualification* in $U \in \mathscr{F}_G$ if

$$N_{\mathscr{F}_G}(U) = \widetilde{N}_G(U).$$

We more generally say that the problem satisfies "a constraint qualification" in U if it satisfies any condition implying the GCQ. ◁

If a GCQ is satisfied, then we can easily demonstrate an exact analogue of the Karush-Kuhn-Tucker theorem.

Theorem 3.2.14 (Karush-Kuhn-Tucker).
Let $U \in \Sigma_\sim$ be locally optimal with respect to Problem (3.19) *such that a constraint qualification holds in U. Then there exist $q \in \mathbb{R}^n_{\geq 0}$ such that*

$$\begin{aligned} \nabla F(U) + \sum_{j=1}^{n} q_j \cdot \nabla G_j(U) &\in M^+(\Sigma_\sim), \\ G_j(U) &\leq 0 \qquad \forall j \in [n], \\ q_j &= 0 \qquad \forall j \in [n] \colon G_j(U) < 0. \end{aligned}$$ ◁

PROOF. First, we note that being locally optimal with respect to Problem (3.19) implies feasibility, i.e., we have

$$G_j(U) \leq 0 \quad \forall j \in [n].$$

We therefore have $U \in \mathscr{F}_G$. We can invoke Proposition 3.2.10 to obtain

$$\nabla F(U) \in -N_{\mathscr{F}_G}(U) \overset{\text{Def. } 3.2.13}{=} -\widetilde{N}_G(U).$$

According to Definition 3.2.12, this implies that there exist $\varphi' \in M^-(\Sigma_\sim)$ and $q \in \mathbb{R}^n_{\geq 0}$ such that

$$\begin{aligned} \nabla F(U) &= -\varphi' - \sum_{i=1}^{n} q_j \nabla G_j(U), \\ q_j &= 0 \quad \forall j \in [n] \colon G_j(U) < 0. \end{aligned}$$

The claim then evidently follows because $-\varphi' \in M^+(\Sigma_\sim)$. □

A GCQ is difficult to prove directly. In conventional nonlinear optimization, there exist several alternative constraint qualifications that imply the GCQ. Instead of proving the equality of $N_{\mathscr{F}_G}(U)$ and $\widetilde{N}_G(U)$ directly, these alternatives generally prove that

$$T_{\mathscr{F}_G}(U) = \widetilde{T}_G(U)$$

and then rely on the fact that $\widetilde{N}_G(U)$ is the polar cone of $\widetilde{T}_G(U)$. Out of these alternative constraint qualifications, the Mangasarian-Fromovitz Constraint Qualification (MFCQ) is of particular interest to us.

Definition 3.2.15 (Mangasarian-Fromovitz Constraint Qualification).
Problem (3.19) is said to satisfy a *Mangasarian-Fromovitz Constraint Qualification* (MFCQ) in $U \in \mathscr{F}_G$ if there exists a similarity class $N \in \Sigma_\sim$ with $\mu(N) > 0$ such that there exists a constant $L > 0$ with

$$\nabla G_j(U)(D) \leq -L \cdot \mu(D) \quad \forall D \subseteq_\mu N, j \in [n] \colon G_j(U) = 0.$$ ◁

We note that this definition differs from the way in which MFCQs are defined in conventional nonlinear optimization. Usually, an MFCQ is satisfied if there exists a single direction along which every active constraint has strictly negative directional derivative. Here, we demand that there is a similarity class on which all derivative measures of active constraints have strictly negative gradient density by a margin of at least $L > 0$ pointwise almost everywhere. We can easily construct a suitable direction from this similarity class. However, the statement is substantially stronger in that we can construct different corrective directions from the class N according to our particular needs.

Our interest in MFCQs in particular is founded in the fact that directions connecting $\emptyset$ with subsets of N point into something that can be thought of as the "local linearized interior" of the feasible set $\mathscr{F}_G$. The existence of such an interior implies that there are no implicit equality constraints within the inequality constraints because equality constraints would locally reduce the dimension of the linearized feasible set and prevent it from having a non-empty interior.

Because equality constraints can be problematic in our setting, so can implicit equality constraints. By focusing on MFCQs, we automatically ensure that no implicit equality constraints exist. An implicit equality constraint is a conical combination of active constraints such that there is no direction in which the combined constraint becomes inactive.

Lemma 3.2.16 (Non-Existence of Implicit Equalities Under MFCQ). *Let $U \in \mathscr{F}_G$ be such that Problem* (3.19) *satisfies an MFCQ in U. Then there exist no coefficients $q \in \mathbb{R}^n_{\geq 0} \setminus \{0\}$ with $q_j = 0$ for all $j \in [n]$ with $G_j(U) < 0$ such that*

$$\sum_{j=1}^{n} q_j \cdot \nabla G_j(U)(D) \geq 0 \quad \forall D \in \Sigma_\sim.$$

◁

Proof. Because Problem (3.19) satisfies an MFCQ in U, there exists a similarity class $N \in \Sigma_\sim$ and a constant $L > 0$ such that $\mu(N) > 0$ and

$$\nabla G_j(U)(N) \leq -L \cdot \mu(N) \quad \forall j \in [n] \colon G_j(U) = 0.$$

We prove that N acts as a counterexample for every coefficent vector q. Let $q \in \mathbb{R}^n_{\geq 0} \setminus \{0\}$ with $q_j = 0$ for all $j \in [n]$ with $G_j(U) < 0$. We have

$$\begin{aligned}
\sum_{j=1}^{n} q_j \cdot \nabla G_j(U)(N) &= \sum_{\substack{j=1 \\ G_j(U)=0}}^{n} q_j \cdot \underbrace{\nabla G_j(U)(N)}_{\leq -L \cdot \mu(N)} \\
&= -L \cdot \mu(N) \cdot \sum_{\substack{j=1 \\ G_j(U)=0}}^{n} q_j \\
&= -L \cdot \mu(N) \cdot \sum_{j=1}^{n} q_j \\
&= -\underbrace{L}_{>0} \cdot \underbrace{\mu(N)}_{>0} \cdot \underbrace{\|q\|_1}_{>0} \\
&< 0.
\end{aligned}$$

□

The existence of such conical combinations would imply that in a linearized sense, if a step makes one of the participating constraints inactive, then another constraint would have to be violated to compensate the decrease in the conical combination of constraint functionals. In this sense, the participating constraints would always be active and would therefore locally act like equality constraints.

Next, we show the primary implication of an MFCQ, namely the equality $T_{\mathscr{F}_G}(U) = \widetilde{T}_G(U)$. This requires local derivative continuity.

Lemma 3.2.17.
Let Problem (3.19) *satisfy an MFCQ in $U \in \Sigma_\sim$, and let the derivative of G_j be continuous in U for each $j \in [n]$. Then*

$$T_{\mathscr{F}_G}(U) = \widetilde{T}_G(U). \qquad \triangleleft$$

PROOF. PART 1 ("⊆"). Let $\gamma \in \big(\widetilde{T}_G(U)\big)^{\complement}$. By definition, this means that there exists $j \in [n]$ with $G_j(U) = 0$ and

$$L \coloneqq \limsup_{\substack{t \to 0 \\ t > 0}} \frac{\nabla G_j(U)\big(\gamma(t)\big)}{t} > 0.$$

Our goal is to prove that $\gamma \notin T_{\mathscr{F}_G}(U)$. Because this is a statement about local behavior around U, we can select a small neighborhood around U within which G_j has desirable properties.

According to Definition 2.4.1, there exists $R_1 \in \operatorname{dom}(\gamma) \setminus \{0\}$ such that

$$\big|G_j\big(U \triangle \gamma(t)\big) - G_j(U) - \nabla G_j(U)\big(\gamma(t)\big)\big| \leq \frac{L}{4} \cdot \mu\big(\gamma(t)\big) = \frac{L}{4} \cdot t \quad \forall t \leq R_1. \tag{3.20}$$

Since G_j is benignly differentiable, $\nabla G_j(U)$ has a density function $g_j \in L^\infty(\Sigma, \mu)$. Because the derivative of G_j is continuous in U, there exists $R_2 > 0$ such that

$$\big(\nabla G_j(V) \underset{U \triangle V}{\ominus} \nabla G_j(U)\big)(W) \leq \frac{\|g_j\|_{L^\infty}}{4} \cdot \mu(W) \quad \forall V \in B_{2R_2}(U), W \in \Sigma_\sim. \tag{3.21}$$

We note that $\|g_j\|_{L^\infty} > 0$ follows from the fact that $L > 0$, which implies that there is a non-nullset N with $\nabla G_j(U)(N) > 0$. We define the aggregate radius

$$R \coloneqq \min\{R_1, R_2\} > 0.$$

Because $0 < R \leq R_1$, we have $R \in \operatorname{dom}(\gamma) \setminus \{0\}$.

We choose a sequence $(t_k)_{k \in \mathbb{N}}$ in $\operatorname{dom}(\gamma) \setminus \{0\}$ such that $t_k \to 0$, $t_k \leq R$ for all $k \in \mathbb{N}$, and

$$\nabla G_j(U)\big(\gamma(t_k)\big) \geq \frac{L}{2} \cdot t_k \quad \forall k \in \mathbb{N}. \tag{3.22}$$

This is possible because $L > 0$ is the limes superior of $\frac{\nabla G_j(U)(\gamma(t))}{t}$ for $t \to 0$. We now demonstrate that there exists a constant $C > 0$ such that

$$\operatorname{dist}\big(U \triangle \gamma(t_k), \mathscr{F}_G\big) \geq C \cdot t_k,$$

which implies that γ is not in the tangent cone $T_{\mathscr{F}_G}(U)$.

First, we establish a lower bound for the value of G_j at $U \mathbin{\triangle} \gamma(t_k)$. Because $t_k \leq R_1$, and because γ is a minimizing geodesic, we have $\mu\big(\gamma(t_k)\big) = t_k \leq R_1$ and therefore

$$\begin{aligned} G_j\big(U \mathbin{\triangle} \gamma(t_k)\big) &\overset{(3.20)}{\geq} G_j(U) + \nabla G_j(U)\big(\gamma(t_k)\big) - \frac{L}{4} \cdot t_k \\ &\overset{(3.22)}{\geq} G_j(U) + \frac{L}{4} \cdot t_k. \end{aligned}$$

This implies that $G_j\big(U \mathbin{\triangle} \gamma(t_k)\big) \geq \frac{L}{4} \cdot t_k > 0$ because $G_j(U) = 0$. Therefore, we have $V_k \coloneqq U \mathbin{\triangle} \gamma(t_k) \notin \mathscr{F}_G$.

Next, we establish an "exclusion radius" around V_k such that an open ball around V_k with that radius does not intersect $\mathscr{F}_G$. This radius must be no less than some fixed, strictly positive multiple of $\mu(U \mathbin{\triangle} V_k) = \mu\big(\gamma(t_k)\big) = t_k$, but must not exceed t_k, because otherwise, the ball would include U itself. We choose the exclusion radius

$$R'_k \coloneqq \min\left\{1,\ \frac{L}{5 \cdot \|g_j\|_{L^\infty}}\right\} \cdot t_k.$$

We will now show that $B_{R'_k}(V_k) \cap \mathscr{F}_G = \emptyset$. It is sufficient to show that $G_j(W) > 0$ for all $W \in B_{R'_k}(V_k)$.

Let $W \in B_{R'_k}(V_k)$. There exists a geodesic $\nu\colon I \to \Sigma_\sim$ with an interval $I \subseteq \mathbb{R}$ that connects V_k with W. Without loss of generality, let ν be minimizing, let $I = \big[0,\ \mu(V_k \mathbin{\triangle} W)\big]$, let $\nu(0) = V_k$, and let $\nu\big(\mu(V_k \mathbin{\triangle} W)\big) = W$. For every $t \in I$, we have

$$\begin{aligned} \mu(U \mathbin{\triangle} \nu(t)) &\leq \mu(U \mathbin{\triangle} V_k) + \mu(V_k \mathbin{\triangle} \nu(t)) \\ &= t_k + t \\ &< 2t_k \\ &< 2R. \end{aligned}$$

This implies that $\nu(t) \in B_{2R}(U)$ for all $t \in I$. Evidently, it follows that $\nu(t) \in B_{2R_2}(U)$ because $R \leq R_2$.

We now perform a polygon chain estimate along ν. Let $\varepsilon > 0$. For every $t \in I$, G_j is benignly differentiable in $\nu(t)$. Therefore, for every $t \in I$ there exists a radius $M(t) > 0$ such that

$$\begin{aligned} \Big|G_j\big(\nu(s)\big) - G_j\big(\nu(t)\big) - \nabla G_j\big(\nu(t)\big)\big(\nu(s) \mathbin{\triangle} \nu(t)\big)\Big| &\leq \varepsilon \cdot \mu\big(\nu(s) \mathbin{\triangle} \nu(t)\big) \\ &= \varepsilon \cdot |s - t| \end{aligned}$$

for all $s \in I$ with $|s - t| < M(t)$.

Because $I \subseteq \mathbb{R}$ is compact and $\big(B_{M(t')}(t')\big)_{t' \in I}$ is an open cover of I, the Heine-Borel theorem guarantees that there exists a finite subcover, i.e., there exists $N \in \mathbb{N}$ and a tuple $(t'_i)_{i \in [N]} \in I^N$ such that

$$I \subseteq \bigcup_{i=1}^{N} B_{M(t'_i)}(t'_i)$$

We assume without loss of generality that $i \mapsto t'_i$ is strictly increasing and that $B_{M(t'_i)}(t'_i) \cap B_{M(t'_{i+1})}(t'_{i+1}) \neq \emptyset$ for all $i \in [N-1]$. As we had argued in the proof

of Proposition 2.4.8, we can achieve the latter by eliminating redundant points. This allows us to choose a support tuple $(s_i)_{i\in[N]_0}$ with

$$\begin{aligned}
&s_0 \coloneqq 0 \in B_{M(t'_1)}(t'_1),\\
&s_i \in B_{M(t'_i)}(t'_i) \cap B_{M(t'_{i+1})}(t'_{i+1}) \cap [t'_i,\ t'_{i+1}] \quad \forall i \in [N-1],\\
&s_N \coloneqq \mu(V_k \mathbin{\triangle} W) \in B_{M(t'_N)}(t'_N).
\end{aligned}$$

This choice guarantees that s_{i-1} and s_i are contained within $B_{M(t'_i)}(t'_i)$ for all $i \in [n]$. It also guarantees that $s_{i-1} \le t'_i \le s_i$ for all $i \in [n]$ because we choose each s_i from the interval $[t'_i,\ t'_{i+1}]$.

We can rewrite the difference $G_j(W) - G_j(V_k)$ as a telescope sum

$$\begin{aligned}
G_j(W) - G_j(V_k) &= G_j\big(\nu(s_N)\big) - G_j\big(\nu(s_0)\big)\\
&= \sum_{i=1}^{N} \Big(G_j\big(\nu(s_i)\big) - G_j\big(\nu(s_{i-1})\big) \Big)\\
&= \sum_{i=1}^{N} \Big(G_j\big(\nu(s_i)\big) - G_j\big(\nu(t'_i)\big) - G_j\big(\nu(s_{i-1})\big) + G_j\big(\nu(t'_i)\big) \Big)\\
&= \sum_{i=1}^{N} \Big(\quad \nabla G_j\big(\nu(t'_i)\big)\big(\nu(s_i) \mathbin{\triangle} \nu(t'_i)\big)\\
&\qquad + G_j\big(\nu(s_i)\big) - G_j\big(\nu(t'_i)\big) - \nabla G_j\big(\nu(t'_i)\big)\big(\nu(s_i) \mathbin{\triangle} \nu(t'_i)\big)\\
&\qquad - \nabla G_j\big(\nu(t'_i)\big)\big(\nu(s_{i-1}) \mathbin{\triangle} \nu(t'_i)\big)\\
&\qquad - G_j\big(\nu(s_{i-1})\big) + G_j\big(\nu(t'_i)\big) + \nabla G_j\big(\nu(t'_i)\big)\big(\nu(s_{i-1}) \mathbin{\triangle} \nu(t'_i)\big) \Big)\\
&= \sum_{i=1}^{N} \Big(\nabla G_j\big(\nu(t'_i)\big)\big(\nu(s_i) \mathbin{\triangle} \nu(t'_i)\big) - \nabla G_j\big(\nu(t'_i)\big)\big(\nu(s_{i-1}) \mathbin{\triangle} \nu(t'_i)\big) \Big)\\
&+ \sum_{i=1}^{N} \Big(G_j\big(\nu(s_i)\big) - G_j\big(\nu(t'_i)\big) - \nabla G_j\big(\nu(t'_i)\big)\big(\nu(s_i) \mathbin{\triangle} \nu(t'_i)\big)\\
&\qquad - G_j\big(\nu(s_{i-1})\big) + G_j\big(\nu(t'_i)\big) + \nabla G_j\big(\nu(t'_i)\big)\big(\nu(s_{i-1}) \mathbin{\triangle} \nu(t'_i)\big) \Big).
\end{aligned}$$

For each $i \in [N]$, we have $\nu(t'_i) \in B_{2R}(U) \subseteq B_{2R_2}(U)$. We can therefore invoke Equation (3.21) to obtain

$$\begin{aligned}
\Big|\nabla G_j\big(\nu(t'_i)\big)\big(\nu(t'_i) \mathbin{\triangle} \nu(s_i)\big)\Big| &= \Big| \quad \nabla G_j(U)\Big(\big(\nu(t'_i) \mathbin{\triangle} \nu(s_i)\big) \setminus \big(U \mathbin{\triangle} \nu(t'_i)\big)\Big)\\
&\qquad + \Big(\nabla G\big(\nu(t'_i)\big) - \nabla G_j(U)\Big)\Big(\big(\nu(t'_i) \mathbin{\triangle} \nu(s_i)\big) \setminus \big(U \mathbin{\triangle} \nu(t'_i)\big)\Big)\\
&\qquad - \nabla G_j(U)\Big(\big(\nu(t'_i) \mathbin{\triangle} \nu(s_i)\big) \cap \big(U \mathbin{\triangle} \nu(t'_i)\big)\Big)\\
&\qquad + \Big(\nabla G\big(\nu(t'_i)\big) + \nabla G_j(U)\Big)\Big(\big(\nu(t'_i) \mathbin{\triangle} \nu(s_i)\big) \cap \big(U \mathbin{\triangle} \nu(t'_i)\big)\Big)\Big|\\
&\le \Big| \quad \nabla G_j(U)\Big(\big(\nu(t'_i) \mathbin{\triangle} \nu(s_i)\big) \setminus \big(U \mathbin{\triangle} \nu(t'_i)\big)\Big)\\
&\qquad - \nabla G_j(U)\Big(\big(\nu(t'_i) \mathbin{\triangle} \nu(s_i)\big) \cap \big(U \mathbin{\triangle} \nu(t'_i)\big)\Big)\Big|\\
&\qquad + \Big(\nabla G\big(\nu(t'_i)\big) \underset{\nu(t'_i) \mathbin{\triangle} U}{\ominus} \nabla G_j(U)\Big)\big(\nu(t'_i) \mathbin{\triangle} \nu(s_i)\big)\\
&\le \int_{\nu(t'_i) \mathbin{\triangle} \nu(s_i)} |g_j| \,\mathrm{d}\mu + \frac{\|g_j\|_{L^\infty}}{4} \cdot \mu\big(\nu(t'_i) \mathbin{\triangle} \nu(s_i)\big)
\end{aligned}$$

$$\begin{aligned}
&\leq \frac{5\cdot\|g_j\|_{L^\infty}}{4}\cdot\mu\big(\nu(t_i')\triangle\nu(s_i)\big)\\
&= \frac{5\cdot\|g_j\|_{L^\infty}}{4}\cdot(s_i-t_i').
\end{aligned}$$

Similarly, we find that

$$\Big|\nabla G_j\big(\nu(t_i')\big)\big(\nu(t_i')\triangle\nu(s_{i-1})\big)\Big| \leq \frac{5\cdot\|g_j\|_{L^\infty}}{4}\cdot(t_i'-s_{i-1}).$$

This yields the aggregate estimate

$$\begin{aligned}
\big|G_j(W)-G_j(V_k)\big| &= \sum_{i=1}^{N}\Big(\Big|\nabla G_j\big(\nu(t_i')\big)\big(\nu(s_i)\triangle\nu(t_i')\big)\Big| + \Big|\nabla G_j\big(\nu(t_i')\big)\big(\nu(s_{i-1})\triangle\nu(t_i')\big)\Big|\Big)\\
&\quad + \sum_{i=1}^{N}\Big(\quad\Big|G_j\big(\nu(s_i)\big)-G_j\big(\nu(t_i')\big)-\nabla G_j\big(\nu(t_i')\big)\big(\nu(s_i)\triangle\nu(t_i')\big)\Big|\\
&\qquad + \Big|G_j\big(\nu(s_{i-1})\big)+G_j\big(\nu(t_i')\big)+\nabla G_j\big(\nu(t_i')\big)\big(\nu(s_{i-1})\triangle\nu(t_i')\big)\Big|\Big)\\
&\leq \sum_{i=1}^{N}\left(\frac{5\cdot\|g_j\|_{L^\infty}}{4}\cdot(s_i-t_i'+t_i'-s_{i-1})+\varepsilon\cdot(s_i-t_i'+t_i'-s_{i-1})\right)\\
&= \left(\frac{5\cdot\|g_j\|_{L^\infty}}{4}+\varepsilon\right)\cdot\sum_{i=1}^{N}(s_i-s_{i-1})\\
&= \left(\frac{5\cdot\|g_j\|_{L^\infty}}{4}+\varepsilon\right)\cdot(s_N-s_0)\\
&= \left(\frac{5\cdot\|g_j\|_{L^\infty}}{4}+\varepsilon\right)\cdot\mu(V_k\triangle W).
\end{aligned}$$

With $\varepsilon\to 0$, we obtain

$$\big|G_j(W)-G_j(V_k)\big| \leq \frac{5\cdot\|g_j\|_{L^\infty}}{4}\cdot\mu(V_k\triangle W).$$

and therefore also

$$\begin{aligned}
G_j(W) &\geq G_j(V_k) - \big|G_j(W)-G_j(V_k)\big|\\
&\geq \underbrace{G_j(V_k)}_{\geq\frac{L}{4}\cdot t_k} - \frac{5\cdot\|g_j\|_{L^\infty}}{4}\cdot\underbrace{\mu(V_k\triangle W)}_{<R_k'\leq\frac{L}{5\cdot\|g_j\|_{L^\infty}}\cdot t_k}\\
&> \left(\frac{L}{4}-\frac{5\cdot\|g_j\|_{L^\infty}}{4}\cdot\frac{L}{5\cdot\|g_j\|_{L^\infty}}\right)\cdot t_k\\
&= \left(\frac{L}{4}-\frac{L}{4}\right)\cdot t_k\\
&= 0.
\end{aligned}$$

We have $G_j(W)>0$ and therefore $W\notin\mathscr{F}_G$. Since this holds for all $W\in B_{R_k'}(V_k)$, we have

$$\operatorname{dist}(V_k,\,\mathscr{F}_G)\geq R_k' = \underbrace{\min\left\{1,\,\frac{L}{5\cdot\|g_j\|_{L^\infty}}\right\}}_{=:C>0}\cdot t_k.$$

However, since $t_k \to 0$ and $\operatorname{dist}\big(U \triangle \gamma(t_k), \mathscr{F}_G\big) \geq C \cdot t_k$, we do not have

$$\lim_{\substack{t \to 0 \\ t>0}} \frac{\operatorname{dist}(U \triangle \gamma(t),\ \mathscr{F}_G)}{t} = 0$$

which demonstrates that $\gamma \notin T_{\mathscr{F}_G}(U)$. By demonstrating this, we have proven that $\big(\widetilde{T}_G(U)\big)^{\complement} \subseteq \big(T_{\mathscr{F}_G}(U)\big)^{\complement}$. This implies that

$$T_{\mathscr{F}_G}(U) \subseteq \widetilde{T}_G(U).$$

PART 2 ("$\supseteq$"). Let $\gamma \in \widetilde{T}_G(U) \subseteq \operatorname{Dir}(\Sigma_\sim)$. Because Problem (3.19) satisfies an MFCQ in U, there exists a similarity class $N \in \Sigma_\sim$ with $\mu(N) > 0$ and a constant $L > 0$ such that

$$\nabla G_j(U)(D) \leq -L \cdot \mu(D) \quad \forall D \subseteq_\mu N, j \in [n] \colon G_j(U) = 0.$$

We will use N as a disposable mass that we can use to "neutralize" constraint violations that occur along the tangential direction γ.

Let $\rho > 0$. Our goal is to show that there exists $R_\rho > 0$ such that $R_\rho \in \operatorname{dom}(\gamma)$ and

$$\operatorname{dist}\big(U \triangle \gamma(t), \mathscr{F}_G\big) \leq \rho \cdot t \quad \forall t \in [0,\ R_\rho].$$

To ensure that $R_\rho \in \operatorname{dom}(\gamma)$, we choose an upper bound $R_\gamma \in \operatorname{dom}(\gamma) \setminus \{0\}$. All $R_\rho \in (0,\ R_\gamma]$ satisfy $R_\rho \in \operatorname{dom}(\gamma)$. We subsequently assume without loss of generality that all geodesic parameters t under discussion satisfy $t \leq R_\gamma$ such that $\gamma(t)$ is always defined.

We will ultimately compensate constraint violations incurred along γ by adding an appropriately sized essential subset $D \subseteq_\nu N$ to $\gamma(t)$. This subset will have a size of $\mu(D) \leq \rho \cdot t$ such that the resulting set realizes the desired distance bound between $U \triangle \gamma(t)$ and $\mathscr{F}_G$. We cannot a priori ensure that $\gamma(t) \cap N$ is a nullset. Therefore, we must ensure that a sufficiently large subset of N remains for us to bridge the distance to $\mathscr{F}_G$, i.e., that

$$\mu\big(N \setminus \gamma(t)\big) \geq \rho \cdot t.$$

This is certainly the case if $\mu\big(\gamma(t)\big) \leq \mu(N) - \rho \cdot t$. Because $\mu\big(\gamma(t)\big) = t$, it is sufficient to demand that

$$t \leq \underbrace{\frac{\mu(N)}{1+\rho}}_{=:R_N}.$$

We have $\mu(N) > 0$ and $\rho > 0$, which ensures that $R_N > 0$. For $t \in [0,\ R_N]$, we have

$$\begin{aligned}
\mu\big(N \setminus \gamma(t)\big) &\geq \mu(N) - \mu\big(\gamma(t)\big) \\
&\geq \mu(N) + \rho \cdot \mu\big(\gamma(t)\big) - (1+\rho) \cdot \mu\big(\gamma(t)\big) \\
&= \mu(N) + \rho \cdot t - (1+\rho) \cdot t \\
&\geq \mu(N) + \rho \cdot t - (1+\rho) \cdot R_N \\
&= \mu(N) + \rho \cdot t - \mu(N) \\
&= \rho \cdot t.
\end{aligned}$$

This means that, because $(X,\ \Sigma,\ \mu)$ is atomless, we can always choose an essential subset $D \subseteq_\mu N \setminus \gamma(t)$ such that $\mu(D) = \rho \cdot t$. We now have to find a radius $R_{j,\rho} > 0$

for each $j \in [n]$ such that $G_j\big(U \mathbin{\triangle} \gamma(t) \mathbin{\triangle} D\big) \leq 0$ for all $t \in [0,\, R_{j,\rho}]$ and $D \subseteq_\mu N \setminus \gamma(t)$ with $\mu(D) = \rho \cdot t$. For this, we distinguish between active and inactive constraints. Let

$$J_< \coloneqq \big\{ j \in [n] \,\big|\, G_j(U) < 0 \big\}$$

be the set indexing all inactive constraint, and let

$$J_= \coloneqq \big\{ j \in [n] \,\big|\, G_j(U) = 0 \big\}$$

be the set indexing all active constraints. By definition, we have $J_< \cap J_= = \emptyset$. Because $U \in \mathscr{F}_G$, we have $J_< \cup J_= = [n]$.

We first consider inactive constraints. Let $j \in J_<$. Because $G_j(U) < 0$, we have $\big|G_j(U)\big| > 0$. Furthermore, $\rho > 0$ implies that $1 + \rho > 0$. Since G_j is continuous in U, there exists $R_{j,\rho} > 0$ such that

$$\big|G_j(V) - G_j(U)\big| \leq \big|G_j(U)\big| \quad \forall V \in \Sigma_\sim \colon \mu(U \mathbin{\triangle} V) \leq (1+\rho) \cdot R_{j,\rho}.$$

For all $t \in [0,\, R_{j,\rho}]$ and $D \in \Sigma_\sim$ with $\mu(D) \leq \rho \cdot t$, we have

$$\begin{aligned}
\mu\Big(U \mathbin{\triangle} \big(U \mathbin{\triangle} \gamma(t) \mathbin{\triangle} D\big)\Big) &= \mu\big(\gamma(t) \mathbin{\triangle} D\big) \\
&\leq \mu\big(\gamma(t) \cup D\big) \\
&\leq \mu\big(\gamma(t)\big) + \mu(D) \\
&\leq t + \rho \cdot t \\
&= (1+\rho) \cdot t \\
&\leq (1+\rho) \cdot R_{j,\rho}
\end{aligned}$$

and therefore

$$\begin{aligned}
G_j\big(U \mathbin{\triangle} \gamma(t) \mathbin{\triangle} D\big) &\leq G_j(U) + \Big|G_j\big(U \mathbin{\triangle} \gamma(t) \mathbin{\triangle} D\big) - G_j(U)\Big| \\
&\leq \underbrace{G_j(U)}_{<0} + \underbrace{\big|G_j(U)\big|}_{=-G_j(U)} \\
&= G_j(U) - G_j(U) \\
&= 0.
\end{aligned}$$

Thus, for every $t \leq R_{j,\rho}$ and every corrective step $D \in \Sigma_\sim$ with $\mu(D) \leq \rho \cdot t$, $U \mathbin{\triangle} \gamma(t) \mathbin{\triangle} D$ is feasible with respect to the j-th constraint.

We now turn our attention towards active constraints. Let $j \in J_=$. Because $G_j(U) = 0$ and $\gamma \in \widetilde{T}_G(U)$, we know that

$$\limsup_{\substack{t \to 0 \\ t > 0}} \frac{\nabla G_j(U)\big(\gamma(t)\big)}{t} \leq 0.$$

Let $R_{j,\rho,1} > 0$ such that

$$\nabla G_j(U)\big(\gamma(t)\big) \leq \frac{L \cdot \rho}{2} \cdot t \quad \forall t \in [0,\, R_{j,\rho,1}].$$

Because G_j is differentiable in U, there exists $R_{j,\rho,2} > 0$ such that

$$\big|G_j(V) - G_j(U) - \nabla G_j(U)(U \mathbin{\triangle} V)\big| \leq \frac{L}{2} \cdot \frac{\rho}{1+\rho} \cdot \mu(U \mathbin{\triangle} V)$$

for all $V \in \Sigma_\sim$ with

$$\mu(U \triangle V) \leq (1+\rho) \cdot R_{j,\rho,2}.$$

We choose $R_{j,\rho} \coloneqq \min\{R_{j,\rho,1}, R_{j,\rho,2}\}$. For all $t \in [0,\ R_{j,\rho}]$ and all corrective steps $D \in \Sigma_\sim$ with $D \subseteq_\mu N \setminus \gamma(t)$ and $\mu(D) = \rho \cdot t$, we have

$$\mu\Big(U \triangle \big(U \triangle \gamma(t) \triangle V\big)\Big) \leq (1+\rho) \cdot t \leq (1+\rho) \cdot R_{j,\rho,2}.$$

Furthermore, we have $\mu\big(\gamma(t)\big) = t \leq R_{j,\rho} \leq R_{j,\rho,1}$. In conjunction, this means that

$$\begin{aligned}
&G_j\big(U \triangle \gamma(t) \triangle D\big) \\
&\quad = \underbrace{G_j(U)}_{\leq 0} + \underbrace{G_j\big(U \triangle \gamma(t) \triangle D\big) - G_j(U) - \nabla G_j(U)\big(\gamma(t) \triangle D\big)}_{\leq \frac{L}{2} \cdot \frac{\rho}{1+\rho} \cdot \mu(\gamma(t) \triangle D)} + \nabla G_j(U)\big(\gamma(t) \triangle D\big) \\
&\quad \leq \frac{L}{2} \cdot \frac{\rho}{1+\rho} \cdot \underbrace{\mu\big(\gamma(t) \triangle D\big)}_{\leq (1+\rho) \cdot t} + \nabla G_j(U)\big(\underbrace{\gamma(t) \triangle D}_{\text{ess. disjoint}}\big) \\
&\quad \leq \frac{L \cdot \rho}{2} \cdot t + \underbrace{\nabla G_j(U)\big(\gamma(t)\big)}_{\leq \frac{L \cdot \rho}{2} \cdot t} + \underbrace{\nabla G_j(U)(D)}_{\leq -L \cdot \mu(D)} \\
&\quad \leq L \cdot \rho \cdot t - L \cdot \underbrace{\mu(D)}_{= \rho \cdot t} \\
&\quad = L \cdot \rho \cdot t - L \cdot \rho \cdot t \\
&\quad = 0.
\end{aligned}$$

Thus, for every $t \in [0,\ R_{j,\rho}]$, and every $D \in \Sigma_\sim$ with $D \subseteq_\mu N \setminus \gamma(t)$ and $\mu(D) = \rho \cdot t$, $U \triangle \gamma(t) \triangle D$ is feasible with respect to the j-th constraint.

We now form the aggregate radius

$$R_\rho \coloneqq \min\Big(\{R_\gamma, R_N\} \cup \{R_{j,\rho} \,|\, j \in [n]\}\Big) > 0.$$

Let $t \in [0,\ R_\rho]$. Because $t \leq R_\gamma$, we have $t \in \operatorname{dom}(\gamma)$, which means that $\gamma(t)$ is defined. Because $t \leq R_N$, there exists a similarity class $D \subseteq_\mu N \setminus \gamma(t)$ with $\mu(D) = \rho \cdot t$. For every $j \in J_<$, we have $t \leq R_{j,\rho}$. Because the corrective step D satisfies $\mu(D) \leq \rho \cdot t$, we have

$$G_j\big(U \triangle \gamma(t) \triangle D\big) \leq 0.$$

For every $j \in J_=$, we have $t \leq R_{j,\rho}$. Along with $D \subseteq_\mu N \setminus \gamma(t)$ and $\mu(D) = \rho \cdot t$, this implies that

$$G_j\big(U \triangle \gamma(t) \triangle D\big) \leq 0.$$

In summary, we have

$$G_j\big(U \triangle \gamma(t) \triangle D\big) \leq 0 \quad \forall j \in J_< \cup J_= = [n],$$

and therefore $U \triangle \gamma(t) \triangle D \in \mathscr{F}_G$. This means that

$$\begin{aligned}
\operatorname{dist}\big(U \triangle \gamma(t), \mathscr{F}_G\big) &\leq \mu\Big(\big(U \triangle \gamma(t)\big) \triangle \big(U \triangle \gamma(t) \triangle D\big)\Big) \\
&= \mu(D) \\
&= \rho \cdot t.
\end{aligned}$$

Thus, we find that for all $\rho > 0$, there exists $R_\rho > 0$ such that

$$\operatorname{dist}\big(U \triangle \gamma(t), \mathscr{F}_G\big) \leq \rho \cdot t \quad \forall t \in [0,\, R_\rho].$$

This implies that

$$\limsup_{\substack{t \to 0 \\ t>0}} \frac{\operatorname{dist}\big(U \triangle \gamma(t), \mathscr{F}_G\big)}{t} \leq \rho \quad \forall \rho > 0$$

and therefore

$$\limsup_{\substack{t \to 0 \\ t>0}} \frac{\operatorname{dist}\big(U \triangle \gamma(t), \mathscr{F}_G\big)}{t} \leq 0.$$

By definition, this means that $\gamma \in T_{\mathscr{F}_G}(U)$. Because this holds for all $\gamma \in \widetilde{T}_G(U)$, we have

$$T_{\mathscr{F}_G}(U) \supseteq \widetilde{T}_G(U). \qquad \square$$

We note that the proof for the subset relation $T_{\mathscr{F}_G}(U) \subseteq \widetilde{T}_G(U)$ is completely independent of the MFCQ. This relation always holds, irrespective of constraint qualification.

Lemma 3.2.17 shows equality of the tangent pseudo-cones of inequality-constrainted feasible sets under assumptions of continuous differentiability and an MFCQ. This is not sufficient to imply a GCQ, unless we can also prove that equality of the tangent pseudo-cones implies equality of the corresponding normal cones. This is only true if the linearized normal cone is shown to be the polar cone of the linearized tangent pseudo-cone. One inclusion is relatively easy to show.

Lemma 3.2.18.
Let $U \in \mathscr{F}_G$. Then we have

$$\widetilde{N}_G(U) \subseteq \big(\widetilde{T}_G(U)\big)^\circ. \qquad \triangleleft$$

Proof. Let $\varphi \in \widetilde{N}_G(U)$. According to Definition 3.2.12 there are $\varphi_0 \in M^-(\Sigma_\sim)$ and $q \in \mathbb{R}^n_{\geq 0}$ such that

$$\varphi = \varphi_0 + \sum_{j=1}^{n} q_j \cdot \nabla G_j(U)$$

and $q_j = 0$ for all $j \in [n]$ with $G_j(U) < 0$.

Let $\gamma \in \widetilde{T}_G(U)$. According to Definition 3.2.11, we have

$$\limsup_{\substack{t \to 0 \\ t>0}} \frac{\nabla G_j(U)\big(\gamma(t)\big)}{t} \leq 0$$

for all $j \in [n]$ with $G_j(U) = 0$. We can then use the fact that the limes superior is subadditive and positive homogeneous to show that

$$\begin{aligned}
\limsup_{\substack{t\to 0\\ t>0}} \frac{\varphi\big(\gamma(t)\big)}{t} &= \limsup_{\substack{t\to 0\\ t>0}} \Bigg(\frac{\varphi_0\big(\gamma(t)\big)}{t} + \sum_{\substack{j\in[n]\\ G_j(U)=0}} \underbrace{q_j}_{\geq 0} \cdot \frac{\nabla G_j(U)\big(\gamma(t)\big)}{t} \Bigg) \\
&\leq \underbrace{\limsup_{\substack{t\to 0\\ t>0}} \frac{\varphi_0\big(\gamma(t)\big)}{t}}_{\leq 0} + \sum_{\substack{j\in[n]\\ G_j(U)=0}} q_j \cdot \underbrace{\limsup_{\substack{t\to 0\\ t>0}} \frac{\nabla G_j(U)\big(\gamma(t)\big)}{t}}_{\leq 0} \\
&\leq 0.
\end{aligned}$$

As this holds for all $\gamma \in \widetilde{T}_G(U)$, we have $\varphi \in \big(\widetilde{T}_G(U)\big)^\circ$. □

The converse inclusion is very difficult to prove. We do not prove the general case here. Instead, we focus on the very limited special case where only one active constraint exists. This is sufficient for the test problem that we solve in Section 4.2, although it is not satisfactory as a general theoretical result. We state a more general conjecture as Hypothesis 3.2.20 on page 280.

Lemma 3.2.19.
Let $U \in \mathscr{F}_G$ such that there exists at most one $j \in [n]$ with $G_j(U) = 0$. Then we have

$$\big(\widetilde{T}_G(U)\big)^\circ \subseteq \widetilde{N}_G(U). \qquad \triangleleft$$

PROOF. PART 1 (OUTLINE AND EDGE CASE). Before we begin, we briefly specify what we have to prove. The cone $\big(\widetilde{T}_G(U)\big)^\circ$ consists of benign measures φ with $\varphi \ll \mu$ such that

$$\limsup_{\substack{t\to 0\\ t>0}} \frac{\varphi\big(\gamma(t)\big)}{t} \leq 0 \quad \forall \gamma \in \widetilde{T}_G(U).$$

If there are no active constraints, then we have to show that $\varphi \in M^-(\Sigma_\sim)$. This case is relatively simple, because $\widetilde{T}_G(U) = \mathrm{Dir}(\Sigma_\sim)$ if no constraints are active. In this case, Lemma 3.2.8 proves that $\big(\widetilde{T}_G(U)\big)^\circ = M^-(\Sigma_\sim) = \widetilde{N}_G(U)$.

We subsequently assume that there exists exactly one $j \in [n]$ such that $G_j(U) = 0$. We have to show that there exists a scalar multiplier $q \in \mathbb{R}_{\geq 0}$ such that

$$\varphi - q \cdot \nabla G_j(U) \in M^-(\Sigma_\sim).$$

Let $f \in L^\infty(\Sigma, \mu)$ be the density function of φ and let $g_j \in L^\infty(\Sigma, \mu)$ be the density function of $\nabla G_j(U)$. Then this is equivalent to finding $q \in \mathbb{R}_{\geq 0}$ such that

$$f - q \cdot g_j \leq 0 \quad \text{a.e. in } X.$$

This relation has to hold almost everywhere on every measurable subset of X. We can therefore partition X based on the values of f and g_j, and establish bounds on q for each part separately. If there exists a multiplier q that simultaneously satisfies the bounds for each part, the q also satisfies the overall condition.

Part 2 ($\{g_j \leq 0\} \cap \{f > 0\}$). Let $\varphi \in \big(\widetilde{T}_G(U)\big)^\circ$, and let $f \in L^\infty(\Sigma, \mu)$ be the density function of φ. Because G_j is benignly differentiable in U, $\nabla G_j(U)$ also has a density function $g_j \in L^\infty(\Sigma, \mu)$.

We begin by showing that $\{g_j \leq 0\} \cap \{f > 0\}$ is a nullset. Let

$$N_\varepsilon := \{g_j \leq 0\} \cap \{f \geq \varepsilon\} \quad \forall \varepsilon > 0.$$

Evidently, we have $\varepsilon \mapsto N_\varepsilon$ is monotonically increasing in ε with

$$\{g_j \leq 0\} \cap \{f > 0\} = \bigcup_{\varepsilon > 0} N_\varepsilon,$$

which means that

$$\mu\big(\{g_j \leq 0\} \cap \{f > 0\}\big) = \sup_{\varepsilon > 0} \mu(N_\varepsilon).$$

Therefore, if we show that $\mu(N_\varepsilon) = 0$ for all $\varepsilon > 0$, then this implies that $\{g_j \leq 0\} \cap \{f > 0\}$ is also a nullset. We prove this indirectly by showing that if there exists $\varepsilon > 0$ with $\mu(N_\varepsilon) > 0$, then $\varphi \notin \big(\widetilde{T}_G(U)\big)^\circ$. We prove this by constructing a direction $\nu \in \widetilde{T}_G(U)$ with

$$\limsup_{\substack{t \to 0 \\ t > 0}} \frac{\varphi\big(\nu(t)\big)}{t} > 0.$$

Let $\varphi \in M(\Sigma_\sim)$ with a density function $f \in L^\infty(\Sigma, \mu)$ such that there exists $\varepsilon > 0$ with $\mu(N_\varepsilon) > 0$. According to Theorem 2.3.71, there is an interval $I \subseteq \mathbb{R}$, $a,\ b \in I$, and a geodesic $\nu_\varepsilon : I \to \Sigma_\sim$ with $\nu_\varepsilon(a) = \emptyset$ and $\nu_\varepsilon(b) = N_\varepsilon$. Without loss of generality, let ν_ε be minimizing, and let $I = [a,\ b]$ with $a = 0$ and $b = \mu(N_\varepsilon) > 0$. By restricting ν_ε to $[0,\ b)$, we obtain $\nu_\varepsilon \in \mathrm{Dir}(\Sigma_\sim)$. However, for all $t \in \mathrm{dom}(\nu_\varepsilon) \setminus \{0\}$, we would have

$$\frac{\nabla G_j(U)\big(\nu_\varepsilon(t)\big)}{t} = \frac{1}{t} \cdot \int_{\nu_\varepsilon(t)} \underbrace{g_j}_{\leq 0} \, \mathrm{d}\mu \leq 0,$$

$$\frac{\varphi\big(\nu_\varepsilon(t)\big)}{t} = \frac{1}{t} \cdot \int_{\nu_\varepsilon(t)} \underbrace{f}_{\geq \varepsilon} \, \mathrm{d}\mu \geq \varepsilon > 0.$$

The former ensures that

$$\limsup_{\substack{t \to 0 \\ t > 0}} \frac{\nabla G_j(U)\big(\nu_\varepsilon(t)\big)}{t} \leq 0,$$

i.e., that $\nu_\varepsilon \in \widetilde{T}_G(U)$. However, the latter implies that

$$\limsup_{\substack{t \to 0 \\ t > 0}} \frac{\varphi\big(\nu_\varepsilon(t)\big)}{t} \geq \varepsilon > 0.$$

Therefore, we have $\varphi \notin (\widetilde{T}_G(U))^\circ$.

Conversely, $\varphi \in \big(\widetilde{T}_G(U)\big)^\circ$ implies $\mu(N_\varepsilon) = 0$ for all $\varepsilon > 0$. We can transfer this result to $N := \{g_j \leq 0\} \cap \{f > 0\}$ by using the equality

$$N = \bigcup_{i=1}^{\infty} N_{2^{-i}}.$$

The sequence $i \mapsto N_{2^{-i}}$ is an increasing sequence. Therefore, we have

$$\mu(N) = \lim_{i \to \infty} \mu(N_{2^{-i}}) = 0.$$

PART 3 ($\{g_j < 0\} \cap \{f \leq 0\}$ AND $\{g_j > 0\} \cap \{f > 0\}$). In the next step, we examine the sets $\{g_j < 0\} \cap \{f \leq 0\}$ and $\{g_j > 0\} \cap \{f > 0\}$. On the former f starts out being non-positive and increasing q subtracts a negative value from it, which risks making $f - q \cdot g_j$ locally positive. Therefore, we can infer an upper bound on q from that set. More precisely, q must be chosen such that

$$f - q \cdot g_j \leq 0 \quad \text{a.e. in } \{g_j < 0\} \cap \{f \leq 0\}.$$

This is equivalent to

$$q \leq \frac{f}{g_j} \quad \text{a.e. in } \{g_j < 0\} \cap \{f \leq 0\}.$$

We note that the relation is reversed because we divide by $g_j < 0$. We can rewrite as a single upper bound on q by using the essential infimum:

$$q \leq \underbrace{\operatorname*{ess\,inf}_{\{g_j<0\}\cap\{f\leq 0\}} \frac{f}{g_j}}_{=:q_{\mathrm{ub}}\geq 0}.$$

We can make a similar argument for $\{g_j > 0\} \cap \{f > 0\}$. Here, we can infer a lower bound on q that specifies to what minimal value q must be set such that

$$f - q \cdot g_j \leq 0 \quad \text{a.e. in } \{g_j > 0\} \cap \{f > 0\}.$$

Because $g_j > 0$ everywhere on $\{g_j > 0\}$, we can equivalently rewrite this as

$$q \geq \frac{f}{g_j} \quad \text{a.e. in } \{g_j > 0\} \cap \{f > 0\}.$$

In terms of the essential supremum, we can write this as

$$q \geq \underbrace{\operatorname*{ess\,sup}_{\{g_j>0\}\cap\{f>0\}} \frac{f}{g_j}}_{=:q_{\mathrm{lb}}\geq 0}.$$

Because $q_{\mathrm{lb}} \geq 0$, any q with $q \geq q_{\mathrm{lb}}$ also satisfies $q \geq 0$. We therefore only need to show that $[q_{\mathrm{lb}},\ q_{\mathrm{ub}}] \neq \emptyset$, i.e., that $q_{\mathrm{lb}} \leq q_{\mathrm{ub}}$. We prove this by contradiction. To improve readability, we divide the proof into multiple parts.

PART 3.1 (INITIAL ASSUMPTION). To generate a contradiction, we begin by assuming that

$$q_{\mathrm{lb}} > q_{\mathrm{ub}}.$$

According to the definition of q_{lb}, this means that there exists a measurable non-nullset $N_+ \subseteq \{g_j > 0\} \cap \{f > 0\}$ such that

$$\inf_{x \in N_+} \frac{f(x)}{g_j(x)} > q_{\mathrm{ub}}.$$

By definition of q_{ub}, there exists a measurable non-nullset $N_- \subseteq \{g_j < 0\} \cap \{f \leq 0\}$ such that

$$\underbrace{\inf_{x \in N_+} \frac{f(x)}{g_j(x)}}_{=:R_+} > \underbrace{\sup_{x \in N_-} \frac{f(x)}{g_j(x)}}_{=:R_-} \geq 0.$$

We define the following symbols for the mean values of g_j over N_+ and N_-, respectively:

$$M_+ := \frac{1}{\mu(N_+)} \cdot \int_{N_+} \underbrace{g_j}_{>0} \, \mathrm{d}\mu > 0,$$
$$M_- := \frac{1}{\mu(N_-)} \cdot \int_{N_-} \underbrace{g_j}_{<0} \, \mathrm{d}\mu < 0.$$

Both are well-defined and strictly different from zero because $\mu(N_+) > 0$ and $\mu(N_-) > 0$.

Part 3.2 (Refinement through Restriction). Our overall plan is to construct a direction $\nu \in \widetilde{T}_G(U)$ with

$$\limsup_{\substack{t \to 0 \\ t>0}} \frac{\varphi\big(\nu(t)\big)}{t} > 0,$$

which would act as a counterexample for $\varphi \in \big(\widetilde{T}_G(U)\big)^\circ$. We will do this by "mixing" subsets of N_+ and N_- in appropriate proportions to achieve a strictly positive slope with respect to φ while maintaining a non-positive slope with respect to $\nabla G_j(U)$.

In their current form, N_+ and N_- are not sufficient to construct this ν We need to ensure that g_j is bounded away from zero on both sets. We define the tightened sets

$$N_+^* := N_+ \cap \{g_j \geq M_+\},$$
$$N_-^* := N_- \cap \{g_j \leq M_-\}.$$

Both of these are non-nullsets because M_+ and M_- are the respective mean values of g_j. If $\mu(N_+^*) = 0$, then we would have

$$M_+ = \frac{1}{\mu(N_+)} \cdot \int_{N_+} \underbrace{g_j}_{<M_+ \text{ a.e.}} \, \mathrm{d}\mu < \frac{M_+ \cdot \mu(N_+)}{\mu(N_+)},$$

which would be contradictory. Similarly, if $\mu(N_-^*) = 0$, then we would have

$$M_- = \frac{1}{\mu(N_-)} \cdot \int_{N_-} \underbrace{g_j}_{>M_- \text{ a.e.}} \, \mathrm{d}\mu > \frac{M_- \cdot \mu(N_-)}{\mu(N_-)},$$

which would also be contradictory. By restricting the sets in this way, the slope of $\nabla G_j(U)$ on subsets of N_+^* is bounded below by $M_+ > 0$. Accordingly, because $\frac{f}{g_j}$ is bounded below by $R_+ > 0$, the slope of φ on those same subsets is bounded below by $R_+ \cdot M_+ > 0$.

On N_-^*, we obtain a similar guarantee. The slope of $\nabla G_j(U)$ on subsets of N_-^* is bounded above by $M_- < 0$. Because $0 \leq \frac{f}{g_j} \leq R_-$ on N_-^*, this means that the slope of φ on those same subsets is bounded above by $R_- \cdot M_- < 0$.

Part 3.3 (Component Geodesics and Mixing). Because N_+^* and N_-^* are non-nullsets, we can construct directions $\nu_+, \nu_- \in \mathrm{Dir}(\Sigma_\sim)$ with origin $\emptyset$ and destination N_+^* and N_-^*, respectively.

Because v_+ and v_- are canonical, the functions $m_+ \colon \operatorname{dom}(v_+) \to \mathbb{R}$ and $m_- \colon \operatorname{dom}(v_-) \to \mathbb{R}$ with

$$m_+(t) \coloneqq \int_{v_+(t)} \underbrace{g_j}_{>0} \, \mathrm{d}\mu \quad \forall t \in \operatorname{dom}(v_+),$$
$$m_-(t) \coloneqq \int_{v_-(t)} \underbrace{g_j}_{<0} \, \mathrm{d}\mu \quad \forall t \in \operatorname{dom}(v_-)$$

are both strictly monotonically increasing and decreasing, respectively, with $m_-(0) = m_+(0) = 0$. Furthermore, due to the continuity of geodesics and the absolute continuity of the Lebesgue integral, they are continuous. We note that this implies that both m_+ and m_- are invertible if their codomain is restricted to their image.

We restrict both geodesics to guarantee that we remain in a part of their domain where v_- can cancel out the linearized constraint violation caused by v_+. Let

$$M \coloneqq \frac{1}{2} \cdot \min\left\{ \sup_{t \in \operatorname{dom}(v_+)} m_+(t), \; - \inf_{t \in \operatorname{dom}(v_-)} m_-(t) \right\} > 0.$$

By halving M, we ensure that M and $-M$ lie within the image of m_+ and m_-, respectively. Let t_+ be such that $m_+(t_+) = M$ and let t_- be such that $m_-(t_-) = -M$. Let subsequently $m_+ \colon [0,\ t_+] \to [0,\ M]$ and $m_- \colon [0,\ t_-] \to [-M,\ 0]$. With this restriction, both m_+ and m_- are bijections.

To "mix" both geodesics, we use v_+ as a guidestone and mix in a prefix of v_- to precisely cancel out the change of $\nabla G_j(U)$ on $v_+(t)$. The parameter function for v_- is $p \colon [0,\ t_+] \to [0,\ t_-]$ with

$$p(t) \coloneqq m_-^{-1}\big(-m_+(t)\big) \quad \forall t \in [0,\ t_+].$$

Because both $-m_+$ and m_- are continuous, strictly monotonically decreasing maps and because m_- maps to a compact interval, p is continuous and strictly monotonically increasing. With this, we define the "proto-geodesic" $N \colon [0,\ t_+] \to \Sigma_\sim$ with

$$N(t) \coloneqq v_+(t) \cup v_-\big(p(t)\big) \quad \forall t \in [0,\ t_+]$$

Because v_+ and v_- are canonical geodesics and because p is strictly monotonically increasing $N(t)$ is μ-essentially increasing in t. N is not yet our desired geodesic v, primarily because it is not parameterized by length. Next, we proceed to correct this.

PART 3.4 (REPARAMETERIZATION). We have $p(0) = 0$ and therefore

$$N(0) = v_+(0) \cup v_-(0) = \emptyset.$$

Because $\operatorname{TV}(v_+) \subseteq_\mu N_+^*$, $\operatorname{TV}(v_-) \subseteq_\mu N_-^*$, and because N_+^* and N_-^* are disjoint, we have

$$\mu\big(N(t)\big) = \mu\big(v_+(t)\big) + \mu\Big(v_-\big(p(t)\big)\Big) = t + p(t) \quad \forall t \in [0,\ t_+].$$

Since p is continuous and strictly monotonically increasing, $\mu \circ N$ is also continuous and strictly monotonically increasing. The image of $\mu \circ N$ is

$$\big[0 + p(0),\ t_+ + p(t_+)\big] = [0,\ t_+ + t_-],$$

which is a compact interval. Therefore $\mu \circ N$ has a continuous and strictly monotonically increasing inverse

$$(\mu \circ N)^{-1} \colon [0,\ t_+ + t_-] \to [0,\ t_+].$$

We define $\nu \colon [0,\ t_+ + t_-] \to \Sigma_\sim$ with

$$\nu(t) \coloneqq N\big((\mu \circ N)^{-1}(t)\big) \quad \forall t \in [0,\ t_+ + t_-].$$

ν is our counterexample. Let $s, t \in [0,\ t_+ + t_-]$. Without loss of generality, let $s \leq t$. Because $(\mu \circ N)^{-1}$ is monotonically increasing and because N is μ-essentially increasing, so is ν, and we have

$$\begin{aligned}
\mu\big(\nu(s) \mathbin{\triangle} \nu(t)\big) &= \mu\big(\nu(t) \setminus \nu(s)\big) \\
&= \mu\big(\nu(t)\big) - \mu\big(\nu(s)\big) \\
&= \mu\Big(N\big((\mu \circ N)^{-1}(t)\big)\Big) - \mu\Big(N\big((\mu \circ N)^{-1}(s)\big)\Big) \\
&= (\mu \circ N)\big((\mu \circ N)^{-1}(t)\big) - (\mu \circ N)\big((\mu \circ N)^{-1}(s)\big) \\
&= t - s \\
&= |s - t|.
\end{aligned}$$

This demonstrates that ν is a minimizing geodesic. In conjunction with $\nu(0) = N(0) = \emptyset$, this means that ν is canonical. We remove the maximum $t_+ + t_-$ from the domain of ν to make ν into a direction.

Part 3.5 (Slopes of φ and $\nabla G_j(U)$). Having completed the construction of ν, we now have to prove that it is a counterexample for $\varphi \in \big(\widetilde{T}_G(U)\big)^\circ$, which generates the contradiction that we ultimately need to prove $q_{\mathrm{lb}} \leq q_{\mathrm{ub}}$. First, we prove that $\nu \in \widetilde{T}_G(U)$. This is relatively simple. For every $t \in [0,\ t_+]$, because ν_+ and ν_- have essentially disjoint total variation, we have

$$\begin{aligned}
\int_{N(t)} g_j \,\mathrm{d}\mu &= \int_{\nu_+(t)} g_j \,\mathrm{d}\mu + \int_{\nu_-(p(t))} g_j \,\mathrm{d}\mu \\
&= m_+(t) + m_-\big(p(t)\big) \\
&= m_+(t) + m_-\Big(m_-^{-1}\big(-m_+(t)\big)\Big) \\
&= m_+(t) - m_+(t) \\
&= 0.
\end{aligned}$$

Accordingly we have

$$\int_{\nu(t)} g_j \,\mathrm{d}\mu = \int_{N((\mu \circ N)^{-1}(t))} g_j \,\mathrm{d}\mu = 0 \quad \forall t \in \mathrm{dom}(\nu),$$

which implies that

$$\limsup_{\substack{t \to 0 \\ t > 0}} \frac{\nabla G_j(U)\big(\nu(t)\big)}{t} = \limsup_{\substack{t \to 0 \\ t > 0}} \frac{\int_{\nu(t)} g_j \,\mathrm{d}\mu}{t} = 0.$$

Because the j-th constraint is the only active constraint, this means that $\nu \in \widetilde{T}_G(U)$. Next, we examine the slope of φ along ν. For every $t \in \mathrm{dom}(\nu)$, we have

$$\int_{\nu(t)} f \,\mathrm{d}\mu = \int_{N((\mu \circ N)^{-1}(t))} f \,\mathrm{d}\mu$$

$$
\begin{aligned}
&= \int_{v_+((\mu\circ N)^{-1}(t))} \underbrace{f}_{\geq R_+\cdot g_j} \,\mathrm{d}\mu + \int_{v_-(p(((\mu\circ N)^{-1}(t))))} \underbrace{f}_{\geq R_-\cdot g_j} \,\mathrm{d}\mu \\
&\geq R_+ \cdot \int_{v_+((\mu\circ N)^{-1}(t))} g_j \,\mathrm{d}\mu + R_- \cdot \int_{v_-(p(((\mu\circ N)^{-1}(t))))} g_j \,\mathrm{d}\mu \\
&= (R_+ - R_-) \cdot \int_{v_+((\mu\circ N)^{-1}(t))} \underbrace{g_j}_{\geq M_+} \,\mathrm{d}\mu + \underbrace{R_- \cdot \int_{N((\mu\circ N)^{-1}(t))} g_j \,\mathrm{d}\mu}_{=0} \\
&= M_+ \cdot (R_+ - R_-) \cdot \mu\Big(v_+\big((\mu\circ N)^{-1}(t)\big)\Big) \\
&= \underbrace{M_+ \cdot (R_+ - R_-)}_{>0} \cdot (\mu\circ N)^{-1}(t)
\end{aligned}
$$

Therefore, φ has strictly positive slope along v if $(\mu\circ N)^{-1}$ has strictly positive slope. We now establish a lower bound on the slope of $(\mu\circ N)^{-1}$. Such a lower bound arises if the slope of $\mu\circ N$ is bounded above. We have previously shown that

$$\mu\big(N(t)\big) = t + p(t).$$

Therefore, we need to establish an upper bound on the slope of p. We have

$$p(t) = m_-^{-1}\big(-m_+(t)\big) \quad \forall t \in [0,\ t_+].$$

$p(t)$ is the unique solution of the equation

$$\int_{v_-(p(t))} \underbrace{g_j}_{\leq M_-} \,\mathrm{d}\mu = -\int_{v_+(t)} \underbrace{g_j}_{\leq \|g_j\|_{L^\infty}} \,\mathrm{d}\mu.$$

Because of the pointwise bounds on g_j, we obtain the inequality

$$M_- \cdot p(t) \geq -\|g_j\|_{L^\infty} \cdot t$$

or, equivalently,

$$p(t) \leq \frac{\|g_j\|_{L^\infty}}{|M_-|} \cdot t.$$

This implies that

$$\mu\big(N(t)\big) \leq \left(1 + \frac{\|g_j\|_{L^\infty}}{|M_-|}\right) \cdot t = \frac{|M_-| + \|g_j\|_{L^\infty}}{|M_-|} \cdot t$$

and therefore

$$(\mu\circ N)^{-1}(t) \geq \frac{|M_-|}{|M_-| + \|g_j\|_{L^\infty}} \cdot t.$$

This gives us the overall estimate

$$
\begin{aligned}
\int_{v(t)} f \,\mathrm{d}\mu &\geq M_+ \cdot (R_+ - R_-) \cdot (\mu\circ N)^{-1}(t) \\
&\geq \underbrace{\frac{M_+ \cdot |M_-| \cdot (R_+ - R_-)}{|M_-| + \|g_j\|_{L^\infty}}}_{>0} \cdot t
\end{aligned}
$$

for all $t \in \operatorname{dom}(v)$. We therefore have

$$\limsup_{\substack{t \to 0 \\ t>0}} \frac{\varphi\big(v(t)\big)}{t} = \limsup_{\substack{t \to 0 \\ t>0}} \frac{\int_{v(t)} f \,\mathrm{d}\mu}{t} \geq \frac{M_+ \cdot |M_-| \cdot (R_+ - R_-)}{|M_-| + \|g_j\|_{L^\infty}} > 0.$$

In conjunction with $v \in \widetilde{T}_G(U)$, this would prove that $\varphi \notin \big(\widetilde{T}_G(U)\big)^\circ$, which would contradict how we had initially chosen φ. To avoid this contradiction, our initial assumption that $q_{\mathrm{lb}} > q_{\mathrm{ub}}$ must have been wrong. Thus, we have $q_{\mathrm{lb}} \leq q_{\mathrm{ub}}$.

Part 4 ($\varphi \in \widetilde{N}_G(U)$). In the previous part, we have shown that

$$\operatorname*{ess\,sup}_{x \in \{g_j > 0\} \cap \{f > 0\}} \frac{f(x)}{g_j(x)} \leq \operatorname*{ess\,inf}_{x \in \{g_j < 0\} \cap \{f \leq 0\}} \frac{f(x)}{g_j(x)}.$$

Because $\frac{f}{g_j}$ is strictly positive on both sets, both sides of this inequality are non-negative. We can therefore choose $q \in \mathbb{R}_{\geq 0}$ such that

$$\operatorname*{ess\,sup}_{x \in \{g_j > 0\} \cap \{f > 0\}} \frac{f(x)}{g_j(x)} \leq q \leq \operatorname*{ess\,inf}_{x \in \{g_j < 0\} \cap \{f \leq 0\}} \frac{f(x)}{g_j(x)}.$$

Let $f_- \coloneqq f - q \cdot g_j \in L^\infty(\Sigma, \mu)$ and let $\varphi_- \in M(\Sigma_\sim)$ be the benign signed measure whose density function with respect to μ is f_-.

We have previously demonstrated that $\{f > 0\} \cap \{g \leq 0\}$ is a nullset. On $\{f \leq 0\} \cap \{g_j \geq 0\}$, we have

$$f_- = \underbrace{f}_{\leq 0} - \underbrace{q \cdot g_j}_{\geq 0} \leq 0$$

everywhere. On $\{f > 0\} \cap \{g_j > 0\}$, we have

$$f_- = f - \underbrace{q}_{\geq \frac{f}{g_j} \text{ a.e.}} \cdot \underbrace{g_j}_{>0} \leq f - \frac{f}{g_j} \cdot g_j = 0$$

almost everywhere. On $\{f \leq 0\} \cap \{g_j < 0\}$, we have

$$f_- = f - \underbrace{q}_{\leq \frac{f}{g_j} \text{ a.e.}} \cdot \underbrace{g_j}_{<0} \leq f - \frac{f}{g_j} \cdot g_j = 0$$

almost everywhere. Because

$$\begin{aligned} X = &\big(\{f > 0\} \cap \{g_j \leq 0\}\big) \\ &\cup \big(\{f > 0\} \cap \{g_j > 0\}\big) \\ &\cup \big(\{f \leq 0\} \cap \{g_j \geq 0\}\big) \\ &\cup \big(\{f \leq 0\} \cap \{g_j < 0\}\big), \end{aligned}$$

this means that $f_- \leq 0$ almost everywhere and therefore $\varphi_- \in M^-(\Sigma_\sim)$. Because

$$\varphi = \varphi_- + q \cdot \nabla G_j(U),$$

this means that $\varphi \in \widetilde{N}_G(U)$. This holds for all $\varphi \in \big(\widetilde{T}_G(U)\big)^\circ$, which means that

$$\big(\widetilde{T}_G(U)\big)^\circ \subseteq \widetilde{N}_G(U). \qquad \square$$

This partial result is not useful in a very large class of problems because it breaks down whenever there are any points in the feasible set where multiple constraints become active at the same time. We hypothesize that this is not a real restriction, but rather a limitation on the method that we use to prove the equality.

Hypothesis 3.2.20.
Let $U \in \mathscr{F}_G$ be such that Problem (3.19) *satisfies an MFCQ in U. Then we have*

$$\left(\widetilde{T}_G(U)\right)^\circ \subseteq \widetilde{N}_G(U). \qquad \triangleleft$$

We sketch roughly how a proof of Hypothesis 3.2.20 might be structured in Section A.2. However, we are unable to present a finished proof at this time. With Lemma 3.2.19, it is relatively easy to show that an MFCQ implies a GCQ under relatively mild continuous differentiability assumptions.

Theorem 3.2.21 (MFCQ implies GCQ).
Let $U \in \mathscr{F}_G$ be such that Problem (3.19) *satisfies an MFCQ in U. Let G_j be continuously and benignly differentiable in U for all $j \in [n]$ and let $G_j(U) = 0$ for no more than one j. Then Problem* (3.19) *also satisfies a GCQ in U, i.e., we have*

$$N_{\mathscr{F}_G}(U) = \widetilde{N}_G(U). \qquad \triangleleft$$

PROOF. We first invoke Lemma 3.2.17 to show that

$$\widetilde{T}_G(U) = T_{\mathscr{F}_G}(U).$$

By taking the polar cone of both sides, we obtain

$$\left(\widetilde{T}_G(U)\right)^\circ = N_{\mathscr{F}_G}(U).$$

We then invoke Lemmas 3.2.18 and 3.2.19 to show that

$$\left(\widetilde{T}_G(U)\right)^\circ = \widetilde{N}_G(U). \qquad \square$$

Of course, if we were to prove Hypothesis 3.2.20, then we could immediately remove the restriction to points with no more than one active constraint from Theorem 3.2.21.

3.2.1.2 SUBOPTIMALITY ESTIMATORS

Necessary optimality criteria are interesting from a theoretical point of view. However, as was the case in unconstrained optimization, they are of limited practical use as long as the optimum does not exist, which is often the case with the type of optimization problem that we are interested in.

Therefore, we must demonstrate that points that are "almost feasible" and "almost stationary" are also almost optimal. We do this by creating a suboptimality estimator. In the case of constrained optimization problems, both slight violations of constraints and slight instationarity should be accounted for in such an underestimator. As was the case with the unconstrained suboptimality estimator in Proposition 2.4.8, we assume Lipschitz continuous differentiability.

Theorem 3.2.22 (Suboptimality Estimator for Scalar Constraints).
Let $U \in \Sigma_\sim$ and let

- *F be Lipschitz continuously differentiable with Lipschitz constant $L_0 \geq 0$;*
- *G_j be Lipschitz continuously differentiable with Lipschitz constant $L_j \geq 0$ for every $j \in [n]$.*

Let further $q \in \mathbb{R}^n_{\geq 0}$ and $\varepsilon > 0$ be such that

$$\nabla F(U)(W) + \sum_{j=1}^{n} q_j \cdot \nabla G_j(U)(W) \geq -\varepsilon \quad \forall W \in \Sigma_{\sim}.$$

Then we have

$$F(V) \geq F(U) + \sum_{j=1}^{n} q_j \cdot G_j(U) - \varepsilon - \frac{L_0 + \sum_{j=1}^{n} q_j \cdot L_j}{2} \cdot \big(\mu(U \triangle V)\big)^2 \quad \forall V \in \mathscr{F}_G. \qquad \lhd$$

Proof. Let $V \in \mathscr{F}_G$. By definition of $\mathscr{F}_G$, we have $G_j(V) \leq 0$ for all $j \in [n]$. According to Proposition 2.4.8, we have

$$\big|G_j(V) - G_j(U) - \nabla G_j(U)(U \triangle V)\big| \leq \frac{L_j}{2} \cdot \big(\mu(U \triangle V)\big)^2 \quad \forall j \in [n].$$

This implies that

$$\begin{aligned} \nabla G_j(U)(U \triangle V) &\leq \underbrace{G_j(V)}_{\leq 0} - G_j(U) + \frac{L_j}{2} \cdot \big(\mu(U \triangle V)\big)^2 \\ &\leq \frac{L_j}{2} \cdot \big(\mu(U \triangle V)\big)^2 - G_j(U) \end{aligned}$$

holds for all $j \in [n]$. Therefore, we have

$$\begin{aligned} \nabla F(U)(U \triangle V) &\geq -\varepsilon - \sum_{j=1}^{n} q_j \cdot \nabla G_j(U)(U \triangle V) \\ &\geq -\varepsilon - \sum_{j=1}^{n} q_j \cdot \left(\frac{L_j}{2} \cdot \big(\mu(U \triangle V)\big)^2 - G_j(U) \right) \\ &\geq -\varepsilon + \sum_{j=1}^{n} q_j \cdot G_j(U) - \left(\sum_{j=1}^{n} \frac{q_j \cdot L_j}{2} \right) \cdot \big(\mu(U \triangle V)\big)^2 \end{aligned}$$

for all $j \in [n]$. Because F is Lipschitz continuously differentiable, Proposition 2.4.8 demonstrates that

$$\big|F(V) - F(U) - \nabla F(U)(U \triangle V)\big| \leq \frac{L_0}{2} \cdot \big(\mu(U \triangle V)\big)^2$$

and therefore

$$\begin{aligned} F(V) &\geq F(U) + \nabla F(U)(U \triangle V) - \frac{L_0}{2} \cdot \big(\mu(U \triangle V)\big)^2 \\ &\geq F(U) + \sum_{j=1}^{n} q_j \cdot G_j(U) - \varepsilon - \frac{L_0 + \sum_{j=1}^{n} q_j \cdot L_j}{2} \cdot \big(\mu(U \triangle V)\big)^2. \end{aligned} \qquad \square$$

We note that the complementarity condition, i.e., the requirement that $q_j = 0$ for all $j \in [n]$ with $G_j(U) < 0$, is absent from the assumptions of Theorem 3.2.22. It is not required for the overall estimate. However, if there existed $j \in [n]$ with $G_j(U) < 0$ and $q_j > 0$, then we would have $q_j \cdot G_j(U) < 0$, which would mean that the instationarity bound ε could no longer meaningfully be considered an upper bound on the optimality gap of the linearized objective function around U. We also note that the result can be generalized to locally Lipschitz continuously differentiable set functionals. This simply requires restricting all statements to a neighborhood of U and using a similarly generalized variant of Proposition 2.4.8.

3.2.1.3 QUADRATIC PENALTY METHOD

The only constrained optimization method that we develop here is a simple quadratic penalty method. The idea behind the quadratic penalty method is that we incorporate the constraint violation into the objective with a *penalty term* that has the form $\frac{m_j}{2} \cdot \max\{0,\, G_j(U)\}^2$ where $m_j \geq 0$ is a penalty parameter. This penalty term is a composition between the benignly differentiable constraint function G_j and the smooth function $t \mapsto \frac{m_j}{2} \cdot \max\{0,\, t\}^2$. According to the chain rule shown in Theorem 2.4.6, such a composition is itself benignly differentiable. We therefore need not be concerned that this might interfere with differentiability.

Definition 3.2.23.
We refer to the parameterized set functional $P\colon \Sigma_\sim \times \mathbb{R}^n_{\geq 0} \to \mathbb{R}$ with

$$P(U,m) \coloneqq F(U) + \sum_{j=1}^{n} \frac{m_j}{2} \cdot \max\big\{0,\, G_j(U)\big\}^2 \quad \forall U \in \Sigma_\sim,\ m \in \mathbb{R}^n_{\geq 0}$$

as the *penalty function* associated with Problem (3.19). ◁

As we have indicated, the penalty function is also benignly differentiable in U for fixed m. It is also evidently affine linear in m for each $j \in [n]$ for fixed U.

Lemma 3.2.24.
Let $U \in \Sigma_\sim$ be such that F as well as G_j for all $j \in [n]$ are benignly differentiable in U. Then the set functional $P(\cdot, m)$ is benignly differentiable in U. Its derivative has the form

$$\nabla_U P(U,m) = \nabla F(U) + \sum_{\substack{j \in [n] \\ G_j(U) \geq 0}} m_j \cdot G_j(U) \cdot \nabla G_j(U) \qquad \forall m \in \mathbb{R}^n_{\geq 0}. \tag{3.23}$$

◁

PROOF. Each summand of the penalty term has the form $h_j \circ G_j$ with $h_j\colon \mathbb{R} \to \mathbb{R}$ being defined by

$$h_j(x) \coloneqq \frac{m_j}{2} \cdot \max\{0,\, x\}^2 \quad \forall x \in \mathbb{R}.$$

According to Theorem 2.4.6, each of these summands is benignly differentiable in U and their respective derivatives have the form

$$\begin{aligned} \nabla(h_j \circ G_j)(U) &= \frac{\mathrm{d}h_j}{\mathrm{d}x}(G_j(U)) \cdot \nabla G_j(U) \\ &= \begin{cases} 0 & \text{if } G_j(U) < 0, \\ m_j \cdot G_j(U) \cdot \nabla G_j(U) & \text{if } G_j(U) \geq 0. \end{cases} \end{aligned}$$

The penalty function is a sum of set functionals that are benignly differentiable in U and is therefore itself benignly differentiable in U with the given derivative. □

Within the derivative of the penalty function, we can already see the familiar structure of the derivative of a Lagrange function. If we allow for slight violations of the constraints and consider a constraint with $G_j(U) > 0$ as "active," then $q \in \mathbb{R}^n$ with

$$q_j := \begin{cases} 0 & \text{if } G_j(U) < 0, \\ m_j \cdot G_j(U) & \text{if } G_j(U) \geq 0 \end{cases}$$

for $j \in [n]$ evidently satisfies $q_j \geq 0$ for all $j \in [n]$ and therefore functions as a multiplier vector for Theorems 3.2.14 and 3.2.22. For fixed penalty parameters $m \in \mathbb{R}^n_{\geq 0}$, it is very easy to show that continuous differentiability transfers from F and $(G_j)_{j\in[n]}$ to $P(\cdot, m)$.

Lemma 3.2.25.
Let $m \in \mathbb{R}^n_{\geq 0}$, and let $U \in \Sigma_\sim$. Let F as well as all G_j for $j \in [n]$ be continuously differentiable in U. Then $P(\cdot, m)$ is continuously differentiable in U. ◁

Proof. Let $\varepsilon > 0$, and let F as well as all G_j be continuously differentiable in U. We define $(C_j)_{j\in[n]}$ via

$$C_j := \max\Big\{1,\ \big|G_j(U)\big|,\ \|g_j\|_{L^\infty}\Big\}$$

where $g_j \in L^\infty(\Sigma, \mu)$ is the density function of $\nabla G_j(U)$. This is a legitimate choice because all G_j are benignly differentiable. We then choose $(\delta_j)_{j\in[2n]_0} \in \mathbb{R}_{>0}^{[2n]_0}$ such that

$$\begin{aligned}
\big(\nabla F(U) \underset{U\triangle V}{\ominus} \nabla F(V)\big)(W) &\leq \frac{\varepsilon}{n+1} \cdot \mu(W) && \forall V, W \in \Sigma_\sim : \mu(U \triangle V) \leq \delta_0, \\
\big|G_j(U) - G_j(V)\big| &\leq \underbrace{\frac{\varepsilon}{3C_j m_j (n+1)}}_{=:M_j} && \forall V \in \Sigma_\sim : \mu(U \triangle V) \leq \delta_{2j-1}, \\
\big(\nabla G_j(U) \underset{U\triangle V}{\ominus} \nabla G_j(V)\big)(W) &\leq \min\{C_j,\ M_j\} \cdot \mu(W) && \forall V, W \in \Sigma_\sim : \mu(U \triangle V) \leq \delta_{2j}
\end{aligned}$$

and select $\delta := \min\{\delta_0, \ldots, \delta_{2n}\} > 0$. Let $V \in \Sigma_\sim$ with $\mu(U \triangle V) \leq \delta$. We have

$$\begin{aligned}
\big|\nabla G_j(V)(W)\big| &\leq \Big|\nabla G_j(V)\big(W \setminus (U \triangle V)\big)\Big| + \Big|\nabla G_j(V)\big(W \cap (U \triangle V)\big)\Big| \\
&\leq \Big|\nabla G_j(U)\big(W \setminus (U \triangle V)\big) + \big(\nabla G_j(V) - \nabla G_j(U)\big)\big(W \setminus (U \triangle V)\big)\Big| \\
&\quad + \Big|-\nabla G_j(U)\big(W \cap (U \triangle V)\big) + \big(\nabla G_j(V) + \nabla G_j(U)\big)\big(W \cap (U \triangle V)\big)\Big| \\
&\leq \big|\nabla G_j(U)\big|(W) + \big(\nabla G_j(V) \underset{U\triangle V}{\ominus} \nabla G_j(U)\big)(W) \\
&\leq 2C_j \cdot \mu(W)
\end{aligned}$$

for all $W \in \Sigma_\sim$. The last estimate in this chain stems from the fact that

$$\big|\nabla G_j(U)\big|(W) = \int_W |g_j| \,\mathrm{d}\mu \leq \|g_j\|_{L^\infty} \cdot \mu(W) \leq C_j \cdot \mu(W).$$

Because the measure μ is σ-additive, we can apply this to partitions, which yields the estimate

$$\begin{aligned}\left|\nabla G_j(V)\right|(W) &= \sup_{\pi \text{ partition of } W} \sum_{A\in\pi} \left|\nabla G_j(V)(A)\right| \\ &\leq \sup_{\pi \text{ partition of } W} \sum_{A\in\pi} 2C_j \cdot \mu(A) \\ &= \sup_{\pi \text{ partition of } W} 2C_j \cdot \mu(W) \\ &= 2C_j \cdot \mu(W).\end{aligned}$$

From this, we can then infer that

$$\begin{aligned}&\Big(\max\{0,\, G_j(U)\} \cdot \nabla G_j(U) \underset{U\triangle V}{\ominus} \max\{0,\, G_j(V)\} \cdot \nabla G_j(V)\Big)(W) \\ &\quad\leq \left|G_j(U)\right| \cdot \big(\nabla G_j(U) \underset{U\triangle V}{\ominus} \nabla G_j(V)\big)(W) \\ &\qquad + \left|G_j(U) - G_j(V)\right| \cdot \left|\nabla G_j(V)\right|(W) \\ &\quad\leq C_j M_j \cdot \mu(W) + M_j \cdot 2C_j \cdot \mu(W) \\ &\quad= 3C_j M_j \cdot \mu(W) \\ &\quad= \frac{\varepsilon}{m_j \cdot (n+1)} \cdot \mu(W).\end{aligned}$$

In aggregate, this yields the estimate

$$\begin{aligned}\left|\big(\nabla_U P(U,m) \ominus_{U\triangle V} \nabla_U P(V,m)\big)(W)\right| &\leq \Big(\frac{\varepsilon}{n+1} + \sum_{j=1}^{n} \frac{\varepsilon}{n+1}\Big) \cdot \mu(W) \\ &= \varepsilon \cdot \mu(W).\end{aligned}$$

Therefore, $P(\cdot, m)$ is continuously differentiable in U. □

These results can also be extended to uniform continuity in cases where the constants C_j can be chosen uniformly. Notably, this is the case on sets $\mathscr{U} \subseteq \Sigma_\sim$ where all G_j are bounded above and their derivatives are bounded. More precisely, we require an L^∞ form of boundedness that takes the form

$$\sup_{U\in\mathscr{U}} \|g_{j,U}\|_{L^\infty} < \infty \quad \forall j \in [n]$$

where $g_{j,U} \in L^\infty(\Sigma, \mu)$ is the density function of $\nabla G_j(U)$ for each $j \in [n]$ and $U \in \mathscr{U}$.

We note that this does not extend to Lipschitz-continuous differentiability because the derivative of the penalty term involves products of $G_j(U)$ and $\nabla G_j(U)$ which can increase the rate of growth. This is not a major issue because Theorem 3.2.22 only requires Lipschitz-continuity of the derivative for each of the component functions, not their aggregate.

Being able to transfer continuity of the derivative from the component functions to the penalty function opens up the possibility of applying the unconstrained descent framework (Algorithm 4 on page 232) to the penalty function.

Preconditions

The quadratic penalty method is identical in many aspects to the unconstrained evaluation loop. As we had previously done in Section 3.1, we aggregate all preconditions for the algorithm in one statement.

Assumption 3.2.26 (Preconditions for the Quadratic Penalty Method). Let

(1) (X, Σ, μ) be a finite atomless measure space with $\mu(X) > 0$;

(2) $\Sigma_\sim := \Sigma/_{\sim_\mu}$, $n \in \mathbb{N}$, $q \in \{1, \infty\}$;

(3) $F \colon \Sigma_\sim \to \mathbb{R}$ be a benignly and uniformly continuously differentiable set functional that is bounded below on $\Sigma_\sim$;

(4) $\widetilde{F} \colon \Sigma_\sim \times \mathbb{R}_{>0} \cup \{\infty\} \to \mathbb{R} \times \mathbb{R}_{\geq 0}$ be an error-controlled evaluation method for F;

(5) $\widetilde{f} \colon \Sigma_\sim \times \mathbb{R}_{>0} \cup \{\infty\} \to L^q(\Sigma, \mu) \times \mathbb{R}_{\geq 0}$ be an L^q-controlled gradient evaluation routine for F;

(6) $G_j \colon \Sigma_\sim \to \mathbb{R}$ for $j \in [n]$ be benignly and uniformly continuously differentiable such that $G_j(U)$ and $\left\| g_j(U) \right\|_{L^\infty}$ are bounded above on $\Sigma_\sim$ where $g_j(U) \in L^\infty(\Sigma, \mu)$ denotes the density function of $\nabla G_j(U)$;

(7) $\widetilde{G}_j \colon \Sigma_\sim \times \mathbb{R}_{>0} \cup \{\infty\} \to \mathbb{R} \times \mathbb{R}_{\geq 0}$ be an error-controlled evaluation method for G_j for each $j \in [n]$;

(8) $\widetilde{g}_j \colon \Sigma_\sim \times \mathbb{R}_{>0} \cup \{\infty\} \to L^\infty(\Sigma, \mu) \times \mathbb{R}_{\geq 0}$ be an L^q-controlled gradient evaluation routine for G_j for each $j \in [n]$;

(9) $\mathscr{S} \colon L^1(\Sigma, \mu) \times \mathbb{R}_{>0} \times \mathbb{R}_{>0} \cup \{\infty\} \to \Sigma_\sim \times \mathbb{R}_{\geq 0}$ be a controlled unconstrained step finding routine with quality $\theta > 0$;

(10) $\sigma_0,\ \sigma_1,\ \sigma_2 \in \mathbb{R}$ such that $0 < \sigma_0 < \sigma_2$ and $\sigma_0 < \sigma_1 < 1$;

(11) $\varepsilon_\tau > 0$, $\varepsilon_v > 0$, and $\overline{\varepsilon}_\tau \geq \varepsilon_\tau$;

(12) $\xi_\tau \in (0,\ 1)$, $\xi_v \in (0,\ 1)$, $\xi_\delta \in (0,\ 1)$, $\xi_g \in (0,\ 1 - \sigma_1)$;

(13) $w_f \in \mathbb{R}^{n+1}_{>0}$, $w_g \in \mathbb{R}^{n+1}_{>0}$, $\beta_{0,f} \in \mathbb{R}_{>0} \cup \{\infty\}$, $\beta_{0,g} \in \mathbb{R}_{>0} \cup \{\infty\}$, $\xi \in (0,\ 1)$, $\zeta \in (0,\ 1)^n$, $\epsilon \in \mathbb{R}^n_{>0}$;

(14) $m_{\text{init}} > 0$, $\overline{m} \geq m_{\text{init}}$, $U_0 \in \Sigma_\sim$, $\Delta_0 \in (0,\ \mu(X)]$. ◁

We note that Assumption 3.2.26 is almost identical to Assumption 3.1.9. This is an intentional choice so that for fixed $m \in \mathbb{R}^n_{\geq 0}$, these parameters satisfy Assumption 3.1.9 with the penalty function $P(\cdot, m)$ as the objective. The primary differences are the following:

- there are additional assumptions for the constraint functionals G_j;
- all functionals must be benignly differentiable;
- we add an infeasibility tolerance ε_v;
- we add a relaxed instationarity tolerance threshold $\overline{\varepsilon}_\tau$;

- we add ξ_v, ζ, and ϵ as error control tuning parameters;
- we add error weight parameters w_f and w_g;
- we add an initial penalty parameter m_{init} and an upper penalty parameter bound $\overline{m}$.

In the following sections, we elaborate on the purpose of these added parameters.

Notes on Error Control

In order to apply Algorithm 4 to the penalty function, we first have to find controlled evaluation methods for the penalty function and its gradient. Luckily, we can derive these from the corresponding evaluation methods for the objective and constraint functions.

First, we break the combined error margin $\eta > 0$ up into individual allocations for each summand of the penalty term. To account for the fact that set functionals may have different degrees of susceptibility to error, this allocation process is controlled by a "weight vector" $w \in \mathbb{R}^{n+1}_{>0}$ where w_j is the weight of the j-th penalty term for $j \in [n]$ and w_{n+1} is the weight of the objective term. For each $j \in [n+1]$, we calculate a term error bound

$$\eta_j \coloneqq \frac{w_j m_j \cdot \eta}{w_{n+1} + \sum_{j=1}^{n} w_j m_j} \quad \forall j \in [n],$$
$$\eta_{n+1} \coloneqq \frac{w_{n+1} \cdot \eta}{w_{n+1} + \sum_{j=1}^{n} w_j m_j}.$$

This choice guarantees that $\eta_j > 0$ for all $j \in [n]$ with $m_j > 0$ and $\eta_{n+1} > 0$. It also ensures that

$$\sum_{j=1}^{n+1} \eta_j = \frac{w_{n+1} + \sum_{j=1}^{n} w_j m_j}{w_{n+1} + \sum_{j=1}^{n} w_j m_j} \cdot \eta = \eta,$$

which ensures that the cumulative error of the penalty function does not exceed η. We note that the denominator is always greater than zero because w_{n+1} appears without a multiplier. If a summand is known ahead of time to evaluate without error and if its evaluator accepts an error bound of zero, then we could set $w_j = 0$ to apportion none of the error bound to the corresponding term. However, in this case, we would have to otherwise ensure that the denominator is non-zero or we would have to provide a fallback mechanism if this occurs. It is easier to assume that $w_j > 0$ for all $j \in [n+1]$.

Once we have broken up our joint error bound η into individual summand error bounds η_j, we need to enforce those individual error bounds. For the objective term, this is straightforward because we can pass η_j as an error bound on to the objective functional evaluator $\widetilde{F}$.

For the penalty terms, this is more complex. If a constraint functional evaluator $\widetilde{G}_j$ yields a result $\widetilde{v} \in \mathbb{R}$ with a deviation $|v - \widetilde{v}| \leq \varepsilon$ from the true value $v \in \mathbb{R}$

for a returned error bound $\varepsilon \geq 0$, then we have

$$\begin{aligned}
&\left| \frac{m_j}{2} \max\{0,\ v\}^2 - \frac{m_j}{2} \max\{0,\ \widetilde{v}\}^2 \right| \\
&\qquad = \frac{m_j}{2} \cdot \left| \max\{0,\ v\}^2 - \max\{0,\ \widetilde{v}\}^2 \right| \\
&\qquad = \frac{m_j}{2} \cdot \underbrace{\left| \max\{0,\ v\} + \max\{0,\ \widetilde{v}\} \right|}_{\leq 2\max\{0,\ \widetilde{v}\} + \varepsilon} \cdot \underbrace{\left| \max\{0,\ v\} - \max\{0,\ \widetilde{v}\} \right|}_{\leq \varepsilon} \\
&\qquad \leq \frac{m_j}{2} \cdot \varepsilon \cdot \left(2\max\{0,\ \widetilde{v}\} + \varepsilon \right).
\end{aligned}$$

Thus, the error of the penalty term depends on both the error estimate returned by the evaluator and the approximate value. The quadratic inequality

$$\varepsilon^2 + 2\max\{0,\ \widetilde{v}\} \cdot \varepsilon \leq \frac{2\eta_j}{m_j}$$

can be solved for ε, yielding

$$\varepsilon \leq \sqrt{\max\{0,\ \widetilde{v}\}^2 + \frac{2\eta_j}{m_j}} - \max\{0,\ \widetilde{v}\}.$$

However, this is problematic because we control ε through the error bound η, which also affects the value of $\widetilde{v}$. We therefore have to use the error controlled evaluation loop (Algorithm 1 on page 215) with $\widetilde{G_j}$ as an evaluator and

$$\omega(x) \coloneqq \sqrt{\max\{0,\ x\}^2 + \frac{2\eta_j}{m_j}} - \max\{0,\ x\}$$

as a bound oracle to obtain sufficiently accurate approximate values of G_j. We can further improve this procedure by reducing the ε returned by $\widetilde{G}_j$ to $\widetilde{v} + \varepsilon$ for $\widetilde{v} < 0$.

Lemma 3.2.27 (Penalty Summand Evaluator).
Let $(X,\ \Sigma,\ \mu)$, $\Sigma_\sim$, F, $\widetilde{F}$, n, $(G_j)_{j\in[n]}$, and $(\widetilde{G}_j)_{j\in[n]}$ satisfy Assumption 3.2.26. Let $j \in [n]$, $\xi \in (0,\ 1)$, and $\beta_0 \in \mathbb{R}_{>0} \cup \{\infty\}$ be fixed. For fixed $m \geq 0$, $U \in \Sigma_\sim$, and $\eta > 0$, let $\phi\colon \mathbb{R}_{>0} \cup \{\infty\} \to \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0}$ be a procedure such that $\phi(\beta)$ obtains an evaluate $(x,\ e)$ from $\widetilde{G}_j(U,\beta)$ and yields

$$\begin{aligned}
\phi_1(\beta) &\coloneqq \max\{0,x\}, \\
\phi_2(\beta) &\coloneqq \min\{e,\ \max\{0,x+e\}\},
\end{aligned}$$

and let $\omega\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0} \cup \{\infty\}$ be given by

$$\omega(x) \coloneqq \sqrt{x^2 + \frac{2\eta}{m}} - x \quad \forall x \geq 0$$

where $\frac{2\eta}{m} = \infty$ for $m = 0$. Let $\widetilde{P}_j\colon \Sigma_\sim \times \mathbb{R}_{\geq 0} \times \mathbb{R}_{>0} \cup \{\infty\} \to \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0}$ be a procedure such that $\widetilde{P}_j(U,m,\eta)$ obtains an output tuple $(\widetilde{x},\ \widetilde{e})$ of Algorithm 1 with evaluator ϕ, bound oracle ω, initial bound β_0, and decay rate ξ, and yields

$$\begin{aligned}
\widetilde{P}_{j,1}(U,m,\eta) &\coloneqq \frac{m}{2} \cdot \widetilde{x}^2, \\
\widetilde{P}_{j,2}(U,m,\eta) &\coloneqq \frac{m}{2} \cdot \widetilde{e} \cdot (2\widetilde{x} + \widetilde{e}).
\end{aligned}$$

Then $(U,\eta) \mapsto \widetilde{P}_j(U,m,\eta)$ *is a controlled evaluation method for the set functional*

$$U \mapsto \frac{m}{2} \max\{0,\ G_j(U)\}^2. \qquad \triangleleft$$

PROOF. PART 1 (ϕ IS A SUITABLE EVALUATOR). We note that $\widetilde{G}_j$ is a controlled evaluator of G_j. Therefore, for every $U \in \Sigma_\sim$ and $\beta \in \mathbb{R}_{>0} \cup \{\infty\}$, any output tuple (x,e) of $\widetilde{G}_j(U,\beta)$ satisfies

$$\left|G_j(U) - x\right| \le e \le \beta.$$

Let $(x',e') \in \mathbb{R}_{\ge 0} \times \mathbb{R}_{\ge 0}$ be an output tuple of $\phi(\beta)$ and let (x,e) be the output tuple internally obtained from $\widetilde{G}_j(U,\beta)$ by ϕ to calculate that output tuple. We have

$$\begin{aligned}\left|\max\{0,\ G_j(U)\} - x'\right| &= \left|\max\{0,\ G_j(U)\} - \max\{0,\ x\}\right| \\ &\le \left|G_j(U) - x\right| \\ &\le e.\end{aligned}$$

For $x < 0$, we have $x' = \max\{0,\ x\} = 0$ and therefore

$$\begin{aligned}\left|\max\{0,\ G_j(U)\} - x'\right| &= \max\{0,\ G_j(U)\} \\ &\le \max\{0,\ x+e\}.\end{aligned}$$

In conjunction, these two inequalities yield

$$\left|\max\{0,\ G_j(U)\} - x'\right| \le \min\{e,\ \max\{0,\ x+e\}\} = e'.$$

In addition, we have

$$e' = \min\{e,\ \max\{0,\ x+e\}\} \le e \le \beta.$$

This means that ϕ satisfies the requirements for an evaluator procedure in Algorithm 1.

PART 2 (ω IS A SUITABLE BOUND ORACLE). The map ω satisfies $\omega(x) > 0$ for all x because $\eta > 0$. We now have to show that ω is bounded from below in the sense that there exists a constant $\overline{e} > 0$ such that for all $\beta^* \in \mathbb{R}_{>0} \cup \{\infty\}$ and any output tuple $(x',\ e')$ of $\phi(\beta^*)$, we have

$$e' \le \overline{e} \implies e' \le \omega(x').$$

We make a case distinction based on the value of m.

Case 1 ($m = 0$). In this case, we have $\omega(x) = \infty$ for all $x \in \mathbb{R}_{\ge 0}$. Because $e' < \infty$ by definition, $e' \le \overline{e}$ implies $e' \le \omega(x')$ for all $\overline{e} \in \mathbb{R}_{>0} \cup \{\infty\}$. $\triangleleft$

Case 2 ($m > 0$). In this case, we have $\omega(x) < \infty$ for all $x \in \mathbb{R}_{\ge 0}$. Because U is fixed, the "true" value

$$x^* \coloneqq \max\{0,\ G_j(U)\} \ge 0$$

that is being approximated by ϕ is constant. The mapping $x \mapsto \omega(x)$ is differentiable in x, and for $x \geq 0$, we have

$$\begin{aligned}\frac{\mathrm{d}\omega}{\mathrm{d}x}(x) &= 2x \cdot \frac{1}{2 \cdot \sqrt{x^2 + \frac{2\eta}{m}}} - 1 \\ &= \underbrace{\sqrt{\frac{x^2}{x^2 + \frac{2\eta}{m}}}}_{\in(0,\,1)} - 1 \\ &< 0.\end{aligned}$$

This shows that ω is strictly decreasing in x. If $e' \leq \overline{e}$ for some $\overline{e} > 0$, then our results from Part 1 show that

$$|x^* - x'| \leq e' \leq \overline{e}$$

and therefore $x' \leq x^* + \overline{e}$, which implies

$$\omega(x') \geq \omega(x^* + \overline{e}).$$

The most straightforward way to show the desired implication is to choose $\overline{e} > 0$ such that

$$\overline{e} = \omega(x^* + \overline{e}) = \sqrt{(x^* + \overline{e})^2 + \frac{2\eta}{m}} - (x^* + \overline{e}),$$

or equivalently

$$2\overline{e} + x^* = \sqrt{(x^* + \overline{e})^2 + \frac{2\eta}{m}}.$$

Both sides of this equation are strictly positive. By squaring both sides, we obtain the quadratic equation

$$4\overline{e}^2 + 4x^*\overline{e} + (x^*)^2 = (x^* + \overline{e})^2 + \frac{2\eta}{m} = (x^*)^2 + 2x^*\overline{e} + \overline{e}^2 + \frac{2\eta}{m}.$$

Through some equivalent reformulation, we obtain the normal form

$$\overline{e}^2 + \frac{2}{3}x^*\overline{e} - \frac{2\eta}{3m} = 0.$$

Using standard solution formulae, we can obtain a positive solution of this equation. We choose

$$\overline{e} := \sqrt{\left(\frac{x^*}{3}\right)^2 + \frac{2\eta}{3m}} - \frac{1}{3}x^* > 0.$$

With this choice, we obtain the desired implication

$$e' \leq \overline{e} \implies e' \leq \underbrace{\omega(x^* + \overline{e})}_{=\overline{e}} \leq \omega(x')$$

which demonstrates that ω is a suitable bound oracle for Algorithm 1. ◁

PART 3. We can now demonstrate that $(U,\eta) \mapsto \widetilde{P}_j(U,m,\eta)$ is a controlled evaluation method for $U \mapsto \frac{m}{2}\max\{0,\ G_j(U)\}^2$. For $U \in \Sigma_\sim$ and $\eta > 0$, let (x,e) be the output tuple returned by $\widetilde{P}_j(U,m,\eta)$ and let $(\widetilde{x},\ \widetilde{e})$ be the output tuple of Algorithm 1 used to calculate $(x,\ e)$.

We first handle the edge case $m = 0$. In this case, we have

$$\begin{aligned}\left|\frac{m}{2}\max\{0,\ G_j(U)\}^2 - x\right| &= \left|\frac{m}{2}\max\{0,\ G_j(U)\}^2 - \frac{m}{2}\cdot\widetilde{x}^2\right| \\ &= \frac{m}{2}\cdot\left|\max\{0,\ G_j(U)\}^2 - \widetilde{x}^2\right| \\ &= 0 \\ &= \frac{m}{2}\cdot\widetilde{e}\cdot(2\widetilde{x}+\widetilde{e}) \\ &= e.\end{aligned}$$

Furthermore, $e = 0$ guarantees $e < \eta$.

Now, we treat the case in which $m > 0$. In this case, Algorithm 1 guarantees that $(\widetilde{x},\ \widetilde{e})$ is an output tuple of ϕ with $e \le \omega(\widetilde{x})$. As we had shown previously, this means that

$$\left|\max\{0,\ G_j(U)\} - \widetilde{x}\right| \le \widetilde{e} \le \omega(\widetilde{x}).$$

This implies

$$\begin{aligned}\left|\frac{m}{2}\max\{0,\ G_j(U)\}^2 - x\right| &= \left|\frac{m}{2}\max\{0,\ G_j(U)\}^2 - \frac{m}{2}\cdot\widetilde{x}^2\right| \\ &= \frac{m}{2}\cdot\left|\max\{0,\ G_j(U)\}^2 - \widetilde{x}^2\right| \\ &= \frac{m}{2}\cdot\left|\max\{0,\ G_j(U)\} - \widetilde{x}\right|\cdot\underbrace{\left|\max\{0,\ G_j(U)\} + \widetilde{x}\right|}_{\le 2|\widetilde{x}|+\widetilde{e}} \\ &\le \frac{m}{2}\cdot\widetilde{e}\cdot\big(2\underbrace{|\widetilde{x}|}_{=\widetilde{x}} + \widetilde{e}\big) \\ &= e.\end{aligned}$$

Because $\widetilde{e} \le \omega(\widetilde{x})$, we also obtain the estimate

$$\begin{aligned}e &= \frac{m}{2}\cdot\widetilde{e}\cdot(2\widetilde{x}+\widetilde{e}) \\ &= \frac{m}{2}\cdot(\widetilde{e}^2 + 2\widetilde{x}\widetilde{e})) \\ &\le \frac{m}{2}\cdot\big(\omega(\widetilde{x})^2 + 2\widetilde{x}\omega(\widetilde{x})\big) \\ &= \frac{m}{2}\cdot\left(\widetilde{x}^2 + \frac{2\eta}{m} - 2\widetilde{x}\sqrt{\widetilde{x}^2+\frac{2\eta}{m}} + \widetilde{x}^2 + 2\widetilde{x}\sqrt{\widetilde{x}^2+\frac{2\eta}{m}} - 2\widetilde{x}^2\right) \\ &= \frac{m}{2}\cdot\frac{2\eta}{m} \\ &= \eta.\end{aligned}$$

This proves that $(U,\eta) \mapsto \widetilde{P}_j(U,m,\eta)$ is a controlled evaluation method for $U \mapsto \frac{m}{2}\cdot\max\{0,\ G_j(U)\}^2$. □

Lemma 3.2.28 (Penalty Function Evaluator).
Let $(X,\ \Sigma,\ \mu)$, $\Sigma_\sim$, F, $\widetilde{F}$, n, $(G_j)_{j\in[n]}$, and $(\widetilde{G}_j)_{j\in[n]}$ satisfy Assumption 3.2.26. For every $j\in[n]$, let $\widetilde{P}_j$ be the evaluation procedure of the same name developed in Lemma 3.2.27 with fixed $\xi\in(0,\ 1)$ and $\omega_0\in\mathbb{R}_{>0}\cup\{\infty\}$.

For $w\in\mathbb{R}^{n+1}_{>0}$, $m\in\mathbb{R}^n_{\geq 0}$, and $\eta>0$ let

$$\eta_{w,j}(\eta) \coloneqq \frac{w_j m_j \eta}{w_{n+1}+\sum_{j=1}^n w_j m_j} \qquad \forall j\in[n]$$
$$\eta_{w,n+1}(\eta) \coloneqq \frac{w_{n+1}\eta}{w_{n+1}+\sum_{j=1}^n w_j m_j}.$$

For $w\in\mathbb{R}^{n+1}_{>0}$, let $\widetilde{P}_w\colon \Sigma_\sim\times\mathbb{R}^n_{\geq 0},\mathbb{R}_{>0}\to\mathbb{R}\times\mathbb{R}_{\geq 0}$ be a procedure such that $\widetilde{P}_w(U,m,\eta)$ obtains an output tuple $(x_j,\ e_j)$ of $\widetilde{P}_j\big(U,m_j,\eta_{w,j}(\eta)\big)$ for each $j\in[n]$ as well as an output tuple (x_{n+1},e_{n+1}) of $\widetilde{F}\big(U,\eta_{w,n+1}(\eta)\big)$ and yields

$$\widetilde{P}_{w,1}(U,m,\eta) \coloneqq x_{n+1}+\sum_{\substack{j\in[n]\\ m_j>0}} x_j,$$
$$\widetilde{P}_{w,2}(U,m,\eta) \coloneqq e_{n+1}+\sum_{\substack{j\in[n]\\ m_j>0}} e_j.$$

Then for every $w\in\mathbb{R}^{n+1}_{>0}$ and $m\in\mathbb{R}^n_{\geq 0}$, $(U,\eta)\mapsto\widetilde{P}_w(U,m,\eta)$ is a controlled evaluation method for the set functional $U\mapsto P(U,m)$. ◁

Proof. Let $w\in\mathbb{R}^{n+1}_{>0}$, $m\in\mathbb{R}^n_{\geq 0}$, $U\in\Sigma_\sim$, and $\eta>0$. Let $(x,\ e)$ be an output tuple generated by $\widetilde{P}_w(U,m,\eta)$ and let $\big((x_j,\ e_j)\big)_{j\in[n]}$ and $(x_{n+1},\ e_{n+1})$ be the output tuples of $\widetilde{P}_j\big(U,m_j,\eta_{w,j}(\eta)\big)$ and $\widetilde{F}\big(U,\eta_{w,n+1}(\eta)\big)$, respectively, from which the output tuple was generated. We have

$$\begin{aligned}
\big|P(U,m)-x\big| &= \Big|F(U)+\sum_{j=1}^n \frac{m_j}{2}\max\{0,\ G_j(U)\}^2 - x_{n+1} - \sum_{\substack{j\in[n]\\ m_j>0}} x_j\Big| \\
&\leq \big|F(U)-x_{n+1}\big| + \sum_{\substack{j\in[n]\\ m_j>0}} \Big|\frac{m_j}{2}\max\{0,\ G_j(U)\}^2 - x_j\Big| \\
&\leq e_{n+1}+\sum_{\substack{j\in[n]\\ m_j>0}} e_j \\
&= e.
\end{aligned}$$

We also obtain the estimate

$$\begin{aligned}
e &= e_{n+1} + \sum_{\substack{j\in[n]\\ m_j>0}} e_j \\
&\le \eta_{w,n+1}(\eta) + \sum_{\substack{j\in[n]\\ m_j>0}} \eta_{w,j}(\eta) \\
&\le \frac{w_{n+1}\eta}{w_{n+1}+\sum_{j=1}^n w_j m_j} + \sum_{\substack{j\in[n]\\ m_j>0}} \frac{w_j m_j \eta}{w_{n+1}+\sum_{k=1}^n w_k m_k} \\
&= \frac{w_{n+1}\eta}{w_{n+1}+\sum_{j=1}^n w_j m_j} + \sum_{j=1}^n \frac{w_j m_j \eta}{w_{n+1}+\sum_{k=1}^n w_k m_k} \\
&= \left(\frac{w_{n+1}}{w_{n+1}+\sum_{j=1}^n w_j m_j} + \sum_{j=1}^n \frac{w_j m_j}{w_{n+1}+\sum_{k=1}^n w_k m_k}\right)\cdot\eta \\
&= \frac{w_{n+1}+\sum_{j=1}^n w_j m_j}{w_{n+1}+\sum_{j=1}^n w_j m_j}\cdot\eta \\
&= \eta.
\end{aligned}$$

This proves that $(U,\eta)\mapsto \widetilde{P}_w(U,m,\eta)$ is a controlled evaluation method for the set functional $U\mapsto P(U,m)$. □

The gradient of the j-th penalty term has the form

$$U\mapsto m_j\cdot\max\{0,\ G_j(U)\}\cdot\nabla G_j(U).$$

This presents a problem because the gradient depends on both the value and the gradient of the underlying constraint functional. Therefore, the error of the gradient cannot be controlled without simultaneously controlling the error of the functional value. Let $j\in[n]$ and let

$$\begin{aligned}
v_f &:= G_j(U), \\
(\widetilde{v}_f,\ e_f) &:= \widetilde{G}_j(U,\eta_f'), \\
(\widetilde{v}_g,\ e_g) &:= \widetilde{g}_j(U,\eta_g')
\end{aligned}$$

for some error bounds $\eta_f'>0$ and $\eta_g'>0$. Let further v_g be the true density function of $\nabla G_j(U)$. Then we have

$$\begin{aligned}
&\left\|m_j\cdot\max\{0,\ v_f\}\cdot v_g - m_j\cdot\max\{0,\ \widetilde{v}_f\}\cdot\widetilde{v}_g\right\| \\
&\qquad = m_j\cdot\left\|\max\{0,\ v_f\}\cdot v_g - \max\{0,\ \widetilde{v}_f\}\cdot\widetilde{v}_g\right\| \\
&\qquad \le m_j\cdot\Big(\quad\left\|\max\{0,\ v_f\}\cdot(v_g-\widetilde{v}_g)\right\| \\
&\qquad\qquad + \left\|\big(\max\{0,\ v_f\}-\max\{\widetilde{v}_f,\ 0\}\big)\cdot\widetilde{v}_g\right\|\Big) \\
&\qquad = m_j\cdot\Big(\quad\max\{0,\ v_f\}\cdot\|v_g-\widetilde{v}_g\| \\
&\qquad\qquad + \left|\max\{0,\ v_f\}-\max\{0,\ \widetilde{v}_f\}\right|\cdot\|\widetilde{v}_g\|\Big) \\
&\qquad \le m_j\cdot\Big(\big(\max\{0,\ \widetilde{v}_f\}+e_f\big)\cdot e_g + e_f\cdot\|\widetilde{v}_g\|\Big)
\end{aligned}$$

where $\|\cdot\|$ is either the L^1 or L^∞ norm, depending on which norm is used to control the gradient error of G_j. We demand that the right hand side of this inequality be bounded above by some desired error bound $\eta_j > 0$. By solving this inequality for e_g, we obtain the error bound

$$e_g \leq \frac{\eta_j - m_j \cdot e_f \cdot \|\widetilde{v}_g\|}{m_j \cdot \big(\max\{0,\ \widetilde{v}_f\} + e_f\big)}.$$

In order to ensure that this error bound is strictly positive, we must first ensure that $\eta_j > m_j \cdot e_f \cdot \|\widetilde{v}_g\|$. If $m_j = 0$, then this is always true and e_g is unbounded. If $m_j > 0$, then we may need to re-evaluate the functional value to enforce the error bound

$$e_f < \frac{\eta_j}{m_j \cdot \|\widetilde{v}_g\|}.$$

To enforce this through a non-strict bound and catch the edge case where $m_j \cdot \|\widetilde{v}_g\| = 0$, we introduce tuning parameters $\zeta_j \in (0,\ 1)$ and $\epsilon_j > 0$ and enforce the tightened error bound

$$e_f \leq \zeta_j \cdot \frac{\eta_j}{\max\{\epsilon_j,\ m_j \cdot \|\widetilde{v}_g\|\}} < \frac{\eta_j}{\max\{\epsilon_j,\ m_j \cdot \|\widetilde{v}_g\|\}} \leq \frac{\eta_j}{m_j \cdot \|\widetilde{v}_g\|}.$$

Lemma 3.2.29 (Penalty Term Gradient Evaluator).
Let (X, Σ, μ), $\Sigma_\sim$, $q \in \{1, \infty\}$, $n \in \mathbb{N}$, $(G_j)_{j\in[n]}$, $(\widetilde{G}_j)_{j\in[n]}$, and $(\widetilde{g}_j)_{j\in[n]}$ satisfy Assumption 3.2.26. Let $j \in [n]$ and let $\omega_0 \in \mathbb{R}_{>0} \cup \{\infty\}$, $\xi \in (0,\ 1)$, $\zeta \in (0,\ 1)$, and $\epsilon > 0$ be fixed algorithmic parameters.

For fixed $U \in \Sigma_\sim$, $m \geq 0$, and $\eta > 0$, let $\phi_{\eta,m} \colon \mathbb{R}_{>0} \cup \{\infty\} \to L^q(\Sigma,\mu) \times \mathbb{R} \times \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0}$ be a procedure such that $\phi_{\eta,m}(\beta)$ obtains an output tuple (x_g, e_g) from $\widetilde{g}_j(U,\beta)$ and an output tuple (x_f, e_f) from $\widetilde{G}_j\Big(U, \zeta \cdot \frac{\eta}{\max\{\epsilon, m\cdot\|x_g\|\}}\Big)$ and yields

$$\phi_{\eta,m}(\beta) \coloneqq (x_g, x_f, e_f, e_g).$$

Let $\omega_{\eta,m} \colon L^q(\Sigma,\mu) \times \mathbb{R} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0} \cup \{\infty\}$ be given by

$$\omega_{\eta,m}(x_g, x_f, e_f) \coloneqq \begin{cases} \frac{\eta - m\cdot e_f \cdot \|x_g\|}{m\cdot(\max\{0,x_f\}+e_f)} & \text{if } m \cdot \big(\max\{0, x_f\} + e_f\big) > 0, \\ \infty & \text{otherwise.} \end{cases}$$

Let $p_j \colon \Sigma_\sim \times \mathbb{R}_{\geq 0} \times \mathbb{R}_{>0} \cup \{\infty\} \to L^q(\Sigma,\mu) \times \mathbb{R}_{\geq 0}$ be a procedure such that $p_j(U,m,\eta)$ obtains an output tuple $(x_{g,i}, x_{f,i}, e_{f,i}, e_{g,i})$ from Algorithm 1 with $\phi_{\eta,m}$ as evaluator, $\omega_{\eta,m}$ as bound oracle, ω_0 as initial bound, and ξ as decay rate, and yields

$$\begin{aligned} p_{j,1}(U,m,\eta) &\coloneqq m \cdot \max\{0, x_{f,i}\} \cdot x_{g,i}, \\ p_{j,2}(U,m,\eta) &\coloneqq m \cdot \Big(\big(\max\{0, x_{f,i}\} + e_{f,i}\big) \cdot e_{g,i} + e_{f,i} \cdot \|x_{g,i}\|\Big). \end{aligned}$$

Then for each $m \geq 0$, $(U,\eta) \mapsto p_j(U,m,\eta)$ is an L^q-controlled gradient evaluation method for the set functional $U \mapsto \frac{m}{2} \cdot \max\big\{0, G_j(U)\big\}^2$. ◁

Proof. According to Theorem 2.4.6, the derivative of the set functional $U \mapsto \frac{m}{2} \cdot \max\big\{0, G_j(U)\big\}^2$ has the form

$$U \mapsto m \cdot \max\big\{0, G_j(U)\big\} \cdot \nabla G_j(U)$$

for all $m \geq 0$.

PART 1 ($\phi_{\eta,m}$ IS A SUITABLE EVALUATOR). Let $U \in \Sigma_{\sim}$, let $m \geq 0$, and let $\eta \in \mathbb{R}_{>0} \cup \{\infty\}$. Let (x_g, x_f, e_f, e_g) be an output tuple of $\phi_{\eta,m}(\beta)$ for $\beta \in \mathbb{R}_{>0} \cup \{\infty\}$. By definition of $\phi_{\eta,m}$, (x_g, e_g) is an output tuple from $\widetilde{g}_j(U, \beta)$ and (x_f, e_f) is an output tuple from $\widetilde{G}_j\Big(U, \zeta \cdot \frac{\eta}{\max\{\epsilon, m \cdot \|x_g\|\}}\Big)$. Let $g \in L^\infty(\Sigma, \mu)$ be the density function of the true gradient $\nabla G_j(U)$. We have

$$\begin{aligned} \|g - x_g\|_{L^q} &\leq e_g \\ &\leq \beta, \\ \big|G_j(U) - x_f\big| &\leq e_f \\ &\leq \zeta \cdot \frac{\eta}{\max\{\epsilon, m \cdot \|x_g\|\}} \\ &\leq \zeta \cdot \frac{\eta}{m \cdot \|x_g\|}. \end{aligned}$$

The former demonstrates that ϕ is a suitable evaluator.

PART 2 ($\omega_{\eta,m}$ IS A SUITABLE BOUND ORACLE). Let $U \in \Sigma_{\sim}$, $m \geq 0$, and let $\eta \in \mathbb{R}_{>0} \cup \{\infty\}$. Let (x_g, x_f, e_f, e_g) be an output tuple of $\phi_{m,\eta}(\beta)$ for some error bound $\beta \in \mathbb{R}_{>0} \cup \{\infty\}$. As we had shown in Part 1, we have

$$e_f \leq \zeta \cdot \frac{\eta}{\max\{\epsilon, m \cdot \|x_g\|\}} \leq \zeta \cdot \frac{\eta}{m \cdot \|x_g\|}.$$

Regardless of whether $m \cdot \|x_g\| = 0$ or not, this implies

$$m \cdot e_f \cdot \|x_g\| \leq \zeta \cdot \eta$$

and therefore

$$\eta - m \cdot e_f \cdot \|x_g\| \geq (1 - \zeta_j) \cdot \eta > 0.$$

If $m \cdot (\max\{0, x_f\} + e_f) > 0$, then this implies that

$$\omega_{\eta,m}(x_g, x_f, e_f) = \frac{\eta - m \cdot e_f \cdot \|x_g\|}{m \cdot (\max\{0, x_f\} + e_f)} > 0.$$

Thus, we have $\omega_{\eta,m}(x_g, x_f, e_f) \in \mathbb{R}_{>0} \cup \{\infty\}$.

Next, we have to prove that there exists a constant $\overline{e}_{\eta,m} > 0$ such that

$$e_g \leq \overline{e}_{\eta,m} \implies e_g \leq \omega_{\eta,m}(x_g, x_f, e_f)$$

for every output tuple (x_g, x_f, e_f, e_g) of $\phi_{m,\eta}(\beta)$, regardless of the particular choice of $\beta \in \mathbb{R}_{>0} \cup \{\infty\}$.

If $m \cdot (\max\{0, x_f\} + e_f) = 0$, then we have

$$\omega_{\eta,m}(x_g, x_f, e_f) = \infty$$

and $e_g \leq \overline{e}_{\eta,m} \implies e_g \leq \omega_{\eta,m}(x_g, x_f, e_f)$ for every $\overline{e}_{\eta,m} > 0$. We therefore only need to further consider cases where $m \cdot \big(\max\{0, x_f\} + e_f\big) > 0$.

Because $e_f \leq \zeta \cdot \frac{\eta}{\max\{\epsilon, m \cdot \|x_g\|\}} \leq \frac{\zeta \cdot \eta}{\epsilon}$ and $\big|G_j(U) - x_f\big| \leq e_f$, we have

$$\max\{0, x_f\} + e_f \leq \max\big\{0, G_j(U)\big\} + 2e_f \leq \max\big\{0, G_j(U)\big\} + 2 \cdot \frac{\zeta \cdot \eta}{\epsilon},$$

which implies that

$$\begin{aligned}\omega_{\eta,m}(x_g, x_f, e_f) &= \frac{\eta - m \cdot e_f \cdot \|x_g\|}{m \cdot (\max\{0, x_f\} + e_f)} \\ &\geq \underbrace{\frac{(1-\zeta)\cdot\eta}{m \cdot \left(\max\{0, G_j(U)\} + 2 \cdot \frac{\zeta\cdot\eta}{\epsilon}\right)}}_{=:\overline{e}_{\eta,m} > 0}.\end{aligned}$$

This choice then ensures that

$$e_g \leq \overline{e}_{\eta,m} \implies e_g \leq \omega_{\eta,m}(x_g, x_f, e_f)$$

for all output tuples of $\phi_{\eta,m}(\beta)$ for all $\beta \in \mathbb{R}_{>0} \cup \{\infty\}$. Therefore, $\omega_{\eta,m}$ is suitable as a bound oracle for Algorithm 1.

Part 3 (Evaluator procedure p_j). Let (x, e) be an output tuple from $p_j(U, m, \eta)$ for $U \in \Sigma_\sim$, $m \geq 0$, and $\eta \in \mathbb{R}_{>0} \cup \{\infty\}$. Let $(x_{g,i}, x_{f,i}, e_{f,i}, e_{g,i})$ be the output tuple from Algorithm 1 used to calculate that output tuple. Let $g \in L^\infty(\Sigma, \mu)$ be the density function of the true derivative $\nabla G_j(U)$

We first handle the case in which $m \cdot \left(\max\{0, x_{f,i}\} + e_{f,i}\right) = 0$. This implies that either $m = 0$ or $\max\{0, x_{f,i}\} + e_{f,i} = 0$. Because the output tuple from Algorithm 1 is forwarded from $\phi_{\eta,m}$, we have $\left|G_j(U) - x_{f,i}\right| \leq e_{f,i}$ and therefore

$$\max\{0, G_j(U)\} \leq \max\{0, x_{f,i}\} + e_{f,i},$$

which means that

$$\max\{0, x_{f,i}\} + e_{f,i} = 0 \implies \max\{0, G_j(U)\} = 0.$$

Therefore, we have

$$m \cdot \max\{0, G_j(U)\} \cdot g = 0 = m \cdot \max\{0, x_{f,i}\} \cdot x_{g,i}$$

and

$$\begin{aligned}e &= m \cdot \left(\left(\max\{0, x_{f,i}\} + e_{f,i}\right) \cdot e_{g,i} + e_{f,i} \cdot \|x_{g,i}\|\right) \\ &= m \cdot \left(\max\{0, x_{f,i}\} + e_{f,i}\right) \cdot e_{g,i} + m \cdot e_{f,i} \cdot \|x_{g,i}\| \\ &= m \cdot e_{f,i} \cdot \|x_{g,i}\| \\ &\geq 0 \\ &= \left\| m \cdot \max\{0, G_j(U)\} \cdot g - m \cdot \max\{0, x_{f,i}\} \cdot x_{g,i} \right\| \\ &= \left\| m \cdot \max\{0, G_j(U)\} \cdot g - x \right\|.\end{aligned}$$

For the case that $m \cdot (\max\{0, x_{f,i}\} + e_{f,i}) > 0$, we find that

$$\begin{aligned}
\left\| m \cdot \max\{0, G_j(U)\} \cdot g - x \right\| &= \left\| m \cdot \max\{0, G_j(U)\} \cdot g - m \cdot \max\{0, x_{f,i}\} \cdot x_{g,i} \right\| \\
&= m \cdot \left\| \max\{0, G_j(U)\} \cdot g - \max\{0, x_{f,i}\} \cdot x_{g,i} \right\| \\
&= m \cdot \Big\| \quad \max\{0, G_j(U)\} \cdot (g - x_{g,i}) \\
&\qquad + \left(\max\{0, G_j(U)\} - \max\{0, x_{f,i}\} \right) \cdot x_{g,i} \Big\| \\
&\leq m \cdot \Big(\quad \max\{0, G_j(U)\} \cdot \|g - x_{g,i}\| \\
&\qquad + \left| \max\{0, G_j(U)\} - \max\{0, x_{f,i}\} \right| \cdot \|x_{g,i}\| \Big) \\
&\leq m \cdot \left(\left(\max\{0, x_{f,i}\} + e_{f,i} \right) \cdot e_{g,i} + e_{f,i} \cdot \|x_{g,i}\| \right) \\
&= e.
\end{aligned}$$

In either case, we have

$$m \cdot \left(\max\{0, x_{f,i}\} + e_{f,i} \right) \cdot e_{g,i} \leq \eta - m \cdot e_{f,i} \cdot \|x_{g,i}\|$$

either because the left hand side of the inequality is zero and the right hand side is strictly positive, or because $e_{g,i} \leq \omega_{\eta,m}(x_{g,i}, x_{f,i}, e_{f,i})$. This then implies that

$$\begin{aligned}
e &= m \cdot \left(\left(\max\{0, x_{f,i}\} + e_{f,i} \right) \cdot e_{g,i} + e_{f,i} \cdot \|x_{g,i}\| \right) \\
&\leq \eta - m \cdot e_{f,i} \cdot \|x_{g,i}\| + m \cdot e_{f,i} \cdot \|x_{g,i}\| \\
&= \eta.
\end{aligned}$$

In summary, we have proven that

$$\left\| m \cdot \max\{0, G_j(U)\} \cdot g - x \right\| \leq e \leq \eta$$

for all output tuples of $p_j(U, m, \eta)$, which proves that $(U, \eta) \mapsto p_j(U, m, \eta)$ is an L^q-controlled gradient evaluation method for the set functional

$$U \mapsto m \cdot \max\{0, G_j(U)\} \cdot g. \qquad \square$$

Lemma 3.2.30 (Penalty Function Gradient Evaluator).
Let (X, Σ, μ), $\Sigma_\sim$, $q \in \{1, \infty\}$, $n \in \mathbb{N}$, F, $\widetilde{F}$, $\widetilde{f}$, $(G_j)_{j\in[n]}$, $(\widetilde{G}_j)_{j\in[n]}$, $(\widetilde{g}_j)_{j\in[n]}$, ζ, and ϵ satisfy Assumption 3.2.26. For $j \in [n]$, let $p_j \colon \Sigma_\sim \times \mathbb{R}_{\geq 0} \times \mathbb{R}_{>0} \cup \{\infty\} \to L^q(\Sigma, \mu) \times \mathbb{R}_{\geq 0}$ be the gradient evaluation method developed in Lemma 3.2.29 with fixed algorithmic parameters $\omega_{0,j} \in \mathbb{R}_{>0} \cup \{\infty\}$, $\xi_j \in (0,\ 1)$, $\zeta_j \in (0,\ 1)$, and $\epsilon_j > 0$. Let $w \in \mathbb{R}^{n+1}_{>0}$ and let

$$\begin{aligned}
\eta_{w,j}(\eta) &:= \frac{w_j m_j \cdot \eta}{w_{n+1} + \sum_{j=1}^n w_j m_j} \quad \forall j \in [n],\ \eta > 0; \\
\eta_{w,n+1}(\eta) &:= \frac{w_{n+1} \cdot \eta}{w_{n+1} + \sum_{j=1}^n w_j m_j} \quad \forall \eta > 0.
\end{aligned}$$

For $w \in \mathbb{R}^{n+1}_{>0}$, let $p_w \colon \Sigma_\sim \times \mathbb{R}^n_{\geq 0} \times \mathbb{R}_{>0} \cup \{\infty\} \to L^q(\Sigma, \mu) \times \mathbb{R}_{\geq 0}$ be a procedure such that $p_w(U, m, \eta)$ obtains an output tuple (x_f, e_f) from $f\big(U, \eta_{w,n+1}(\eta)\big)$ as well as

an output tuple $(x_{g,j}, e_{g,j})$ *from* $p_j\big(U, m_j, \eta_{w,j}(\eta)\big)$ *for each* $j \in [n]$ *with* $m_j > 0$*, and then yields*

$$p_{w,1}(U,m,\eta) \coloneqq x_f + \sum_{\substack{j\in[n]\\ m_j>0}} x_{g,j},$$
$$p_{w,2}(U,m,\eta) \coloneqq e_f + \sum_{\substack{j\in[n]\\ m_j>0}} e_{g,j}.$$

Then for each $m \in \mathbb{R}^n_{\geq 0}$ *and* $w \in \mathbb{R}^{n+1}_{>0}$*,* $(U,\eta) \mapsto p_w(U,m,\eta)$ *is an* L^q*-controlled gradient evaluation method for the set functional* $U \mapsto P(U,m)$*.* ◁

Proof. Let $w \in \mathbb{R}^{n+1}_{>0}$, $m \in \mathbb{R}^n_{\geq 0}$, $U \in \Sigma_{\sim}$, and $\eta > 0$. Let (x,e) be an output tuple of $p_w(U,m,\eta)$, and let (x_f, e_f) and $\big((x_{g,j}, e_{g,j})\big)_{j\in[n]}$ be the output tuples of $f\big(U, \eta_{w,n+1}(\eta)\big)$ and $\big(g_j\big(U, \eta_{w,j}(\eta)\big)\big)_{j\in[n]}$, respectively, that are used to calculate that output tuple.

Let $p \in L^\infty(\Sigma,\mu) \subseteq L^q(\Sigma,\mu)$ be the density function of $\nabla_U P(U,m)$, and let f' and $(g'_j)_{j\in[n]}$ be the respective density functions of $\nabla F(U)$ and $\big(\nabla G_j(U)\big)_{j\in[n]}$. We have

$$\begin{aligned}
\big\|p - x\big\| &= \Big\| f' - x_f + \sum_{\substack{j\in[n]\\ G_j(U)>0}} m_j \cdot G_j(U) \cdot g'_j - \sum_{\substack{j\in[n]\\ m_j>0}} x_{g,j} \Big\| \\
&= \Big\| f' - x_f + \sum_{\substack{j\in[n]\\ m_j>0}} \Big(m_j \cdot \max\{0, G_j(U)\} \cdot g'_j - x_{g,j} \Big) \Big\| \\
&\leq \|f' - x_f\| + \sum_{\substack{j\in[n]\\ m_j>0}} \Big\| m_j \cdot \max\{0, G_j(U)\} \cdot g'_j - x_{g,j} \Big\| \\
&\leq e_f + \sum_{\substack{j\in[n]\\ m_j>0}} e_{g,j} \\
&= e.
\end{aligned}$$

We also have

$$\begin{aligned}
e &= e_f + \sum_{\substack{j\in[n]\\ m_j>0}} e_{g,j} \\
&\leq \eta_{w,n+1}(\eta) + \sum_{\substack{j\in[n]\\ m_j>0}} \eta_{w,j}(\eta) \\
&= \frac{w_{n+1}\eta}{w_{n+1} + \sum_{j=1}^n w_j m_j} + \sum_{\substack{j\in[n]\\ m_j>0}} \frac{w_j m_j \eta}{w_{n+1} + \sum_{j=1}^n w_j m_j} \\
&= \frac{w_{n+1}\eta}{w_{n+1} + \sum_{j=1}^n w_j m_j} + \sum_{j=1}^n \frac{w_j m_j \eta}{w_{n+1} + \sum_{j=1}^n w_j m_j} \\
&= \frac{w_{n+1} + \sum_{j=1}^n w_j m_j}{w_{n+1} + \sum_{j=1}^n w_j m_j} \cdot \eta \\
&= \eta.
\end{aligned}$$

This proves that $(U, \eta) \mapsto p_w(U, m, \eta)$ is an L^q-controlled gradient evaluation method for the set functional $U \mapsto P(U, m)$. □

Wrapped Evaluators

To simplify the formulation of the quadratic penalty method, we wrap the evaluators that we have formulated thus far into evaluators for our quantities of interest: instationarity, infeasibility, projected descent, actual descent, and step quality.

We have already devoted a substantial amount of effort to the tuning of error bounds in Section 3.1.2. We reuse the step quality evaluator from that section. We had previously used the symbol τ for instationarity. In constrained optimization, constraint violation becomes a second quantity of interest in termination criteria. For the constraint violation, we use the symbol

$$v(U) \coloneqq \sum_{j=1}^{n} \max\{0,\ G_j(U)\} \geq 0 \qquad \forall U \in \Sigma_{\sim}.$$

Algorithm 6 Evaluate constraint violation

Require: $U \in \Sigma_{\sim}$, $w \in \mathbb{R}^n_{>0}$, $\varepsilon_v > 0$, $\xi_v \in (0,\ 1)$, $\beta_0 \in \mathbb{R}_{>0} \cup \{\infty\}$, $\xi \in (0,\ 1)$, $\widetilde{G}_j$ controlled evaluator for G_j for each $j \in [n]$.

Ensure: Yields $(v,\ e_v) \in \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0}$ such that

$$\left|\sum_{j=1}^{n} \max\{0,\ G_j(U)\} - v\right| \leq e_v \leq \xi_v \cdot \max\{\varepsilon_v,\ v - \varepsilon_v\}.$$

function $\omega(v)$ ▷ Parameter oracle
  return $\xi_v \cdot \max\{\varepsilon_v,\ v - \varepsilon_v\}$
end function

procedure $\phi(\beta)$ ▷ Inner evaluator
  for all $j \in [n]$ **do**
    $(v_j,\ e_j) \leftarrow \widetilde{G}_j\left(U,\ \frac{w_j \beta}{\sum_{k=1}^{n} w_k}\right)$
    $(\widetilde{v}_j,\ \widetilde{e}_j) \leftarrow \left(\max\{0, v_j\},\ \min\{e_j,\ \max\{0,\ v_j + e_j\}\}\right)$
  end for
  return $\left(\sum_{j=1}^{n} \widetilde{v}_j,\ \sum_{j=1}^{n} \widetilde{e}_j\right)$
end procedure

procedure EVALINFEAS(U, w, ξ_v, ε_v; β_0, ξ) ▷ Main procedure
  return CONTROLLEDEVAL(ϕ, ω, β_0, ξ)
end procedure

In Algorithm 6, we state an evaluation procedure for the constraint violation. This is a straightforward application of the controlled evaluation loop (Algorithm 1 on page 215) that yields an output tuple $(\widetilde{v},\ e_v)$ such that

$$\left|v(U) - \widetilde{v}\right| \leq e_v \leq \xi_v \cdot \max\{\varepsilon_v,\ \widetilde{v} - \varepsilon_v\}$$

where ε_v is the infeasibility tolerance and ξ_v is a tuning parameter. This gives us a similar guarantee for our feasibility criterion as we had previously established

for our stationarity criterion:

$$\widetilde{\nu} \le \varepsilon_\nu \Longrightarrow \nu(U) \le (1+\xi_\nu)\cdot\varepsilon_\nu,$$
$$\widetilde{\nu} \ge \varepsilon_\nu \Longrightarrow \nu(U) \ge (1-\xi_\nu)\cdot\varepsilon_\nu.$$

The bound oracle ω is clearly bounded below by $\xi_\nu\cdot\varepsilon_\nu > 0$. The inner evaluator ϕ obtains output tuples (v_j, e_j) from $\widetilde{G}_j$ and calculates

$$\widetilde{v}_j \coloneqq \max\{0, v_j\},$$
$$\widetilde{e}_j \coloneqq \min\bigl\{e_j, \max\{0,\ v_j + e_j\}\bigr\},$$

which satisfies

$$\begin{aligned}
\widetilde{e}_j &= \min\bigl\{e_j, \max\{0, v_j + e_j\}\bigr\} \\
&= \begin{cases} e_j & \text{if } v_j \ge 0, \\ \max\{0, v_j + e_j\} & \text{if } v_j < 0 \end{cases} \\
&\ge \begin{cases} \bigl|G_j(U) - v_j\bigr| & \text{if } v_j \ge 0, \\ \max\{0, G_j(U)\} & \text{if } v_j < 0 \end{cases} \\
&= \Bigl|\max\bigl\{0, G_j(U)\bigr\} - \max\bigl\{0, v_j\bigr\}\Bigr| \\
&= \Bigl|\max\bigl\{0, G_j(U)\bigr\} - \widetilde{v}_j\Bigr|.
\end{aligned}$$

Hence, the output tuple of ϕ satisfies

$$\begin{aligned}
\Bigl|\sum_{j=1}^{n} \max\bigl\{0,\ G_j(U)\bigr\} - \sum_{j=1}^{n} \widetilde{v}_j\Bigr| &\le \sum_{j=1}^{n} \widetilde{e}_j \\
&= \sum_{j=1}^{n} \min\bigl\{e_j,\ \max\{0,\ v_j + e_j\}\bigr\} \\
&\le \sum_{j=1}^{n} e_j \\
&\le \sum_{j=1}^{n} \frac{w_j \beta}{\sum_{k=1}^{n} w_k} \\
&= \beta.
\end{aligned}$$

This proves that the controlled evaluation loop terminates and yields an output tuple $(\widetilde{\nu}, e_\nu)$ with

$$\Bigl|\sum_{j=1}^{n} \max\bigl\{0,\ G_j(U)\bigr\} - \widetilde{\nu}\Bigr| \le e_\nu \le \omega(\widetilde{\nu}) = \xi_\nu\cdot\max\{\varepsilon_\nu,\ \widetilde{\nu} - \varepsilon_\nu\}.$$

Once we have established whether a solution is within feasibility tolerance, we have to evaluate the gradient from which instationarity and step are determined. Because the penalty method is essentially an unconstrained optimization method with the penalty function as its objective, we can reuse Algorithm 2 on page 225. However, we have to apply a small modification. If a solution appears stationary but infeasible, then we have to raise the penalty parameter m until the step either stops appearing stationary or an upper bound $\overline{m} > 0$ for m is reached.

Algorithm 7 Penalty parameter and gradient update

Require: $U \in \Sigma_{\sim}$, $w \in \mathbb{R}_{>0}^{n+1}$ where $n \in \mathbb{N}_0$ represents the number of constraints, $\overline{\varepsilon}_\tau \geq \varepsilon_\tau > 0$, $\overline{m} \geq m_0 > 0$, $\xi_\tau \in (0,\ 1)$, $\xi_\delta \in (0,\ 1)$, $\xi_g \in (0,\ 1)$, $\Delta > 0$, $\widetilde{\nu} \geq 0$, $\varepsilon_\nu > 0$, $\xi \in (0,\ 1)$, $\beta_0 \in \mathbb{R}_{>0} \cup \{\infty\}$, $(U, \beta) \mapsto \widetilde{p}_w(U, m, \beta)$ L^q-controlled gradient density evaluator for $U \mapsto P(U, m)$ for all $w \in \mathbb{R}_{>0}^{n+1}$ and $m \in \mathbb{R}_{>0}^{n}$ with $q \in \{1, \infty\}$.

Ensure: Let $p_m \in L^1(\Sigma, \mu)$ be the gradient density function of $U \mapsto P(U, m \cdot \mathbb{1}_n)$ in U, and let $\tau(f) := -\int_X \min\{0, f\}\,\mathrm{d}\mu$ for all $f \in L^1(\Sigma, \mu)$. Then the algorithm yields an output tuple $(m, g, e) \in \mathbb{R}_{>0} \times L^1(\Sigma, \mu) \times \mathbb{R}_{\geq 0}$ such that $m = 2^k m_0$ for some $k \in \mathbb{N}_0$ and either $m > \overline{m}$, or

$$\begin{aligned} \|p_m - g\|_{L^q} &\leq e, \\ \big|\tau(p_m) - \tau(g)\big| &\leq \xi_\tau \cdot \max\big\{\varepsilon_\tau, \tau(g) - \varepsilon_\tau\big\}, \end{aligned}$$

and

$$\left|\int_D p_m\,\mathrm{d}\mu - \int_D g\,\mathrm{d}\mu\right| \leq -\xi_g \cdot \int_D g\,\mathrm{d}\mu$$

for all $D \in \Sigma_{\sim}$ with $\mu(D) \leq \Delta$ for which

$$-\int_D g\,\mathrm{d}\mu \geq (1 - \xi_\delta) \cdot \theta \cdot \frac{\Delta}{\max\{\Delta,\ \mu(\{g < 0\})\}} \cdot \max\big\{\tau(g), \varepsilon_\tau\big\},$$

holds. If $m \leq \overline{m}$ and $\widetilde{\nu} > \varepsilon_\nu$, then

$$\tau(g) > \varepsilon_\tau.$$

▷ Let $\phi_{w,m} := (U, \beta) \mapsto \widetilde{p}_w(U, m \cdot \mathbb{1}_n, \beta)$ for $w \in \mathbb{R}_{>0}^{n+1}$ and $m > 0$.

procedure EVALPENGRAD(U, w, ε_τ, $\overline{\varepsilon}_\tau$, $\widetilde{\nu}$, m_0, $\overline{m}$, Δ; β_0, ξ)
  $(g_0, e_0) \leftarrow$ EVALGRAD$\big(\phi_{w,m_0}, U, \varepsilon_\tau, \Delta;\ \xi, \beta_0\big)$
  if $\widetilde{\nu} \leq \varepsilon_\nu$ **then**
    return (m_0, g_0, e_0)
  else
    $k \leftarrow 0$
    if $m_k \leq \frac{\overline{m}}{2} \wedge -\int_{\{g_k < 0\}} g_k\,\mathrm{d}\mu \leq \overline{\varepsilon}_\tau$ **then** ▷ Nearly stationary solutions
      $m_{k+1} \leftarrow 2 \cdot m_k$
      $(g_{k+1}, e_{k+1}) \leftarrow$ EVALGRAD$\big(\phi_{w,m_{k+1}}, U, \varepsilon_\tau, \Delta;\ \xi, \beta_0\big)$
      $k \leftarrow k + 1$
    end if
    while $m_k \leq \overline{m} \wedge -\int_{\{g_k < 0\}} g_k\,\mathrm{d}\mu \leq \varepsilon_\tau$ **do** ▷ Stationary solutions
      $m_{k+1} \leftarrow 2 \cdot m_k$
      $(g_{k+1}, e_{k+1}) \leftarrow$ EVALGRAD$\big(\phi_{w,m_{k+1}}, U, \varepsilon_\tau, \Delta;\ \xi, \beta_0\big)$
      $k \leftarrow k + 1$
    end while
    return (m_k, g_k, e_k)
  end if
end procedure

If the latter is the case, then we take this as an indicator that the solution is infeasible and that its infeasibility cannot be improved locally. We then consider the problem locally infeasible. The penalty and gradient update routine is stated in Algorithm 7 on the facing page.

The algorithm is comparatively simple because it makes use of the unconstrained gradient evaluation procedure EVALGRAD from Algorithm 2. Therefore, most of the guarantees derive from the correctness of that algorithm. It is easy to see that the loop increasing m terminates because $m_0 > 0$ and because $m_k = m_0 \cdot 2^k$ is easy to show via induction. Together, this implies that there exists $k_0 \in \mathbb{N}$ such that $m_{k_0} > \overline{m}$ always holds if k_0 iterations are reached.

If $\widetilde{v} \leq \varepsilon_v$, then the algorithm simply invokes Algorithm 2 to obtain an output tuple that satisfies the required guarantees with an unchanged penalty parameter m_0. If $\widetilde{v} > \varepsilon_v$, then the loop terminates with either $m_k > \overline{m}$ or

$$\tau(g_k) = -\int_X \min\{0, g_k\}\,\mathrm{d}\mu = -\int_{\{g_k<0\}} g_k\,\mathrm{d}\mu > \varepsilon_\tau.$$

We add a secondary mechanism to increase m. If doubling m_k does not move m_k above the threshold $\overline{m}$ and $\tau(g_k) \leq \overline{\varepsilon}_\tau$ for some relaxed instationarity tolerance $\overline{\varepsilon}_\tau \geq \varepsilon_\tau$, then we double m_k once. This is not strictly necessary. We add it so that the algorithm does not need to achieve full ε_τ-stationarity before raising the penalty parameter. Otherwise, it could take a very long time to increase m to the necessary value. This if clause has no impact on the overall correctness of the penalty method because it does not trigger an unwarranted local infeasibility detection.

Quadratic Penalty Method

With the error-controlled evaluation loops in place, we formulate the trust region penalty method in Algorithm 8 on the next page. To simplify notation, we write

$$\mathscr{C}_2(G,U) \coloneqq \sum_{j=1}^{n} \max\{0,\, G_j(U)\}.$$

The correctness proof for Algorithm 8 on the next page relies mostly on the guarantees given by the evaluators. Termination can be guaranteed because the penalty parameter m is bounded above by a finite threshold $\overline{m}$. Because EVALPENGRAD can only change the penalty parameter by doubling it, it is guaranteed that this threshold will be exceeded after a finite number of changes in m. Therefore, after an unknown finite number of iterations, m will remain constant.

We note that we give every controlled evaluation loop the same tolerance decay rate ξ and that we use consistent initial bounds and weight vectors for all functional and gradient evaluations. This is not technically required. However, introducing different initial bounds, weight vectors, and decay rates for each evaluation would clutter the symbol pool of Algorithm 8 even more. Allowing for different parameterization between functional and gradient evaluators is reasonable because functional value and gradient are evaluated using different numerical algorithms. This strikes a good balance between notational complexity and practicality.

With these preliminaries out of the way, we can prove the correctness of Algorithm 8. The proof is very similar to that of Theorem 3.1.10. We first prove

Algorithm 8 Trust-Region Quadratic Penalty Method

Require: Let $(X,\ \Sigma,\ \mu)$, $\Sigma_\sim$, q, F, $\widetilde{F}$, $\widetilde{f}$, n, G, $\widetilde{G}$, $\widetilde{g}$, $\mathscr{S}$, σ_0, σ_1, σ_2, ε_τ, ε_v, $\overline{\varepsilon}_\tau$, ξ_τ, ξ_v, ξ_δ, ξ_g, w_f, w_g, $\beta_{0,f}$, $\beta_{0,g}$, ξ, ζ, ϵ, m_{init}, $\overline{m}$, U_0, Δ_0 satisfy Assumption 3.2.26 on page 285. Let $K \in \mathbb{N}_0 \cup \{\infty\}$.

Ensure: $k \in \mathbb{N}_0$ with $k \leq K$, $U^* \in \Sigma_\sim$, and $m_k > 0$ such that $k < K \wedge m_k \leq \overline{m}$ implies

$$\mathscr{C}_1\big(U \mapsto P(U, m_k),\ U^*\big) \geq -(1+\xi_\tau)\cdot\varepsilon_\tau,$$
$$\mathscr{C}_2(G,\ U^*) \leq (1+\xi_v)\cdot\varepsilon_v.$$

▷ Let $\tau(h) := -\int_X \min\{0,\ h\}\,\mathrm{d}\mu$ for all integrable functions h.

$(j,\ k) \leftarrow (0,\ 0)$
$(\widetilde{v}_0,\ e_{v,0}) \leftarrow$ EVALINFEAS$(U_0,\ w_f,\ \xi_v,\ \varepsilon_v;\ \beta_{0,f},\ \xi)$
$(m_0,\ g_0,\ e_{g,0}) \leftarrow$ EVALPENGRAD$(U_0,\ w_g,\ \overline{\varepsilon}_\tau,\ \widetilde{v}_0,\ m_{\text{init}},\ \overline{m},\ \Delta_0;\ \beta_{0,g},\ \xi)$
while $k < K \wedge m_j \leq \overline{m} \wedge \tau(g_j) > \varepsilon_\tau$ **do** ▷ Main loop
  $(D_j, \delta_j) \leftarrow \mathscr{S}\Big(g_j, \Delta_j, \xi_\delta \cdot \theta \cdot \frac{\Delta_j}{\max\{\Delta_j,\ \mu(\{g_j<0\})\}} \cdot \tau(g_j)\Big)$
  $(\widetilde{\rho}_j,\ e_{\rho,j}) \leftarrow$ EVALRHO$(g_j,\ U_j,\ D_j;\ \beta_{0,f},\ \xi)$
  if $\widetilde{\rho}_j - e_{\rho,j} \geq \sigma_0$ **then** ▷ Accept step
    $(U_{j+1},\ k) \leftarrow (U_j \vartriangle D_j,\ k+1)$
    $\Delta_{j+1} \leftarrow \min\{2\Delta_j, \mu(X)\}$ **if** $\widetilde{\rho}_j \geq \sigma_2$ **else** Δ_j
    $(\widetilde{v}_{j+1},\ e_{v,j+1}) \leftarrow$ EVALINFEAS$(U_{j+1},\ w_f,\ \xi_v,\ \varepsilon_v;\ \beta_{0,f},\ \xi)$
    $m_{j+1} \leftarrow m_j$
  else ▷ Reject step
    $(U_{j+1},\ \Delta_{j+1}) \leftarrow \Big(U_j,\ \frac{\Delta_j}{2}\Big)$
    $(\widetilde{v}_{j+1},\ e_{v,j+1}) \leftarrow (\widetilde{v}_j,\ e_{v,j})$
  end if
  $j \leftarrow j+1$
  $(m_j,\ g_j,\ e_{g,j}) \leftarrow$ EVALPENGRAD$(U_j,\ w_g,\ \overline{\varepsilon}_\tau,\ \widetilde{v}_j,\ m_{j-1},\ \overline{m},\ \Delta_j;\ \beta_{0,g},\ \xi)$
end while
$U^* \leftarrow U_j$

certain invariant error guarantees. Then we use those guarantees along with the step-finding guarantees and the continuity of the derivative to argue that there is a minimal trust region radius. This then guarantees a minimal objective decrease per accepted step, which guarantees termination of the algorithm if the objective is bounded below.

Theorem 3.2.31 (Correctness of Algorithm 8).
Let (X, Σ, μ), $\Sigma_\sim$, q, F, $\widetilde{F}$, $\widetilde{f}$, n, G, $\widetilde{G}$, $\widetilde{g}$, $\mathscr{S}$, σ_0, σ_1, σ_2, ε_τ, ε_v, $\overline{\varepsilon}_\tau$, ξ_τ, ξ_v, ξ_δ, ξ_g, w_f, w_g, $\beta_{0,f}$, $\beta_{0,g}$, ξ, ζ, ϵ, m_{init}, $\overline{m}$, U_0, Δ_0 satisfy Assumption 3.2.26 on page 285. Let $K \in \mathbb{N}_0 \cup \{\infty\}$.

Then Algorithm 8 terminates in finite time and yields $k \in \mathbb{N}_0$ with $k \leq K$, $U^* \in \Sigma_\sim$, and $m_k > 0$ such that either

$$k \geq K \vee m_k > \overline{m}$$

or

$$\begin{aligned}\mathscr{C}_1\big(U \mapsto P(U, m_k), U^*\big) &\geq -(1+\xi_\tau)\cdot\varepsilon_\tau,\\ \mathscr{C}_2(G, U^*) &\leq (1+\xi_v)\cdot\varepsilon_v.\end{aligned}$$ ◁

PROOF. To simplify notation, we adopt the notation from Algorithm 8 and write

$$\tau(h) \coloneqq \int_X \min\{0, h\}\,\mathrm{d}\mu \quad \forall h \in L^1(\Sigma, \mu).$$

For every $U \in \Sigma_\sim$ and $m > 0$, let $p_{U,m} \in L^\infty(\Sigma, \mu)$ be the density function of $\nabla_U P(U, m)$.

PART 1 (GUARANTEES FOR $\widetilde{v}_j$). First, we note that the output tuple $(\widetilde{v}_0, e_{v,0})$ obtained from EVALINFEAS satisfies

$$\big|\mathscr{C}_2(G, U_0) - \widetilde{v}_0\big| \leq e_{v,0} \leq \xi_v \cdot \max\{\varepsilon_v, \widetilde{v}_0 - \varepsilon_v\}.$$

Let $j \in \mathbb{N}_0$ be such that

$$\big|\mathscr{C}_2(G, U_j) - \widetilde{v}_j\big| \leq e_{v,j} \leq \xi_v \cdot \max\{\varepsilon_v, \widetilde{v}_j - \varepsilon_v\}.$$

The tuple $(\widetilde{v}_{j+1}, e_{v,j+1})$ is calculated differently based on whether the step is rejected or not. If the step is accepted, then $(\widetilde{v}_{j+1}, e_{v,j+1})$ is an output tuple of EVALINFEAS and therefore satisfies

$$\big|\mathscr{C}_2(G, U_{j+1}) - \widetilde{v}_{j+1}\big| \leq e_{v,j+1} \leq \xi_v \cdot \max\{\varepsilon_v, \widetilde{v}_{j+1} - \varepsilon_v\}.$$

If the step is rejected, then we have $U_{j+1} = U_j$ and therefore

$$\begin{aligned}\big|\mathscr{C}_2(G, U_{j+1}) - \widetilde{v}_{j+1}\big| &= \big|\mathscr{C}_2(G, U_j) - \widetilde{v}_j\big|\\ &\leq \underbrace{e_{v,j}}_{=e_{v,j+1}}\\ &\leq \xi_v \cdot \max\{\varepsilon_v, \widetilde{v}_j - \varepsilon_v\}\\ &= \xi_v \cdot \max\{\varepsilon_v, \widetilde{v}_{j+1} - \varepsilon_v\}.\end{aligned}$$

Therefore, no re-evaluation is necessary. This is because the output of EVALINFEAS does not depend on the value of Δ_j or m_j.

We therefore know inductively that

$$\big|\mathscr{C}_2(G, U_j) - \widetilde{v}_j\big| \leq e_{v,j} \leq \xi_v \cdot \max\{\varepsilon_v, \widetilde{v}_j - \varepsilon_v\}$$

for all $j \in \mathbb{N}_0$ prior to main loop termination. This implies that

$$\begin{aligned}\widetilde{v}_j \leq \varepsilon_v &\Longrightarrow \mathscr{C}_2(G, U_j) \leq (1+\xi_v)\cdot\varepsilon_v,\\ \widetilde{v}_j > \varepsilon_v &\Longrightarrow \mathscr{C}_2(G, U_j) > (1-\xi_v)\cdot\varepsilon_v\end{aligned}$$

also holds for all $j \in \mathbb{N}_0$ prior to termination.

PART 2 (VALUE OF m_j AND GUARANTEES FOR g_j). Due to the output guarantees of EVALPENGRAD (see Algorithm 7 on page 300), we have $m_0 = 2^{l_0} \cdot m_{\text{init}}$

for some $l_0 \in \mathbb{N}_0$. If $m_0 \leq \overline{m}$, then the algorithm also guarantees that we have $\|p_{U_0,m_0} - g_0\|_{L^q} \leq e_{g,0}$ and

$$\left|\tau(p_{U_0,m_0}) - \tau(g_0)\right| \leq \xi_\tau \cdot \max\{\varepsilon_\tau, \tau(g_0) - \varepsilon_\tau\}.$$

In addition to that, for every $D \in \Sigma_\sim$ with $\mu(D) \leq \Delta_0$ and

$$-\int_D g_0 \,\mathrm{d}\mu \geq (1-\xi_\delta)\cdot\theta\cdot\frac{\Delta_0}{\max\{\Delta_0, \mu(\{g_0 < 0\})\}} \cdot \max\{\tau(g_0), \varepsilon_\tau\},$$

we have

$$\left|\int_D p_{U_0,m_0} \,\mathrm{d}\mu - \int_D g_0 \,\mathrm{d}\mu\right| \leq -\xi_g \cdot \int_D g_0 \,\mathrm{d}\mu.$$

Let $j \in \mathbb{N}_0$ be such that $m_j = 2^{l_j} \cdot m_{\text{init}} \leq \overline{m}$ for some $l_j \in \mathbb{N}_0$. Then the tuple $(m_{j+1}, g_{j+1}, e_{g,j+1})$ consisting of the penalty parameter, gradient, and gradient error for the subsequent iteration is also an output tuple of EVALPENGRAD and therefore satisfies

$$m_{j+1} = 2^{l'_j} \cdot m_j = 2^{l'_j + l_j} \cdot m_{\text{init}}$$

for some $l'_j \in \mathbb{N}_0$. Therefore, we have $m_{j+1} = 2^{l_{j+1}} \cdot m_{\text{init}}$ with $l_{j+1} \coloneqq l_j + l'_j \in \mathbb{N}_0$, which implies that $l_{j+1} \geq l_j$. Because the sequence $(l_j)_{j\in\mathbb{N}_0}$ is monotonically increasing in $\mathbb{N}_0$ and because the main loop terminates once m_j exceeds $\log_2\left(\frac{\overline{m}}{m_{\text{init}}}\right)$, there exists an iteration index $j^* \in \mathbb{N}_0$ such that $m_j = m_{j^*}$ for all $j \geq j^*$ prior to main loop termination.

Additionally, if $m_{j+1} \leq \overline{m}$, then we have $\|p_{U_{j+1},m_{j+1}} - g_{j+1}\|_{L^q} \leq e_{g,j+1}$,

$$\left|\tau(p_{U_{j+1},m_{j+1}}) - \tau(g_{j+1})\right| \leq \xi_\tau \cdot \max\{\varepsilon_\tau, \tau(g_{j+1}) - \varepsilon_\tau\},$$

and for every $D \in \Sigma_\sim$ with $\mu(D) \leq \Delta_{j+1}$ and

$$-\int_D g_{j+1} \,\mathrm{d}\mu \geq (1-\xi_\delta)\cdot\theta\cdot\frac{\Delta_{j+1}}{\max\{\Delta_{j+1}, \mu(\{g_{j+1} < 0\})\}} \cdot \max\{\tau(g_{j+1}), \varepsilon_\tau\},$$

we have

$$\left|\int_D p_{U_{j+1},m_{j+1}} \,\mathrm{d}\mu - \int_D g_{j+1} \,\mathrm{d}\mu\right| \leq -\xi_g \cdot \int_D g_{j+1} \,\mathrm{d}\mu.$$

In summary, this means that for every j prior to main loop termination with $m_j \leq \overline{m}$, we have $\|p_{U_j,m_j} - g_j\|_{L^q} \leq e_{g,j}$ and

$$\left|\tau(p_{U_j,m_j}) - \tau(g_j)\right| \leq \xi_\tau \cdot \max\{\varepsilon_\tau, \tau(g_j) - \varepsilon_\tau\},$$

which guarantees that

$$\begin{aligned}
\tau(g_j) \leq \varepsilon_\tau &\Longrightarrow \tau(p_{U_j,m_j}) \leq (1+\xi_\tau)\cdot\varepsilon_\tau \quad \forall j \in \mathbb{N}_0 \colon m_j \leq \overline{m},\\
\tau(g_j) > \varepsilon_\tau &\Longrightarrow \tau(p_{U_j,m_j}) > (1-\xi_\tau)\cdot\varepsilon_\tau \quad \forall j \in \mathbb{N}_0 \colon m_j \leq \overline{m}.
\end{aligned}$$

Furthermore, for all such j and for every $D \in \Sigma_\sim$ with $\mu(D) \leq \Delta_j$ and

$$-\int_D g_j \,\mathrm{d}\mu \geq (1-\xi_\delta)\cdot\theta\cdot\frac{\Delta_j}{\max\{\Delta_j, \mu(\{g_j < 0\})\}} \cdot \max\{\tau(g_j), \varepsilon_\tau\},$$

we have

$$\left|\int_D p_{U_j,m_j} \,\mathrm{d}\mu - \int_D g_j \,\mathrm{d}\mu\right| \leq -\xi_g \cdot \int_D g_j \,\mathrm{d}\mu.$$

Part 3 (Upper and lower bounds on Δ). For $j = 0$, Assumption 3.2.26 ensures that $\Delta_j \in \big(0,\ \mu(X)\big]$. In each iteration of the main loop, the value of Δ_{j+1} is equal to either Δ_j, $\frac{\Delta_j}{2}$, or the minimum of $2\Delta_j$ and $\mu(X)$. Regardless of which of the three it is, $\Delta_j \in \big(0,\ \mu(X)\big]$ always implies $\Delta_{j+1} \in \big(0,\ \mu(X)\big]$.

This is not yet sufficient. We have to find a lower bound for Δ_j that is strictly greater than zero. As we had discussed in our discussion of Lemma 3.2.25 on page 283, the fact that all G_j are bounded above and that their derivative's density functions are bounded in L^∞ implies that the derivative of the penalty functional as a whole is uniformly continuous on $\Sigma_\sim$ for fixed penalty parameter m.

We use the fact that m does not increase after a finite iteration index j^*. For all $j \geq j^*$, we have $m_j = m_{j^*}$ and the penalty functional can be considered a fixed, uniformly continuously differentiable functional of U.

We define the constant

$$\overline{M} \coloneqq (1-\xi_g)\cdot(1-\xi_\delta)\cdot\theta\cdot\frac{\varepsilon_\tau}{\mu(X)} > 0$$

Because the acceptance or rejection of steps is only at issue in cases where the termination criteria are not yet satisfied, we have $\tau(g_j) > \varepsilon_\tau$. The step D_j is taken from an output tuple of $\mathscr{S}$ and satisfies $\mu(D_j) \leq \Delta_j$ as well as

$$-\int_{D_j} g_j \,\mathrm{d}\mu \geq (1-\xi_\delta)\cdot\theta\cdot\frac{\Delta_j}{\max\{\Delta_j,\, \mu(\{g_j<0\})\}}\cdot\tau(g_j).$$

By combining this with the guarantees that we have demonstrated in the previous part of this proof, we obtain

$$\begin{aligned}
-\int_{D_j} p_{U_j,m_j}\,\mathrm{d}\mu &\geq (1-\xi_g)\cdot\left(-\int_{D_j} g_j\,\mathrm{d}\mu\right)\\
&\geq (1-\xi_g)\cdot(1-\xi_\delta)\cdot\theta\cdot\frac{\Delta_j}{\max\{\Delta_j,\, \mu(\{g_j<0\})\}}\cdot\tau(g_j)\\
&\geq (1-\xi_g)\cdot(1-\xi_\delta)\cdot\theta\cdot\frac{\Delta_j}{\max\{\Delta_j,\, \mu(\{g_j<0\})\}}\cdot\varepsilon_\tau\\
&\geq (1-\xi_g)\cdot(1-\xi_\delta)\cdot\theta\cdot\frac{\Delta_j}{\mu(X)}\cdot\varepsilon_\tau\\
&\geq (1-\xi_g)\cdot(1-\xi_\delta)\cdot\theta\cdot\frac{\mu(D_j)}{\mu(X)}\cdot\varepsilon_\tau\\
&= (1-\xi_g)\cdot(1-\xi_\delta)\cdot\theta\cdot\frac{\varepsilon_\tau}{\mu(X)}\cdot\mu(D_j)\\
&= \overline{M}\cdot\mu(D_j).
\end{aligned}$$

Because $U \mapsto P(U, m_{j^*})$ is a uniformly continuously differentiable set functional, there exists $\delta > 0$ such that

$$\big(\nabla_U P(U \mathbin{\ominus}_{U\triangle V} m_{j^*}), \nabla_U P(V, m_{j^*})\big)(W) \leq \underbrace{\frac{\left(1-\frac{\sigma_1}{1-\xi_g}\right)\cdot\overline{M}}{2}}_{>0}\cdot\mu(W)$$

holds for all $U, V, W \in \Sigma_\sim$ with $\mu(U \triangle V) \leq \delta$. Here, we make use of the fact that $\xi_g < 1-\sigma_1$ and therefore $1-\xi_g > \sigma_1$.

Let $j \geq j^*$ be such that $\Delta_j \leq \delta$. Let $\gamma\colon I \to \Sigma_\sim$ be a minimizing geodesic connecting U_j and $U_j \triangle D_j$. Without loss of generality, let $I = [0,\ \mu(D_j)]$, $\gamma(0) = U_j$, and $\gamma\big(\mu(D_j)\big) = U_j \triangle D_j$. We have

$$\mu\big(U \triangle \gamma(t)\big) = \mu\big(\gamma(0) \triangle \gamma(t)\big) = t \leq \mu(D_j) \leq \delta \quad \forall t \in I.$$

We now make the same polygon chain approximation argument that we have made multiple times before. Let $\varepsilon > 0$. For each $t \in I$, there exists a radius $R(t) > 0$ such that

$$\Big|P\big(\gamma(s), m_j\big) - P\big(\gamma(t), m_j\big) - \nabla_U P\big(\gamma(t), m_j\big)\big(\gamma(s) \triangle \gamma(t)\big)\Big| \leq \frac{\varepsilon}{\mu(D_j)} \cdot \mu\big(\gamma(s) \triangle \gamma(t)\big)$$

for all $s \in I$ with $|s - t| \leq R(t)$. We choose a strictly increasing sequence $(t_l)_{l \in [N]}$ in I with $N \in \mathbb{N}$ such that $I \subseteq \bigcup_{l=1}^{N} B_{R(t_l)}(t_l)$. We then define support points $(s_l)_{l \in [N]_0}$ via

$$\begin{aligned} &s_0 \coloneqq 0, \\ &s_l \in B_{R(t_{l-1})}(t_{l-1}) \cap B_{R(t_l)}(t_l) \cap [t_{l-1},\ t_l] \quad \forall l \in [N-1], \\ &s_N \coloneqq \mu(D_j). \end{aligned}$$

As we have discussed before, this ensures that $s_{l-1} \in B_{R(t_l)}(t_l)$, $s_l \in B_{R(t_l)}(t_l)$, and $s_{l-1} \leq t_l \leq s_l$ for all $l \in [N]$. We then have

$$\begin{aligned}
&\big|P(U_j \triangle D_j, m_j) - P(U_j, m_j) - \nabla_U P(U_j, m_j)(D_j)\big| \\
&\quad\leq \sum_{l=1}^{N} \Big|P\big(\gamma(s_l), m_j\big) - P\big(\gamma(s_{l-1}), m_j\big) - \nabla_U P(U_j, m_j)\big(\gamma(s_l) \triangle \gamma(s_{l-1})\big)\Big| \\
&\quad= \sum_{l=1}^{N} \Big|P\big(\gamma(s_l), m_j\big) - P\big(\gamma(t_l), m_j\big) - \nabla_U P(U_j, m_j)\big(\gamma(s_l) \triangle \gamma(t_l)\big) \\
&\qquad\quad -P\big(\gamma(s_{l-1}), m_j\big) + P\big(\gamma(t_l), m_j\big) - \nabla_U P(U_j, m_j)\big(\gamma(s_{l-1}) \triangle \gamma(t_l)\big)\Big| \\
&\quad= \sum_{l=1}^{N} \Big|P\big(\gamma(s_l), m_j\big) - P\big(\gamma(t_l), m_j\big) - \nabla_U P\big(\gamma(t_l), m_j\big)\big(\gamma(s_l) \triangle \gamma(t_l)\big) \\
&\qquad\quad +\Big(\nabla_U P\big(\gamma(t_l), m_j\big) - \nabla_U P\big(U_j, m_j\big)\Big)\big(\gamma(s_l) \triangle \gamma(t_l)\big) \\
&\qquad\quad -P\big(\gamma(s_{l-1}), m_j\big) + P\big(\gamma(t_l), m_j\big) + \nabla_U P\big(\gamma(t_l), m_j\big)\big(\gamma(s_{l-1}) \triangle \gamma(t_l)\big) \\
&\qquad\quad -\Big(\nabla_U P\big(\gamma(t_l), m_j\big) + \nabla_U P(U_j, m_j)\Big)\big(\gamma(s_{l-1}) \triangle \gamma(t_l)\big)\Big| \\
&\quad\leq \sum_{l=1}^{N} \Bigg(\Big|P\big(\gamma(s_l), m_j\big) - P\big(\gamma(t_l), m_j\big) - \nabla_U P\big(\gamma(t_l), m_j\big)\big(\gamma(s_l) \triangle \gamma(t_l)\big)\Big| \\
&\qquad\quad +\Big|P\big(\gamma(s_{l-1}), m_j\big) - P\big(\gamma(t_l), m_j\big) - \nabla_U P\big(\gamma(t_l), m_j\big)\big(\gamma(s_{l-1}) \triangle \gamma(t_l)\big)\Big| \\
&\qquad\quad +\Big|\Big(\nabla_U P\big(\gamma(t_l), m_j\big) - \nabla_U P(U_j, m_j)\Big)\big(\gamma(s_l) \triangle \gamma(t_l)\big)\Big| \\
&\qquad\quad +\Big|\Big(\nabla_U P\big(\gamma(t_l), m_j\big) + \nabla_U P(U_j, m_j)\Big)\big(\gamma(s_{l-1}) \triangle \gamma(t_l)\big)\Big|\Bigg) \\
&\quad\leq \sum_{l=1}^{N} \frac{\varepsilon}{\mu(D_j)} \cdot \mu\big(\gamma(s_l) \triangle \gamma(s_{l-1})\big) \\
&\qquad + \sum_{l=1}^{N} \Big(\nabla_U P\big(\gamma(t_l), m_j\big) \underset{U_j \triangle \gamma(t_l)}{\ominus} \nabla_U P(U_j, m_j)\Big)\big(\gamma(s_l) \triangle \gamma(s_{l-1})\big)
\end{aligned}$$

$$\leq \sum_{l=1}^{N} \left(\frac{\varepsilon}{\mu(D_j)} + \frac{\left(1 - \frac{\sigma_1}{1-\xi_g}\right) \cdot \overline{M}}{2} \right) \cdot \mu\big(\gamma(s_l) \triangle \gamma(s_{l-1})\big)$$
$$= \varepsilon + \frac{1 - \frac{\sigma_1}{1-\xi_g}}{2} \cdot \overline{M} \cdot \mu(D_j).$$

With $\varepsilon \to 0$, we obtain the estimate

$$\left|P(U_j \triangle D_j, m_j) - P(U_j, m_j) - \nabla_U P(U_j, m_j)(D_j)\right| = \frac{1 - \frac{\sigma_1}{1-\xi_g}}{2} \cdot \overline{M} \cdot \mu(D_j).$$

For the true step quality ρ, this implies that

$$\begin{aligned}
\rho_j &= \frac{P(U_j \triangle D_j, m_j) - P(U_j, m_j)}{\int_{D_j} p_{U_j,m_j} \,\mathrm{d}\mu} \\
&\geq 1 - \frac{1 - \frac{\sigma_1}{1-\xi_g}}{2} \cdot \frac{\overline{M} \cdot \mu(D_j)}{-\int_{D_j} p_{U_j,m_j} \,\mathrm{d}\mu} \\
&\geq 1 - \frac{1 - \frac{\sigma_1}{1-\xi_g}}{2} \cdot \frac{\overline{M} \cdot \mu(D_j)}{\overline{M} \cdot \mu(D_j)} \\
&= 1 - \frac{1 - \frac{\sigma_1}{1-\xi_g}}{2} \\
&= \frac{1}{2} \cdot \left(1 + \frac{\sigma_1}{1-\xi_g}\right) \\
&> \frac{\sigma_1}{1-\xi_g}.
\end{aligned}$$

This demonstrates that for $j \geq j^*$ and $\Delta_j \leq \delta$, the true step quality ρ_j is strictly greater than $\frac{\sigma_1}{1-\xi_g}$. As we have discussed in Section 3.1.2, this implies that the semi-approximate step quality satisfies

$$\begin{aligned}
\overline{\rho}_j &= \frac{P(U_j \triangle D_j, m_j) - P(U_j, m_j)}{\int_{D_j} g_j \,\mathrm{d}\mu} \\
&= \rho_j \cdot \frac{\int_{D_j} g_j \,\mathrm{d}\mu}{\int_{D_j} p_{U_j,m_j} \,\mathrm{d}\mu} \\
&\geq \rho_j \cdot \frac{(1-\xi_g) \cdot \int_{D_j} p_{U_j,m_j} \,\mathrm{d}\mu}{\int_{D_j} p_{U_j,m_j} \,\mathrm{d}\mu} \\
&= \rho_j \cdot (1-\xi_g) \\
&> \sigma_1.
\end{aligned}$$

Because $(\widetilde{\rho}_j, e_{\rho,j})$ is an output tuple of EVALRHO, Theorem 3.1.8 guarantees that

$$|\widetilde{\rho}_j - \overline{\rho}_j| \leq e_{\rho,j} \leq \max\{\widetilde{\rho}_j - \sigma_0, \sigma_1 - \widetilde{\rho}_j\}.$$

If $\widetilde{\rho}_j - \sigma_0 < \sigma_1 - \widetilde{\rho}_j$, then we would have

$$\widetilde{\rho}_j - e_{\rho_j} \geq \widetilde{\rho}_j - (\underbrace{\sigma_1}_{<\overline{\rho}_j} - \widetilde{\rho}_j) > \widetilde{\rho}_j - \underbrace{(\overline{\rho}_j - \widetilde{\rho}_j)}_{\leq e_{\rho,j}} \geq \widetilde{\rho}_j,$$

which would generate a contradiction. Therefore, we must have $\widetilde{\rho}_j - \sigma_0 \geq \sigma_1 - \widetilde{\rho}_j$. From this, we can infer that

$$\widetilde{\rho}_j - e_{\rho,j} \geq \widetilde{\rho}_j - \widetilde{\rho}_j + \sigma_0 \geq \sigma_0$$

and that the step is accepted. In summary, for $j \geq j^*$ and $\Delta_j \leq \delta$, the step is always accepted and can never be rejected. As the trust region radius is only reduced upon rejection, we obtain the lower bound

$$\Delta_j \geq \Delta^* := \min\left\{\Delta_{j^*}, \frac{\delta}{2}\right\} > 0 \quad \forall j \geq j^*.$$

Furthermore, because the trust region radius is halved on every step rejection, this demonstrates that there can never be an infinite sequence of successive step rejections past j^*. As j^* is finite, this applies throughout the entire main loop iteration.

PART 4 (TERMINATION). We have demonstrated that there is a finite index $j^* \in \mathbb{N}_0$ such that $m_j = m_{j^*}$ for all $j \geq j^*$. Furthermore, we have demonstrated that $\Delta_j \geq \Delta^* > 0$ for all $j \geq j^*$ and that, therefore, there can never be an infinite sequence of successive step rejections.

The penalty functional value $P(U_j, m_j) = P(U_j, m_{j^*})$ does not change on rejected steps if $j \geq j^*$, because the penalty parameter no longer changes between main loop iterations. We now consider $j \geq j^*$ such that the step D_j is accepted.

In this case, we have $\widetilde{\rho}_j - e_{\rho,j} \geq \sigma_0$ and therefore $\overline{\rho}_j \geq \sigma_0$. This then implies

$$\rho_j \geq \frac{\sigma_0}{1+\xi_g} > 0.$$

From this, we can infer that

$$\begin{aligned}
P(U_j \triangle D_j, m_j) &= P(U_j, m_j) + \rho_j \cdot \int_{D_j} p_{U_j,m_j} \,\mathrm{d}\mu \\
&\leq P(U_j, m_j) + \rho_j \cdot (1-\xi_g) \cdot \underbrace{\int_{D_j} g_j \,\mathrm{d}\mu}_{<0} \\
&\leq P(U_j, m_j) - \rho_j \cdot (1-\xi_g) \cdot (1-\delta) \cdot \theta \cdot \frac{\Delta_j}{\max\{\Delta_j, \mu(\{g_j < 0\})\}} \cdot \tau(g_j) \\
&\leq P(U_j, m_j) - \rho_j \cdot (1-\xi_g) \cdot (1-\delta) \cdot \theta \cdot \frac{\Delta_j}{\mu(X)} \cdot \varepsilon_\tau \\
&\leq P(U_j, m_j) - \rho_j \cdot (1-\xi_g) \cdot (1-\delta) \cdot \theta \cdot \frac{\Delta^*}{\mu(X)} \cdot \varepsilon_\tau \\
&\leq P(U_j, m_j) - \frac{\sigma_0}{1+\xi_g} \cdot (1-\xi_g) \cdot (1-\delta) \cdot \theta \cdot \frac{\Delta^*}{\mu(X)} \cdot \varepsilon_\tau.
\end{aligned}$$

Therefore, on every accepted step, the true value of the penalty functional decreases by at least

$$\frac{\sigma_0}{1+\xi_g} \cdot (1-\xi_g) \cdot (1-\delta) \cdot \theta \cdot \frac{\Delta^*}{\mu(X)} \cdot \varepsilon_\tau > 0.$$

Because the true value of the penalty function can only increase when the penalty parameter changes, this guarantees a steady decrease of the penalty functional value past iteration j^*.

Because F is bounded below on $\Sigma_\sim$ and because all penalty terms are bounded below by 0, the penalty function is bounded below on $\Sigma_\sim$ for all penalty parameters. Because every accepted step past j^* achieves at least the aforementioned decrease and because the value cannot increase past iteration j^*, there can only be a finite number of accepted steps. In conjunction with the fact that there can never be an infinite sequence of successive rejections and the fact that j^* is finite, this means that the algorithm terminates.

Upon termination, the termination criterion must be met, which means that either the step limit K is exceeded, the penalty parameter limit is exceeded, or we have $\tau(g_j) \leq \varepsilon_\tau$. If the latter is the case, then

$$\begin{aligned}\mathscr{C}_1(U \mapsto P(U, m_j), U_j) &= -\tau(p_{U_j, m_j}) \\ &\geq -(1-\xi_\tau)\cdot\tau(g_j) \\ &\geq -(1-\xi_\tau)\cdot\varepsilon_\tau\end{aligned}$$

because of the guarantees given by EVALPENGRAD. Furthermore, EVALPENGRAD ensures that $\tau(g_j) \leq \varepsilon_\tau$ implies

$$\widetilde{\nu}_j \leq \varepsilon_\nu.$$

The guarantees given by EVALINFEAS then ensure that

$$\mathscr{C}_2(G, U_j) \leq (1+\xi_\nu)\cdot\widetilde{\nu}_j \leq (1+\xi_\nu)\cdot\varepsilon_\nu,$$

which completes the desired guarantees on the solution $U^* = U_j$. □

3.2.1.4 SCALAR-VALUED EQUALITY CONSTRAINTS

Problem (3.19) has no equality constraints and the Mangasarian-Fromovitz Constraint Qualification prohibits inequality constraints from being used to implicitly add equality constraints to a problem.

This is a major limitation of the theory developed in this section. We choose to limit our theory in this way because scalar equality constraints are fundamentally difficult to deal with in similiarity spaces. It is easy to find simple equality constraints that appear feasible but are not.

Example 3.2.32 (Almost Feasible Equality Constraint).
For instance, let $X := [0,\ 1]$, let $\Sigma := \mathscr{B}(X)$ be the σ-algebra of Borel-measurable subsets of X, and let λ be the Lebesgue measure. The constraint

$$\int_0^1 \left|\lambda(U \cap [0,\ t]) - \frac{t}{2}\right|^2 \mathrm{d}t = 0$$

for $U \in \Sigma_\sim$ is an easy example. It is of the form $H(U) = 0$ with

$$H(U) := \int_0^1 \left|y(\chi_U) - \frac{t}{2}\right|^2 \mathrm{d}t$$

where $y(\chi_U)$ is the solution of the initial value problem

$$y(0) = 0, \quad \dot{y}(t) = \chi_U(t) \quad \text{a.e. in } [0,\ 1].$$

This satisfies Assumption 2.4.17, which implies that H is benignly Lipschitz-continuously differentiable according to Theorem 2.4.23.

For each $i \in \mathbb{N}$, we choose the set

$$U_i \coloneqq \bigcup_{j=1}^{2^{i-1}} \left[\frac{2j-1}{2^i}, \frac{2j}{2^i} \right] \subseteq [0,\ 1].$$

Intuitively, this recursively subdivides X into 2^{i-1} equally sized sub-intervals and activates the control only on the upper half of each such sub-interval. The function $y_i \coloneqq y(\chi_{U_i})$ is absolutely continuous and piecewise linear with slope 0 on the "left" half and 1 on the "right" half of each sub-interval, meaning that the mean slope over each full interval of the partition is $\frac{1}{2}$.

The ODE solution y_i reaches its maximal deviation from $t \mapsto \frac{t}{2}$ in the middle of each sub-interval, where it deviates by $\frac{1}{2^i}$. Therefore, we have

$$H(U_i) = \int_0^1 \left| y_i - \frac{t}{2} \right|^2 \mathrm{d}t \leq \int_0^1 \frac{1}{2^{2i}} \,\mathrm{d}t = \frac{1}{2^{2i}} \xrightarrow{i\to\infty} 0.$$

However, there can never be an exact solution. Assume that $U^* \in \Sigma$ existed such that $H(U^*) = 0$. Then we would have

$$\lambda\big(U^* \cap [0,\ t]\big) = \frac{t}{2} \quad \text{for a.a. } t \in [0,\ 1].$$

Because both sides of the equation are continuous in t, the equality would hold for all $t \in [0,\ 1]$. By forming differences between the values of both sides for $t_1, t_2 \in [0,\ 1]$ with $t_1 \leq t_2$, we would obtain

$$\begin{aligned}
\lambda\big(U^* \cap (t_1,\ t_2]\big) &= \lambda\big(U^* \cap [0,\ t_2]\big) - \lambda\big(U^* \cap [0,\ t_1]\big) \\
&= \frac{t_2 - t_1}{2}, \\
\lambda\big(U^* \cap (t_1,\ t_2)\big) &= \lambda\big(U^* \cap (t_1,\ t_2]\big) - \lambda\big(U^* \cap [t_2,\ t_2]\big) \\
&= \lambda\big(U^* \cap (t_1,\ t_2]\big) \\
&= \frac{t_2 - t_1}{2}, \\
\lambda\big(U^* \cap [t_1,\ t_2]\big) &= \lambda\big(U^* \cap (t_1,\ t_2]\big) + \lambda\big(U^* \cap [t_1,\ t_1]\big) \\
&= \frac{t_2 - t_1}{2}, \\
\lambda\big(U^* \cap [t_1,\ t_2)\big) &= \lambda\big(U^* \cap [t_1,\ t_2]\big) - \lambda\big(U^* \cap [t_2,\ t_2]\big) \\
&= \frac{t_2 - t_1}{2}.
\end{aligned}$$

In summary, we would find that for all intervals $I \subseteq [0,\ 1]$,

$$\lambda(U^* \cap I) = \frac{\lambda(I)}{2}.$$

We had already stated in Lemma 2.3.30 that every open subset of R can be written as a countable union of pairwise disjoint open intervals. By intersecting each interval with $[0,\ 1]$, we can infer that every relatively open subset of $[0,\ 1]$ can be written as a countable union of pairwise disjoint intervals.

Let $(V_i)_{i\in\mathbb{N}}$ be a sequence of pairwise disjoint sets in $\mathscr{B}(I)$ such that $\lambda(U^* \cap V_i) = \frac{\lambda(V_i)}{2}$ for all $i \in \mathbb{N}$. We would have

$$\begin{aligned}\lambda\Big(U^* \cap \bigcup_{i=1}^{\infty} V_i\Big) &= \lambda\Big(\underbrace{\bigcup_{i=1}^{\infty}(U^* \cap V_i)}_{\text{pairwise disjoint}}\Big)\\ &= \sum_{i=1}^{\infty} \lambda(U^* \cap V_i)\\ &= \sum_{i=1}^{\infty} \frac{\lambda(V_i)}{2}\\ &= \frac{\lambda(\bigcup_{i=1}^{\infty} V_i)}{2}.\end{aligned}$$

This proves that every relatively open subset $V \subseteq [0,\ 1]$ would satisfy $\lambda(U^* \cap V) = \frac{\lambda(V)}{2}$.

Finally, let $V \in \mathscr{B}([0,\ 1])$. By definition, there exists $V' \in \mathscr{B}(\mathbb{R})$ such that $V = V' \cap [0,\ 1]$. According to Lemma 2.3.32, for $\varepsilon > 0$, there exists $U'_\varepsilon \subseteq \mathbb{R}$ open and $F'_\varepsilon \subseteq \mathbb{R}$ closed such that $U'_\varepsilon \subseteq V' \subseteq F'_\varepsilon$ and $\lambda(F'_\varepsilon \setminus U'_\varepsilon) \leq \varepsilon$.

Let $U_\varepsilon \coloneqq U'_\varepsilon \cap [0,\ 1]$ and $F_\varepsilon \coloneqq F'_\varepsilon \cap [0,\ 1]$. U_ε and F_ε are relatively open and closed, respectively, in $[0,\ 1]$. We have

$$\lambda(\underbrace{F_\varepsilon \setminus U_\varepsilon}_{\subseteq F'_\varepsilon \setminus U'_\varepsilon}) \leq \lambda(F'_\varepsilon \setminus U'_\varepsilon) \leq \varepsilon.$$

For every $\varepsilon > 0$, we would have

$$\lambda(\underbrace{U^* \cap U_\varepsilon}_{\subseteq U^* \cap V}) \leq \lambda(\underbrace{U^* \cap V}_{\subseteq U^* \cap F_\varepsilon}) \leq \lambda(U^* \cap F_\varepsilon),$$

as well as

$$\begin{aligned}\lambda(U^* \cap U_\varepsilon) &= \frac{\lambda(U_\varepsilon)}{2},\\ \lambda(U^* \cap F_\varepsilon) &= \lambda(U^* \cap [0,\ 1]) - \lambda(U^* \cap F_\varepsilon^{\complement})\\ &= \frac{\lambda([0,\ 1]) - \lambda(F_\varepsilon^{\complement})}{2}\\ &= \frac{\lambda(F_\varepsilon)}{2}\end{aligned}$$

because U_ε and $F_\varepsilon^{\complement} = [0,\ 1] \setminus F_\varepsilon$ are relatively open.

We therefore have

$$\frac{\lambda(U_\varepsilon)}{2} \leq \lambda(U^* \cap V) \leq \frac{\lambda(F_\varepsilon)}{2}.$$

However, we have

$$\lambda(F_\varepsilon) - \lambda(U_\varepsilon) = \lambda(F_\varepsilon \setminus U_\varepsilon) \leq \varepsilon.$$

which means that both $\frac{\lambda(U_\varepsilon)}{2}$ and $\frac{\lambda(F_\varepsilon)}{2}$ must converge to the fixed value $\lambda(U^* \cap V)$. Furthermore, we have $U_\varepsilon \subseteq V \subseteq F_\varepsilon$ and therefore

$$\lambda(U_\varepsilon) \leq \lambda(V) \leq \lambda(F_\varepsilon).$$

Again, this implies that $\lambda(U_\varepsilon)$ and $\lambda(F_\varepsilon)$ converge to the fixed value $\lambda(V)$. The conjunction between both convergence results yields

$$\lambda(U^* \cap V) = \lim_{\varepsilon \to 0} \frac{\lambda(U_\varepsilon)}{2} = \frac{\lambda(V)}{2}$$

for all $V \in \mathscr{B}([0,\ 1])$.

Finally, we could apply this result to show a contradiction because it implies that

$$\lambda(U^*) = \lambda(U^* \cap U^*) = \frac{\lambda(U^*)}{2}$$

and therefore $\lambda(U^*) = 0$ which would mean that $y(\chi_{U^*})$ is equal to zero almost everywhere. However, we would then have

$$H(U^*) = \int_0^1 \left| y(\chi_{U^*}) - \frac{t}{2} \right|^2 \mathrm{d}t = \int_0^1 \frac{1}{4} t^2 \,\mathrm{d}t = \frac{1}{12} > 0$$

which would contradict our assumption that $H(U^*) = 0$. ◁

It is difficult to pinpoint exactly why the constraint in Example 3.2.32 behaves in this way. It is tempting to assume that the problem lies with a lack of completeness. However, this is false because similarity spaces are actually complete.

Lemma 3.2.33.
Let (X, Σ, μ) be an atomless measure space. Then the similarity space $\Sigma/_{\sim_\mu}$ with the metric

$$(A,\ B) \mapsto \mu(A \triangle B)$$

is a complete metric space. ◁

PROOF. Let

$$\mathscr{S} := \left\{ F \in \Sigma/_{\sim_\mu} \,\middle|\, \mu(F) < \infty \right\}$$

be the family of all similarity classes of finite measure and let

$$\mathscr{F} := \left\{ f \in L^1(\Sigma,\ \mu) \,\middle|\, f(x) \in \{0,\ 1\} \text{ a.e.} \right\}$$

be the set of all μ-integrable functions that are binary-valued almost everywhere. The map $\chi_\cdot : \mathscr{S} \to \mathscr{F}$ that maps a similarity class to its indicator function is straightforwardly a bijective isometry.

It is important to stress that this isometric embedding is only valid for similarity classes of finite measure because the indicator functions of sets of infinite measure are not in $L^1(\Sigma, \mu)$. We note that $\chi_\cdot$ is not bijective as a map from the original measure space to $L^1(\Sigma, \mu)$ because functions in $L^1(\Sigma, \mu)$ are only well-defined up to differences on μ-nullsets and therefore, set differences of measure zero are lost during the embedding. Working with similarity spaces removes this problem.

The isometric embedding allows us to transfer the completeness of $L^1(\Sigma, \mu)$ to the similarity space. We first show that $\mathscr{F}$ is a closed subset of $L^1(\Sigma, \mu)$. Let

$f \in \mathscr{F}^{\complement}$. Then there exists a set $N \in \Sigma$ with $\mu(N) > 0$ such that $f(x) \notin \{0,\ 1\}$ for all $x \in N$. For every $g \in \mathscr{F}$, we have

$$\begin{aligned}
\|f - g\|_{L^1} &= \int_X |f - g| \,\mathrm{d}\mu \\
&\geq \int_N |f - \underbrace{g}_{\in\{0,\ 1\}}| \,\mathrm{d}\mu \\
&\geq \underbrace{\int_N \overbrace{\operatorname{dist}\big(f(x),\ \{0,\ 1\}\big)}^{>0} \,\mathrm{d}\mu}_{=:R} \\
&> 0.
\end{aligned}$$

Therefore, the open L^1 ball of radius $R > 0$ around f does not intersect $\mathscr{F}$. This implies that the complement of $\mathscr{F}$ is open and therefore that $\mathscr{F}$ is closed.

As a closed subset of a complete vector space, $\mathscr{F}$ is itself complete. Because $\chi_{\cdot}$ is a bijective isometry, it follows that $\mathscr{S}$ is a complete subset of the similarity space $\Sigma/_{\sim_\mu}$.

If the underlying measure space is finite, then this completes the argument because $\mathscr{S}$ is the entire similarity space. If there exist similarity classes of infinite measure, then an additional step is required.

Let $(A_i)_{i\in\mathbb{N}}$ be a Cauchy sequence in $\Sigma/_{\sim_\mu}$. Then there exists $i_0 \in \mathbb{N}$ such that

$$\mu(A_i \mathbin{\triangle} A_j) < \infty \quad \forall i, j \geq i_0.$$

We now define $B_i \coloneqq A_i \mathbin{\triangle} A_{i_0}$ for all $i \in \mathbb{N}$. This translated sequence satisfies

$$\begin{aligned}
\mu(B_i \mathbin{\triangle} B_j) &= \mu(A_i \mathbin{\triangle} A_{i_0} \mathbin{\triangle} A_j \mathbin{\triangle} A_{i_0}) \\
&= \mu(A_i \mathbin{\triangle} A_j)
\end{aligned}$$

for all $i, j \in \mathbb{N}$, which implies that $(B_i)_{i\in\mathbb{N}}$ is also a Cauchy sequence. In addition to this, we have

$$\mu(B_i) = \mu(A_i \mathbin{\triangle} A_{i_0}) < \infty \qquad \forall i \geq i_0.$$

Therefore, the sequence $(B_i)_{i\geq i_0}$ is a Cauchy sequence in $\mathscr{S}$. Because $\mathscr{S}$ is complete, there exists $B^* \in \mathscr{S}$ such that $B_i \xrightarrow{i\to\infty} B^*$. Let $A^* \coloneqq B^* \mathbin{\triangle} A_{i_0} \in \Sigma/_{\sim_\mu}$. Then we have

$$\begin{aligned}
\mu(A_i \mathbin{\triangle} A^*) &= \mu(A_i \mathbin{\triangle} A_{i_0} \mathbin{\triangle} A^* \mathbin{\triangle} A_{i_0}) \\
&= \mu(B_i \mathbin{\triangle} B^*) \\
&\xrightarrow{i\to\infty} 0
\end{aligned}$$

which proves that $(A_i)_{i\in\mathbb{N}}$ converges to a similarity class $A^* \in \Sigma/_{\sim_\mu}$. Because this applies to all Cauchy sequences in $\Sigma/_{\sim_\mu}$, $\Sigma/_{\sim_\mu}$ is complete. □

We note that completeness implies sequential closedness because every convergent sequence is also a Cauchy sequence.

Another intuitive explanation would be that even bounded similarity spaces, although complete, are neither themselves compact, nor do they exhibit a weaker compactness property such as local compactness. This is demonstrated by the

Table 3.1: Equivalence between set and logical operations

Set operation	Logical operation	Correspondence
$A \cup B$	$x \vee y$	$x \in A \cup B \iff \big((x \in A) \vee (x \in B)\big)$
$A \cap B$	$x \wedge y$	$x \in A \cap B \iff \big((x \in A) \wedge (x \in B)\big)$
$A \triangle B$	$x \oplus y$	$x \in A \triangle B \iff \big((x \in A) \oplus (x \in B)\big)$
$A^{\complement}$	$\neg x$	$x \in A^{\complement} \iff \neg(x \in A)$

approximating sequence that we had generated in Example 3.2.32, which is designed such that all points have distance $\frac{1}{2}$ from one another. Similar sequences are easily generated in arbitrary atomless measure spaces as long as they contain any set of non-zero finite measure.

However, optimization problems with equality constraints are routinely solved in search spaces that are neither compact nor locally compact. A requirement like that would, among others, preclude all L^p spaces and all infinite dimensional Hilbert spaces from consideration in mathematical optimization.

In infinite dimensional Banach spaces, the inverse and implicit function theorems (see, e.g., [Pat18, Sec. 3.4.2], [Zei95, Sec. 4.8], [BS20, Sec. 12.3]) form the backbone of the theory of equality-constrained optimization. Both are generally stated in Banach spaces and derive from the Banach fixed point theorem. Our setting technically satisfies the requirements for the fixed point theorem. However, all instances of inverse and implicit function theorems generally add the additional requirement that the space in question must be a vector space.

At this time, we can unfortunately not give an accurate account of why solving nonlinear equations is difficult in our search space. Our search space is a complete metric space with a commutative group structure and a translation invariant metric. It therefore satisfies all Banach space axioms except for those that concern scaling.

We suspect that scaling may be necessary for inverse and implicit function theorems in a manner that cannot be substituted with geodesics. However, we cannot point to a specific axiom whose absence is the specific reason why such theorems would be impossible to formulate in a similarity space setting. We leave this as a question for future research.

3.2.2 Logical Constraints

A second conceivable type of constraint is the "logical" constraint. Logical constraints are constraints that enforce relations between different set variables in the same measure space. They are therefore specific to problems that follow the "layering" approach laid out in Section 2.1.3. Moreover, they only apply when multiple of the "layers" are drawn from the same underlying measure space. We will not discuss logical constraints to the same extent to which we have discussed scalar constraints, but we will make some remarks on how they could be worked with.

We use the term "logical constraint" because the constraint is stated in terms of elementary set operations, which are equivalent to the basic boolean operators. We enumerate these well-known equivalences in Table 3.1.

If we refer to $n \in \mathbb{N}$ layered set variables as if they were components in a vector $(A_i)_{i \in [n]}$ where all A_i are drawn from the same measure space (X, Σ, μ),

then we can translate any n-ary boolean function φ into a function mapping multiple set-valued inputs to one set-valued output by applying the boolean function pointwise:

$$\begin{aligned}\phi\colon \quad & \Sigma^n \to \Sigma \\ & (A_i)_{i\in[n]} \mapsto \left\{x \in X \,\middle|\, \varphi\big((x \in A_i)_{i\in[n]}\big)\right\}.\end{aligned}$$

Every boolean function can be expressed in terms of the basic boolean operators $\vee$, $\wedge$, and $\neg$, and there exists a variety of normalized representations of boolean functions. We can similarly express the set-valued mapping ϕ as a combination of basic set operations.

Among the normal forms into which all boolean functions can be translated, the *disjunctive normal form* (DNF) is likely the most convenient for algorithmic processing. Therefore, we will focus on the DNF as our preferred encoding for boolean functions.

Definition 3.2.34 (Disjunctive Normal Form).
Let (X, Σ, μ) be a measure space, let $n \in \mathbb{N}$. A boolean function $\varphi\colon \{0,\ 1\}^n \to \{0,\ 1\}$ is said to be in *disjunctive normal form* (or *DNF*) if there exist $m \in \mathbb{N}_0$, constants $(c_j)_{j\in[m]} \in \{0,\ 1\}^m$, and disjoint index sets $K_j^+ \subseteq [n]$ and $K_j^- \subseteq [n]$ for all $j \in [m]$ such that

$$\varphi(x) = \bigvee_{j=1}^{m} \underbrace{\left(c_j \wedge \left(\bigwedge_{k\in K_j^+} x_k\right) \wedge \left(\bigwedge_{k\in K_j^-} \neg x_k\right)\right)}_{=:\varphi_j(x)} \quad \forall x \in \{0,\ 1\}^n.$$

The conjunctions φ_j are called *clauses* of φ.

Analogously, we refer to a mapping $\phi\colon \left(\Sigma/_{\sim_\mu}\right)^n \to \Sigma/_{\sim_\mu}$ as being in disjunctive normal form if there exist $m \in \mathbb{N}_0$, constants $C_j \in \Sigma/_{\sim_\mu}$ for $j \in [m]$, and disjoint index sets $K_j^+ \subseteq [n]$ and $K_j^- \subseteq [n]$ for $j \in [m]$ such that

$$\phi(A) = \bigcup_{j=1}^{m} \underbrace{\left(C_j \cap \left(\bigcap_{k\in K_j^+} A_k\right) \cap \left(\bigcap_{k\in K_j^-} A_k^{\complement}\right)\right)}_{=:\phi_j(A)} \quad \forall A \in \left(\Sigma/_{\sim_\mu}\right)^n.$$

In this case, we also refer to the ϕ_j as *clauses* of ϕ.

We refer to a constraint in a set-valued optimization problem as being a *logical constraint in DNF* if there exists a mapping $\phi\colon \left(\Sigma/_{\sim_\mu}\right)^n \to \Sigma/_{\sim_\mu}$ in DNF such that the constraint has the form $\phi(A) = X$. ◁

Remark 3.2.35.
Using the DNF as our encoding for boolean functions is a significant restriction. Forming the conjunction of two functions in DNF is not a simple task. Therefore, it is not easy to apply two logical constraints in DNF in the same domain at the same time.

This would be much easier if we chose the *conjunctive normal form* (CNF) as our representation. However, determining the feasibility of a constraint in CNF would require that we solve a general boolean satisfiability (SAT) problem. Since the SAT problem is generally an NP-hard problem, this would likely make constraints in CNF prohibitively expensive to work with. ◁

What sets the DNF apart is the ease with which the feasibility problem can be solved. For any clause, the constant C_j dictates a superset to which the output of that clause is restricted. However, because K_j^+ and K_j^- are disjoint, clauses cannot be self-contradictory and we can always find a configuration of A to make the output any desired subset of C_j.

Lemma 3.2.36.
Let (X, Σ, μ) be a measure space, and let $\phi\colon (\Sigma/_{\sim_\mu})^n \to \Sigma/_{\sim_\mu}$ be a mapping in DNF with $m \in \mathbb{N}_0$ clauses ϕ_j and constants $C_j \in \Sigma/_{\sim_\mu}$ for $j \in [m]$. Then the logical constraint

$$\phi(A) = X$$

is feasible if and only if

$$\bigcup_{j=1}^{m} C_j = X.$$

◁

PROOF. By Definition 3.2.34, each clause ϕ_j intersects its output with C_j. Therefore, we have $\phi_j(A) \subseteq_\mu C_j$ for all A and j. It follows that

$$\phi(A) = \bigcup_{j=1}^{n} \phi_j(A) \subseteq_\mu \bigcup_{j=1}^{n} C_j \quad \forall A \in (\Sigma/_{\sim_\mu})^n.$$

If there exists $A \in (\Sigma/_{\sim_\mu})^n$ with $\phi(A) = X$, then $X \subseteq_\mu \bigcup_{j=1}^n C_j$. As $C_j \subseteq_\mu X$ for all $j \in [m]$, this implies $X = \bigcup_{j=1}^n C_j$.

Conversely, if $\bigcup_{j=1}^n C_j = X$, then we can construct $A \in (\Sigma/_{\sim_\mu})^n$ such that $\phi(A) = X$. To do this, we define

$$S_j \coloneqq C_j \setminus \bigcup_{k=1}^{j-1} S_k \in \Sigma/_{\sim_\mu} \quad \forall j \in [m].$$

It is evident that $S_j \subseteq_\mu C_j$ for all j and that the S_j are pairwise essentially disjoint and satisfy

$$\bigcup_{j=1}^{m} S_j = \bigcup_{j=1}^{m} C_j = X.$$

We then define

$$A_i \coloneqq \bigcup_{\substack{j \in [m] \\ i \in K_j^+}} S_j \quad \forall i \in [n]$$

where K_j^+, $K_j^- \subseteq [n]$ are disjoint index sets that indicate which components of A participate in the clause themselves or via their complement, respectively. For each $j \in [m]$, this choice satisfies

$$\begin{aligned} S_j \cap \phi_j(A) &= S_j \cap C_j \cap \bigcap_{k \in K_j^+} A_k \cap \bigcap_{k \in K_j^-} A_k^\complement \\ &= S_j \cap \bigcap_{k \in K_j^+} A_k \cap \bigcap_{k \in K_j^-} A_k^\complement. \end{aligned}$$

Due to the way that we had defined A, we have $S_j \subseteq_\mu A_k$ for all $k \in K_j^+$. Because the S_j are pairwise disjoint, we also have $S_j \subseteq_\mu A_k^\complement$ for all $k \in K_j^-$. This means that

$$\begin{aligned} S_j \cap \phi_j(A) &= S_j \cap \bigcap_{k \in K_j^+} A_k \cap \bigcap_{k \in K_j^-} A_k^\complement \\ &= S_j. \end{aligned}$$

By uniting the outputs of all clauses, we obtain $\phi(A)$. By restricting each component, we obtain

$$\begin{aligned} \phi(A) &= \bigcup_{j=1}^{n} \phi_j(A) \\ &\supseteq_\mu \bigcup_{j=1}^{n} \big(S_j \cap \phi_j(A)\big) \\ &= \bigcup_{j=1}^{n} S_j \\ &= X. \end{aligned}$$

In conjunction with $\phi(A) \subseteq_\mu X$, which is the case because X is the universal set of the underlying measure space, we obtain $\phi(A) = X$. □

As Lemma 3.2.36 shows, a logical constraint in DNF is feasible if and only if the union of all clause constants is the universal set. The feasibility problem for constraints in DNF is therefore trivial. Because the proof is constructive, it also shows that we can find a feasible configuration relatively easily. If we use such a configuration as our starting point, we can enforce the constraint by restricting step finding to steps that preserve feasibility.

Definition 3.2.37 (Feasible Space under Logical Constraints).
Let (X, Σ, μ) be a measure space, let $n \in \mathbb{N}$, and let $\phi\colon (\Sigma/_{\sim_\mu})^n \to \Sigma/_{\sim_\mu}$ be a mapping in DNF. We refer to

$$\mathscr{F}_\phi := \big\{A \in (\Sigma/_{\sim_\mu})^n \,\big|\, \phi(A) = X\big\}$$

as the *feasible space* of the constraint $\phi(A) = X$. Let further $U \in (\Sigma/_{\sim_\mu})^n$. Then we refer to

$$\mathscr{F}_\phi(U) := \big\{D \in (\Sigma/_{\sim_\mu})^n \,\big|\, \phi(U \mathbin{\triangle} D) = X\big\}$$

as the *space of admissible steps* or *admissible step space* of the constraint $\phi(A) = X$ in U. ◁

In contrast to scalar inequality constraints, where we rely on relaxation, logical constraints are can likely be solved with methods that use a modified step-finding procedure. We have already shown in Lemma 3.2.36 that it is easy to find an initial feasible solution. If we can find a step-finding routine that optimizes projected descent over all steps in $\mathscr{F}_\phi(U)$, then the unconstrained framework that we had developed in Section 3.1 could be made to optimize over $\mathscr{F}_\phi$.

Of course, this would need to be properly theoretically argued. Optimality criteria would have to be proven for this case. This would likely not be a major obstacle. However, doing so would significantly expand the theoretical scope of this discussion. Rather than engaging in further theory building, we construct a proof of concept for a special type of logical constraint, thus demonstrating that developing logically constrained step-finding routines is a very feasible endeavor.

3.2.2.1 PARTITION CONSTRAINTS

In combinatorial optimization, the *special ordered set of type 1* ("SOS1") is a type of constraint that is amenable to more intelligent search algorithms than regular binary Branch and Bound. It is therefore a popular structure in combinatorial modelling. An SOS1 is a subset of binary optimization variables, out of which exactly one must be active in order for the SOS1 constraint to be satisfied.

The SOS1 constraint has its counterpart in our setting in the partition constraint, which dictates that $n \in \mathbb{N}$ set variables residing in a shared measure space (X, Σ, μ) must form a partition of the universal set X. In DNF, the constraint has the form

$$X = \phi(A) \coloneqq \bigcup_{i=1}^{n} \Big(A_i \cap \bigcap_{\substack{j \in [n] \\ j \neq i}} A_j^{\complement} \Big).$$

In other words, it is satisfied if and only if almost all points are members of one set while simultaneously not being members of any of the other sets. In terms of step-finding, the partition constraint is particularly benign because transitioning from one valid configuration to another always requires two set membership changes: removal from the old set and addition to the new set.

This means that, when we assess the "steepness" of the objective descent associated with a step, we do not need to explicitly account for locally differing numbers of individual set transitions. We also need not account for the possibility that the optimal configuration to transition a given point to can change based on the amount of potential step measure that remains.

This could, for instance, happen if there were two valid transitions for a point such that one requires more individual set membership changes but promises greater absolute descent, while the other requires less individual changes, which makes it shorter, but is also steeper. For the partition constraint, this is not possible because all transitions require the same number of individual set membership changes.

Algorithm 9 on the facing page describes how to find a steepest descent step under a partition constraint. The algorithm is fairly straightforward because it derives from the approximate unconstrained minimal mean step, which already gives all necessary guarantees.

The algorithm first calculates an "adjusted gradient" which gives us the aggregate cost of switching from the current set to the i-th set in a given point. This is simply done by adding the gradient from the i-th layer to the gradient for the currently active layer. Because the layers of U form a partition of X, the result is a sum of only two gradients almost everywhere. Special care must be taken for points where the i-th layer is already active. Here, the gradient inverts for re-activation, yielding an adjusted gradient density of zero.

Next, the algorithm finds the pointwise minimum of these adjusted gradient densities. This yields a piecewise function $g' \in L^1(\Sigma, \mu)$ that is patched together from the adjusted gradient densities g'_i, and a measurable index function k that indicates which layer the minimum corresponds to in each point.

This pointwise minimum is then used as an input for the unconstrained minimal mean step procedure (Algorithm 5). The trust region radius is halved because we have to make two set membership changes in each point of the step. This is inaccurate if the minimum corresponds to the currently active layer, in which case zero changes are necessary. However, in such points, the minimum

Algorithm 9 Steepest descent step under partition constraint

Require: (X, Σ, μ) atomless measure space, $n \in \mathbb{N}$, g integrable over the layered space consisting of n Σ-layers, U from the same layered space such that $\big(\{x \in X \,|\, (i,\, x) \in U\}\big)_{i\in[n]}$ is a μ-essential partition of X, $\Delta > 0$ and $\overline{\delta} > 0$.

Ensure: Yields V from layered space such that $\big(\{x \in X \,|\, (i,\, x) \in V\}\big)_{i\in[n]}$ is a μ-essential partition of X and $\mu'(U \mathbin{\triangle} V) \le \Delta$ where μ' is the sum measure in the layered space, as well as $\delta \in [0,\, \overline{\delta}]$ such that

$$\int_{U\triangle V} g \,\mathrm{d}\mu' \le \int_{U\triangle W} g \,\mathrm{d}\mu' + \delta$$

for all W from layered space for which $\big(\{x \in X \,|\, (i,\, x) \in V\}\big)_{i\in[n]}$ is a μ-essential partition of X and $\mu'(U \mathbin{\triangle} W) \le \Delta$.

function AdjustedGradient(i, g, U) ▷ "Gradient" for switch to i-th set
  $g' \leftarrow x \mapsto g(i,x)$
  $g' \leftarrow \begin{cases} 0 & \text{on } \{x \in X \,|\, (i,\, x) \in U\} \\ g' & \text{on } \{x \in X \,|\, (i,\, x) \notin U\} \end{cases}$
  for all $j \in [n] \setminus \{i\}$ **do**
    $g' \leftarrow \begin{cases} g' + g(j,\cdot) & \text{on } \{x \in X \,|\, (j,\, x) \in U\} \\ g' & \text{on } \{x \in X \,|\, (j,\, x) \notin U\} \end{cases}$
  end for
  return g'
end function

function PointwiseMinimum(g_1, …, g_n)
  $(g'_1,\, k_1) \leftarrow \big(x \mapsto g_1,\, x \mapsto 1\big)$
  for all $i \in [n] \setminus \{1\}$ **do**
    $N_i \leftarrow \big\{g_i < g'_{i-1}\big\}$
    $g'_i \leftarrow \begin{cases} g'_{i-1} & \text{on } N_i^{\complement}, \\ g_i & \text{on } N_i \end{cases}$
    $k_i \leftarrow \begin{cases} k_{i-1} & \text{on } N_i^{\complement}, \\ i & \text{on } N_i \end{cases}$
  end for
  return $(g'_n,\, k_n)$
end function

procedure AMMStepPartition(g, U, Δ, $\overline{\delta}$)
  for all $i \in [n]$ **do**
    $g'_i \leftarrow$ AdjustedGradient(i, g, U)
  end for
  $(g',\, k) \leftarrow$ PointwiseMinimum(g'_1, …, g'_n)
  $(D,\, \delta) \leftarrow \mathrm{AMMStep}\Big(g',\, \frac{\Delta}{2},\, \overline{\delta}\Big)$ ▷ Invoke Algorithm 5
  $V \leftarrow \bigcup_{i=1}^{n} \Big(\big(\{i\} \times \big(D \cap \{k = i\}\big)\big) \cup \big\{(i,\, x) \in U \,\big|\, x \notin D\big\}\Big)$
  return $(V,\, \delta)$
end procedure

g' would have the exact value zero and Algorithm 5 never selects points with gradient density zero.

The result is a step set $D \in \Sigma$ and $\delta \in [0, \overline{\delta}]$ such that $\mu(D) \leq \frac{\Delta}{2}$ and

$$\int_D g' \,\mathrm{d}\mu \leq \int_W g' \,\mathrm{d}\mu + \delta$$

for all $W \in \Sigma$ with $\mu(W) \leq \frac{\Delta}{2}$.

To simplify notation, let $A_i \coloneqq \{x \in X \,|\, (i,\, x) \in A\}$ refer to the i-th layer of a set A in layered space. The endpoint of the step satisfies

$$V_i = \big(\{k = i\} \cap D\big) \cup (U_i \setminus D) \quad \forall i \in [n].$$

For each $i \in [n]$, we have

$$\begin{aligned}
U_i \setminus V_i &= (U_i \cap D) \setminus \{k = i\} \\
&= D \cap U_i \cap \{k \neq i\}, \\
V_i \setminus U_i &= \big(\{k = i\} \cap D\big) \setminus U_i \\
&= D \cap U_i^{\complement} \cap \{k = i\}, \\
U_i \mathbin{\triangle} V_i &= (U_i \setminus V_i) \cup (V_i \setminus U_i) \\
&= D \cap \Big(\big(U_i \cap \{k \neq i\}\big) \cup \big(U_i^{\complement} \cap \{k = i\}\big)\Big).
\end{aligned}$$

We know that the layers of U form an essential partition of X. The level sets $\{k = i\}$ for $i \in [n]$ similarly form an essential partition of X. Therefore, we have

$$\begin{aligned}
\mu'(U \mathbin{\triangle} V) &= \sum_{i=1}^{n} \mu(U_i \mathbin{\triangle} V_i) \\
&\leq \sum_{i=1}^{n} \Big(\mu\big(D \cap U_i \cap \{k \neq i\}\big) + \mu\big(D \cap U_i^{\complement} \cap \{k = i\}\big)\Big) \\
&\leq \mu\Big(\bigcup_{i=1}^{n} D \cap U_i\Big) + \mu\Big(\bigcup_{i=1}^{n} D \cap \{k = i\}\Big) \\
&\leq \mu(D \cap X) + \mu(D \cap X) \\
&\leq 2\mu(D) \\
&\leq \Delta.
\end{aligned}$$

On $D^{\complement}$, the layers of V are unaltered from the layers of U and therefore still form an essential partition. On D, they are the level sets of k and therefore form a partition. Thus, the layers of V still form an essential partition of the universal set X.

Let W be any set in layered space whose layers form an essential partition of X. The difference in projected descent can be written as follows:

$$\begin{aligned}
\int_{U \triangle W} g \,\mathrm{d}\mu' - \int_{U \triangle V} g \,\mathrm{d}\mu' &= \sum_{i=1}^{n} \left(\int_{U_i \triangle W_i} g(i,\, x) \,\mathrm{d}\mu(x) - \int_{U_i \triangle V_i} g(i,\, x) \,\mathrm{d}\mu(x)\right) \\
&= \sum_{i=1}^{n} \Bigg(\int_{U_i \setminus W_i} g(i,\, x) \,\mathrm{d}\mu(x) + \int_{W_i \setminus U_i} g(i,\, x) \,\mathrm{d}\mu(x) \\
&\qquad - \int_{U_i \setminus V_i} g(i,\, x) \,\mathrm{d}\mu(x) - \int_{V_i \setminus U_i} g(i,\, x) \,\mathrm{d}\mu(x)\Bigg).
\end{aligned}$$

Because the layers of U, V, and W all respectively form essential partitions of X, we can rewrite

$$U_i \setminus W_i = \bigcup_{\substack{j \in [n] \\ j \neq i}} U_i \cap W_j.$$

Similar equalities also hold for $W_i \setminus U_i$, $U_i \setminus V_i$, and $V_i \setminus U_i$. This then yields

$$\int_{U_i \setminus W_i} g(i,\, x)\,\mathrm{d}\mu(x) = \sum_{\substack{j \in [n] \\ j \neq i}} \int_{U_i \cap W_j} g(i,\, x)\,\mathrm{d}\mu(x).$$

Again, similar reformulations are possible for the remaining integrals. By rearranging the summands, we obtain

$$\begin{aligned}
\int_{U \triangle W} g\,\mathrm{d}\mu' - \int_{U \triangle V} g\,\mathrm{d}\mu' &= \sum_{\substack{i,\, j \in [n] \\ i \neq j}} \Bigg(\int_{U_i \cap W_j} g(i,\, x) + g(j, x)\,\mathrm{d}\mu(x) \\
&\qquad - \int_{U_i \cap V_j} g(i,\, x) + g(j,\, x)\,\mathrm{d}\mu(x) \Bigg) \\
&= \sum_{\substack{i,\, j \in [n] \\ i \neq j}} \left(\int_{U_i \cap W_j} g'_j\,\mathrm{d}\mu - \int_{U_i \cap V_j} g'_j\,\mathrm{d}\mu \right).
\end{aligned}$$

Because the aggregate g' is chosen as a pointwise minimum over all g'_j, we know that $g'_j \geq g'$. What sets $U \triangle V$ apart from all other steps is that $U \triangle V$ realizes the pointwise minimum, i.e., we have

$$U_i \cap V_j = D \cap U_i \cap \{k = j\}$$

and therefore

$$\int_{U_i \cap V_j} g'_j\,\mathrm{d}\mu = \int_{D \cap U_i \cap \{k=j\}} g'_j\,\mathrm{d}\mu = \int_{D \cap U_i \cap \{k=j\}} g'\,\mathrm{d}\mu$$

because $\{k = j\}$ is precisely the set where $g' = g'_j$. For the sum, this means that

$$\begin{aligned}
\sum_{\substack{i,\, j \in [n] \\ i \neq j}} \int_{U_i \cap V_j} \underbrace{g'_j}_{=0 \text{ for } i=j}\,\mathrm{d}\mu &= \sum_{i=1}^{n} \sum_{j=1}^{n} \int_{U_i \cap V_j} g'_j\,\mathrm{d}\mu \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \int_{D \cap U_i \cap \{k=j\}} g'\,\mathrm{d}\mu \\
&= \sum_{i=1}^{n} \int_{D \cap U_i} g'_j\,\mathrm{d}\mu \\
&= \int_D g'\,\mathrm{d}\mu.
\end{aligned}$$

For $U \triangle W$, the reformulation works slightly differently because $g'_j \neq g'$ is possible on $U_i \cap W_j$. Here, we can use the fact that the layers of U and W form essential partitions of X. Aside from the pairwise essential disjointness of the intersections

$U_i \cap W_j$, this also implies that

$$\begin{aligned}\sum_{\substack{i,\, j\in[n]\\ i\neq j}} \int_{U_i\cap W_j} \underbrace{g'_j}_{\geq g'} \,\mathrm{d}\mu &\geq \sum_{\substack{i,\, j\in[n]\\ i\neq j}} \int_{U_i\cap W_j} g' \,\mathrm{d}\mu \\ &= \sum_{i=1}^{n} \sum_{\substack{j\in[n]\\ i\neq j}} \int_{U_i\cap W_j} g' \,\mathrm{d}\mu \\ &= \sum_{i=1}^{n} \int_{U_i\setminus W_i} g' \,\mathrm{d}\mu \\ &= \int_{\bigcup_{i=1}^{n} U_i\setminus W_i} g' \,\mathrm{d}\mu.\end{aligned}$$

The last equality is justified because the U_i are pairwise disjoint. Because the U_i form an essential partition of X, almost every point $x \in W_i \setminus U_i$ is an element of some U_j. Simultaneously, almost all $x \in W_i \setminus U_i$ would satisfy $x \notin W_j$. Therefore, for almost every $x \in W_i \setminus U_i$, there exists $j \in [n]$ such that $x \in U_j \setminus W_j$. From this, we can infer that

$$\bigcup_{i=1}^{n}(U_i \setminus W_i) \sim_\mu \bigcup_{i=1}^{n}(W_i \setminus U_i) \sim_\mu \bigcup_{i=1}^{n}(U_i \triangle W_i),$$

which implies that

$$\begin{aligned}\mu\Big(\bigcup_{i=1}^{n}(U_i \triangle W_i)\Big) &= \frac{1}{2}\cdot\Big(\mu\Big(\bigcup_{i=1}^{n}(U_i \setminus W_i)\Big) + \mu\Big(\bigcup_{i=1}^{n}(W_i \setminus U_i)\Big)\Big) \\ &= \frac{1}{2}\cdot\Big(\sum_{i=1}^{n}\mu(U_i \setminus W_i) + \sum_{i=1}^{n}\mu(W_i \setminus U_i)\Big) \\ &= \frac{1}{2}\cdot\sum_{i=1}^{n}\mu(U_i \triangle W_i) \\ &= \frac{1}{2}\cdot\mu'(U \triangle W).\end{aligned}$$

Therefore, $\mu'(U \triangle W) \leq \Delta$, implies

$$\mu\Big(\bigcup_{i=1}^{n} U_i \triangle W_i\Big) \leq \frac{\Delta}{2}.$$

Because D is an output from Algorithm 5, we have

$$\begin{aligned}\sum_{\substack{i,\, j\in[n]\\ i\neq j}} \int_{U_i\cap W_j} \underbrace{g'_j}_{\geq g'} \,\mathrm{d}\mu &\geq \int_{\bigcup_{i=1}^{n} U_i\setminus W_i} g' \,\mathrm{d}\mu \\ &= \int_{\bigcup_{i=1}^{n} U_i\triangle W_i} g' \,\mathrm{d}\mu \\ &\geq \int_D g' \,\mathrm{d}\mu - \delta.\end{aligned}$$

By aggregating this estimate with the prior results, we obtain

$$\begin{aligned}\int_{U\triangle W} g\,\mathrm{d}\mu' - \int_{U\triangle V} g\,\mathrm{d}\mu' &= \sum_{\substack{i,\ j\in[n]\\ i\neq j}} \left(\int_{U_i\cap W_j} g'_j\,\mathrm{d}\mu - \int_{U_i\cap V_j} g'_j\,\mathrm{d}\mu\right)\\ &\geq \int_D g'\,\mathrm{d}\mu - \delta - \int_D g'\,\mathrm{d}\mu\\ &= -\delta\end{aligned}$$

which yields the desired guarantee

$$\int_{U\triangle V} g\,\mathrm{d}\mu' \leq \int_{U\triangle W} g\,\mathrm{d}\mu' + \delta$$

with $\delta \leq \overline{\delta}$. Thus, within a certain margin of error, Algorithm 9 yields the greatest descent among all steps respecting the partition constraint. This would make it suitable as a steepest descent step for optimization under partition constraints.

3.2.2.2 ROUNDING SCHEMES

Partition constraints also open up the problem for a number of relaxation schemes from optimal control. In particular, we refer to sum up rounding (SUR) and combinatorial integral approximation (CIA) [Sag06; SJK11], as well as next-forced rounding (NFR) [Jun14]. All of these methods and their countless improvements approximate the behavior of a "relaxed" solution with a binary-valued one.

Our setting does not account for "relaxed" solutions. There is no immediate way to relax set membership. For the time being, we will assume that a relaxed solution takes the form of an integrable function u in layered space such that $u(i,x) \in [0,\ 1]$ and almost everywhere and

$$\sum_{i=1}^{n} u(i,x) = 1 \quad \text{almost everywhere on } X.$$

This is the approach followed by [MK20] in their approach to transferring SUR to a multi-dimensional setting.

Discussing the theory behind why the rounded solution behaves like the original solution is problem-specific and goes beyond the scope of our discussion here. However, the rounding process itself can benefit from our discussion in Section 2.3, because geodesics can be used to formulate rounding schemes with adaptive control grids in multiple dimensions.

The key to this lies in the concept of generator geodesics that we have introduced in Section 2.3.5. More specifically, we have shown in Theorem 2.3.62 on page 141 that every countable generator of an atomless measure space that consists entirely of sets of finite measure can be made into a generator geodesic.

The class of countably generated measure spaces notably includes $\mathbb{R}^n$ with the Borel σ-algebra $\mathscr{B}(\mathbb{R}^n)$ and the Lebesgue measure, which is one of the most frequently occurring settings in ODE- and PDE-constrained optimization. It also includes all finite subspaces of that measure space and all layerings of such spaces.

Let subsequently (X, Σ, μ) be a countably generated finite atomless measure space, let $\big([n] \times X, \Sigma^n, \mu^n\big)$ with

$$\Sigma^n \coloneqq \Big\{ U \in [n] \times X \;\Big|\; \{x \in X \,|\, (i,\, x) \in U\} \in \Sigma \;\forall i \in [n] \Big\},$$
$$\mu^n(U) \coloneqq \sum_{i=1}^{n} \mu\Big(\{x \in X \,|\, (i,\, x) \in U\}\Big)$$

be the layered measure space consisting of n Σ-layers. Let $u \in L^1(\Sigma^n,\, \mu^n)$ with $u(i,x) \in [0,\, 1]$ and

$$\sum_{i=1}^{n} u(i,x) = 1$$

for almost all $x \in X$. The function u will act as our "relaxed" set membership indicator.

According to Theorem 2.3.62, there exists a generator geodesic $\gamma\colon I \to \Sigma/_{\sim_\mu}$. Without loss of generality, let γ be minimizing. Let $g \in \mathscr{G}(X, \Sigma, \mu, I)$ be the GLSF of γ. For each $i \in [n]$, let

$$\widetilde{u}_i\colon X \ni x \mapsto u(i,\, x) \in \mathbb{R}$$

be the i-th layer of u. The σ-algebra generated by $\widetilde{u}_i$ is a sub-σ-algebra of Σ. Because γ is a generator geodesic, we have

$$\sigma(\widetilde{u}_i)/_{\sim_\mu} \subseteq \Sigma/_{\sim_\mu} = \sigma(g)/_{\sim_\mu} \quad \forall i \in [n].$$

We can therefore use the pushforward operator (see Definition 2.3.43 on page 113) to "unravel" each $\widetilde{u}_i$ into a function

$$u_i \coloneqq \widetilde{u}_i \otimes g^{-1} \in L^1(\mathscr{B}(I),\, \lambda) \quad \forall i \in [n].$$

This transforms our rounding problem in the layered space Σ^n into a more traditional one-dimensional rounding problem. We note that Manns and Kirches [MK20] follow a very similar approach in their multi-dimensional rounding scheme. They refer to so-called "order-conserving domain dissections," which are essentially hierarchial discretizations of generator geodesics. However, discretization occurs prior to rounding in their scheme. With the pushforward operator, we have the option of discretizing after our transition to the one-dimensional setting, which makes it easier to define adaptive rounding schemes. We begin with a simple meshless variant of the SUR algorithm which is stated in Algorithm 10 on the next page.

Algorithm 10 is relatively straightforward. Compared to the regular SUR algorithm presented by Sager [Sag06] (under the name SUR-SOS1), the primary difference is the dynamic determination of time steps, which is not necessary in the original algorithm because a time grid is given. The meshless algorithm has to choose its own time time grid, which consists of a dynamically chosen number of time points $t_k \in I$. Initially, it starts at $t_0 = a$.

On each time interval $[t_k,\, t_{k+1})$ for $k \in \mathbb{N}_0$, the index of the active variable is given by $i_k \in [n]$. For each time interval, the algorithm first determines i_k and then calculates t_{k+1} dependent on its choice of i_k. For each $i \in [n]$, the *cumulative error* of the i-th variable is $e_i\colon I \to \mathbb{R}$

$$e_i(t) \coloneqq \int_a^t u_i(s) - w_i(s)\,\mathrm{d}s \quad \forall i \in [n],\; t \in I.$$

Algorithm 10 Meshless sum up rounding

Require: $I = [a,\ b]$ non-empty compact interval, $n \in \mathbb{N}$, $u_i \in L^1\big(\mathscr{B}(I), \lambda\big)$ such that $u_i(t) \in [0,\ 1]$ almost everywhere for all $i \in [n]$, $\sum_{i=1}^{n} u_i(t) = 1$ almost everywhere, $\varepsilon > 0$.

Ensure: Yields n functions $w_i \in L^1\big(\mathscr{B}(I), \lambda\big)$ with

$$
\begin{aligned}
w_i(t) &\in \{0,\ 1\} && \forall i \in [n],\ t \in I,\\
\sum_{i=1}^{n} w_i(t) &= 1 && \forall t \in I,\\
\left| \int_0^t u_i(s) - w_i(s)\,\mathrm{d}s \right| &\leq \varepsilon && \forall i \in [n],\ t \in I.
\end{aligned}
$$

function MESHLESSSUR(a, b, $(u_i)_{i\in[n]}$, ε)
  $(k, t_0, e_{1,0}, \ldots, e_{n,0}) \leftarrow (0, a, 0, \ldots, 0)$ ▷ Time and error accumulators
  for all $i \in [n]$ **do** ▷ Calculate switching times for each choice
    for all $j \in [n] \setminus \{i\}$ **do** ▷ Project time to first switch
      $\widetilde{t}_{i,j,0} \leftarrow \sup\big\{t \in [t_0,\ b] \,\big|\, \int_{t_0}^{t} 1 + u_j(s) - u_i(s)\,\mathrm{d}s < \frac{\varepsilon}{n-1}\big\}$
    end for
    $\overline{t}_{i,0} \leftarrow \min\big(\{b\} \cup \{\widetilde{t}_{i,j,0} \,|\, j \in [n] \setminus \{i\}\}\big)$
  end for
  $i_0 \leftarrow \min\big\{i \in [n] \,\big|\, \overline{t}_{i,0} = \max\{\overline{t}_{j,0} \,|\, j \in [n]\}\big\}$
  $t_1 \leftarrow \overline{t}_{i_0,0}$
  while $t_{k+1} < b$ **do** ▷ Main rounding loop
    for all $i \in [n] \setminus \{i_k\}$ **do** ▷ Update accumulators
      $e_{i,k+1} \leftarrow e_{i,k} + \int_{t_k}^{t_{k+1}} u_i(s)\,\mathrm{d}s$
    end for
    $e_{i_k,k+1} \leftarrow e_{i_k,k} + \int_{t_k}^{t_{k+1}} u_{i_k}(s) - 1\,\mathrm{d}s$
    $k \leftarrow k + 1$
    $i_k \leftarrow \min\big\{i \in [n] \,\big|\, e_{i,k} = \max\{e_{j,k} \,|\, j \in [n]\}\big\}$
    for all $j \in [n] \setminus \{i_k\}$ **do** ▷ Find time to next switch
      $\widetilde{t}_{i_k,j,k} \leftarrow \sup\big\{t \in [t_k,\ b] \,\big|\, (e_{j,k} - e_{i_k,k}) + \int_{t_k}^{t} 1 + u_j(s) - u_i(s)\,\mathrm{d}s < \frac{\varepsilon}{n-1}\big\}$
    end for
    $t_{k+1} \leftarrow \min\big(\{b\} \cup \{\widetilde{t}_{i_k,j,k} \,|\, j \in [n] \setminus \{i_k\}\}\big)$
  end while
  for all $i \in [n]$ **do** ▷ Assemble output functions
    if $i = i_k$ **then**
      $w_i \leftarrow t \mapsto \chi_{\{b\}}(t) + \sum_{\substack{j=0 \\ i=i_j}}^{k} \chi_{[t_j,\ t_{j+1})}(t)$
    else
      $w_i \leftarrow t \mapsto \sum_{\substack{j=0 \\ i=i_j}}^{k} \chi_{[t_j,\ t_{j+1})}(t)$
    end if
  end for
  return $(w_i)_{i\in[n]}$
end function

We will see that $e_{i,k} = e_i(t_k)$ for all $i \in [n]$ and k such that t_k is defined.

The active control index i_k is determined using expressions of the form

$$\min\{i \in [n] \,|\, \bar{t}_{i,k} = \min\{\bar{t}_{j,k} \,|\, j \in [n]\}\}.$$

This may seem complicated, but it is simply a tie-broken variant of "arg min" where the tie is always broken by choosing the smallest index. We choose this over a bare "arg min" because it makes the algorithm fully deterministic.

Before we start the main induction argument, we should clarify the meaning of the expression

$$(e_{j,k} - e_{i,k}) + \int_{t_k}^{t} 1 + u_j(s) - u_i(s)\,\mathrm{d}s$$

for $t \geq t_k$ and $i, j \in [n]$ with $i \neq j$, which forms the basis of our time grid choice. Let us assume that $t_k \in I$ and that it has been shown that $e_{j,k} = e_j(t_k)$ and $e_{i,k} = e_i(t_k)$. If we were to extend w_i and w_j such that $w_i(s) = 1$ and $w_j(s) = 0$ for all $s \in [t_k,\ t)$, then we would have

$$\begin{aligned}
e_j(t) - e_i(t) = &\int_a^t u_j(s) - w_j(s)\,\mathrm{d}s - \int_a^t u_i(s) - w_i(s)\,\mathrm{d}s \\
= &\underbrace{\int_a^{t_k} u_j(s) - w_j(s)\,\mathrm{d}s}_{=e_j(t_k)} + \int_{t_k}^t u_j(s) - w_j(s)\,\mathrm{d}s \\
&- \underbrace{\int_a^{t_k} u_i(s) - w_i(s)\,\mathrm{d}s}_{=e_i(t_k)} - \int_{t_k}^t u_i(s) - w_i(s)\,\mathrm{d}s \\
= &\,(e_{j,k} - e_{i,k}) + \int_{t_k}^t \big(u_j(s) - u_i(s)\big) - \big(\underbrace{w_j(s)}_{=0} - \underbrace{w_i(s)}_{=1}\big)\,\mathrm{d}s \\
= &\,(e_{j,k} - e_{i,k}) + \int_{t_k}^t 1 + u_j(s) - u_i(s)\,\mathrm{d}s.
\end{aligned}$$

Thus, this expression is simply the difference between e_j and e_i extrapolated forward in time under the assumption that the i-th variable becomes active. Control switches are triggered when there is a $j \in [n]$ such that e_j is greater than e_{i_k} by a margin of at least $\frac{\varepsilon}{n-1}$. This margin exists so that a minimal difference between time steps can be guaranteed. This margin is infinite for $n = 1$, which is not a problem because there is only one variable to choose and no error can accumulate over any period of time.

The mapping

$$t \mapsto \int_{t_k}^t 1 + u_j(s) - u_i(s)\,\mathrm{d}s$$

has some useful properties. The integrand satisfies

$$0 \leq 1 + u_j(s) - u_i(s) \leq 2$$

for all $s \in [t_k,\ t)$. Therefore the difference in cumulative error is monotonically increasing, but can never increase with a rate of more than 2, i.e., we have

$$e_j(t) - e_i(t) \leq (e_{j,k} - e_{i,k}) + 2 \cdot (t - t_k)$$

for $t \geq t_k$ assuming that the i-th control is active between t_k and t. The monotonicity and continuity of the extrapolated cumulative error difference is also the reason why the algorithm is not problematic to implement. The times $\widetilde{t}_{i,j,k}$ can be approximated to arbitrary precision with a bisection scheme.

Theorem 3.2.38 (Correctness of Algorithm 10).
Let $I \coloneqq [a,\ b]$ be a non-empty compact interval, let $n \in \mathbb{N}$, let $u_i \in L^1\big(\mathscr{B}(I), \lambda\big)$ for each $i \in [n]$ where λ is the Lebesgue measure on $\mathscr{B}(I)$ such that $u_i(t) \in [0,\ 1]$ for all i and $\sum_{i=1}^n u_i(t) = 1$ almost everywhere in I. Let $\varepsilon > 0$.

Then Algorithm 10 terminates in finite time and returns integrable functions $w_i : I \to \{0,\ 1\}$ for $i \in [n]$ such that

$$\sum_{i=1}^{n} w_i(t) = 1 \quad \forall t \in I,$$
$$\left| \int_0^t u_i(s) - w_i(s)\,\mathrm{d}s \right| \leq \varepsilon \quad \forall i \in [n],\ t \in I. \qquad \triangleleft$$

Proof. Part 1 (Simplifications). For all $k \in \mathbb{N}_0$ prior to termination of the main rounding loop, t_{k+1} is chosen as a minimum over a non-empty set of numbers from $[t_k,\ b]$. Because $t_0 = a \leq b$, this guarantees inductively that the sequence $k \mapsto t_k$ is monotonically increasing and bounded above by b. This notably implies that the intervals $[t_k,\ t_{k+1})$ are pairwise disjoint. Once i_k is fixed, the value of each w_i is fixed on $[t_0,\ t_{k+1})$. We will not introduce separate symbols for partial sums up to those grid points. Instead, once i_j has been fixed for all $j \leq k$, we can argue, for instance, that

$$\sum_{i=1}^{n} w_i(t) = \sum_{i=1}^{n} \sum_{\substack{j=0 \\ i_j = i}}^{K-1} \underbrace{\chi_{[t_j,\ t_{j+1})}(t)}_{=0 \text{ for } j>k} = \sum_{j=0}^{k-1} \sum_{\substack{i=1 \\ i_j = i}}^{n} \chi_{[t_j,\ t_{j+1})}(t) = \chi_{[t_0,\ t_{k+1})}(t) = 1$$

for all $t \in [t_0,\ t_{k+1})$ where K is some hypothetical index – possibly ∞ – at which the loop terminates. Similarly, we can argue that the cumulative error functions

$$e_i(t) \coloneqq \int_a^t u_i(s) - w_i(s)\,\mathrm{d}s$$

are well defined on $[t_0,\ t_{k+1}]$. They satisfy

$$\begin{aligned}
\sum_{i=1}^{n} e_i(t) &= \sum_{i=1}^{n} \int_a^t u_i(s) - w_i(s)\,\mathrm{d}s \\
&= \int_a^t \underbrace{\Big(\sum_{i=1}^{n} u_i(s)\Big)}_{=1 \text{ a.e.}} - \underbrace{\Big(\sum_{i=1}^{n} w_i(s)\Big)}_{=1 \text{ a.e.}}\,\mathrm{d}s \\
&= \int_a^t 1 - 1\,\mathrm{d}s \\
&= 0
\end{aligned}$$

for all $t \in [t_0,\ t_{k+1}]$. This latter result is significant because the integrand of the cumulative error function satisfies

$$u_i(t) - w_i(t) \begin{cases} \leq 0 & \text{if } w_i(t) = 1, \\ \geq 0 & \text{if } w_i(t) = 0. \end{cases}$$

Thus, the cumulative error function is monotonically decreasing for the active control and monotonically increasing for all inactive controls. We will use this to show that none of the cumulative error functions can ever fall below $-\frac{\varepsilon}{n-1}$. Conversely, the upper bound on all e_i then becomes

$$e_i(t) = -\sum_{\substack{j=1 \\ j\neq i}}^{n} \underbrace{e_j(t)}_{\geq -\frac{\varepsilon}{n-1}} \leq (n-1)\cdot\frac{\varepsilon}{n-1} = \varepsilon,$$

which is the central rounding guarantee. It is also important to note that, because the sum of all cumulative errors is always zero, there are always at least one $i \in [n]$ and $j \in [n]$ such that $e_i(t) \leq 0$ and $e_j(t) \geq 0$. This will be at the root of our argument for the global lower bound on the cumulative error.

PART 2 (MAIN INDUCTION ARGUMENT). We prove inductively that the following claims hold for all $k \in \mathbb{N}_0$ prior to the termination of the main rounding loop.

- $e_{i,k} = e_i(t_k)\ \forall i \in [n]$;
- $t_{k+1} = b$ or $t_{k+1} \geq t_k + \frac{\varepsilon}{2(n-1)}$;
- for all $t \in [t_k,\ t_{k+1})$ and $j \in [n]$, we have $e_j(t) < e_{i_k}(t) + \frac{\varepsilon}{n-1}$.

We begin the inductive argument with $k = 0$. We have

$$e_i(t_k) = \int_a^{t_0} u_i(s) - w_i(s)\,\mathrm{d}s = \int_a^a u_i(s) - w_i(s)\,\mathrm{d}s = 0 = e_{i,k} \quad \forall i \in [n].$$

For each $i, j \in [n]$ with $i \neq j$, we have $t_0 \leq \widetilde{t}_{i,j,0} \leq b$ by definition. For all $t \in I$ with $t < t_0 + \frac{\varepsilon}{2(n-1)}$, we have

$$\int_{t_0}^{t} 1 + u_j(s) - u_i(s)\,\mathrm{d}s \leq 2(t - t_0) < \frac{\varepsilon}{n-1}.$$

This implies that $\widetilde{t}_{i,j,0} \geq t_0 + \frac{\varepsilon}{2(n-1)}$. For all $i \in \mathbb{N}_0$ we have $\overline{t}_{i,0} \leq b$ with either $\overline{t}_{i,0} = b$ or

$$\overline{t}_{i,0} = \min\left\{\widetilde{t}_{i,j,0} \,\middle|\, j \in [n] \setminus \{i\}\right\} \geq t_0 + \frac{\varepsilon}{2(n-1)}.$$

From this, we can then infer that $t_{k+1} \leq b$ and that we have either $t_{k+1} = b$ or $t_{k+1} \geq t_0 + \frac{\varepsilon}{2(n-1)}$.

Because t_{k+1} is no greater than the infimum over all $\widetilde{t}_{i_k,j,0}$ and because

$$e_j(t) - e_{i_k}(t) = \int_a^t 1 + u_j(s) - u_{i_k}(s)\,\mathrm{d}s$$

is monotonically increasing, we have

$$e_j(t) - e_{i_k}(t) = \int_a^t 1 + u_j(s) - u_{i_k}(s)\,\mathrm{d}s < \frac{\varepsilon}{n-1}$$

for all $j \in [n]$ and $t \in [t_0,\ t_{k+1})$.

Next, we turn to the induction step. Let $k \in \mathbb{N}_0$ be such that the three induction claims hold for all prior k. The accumulator update sets

$$\begin{aligned}
e_{i,k+1} &= e_{i,k} + \begin{cases} \int_{t_k}^{t_{k+1}} u_i(s)\,\mathrm{d}s & \text{if } i \neq i_k, \\ \int_{t_k}^{t_{k+1}} u_i(s) - 1\,\mathrm{d}s & \text{if } i = i_k \end{cases} \\
&= \int_a^{t_k} u_i(s) - w_i(s)\,\mathrm{d}s + \begin{cases} \int_{t_k}^{t_{k+1}} u_i(s) - 0\,\mathrm{d}s & \text{if } i \neq i_k, \\ \int_{t_k}^{t_{k+1}} u_i(s) - 1\,\mathrm{d}s & \text{if } i = i_k \end{cases} \\
&= \int_a^{t_k} u_i(s) - w_i(s)\,\mathrm{d}s + \int_{t_k}^{t_{k+1}} u_i(s) - w_i(s)\,\mathrm{d}s \\
&= \int_a^{t_{k+1}} u_i(s) - w_i(s)\,\mathrm{d}s \\
&= e_i(t_{k+1})
\end{aligned}$$

for all $i \in [n]$.

The active control i_{k+1} is specifically chosen to maximize $e_{i,k+1} = e_i(t_{k+1})$, which means that

$$e_{j,k+1} - e_{i_{k+1},k+1} \leq 0$$

for all $j \in [n]$. Because the expression

$$e_j(t) - e_{i_{k+1}}(t) = (e_{j,k+1} - e_{i_{k+1},k+1}) + \int_{t_{k+1}}^{t} 1 + u_j(s) - u_{i_{k+1}}(s)\,\mathrm{d}s$$

is monotonically increasing for $t \geq t_{k+1}$ with a non-positive starting value and a slope of no more than 2, we have

$$\widetilde{t}_{i_{k+1},j,k+1} \geq t_{k+1} + \frac{\varepsilon}{2(n-1)} \quad \forall j \neq i_{k+1},$$

which then implies that $t_{k+2} = b$ or $t_{k+2} \geq t_{k+1} + \frac{\varepsilon}{2(n-1)}$. Because

$$(e_{j,k+1} - e_{i_{k+1},k+1}) + \int_{t_{k+1}}^{t} 1 + u_j(s) - u_{i_{k+1}}(s)\,\mathrm{d}s$$

is monotonically increasing on $[t_{k+1},\ b)$ and $t_{k+2} \leq \widetilde{t}_{i_{k+1},j,k+1}$ for all $j \neq i_{k+1}$, we have

$$\begin{aligned}
e_j(t) - e_{i_{k+1}}(t) &= (e_{j,k+1} - e_{i_{k+1},k+1}) + \int_{t_{k+1}}^{t} 1 + u_j(s) - u_{i_{k+1}}(s)\,\mathrm{d}s \\
&< \frac{\varepsilon}{n-1}
\end{aligned}$$

for all $j \neq i_{k+1}$ and $t \in [t_{k+1},\ t_{k+2})$ by definition of $\widetilde{t}_{i_{k+1},j,k+1}$ as the supremum over all times where the extrapolated difference in cumulative error is less than $\frac{\varepsilon}{n-1}$.

This proves the induction claims for all k until $t_{k+1} = b$, at which point the loop terminates. Termination occurs after at most $\left\lceil \frac{2(n-1)(b-a)}{\varepsilon} \right\rceil$ iterations because we have $t_{k+1} = b$ or $t_{k+1} - t_k \geq \frac{\varepsilon}{2(n-1)}$ for all k.

As we had noted before, the cumulative error $e_i(t)$ can only decrease if the i-th variable is active. However, there is always a variable for which $e_i(t) \geq 0$. The third induction claim therefore implies that none of the cumulative errors

can ever fall below $-\frac{\varepsilon}{n-1}$. As we had shown previously, this then implies that no cumulative error can rise above ε. In summary, we have

$$-\frac{\varepsilon}{n-1} \leq e_i(t) \leq \varepsilon \quad \forall i \in [n],\ t \in I.$$

This is sufficient to prove the approximation guarantee for $n > 1$. If $n = 1$, then we have $e_i(t) = 0\ \forall t$ by default because there is only one possible active variable. By extending the last active variable index into the nullset $\{b\}$, we additionally ensure that

$$\sum_{i=1}^{n} w_i(t) = 1 \quad \forall t \in I.$$

□

The layered set can then be reconstructed by taking the pullback functions $w_i \otimes g$ (according to Definition 2.3.43 on page 113) and defining $w\colon [n] \times X \to \mathbb{R}$ with

$$w(i,x) \coloneqq (w_i \otimes g)(x) \quad \forall i \in [n],\ x \in X.$$

Here, w is a layered version of the rounded control functions w_i and g is the GLSF of the generator geodesic γ. According to Theorem 2.3.41 on page 106, $w_i \otimes g$ has values in $\{0,\ 1\}$ because γ is minimizing. Theorem 2.3.41 also proves that $w_i \otimes g$ is integrable.

Because $\mathrm{TV}(\gamma) = X$, we have $g(x) < \infty$ almost everywhere in X, from which we can infer that

$$\sum_{i=1}^{n} w(i,x) = \sum_{i=1}^{n} (w_i \otimes g)(x) = \sum_{i=1}^{n} w_i\big(g(x)\big) = 1 \quad \text{almost everywhere in } X.$$

This guarantees that the layers of the set $W \coloneqq \{w = 1\}$ form an essential partition of X. Further, Theorem 2.3.41 guarantees the integrability of each $w_i \otimes g$ over all preimages of Borel-measurable subsets of I under g as well as

$$\int_B w_i \,\mathrm{d}\lambda = \int_{g^{-1}(B)} w_i \otimes g \,\mathrm{d}\mu = \int_{\gamma(B)} w_i \otimes g \,\mathrm{d}\mu \quad \forall B \in \mathscr{B}(I),$$

which specifically gives us the approximation guarantee

$$\begin{aligned}\left|\int_{\gamma(t)} \widetilde{u}_i - (w_i \otimes g)\,\mathrm{d}\mu\right| &= \left|\int_0^t (\widetilde{u}_i \otimes g^{-1}) - w_i \,\mathrm{d}s\right| \\ &= \left|\int_0^t u_i - w_i \,\mathrm{d}s\right| \\ &\leq \varepsilon\end{aligned}$$

for all $i \in [n]$ and $t \in I$. As we had noted, it is not easy to prove that this approximation guarantee translates into a useful approximation in the optimization objective. Such arguments are somewhat problem-specific and must likely be based on the fact that γ generates the entire similarity space $\Sigma/_{\sim_\mu}$. We point to the extensive body of work of Paul Manns with various co-authors[1] [MK20; KMU21; LMW21; Man+23] for a variety of weak-$*$ convergence arguments for rounding methods in multi-dimensional spaces.

[1] The author has participated in [Man+23], but considers his contributions to be limited to the formulation of the binary trust-region framework.

3.2.2.3 MESH-AWARE ROUNDING

Algorithm 10 works completely without a pre-defined mesh. The minimal distance $\frac{\varepsilon}{2(n-1)}$ means that the algorithm works with a fixed *resolution*, but the chosen time grid points t_k are not guaranteed to align with any particular pre-defined time grid. This is acceptable for problems that work in a one-dimensional domain. In fact, some numerical integration software, such as the ODE and DAE solvers of the SUNDIALS suite [Gar+22; Hin+05], offer special facilities to dynamically stop integration at the approximate time point when an integrated quantity crosses a certain threshold value. Hairer [HWN93, ch. II.6] gives an account of how such a system can work in practice under the term "event location."

In multi-dimensional problems, this approach of leaving the rounding algorithm completely free in its choice of time grid is problematic. Multi-dimensional grids (or "meshes" more generally) are usually limited in how they can be refined due to restrictions in the underlying data structures or design choices in the refinement procedure. Often, such design choices are informed by significant numerical concerns. For instance, they can be designed to avoid degeneration of mesh cells, which can, for instance, lead to near singularities in relevant linear equation systems, greatly damaging solver performance and solution quality.

This is a topic which greatly exceeds the scope of this thesis, as well as the expertise of its author. We will therefore assume the following simplified setting:

1. There is a dense subset of "possible" grid points $T \subseteq I =: \mathrm{dom}(\gamma)$ where γ is a generator geodesic;

2. There is an "order of preference" $\preceq$, which is a well-order on T.

The order of preference may require some explanation. We intend the order of preference to encode the difficulty of realizing a certain grid point. For instance, in a hierarchially refined mesh, some grid points may require advancing by multiple levels of refinement while others may be reachable without any further refinement at all. In almost all cases, the order of preference should be defined according to the following principle:

> For $s, t \in T$, "$s \preceq t$" should mean that realizing s as a switching time is no more computationally demanding than realizing t. If both are equally demanding to realize, then it means that $s \geq t$.

Breaking the tie in favor of the greater number makes it easier to ensure that $\preceq$ is a well-order and will make it so that the modified SUR algorithm always chooses the largest possible step among those easiest to computationally realize.

We should briefly define what a "well-order" is. A well-order $\preceq$ is a total order on the set T such that every non-empty subset $S \subseteq T$ has a minimum according to $\preceq$, i.e., an element $s \in S$ such that $s \preceq t \ \forall t \in S$. Because $\preceq$ is a total order, this minimum is also always unique.

Demanding that $\preceq$ be a well-order on a countable subset $T \subset I$ appears to be a very strong restriction. It is easy to define a countable subset of even the bounded interval $I \subset \mathbb{R}$ that has no minimum or maximum according to intuitive order relations such as $\leq$. This is where our additional suggestion to prefer time points that require smaller effort to realize becomes useful. Our assumption is that this will naturally bound the level of refinement that a minimum according to $\preceq$ can

require. Once the level of refinement is bounded, the number of time points is much more likely to be finite and should therefore have a unique maximum and minimum.

Hence, the well-order is likely not a strong restriction and provides a very elegant way for problem experts to pass their technical knowledge about problem implementation details into the algorithm. The question then becomes how we can incorporate the grid T and the order $\preceq$ into Algorithm 10.

We do so by splitting the cumulative error bound ε into two error bounds ε_1 and ε_2 such that $0 < \varepsilon_1 < \varepsilon_2$. Because the difference between the cumulative error function $e_j(t)$ of some inactive variable and $e_{i_k}(t)$ monotonically grows at a rate of no more than 2 and starts at $t = t_k$ with a non-positive initial value, we have guarantees that, after reaching $\frac{\varepsilon_1}{n-1}$, the difference requires at least $\frac{\varepsilon_2-\varepsilon_1}{2(n-1)}$ to cross the gap between the lower and upper error bounds. This gives us an interval with a non-empty interior to choose the next time point from. Because T is dense in I, the intersection between T and the interval of allowed switching times is always non-empty and we can choose its unique minimum according to the order of preference $\preceq$ as the next time point. The resulting rounding procedure is stated in Algorithm 11 on the facing page.

Algorithm 11 is largely identical to Algorithm 10. The only point where the two differ is the selection of the time grid. We will therefore not replicate the entire proof of Theorem 3.2.38 and instead only briefly note the differences.

For each $i, j \in [n]$ with $i \neq j$ and $k \in \mathbb{N}_0$ prior to termination, we have

$$\begin{aligned}
\check{t}_{i,j,0} &\geq t_0 + \frac{\varepsilon_1}{2(n-1)}, \\
\hat{t}_{i,j,0} &\geq \check{t}_{i,j,0} + \frac{\varepsilon_2 - \varepsilon_1}{2(n-1)}, \\
\check{t}_{i_k,j,k} &\geq t_k + \frac{\varepsilon_1}{2(n-1)} && \forall k > 0, \\
\hat{t}_{i_k,j,k} &\geq \check{t}_{i_k,j,k} + \frac{\varepsilon_2 - \varepsilon_1}{2(n-1)} && \forall k > 0.
\end{aligned}$$

The bound for $\check{t}$ can be inferred using the same arguments as we had used for $\widetilde{t}$ in Theorem 3.2.38. The bound for $\hat{t}$ follows from the bounded growth rate of the cumulative error difference.

This means that the lower bound for differences between time points now derives from ε_1, i.e., $t_{k+1} \geq t_k + \frac{\varepsilon_1}{2(n-1)}$, while the upper bound on inactive variables cumulative error is

$$e_j(t) \leq e_{i_k}(t) + \frac{\varepsilon_2}{n-1} \quad \forall j \neq i_k,\ t \in [t_k,\ t_{k+1}).$$

This then gives us the guarantee that

$$e_i(t) \in \left[-\frac{\varepsilon_2}{n-1},\ \varepsilon_2\right] \quad \forall i \in [n],\ t \in I.$$

Bearing in mind that $n = 1$ implies that the cumulative error is always zero, this ensures that the absolute value of the cumulative error of any of the control variables never exceeds ε_2. The output functions w_i can then be turned into a layered control set by using the same process that we had described for solutions of the meshless rounding algorithm.

Algorithm 11 Mesh-aware sum up rounding

Require: $I = [a,\ b]$ non-empty compact interval, $n \in \mathbb{N}$, $u_i \in L^1\big(\mathscr{B}(I), \lambda\big)$ such that $u_i(t) \in [0,\ 1]$ almost everywhere for all $i \in [n]$, $\sum_{i=1}^n u_i(t) = 1$ almost everywhere, $0 < \varepsilon_1 < \varepsilon_2$, $T \subseteq I$ dense in I with $a \in T$ and $b \in T$, $\preceq$ well-order relation on T.

Ensure: Yields n functions $w_i \in L^1\big(\mathscr{B}(I), \lambda\big)$ with $\sum_{i=1}^n w_i(t) = 1\ \forall t \in I$ and

$$\left| \int_0^t u_i(s) - w_i(s) \mathrm{d}s \right| \leq \varepsilon_2 \quad \forall i \in [n],\ t \in I.$$

function MeshAwareSUR(a, b, $(u_i)_{i\in[n]}$, ε_1, ε_2, T, $\preceq$)
  $(k,\ t_0,\ e_{1,0},\ \ldots,\ e_{n,0}) \leftarrow (0,\ a,\ 0,\ \ldots,\ 0)$ ▷ Time and error accumulators
  for all $i \in [n]$ **do** ▷ Calculate switching times for each choice
    for all $j \in [n] \setminus \{i\}$ **do** ▷ Project time to first switch
      $\check{t}_{i,j,0} \leftarrow \sup\big\{t \in [t_0,\ b] \,\big|\, \int_{t_0}^t 1 + u_j(s) - u_i(s) \mathrm{d}s < \frac{\varepsilon_1}{n-1}\big\}$
      $\hat{t}_{i,j,0} \leftarrow \sup\big\{t \in [t_0,\ b] \,\big|\, \int_{t_0}^t 1 + u_j(s) - u_i(s) \mathrm{d}s < \frac{\varepsilon_2}{n-1}\big\}$
      $\widetilde{t}_{i,j,0} \leftarrow \min_{\preceq}\big(T \cap [\check{t}_{i,j,0},\ \hat{t}_{i,j,0}]\big)$
    end for
    $\overline{t}_{i,0} \leftarrow \min\big(\{b\} \cup \big\{\widetilde{t}_{i,j,0} \,\big|\, j \in [n] \setminus \{i\}\big\}\big)$
  end for
  $i_0 \leftarrow \min\big\{i \in [n] \,\big|\, \overline{t}_{i,0} = \max\{\overline{t}_{j,0} \,|\, j \in [n]\}\big\}$
  $t_1 \leftarrow \overline{t}_{i_0,0}$
  while $t_{k+1} < b$ **do** ▷ Main rounding loop
    for all $i \in [n] \setminus \{i_k\}$ **do** ▷ Update accumulators
      $e_{i,k+1} \leftarrow e_{i,k} + \int_{t_k}^{t_{k+1}} u_i(s) \mathrm{d}s$
    end for
    $e_{i_k,k+1} \leftarrow e_{i_k,k} + \int_{t_k}^{t_{k+1}} u_{i_k}(s) - 1 \mathrm{d}s$
    $k \leftarrow k + 1$
    $i_k \leftarrow \min\big\{i \in [n] \,\big|\, e_{i,k} = \max\{e_{j,k} \,|\, j \in [n]\}\big\}$
    for all $j \in [n] \setminus \{i_k\}$ **do** ▷ Find time to next switch
      $\check{t}_{i_k,j,k} \leftarrow \sup\big\{t \in [t_k,\ b] \,\big|\, (e_{j,k} - e_{i_k,k}) + \int_{t_k}^t 1 + u_j(s) - u_i(s) \mathrm{d}s < \frac{\varepsilon_1}{n-1}\big\}$
      $\hat{t}_{i_k,j,k} \leftarrow \sup\big\{t \in [t_k,\ b] \,\big|\, (e_{j,k} - e_{i_k,k}) + \int_{t_k}^t 1 + u_j(s) - u_i(s) \mathrm{d}s < \frac{\varepsilon_2}{n-1}\big\}$
      $\widetilde{t}_{i_k,j,k} \leftarrow \min_{\preceq}\big(T \cap [\check{t}_{i_k,j,k},\ \hat{t}_{i_k,j,k}]\big)$
    end for
    $t_{k+1} \leftarrow \min\big(\{b\} \cup \big\{\widetilde{t}_{i_k,j,k} \,\big|\, j \in [n] \setminus \{i_k\}\big\}\big)$
  end while
  $w_{i_k} \leftarrow t \mapsto \chi_{\{b\}}(t) + \sum_{\substack{j=0 \\ i_j = i_k}}^{k} \chi_{[t_j,\ t_{j+1})}(t)$
  for all $i \in [n]$ **do**
    $w_i \leftarrow t \mapsto \sum_{\substack{j=0 \\ i_j = i}}^{k} \chi_{[t_j,\ t_{j+1})}(t)$
  end for
  return $(w_i)_{i\in[n]}$
end function

Chapter 4

Numerical Experiments

In the previous chapters, we have devoted considerable time to developing a framework for optimization by iterative local improvement in similarity spaces. In Chapter 3, we have developed two concrete optimization algorithms: a steepest descent method for unconstrained problems, and a quadratic penalty method for constrained problems. In this chapter, we demonstrate that these methods, which we had previously described in abstract form, are sufficiently concrete to be implemented and used in practice. To this end, we discuss two different optimization problems.

In Section 4.1, we discuss an instance of the Lotka-Volterra fishing problem. This is a classic example of an ODE-constrained optimal control problem with a single binary control variable that varies over time. It does not have any constraints aside from the ODE itself and is therefore suitable for the unconstrained steepest descent method that we have developed in Section 3.1.

In Section 4.2, we discuss a topology design problem based on the Poisson equation. This is a PDE-constrained optimal design problem with one binary design variable that varies over a spatial coordinate, and one simple scalar inequality constraint. This problem is suitable to test the quadratic penalty problem that we have discussed in Section 3.2.

We solve both problems with a custom software package that we have implemented in the Python programming language. The package is named PyCoimset (short for "**Co**ntinuous **Im**provement of **Set**s in **Py**thon"). The algorithm implementations in that package are designed to be applied to arbitrary user-defined optimization problems. Additional problems can be implemented by using an interface that we describe in broad strokes in Chapter C. As part of this thesis, we make the source code of PyCoimset available under the Apache 2.0 License at `https://github.com/mirhahn/pycoimset`. The results that we present in this chapter are generated with release version 0.1.7[1], which has been archived at [Hah25a]. PyCoimset uses basic functionality from the NumPy package [Har+20]. We cite additional packages used for the problem implementations in the respective subsections of this chapter. We make the run data from which all plots and tables in this chapter are generated available separately [Hah25b].

[1] Git commit `acf82bf85a10db37a97b7a0825a4345c794f2051`.

A pool of two test problems is likely too small to make an informed judgement about the quality of our methods. However, as we will see, each problem requires a non-negligible amount of theoretical consideration and implementation effort. We limit the problem pool to keep this chapter reasonably short.

We will also not compare the performance of our solvers with alternative methods by which the problems could be solved. There are multiple reasons for this. Most off-the-shelf optimization solvers require the search space to be of fixed, finite dimension and do not support adaptive error control. Comparing such a fixed-discretization solver with a variable-discretization solver is inherently difficult because the former does not incur any performance penalty from mesh refinement and error control while the latter does. The comparison is further complicated by the fact that it is not evident what mesh resolution to select for the fixed-discretiztion solver to obtain a "fair" comparison. Especially in mixed-integer problems, where the effort required to solve a problem can rise exponentially with increasing resolution, picking a high resolution may bias the comparison unfairly in favor of the variable-discretization solver. Even though the variable-discretization solver needs to expend more effort for a single function or gradient evaluation at higher resolutions, it can benefit from the greater accuracy that comes with higher resolution. Furthermore, because the algorithms in Chapter 3 are stated mesh-independently, it is not evident that their iteration count would increase at all when the mesh resolution increases. On the other hand, low resolutions can unfairly bias the comparison in favor of the fixed-discretization solver because it is unconcerned with the lesser accuracy of its function value and gradients, while the variable-discretization solver will have to perform more mesh refinement.

Even if we were capable of selecting a perfectly "fair" fixed resolution mesh for a comparison, the informative quality of the comparison would still be questionable unless we could be reasonably sure that both solvers are close to being "optimally implemented" in the sense that we could rely on any difference in performance being due to algorithmic properties rather than inadequacies in implementation[2]. Some performance comparisons avoid these problems by comparing more high-level performance metrics such as iteration counts. However, our method is so dissimilar from traditional integer optimization algorithms that it is questionable whether any such comparison would be meaningful.

In light of these problems, and bearing in mind the additional effort associated with writing additional problem implementations, we forego performance comparisons and provide these problems as mere proofs of concept in order to demonstrate that our algorithms can be applied in practice.

4.1 LOTKA-VOLTERRA FISHING PROBLEM

Our first problem is a standard instance of the Lotka-Volterra fishing problem. The problem is based on the well-known Lotka-Volterra system of ordinary differential equations. The Lotka-Volterra system was first put forward in [Lot25; Vol26b; Vol26a] to describe the population dynamics of multiple co-existing species.

A notable aspect of the Lotka-Volterra system is that it is conservative in that its initial state imbues the system with a fixed "energy" and that, assuming

[2]The author would like to emphasize that he is not, in fact, an expert in numerical ODE and PDE solution methods, or error control.

no outside influence, the system remains confined to fixed orbits around an equilibrium. The goal of the Lotka-Volterra fishing problem is to force the system into its equilibrium state by using a binary time-distributed control that increases the mortality of both species when activated.

We adopt the specific problem instance presented in [Sag+06; Sag06]:

$$\begin{aligned} &\inf_{w,x} \int_{t_0}^{t_f} \left\| x - (1,\ 1) \right\|^2 \mathrm{d}t \\ &\text{s.t.} \quad \dot{x}_1 = x_1 - x_1 x_2 - c_1 x_1 w \quad \text{a.e. in } [t_0,\ t_f], \\ &\qquad \dot{x}_2 = x_1 x_2 - x_2 - c_2 x_2 w \quad \text{a.e. in } [t_0,\ t_f], \\ &\qquad x(t_0) = x_0, \\ &\qquad w(t) \in \{0,\ 1\} \qquad\qquad \text{a.e. in } [t_0,\ t_f] \end{aligned} \tag{4.1}$$

with parameters

$$\begin{aligned} (t_0,\ t_f) &= (0,\ 12), \\ (c_1,\ c_2) &= (0.4,\ 0.2), \\ x_0 &= (0.5,\ 0.7). \end{aligned}$$

By fixing all coefficients of the regular Lotka-Volterra system to 1, the equilibrium state is fixed to (1, 1). The Lotka-Volterra fishing problem is of theoretical interest because the optimal solution of its canonical relaxation is known to include a singular arc, i.e., a period of time where the control function assumes intermediate values between 0 and 1. This provides a challenge to binary solvers because the behavior of the relaxed solution has to be approximated by switching between 0 and 1.

4.1.1 Theoretical Discussion

Before we discuss implementation details and results, we must first demonstrate that Problem (4.1) satisfies the theoretical preconditions for our algorithm. As a set-valued optimization problem, the problem is unconstrained. Therefore, the preconditions to satisfy are enumerated in Assumption 3.1.9 and Theorem 3.1.10 on page 231 and on page 232. The solution algorithm at issue will be Algorithm 4 on page 232 with Algorithm 5 on page 242 as its step-finding method.

The underlying measure space is, quite naturally, the space of Borel-measurable subsets of $[t_0,\ t_f]$ with the Lebesgue measure. Applying our usual notation, we are working in $(X,\ \Sigma,\ \mu)$ with $X := [t_0,\ t_f]$, $\Sigma := \mathscr{B}(X)$, and $\mu := \lambda$. We note that restricting ourselves to Borel-measurable sets is not a significant restriction from working with general Lebesgue-measurable sets. Every Lebesgue-measurable set is equal to a Borel-measurable set up to a Lebesgue-nullset, which means that the Lebesgue-σ-algebra and the Borel-σ-algebra generate the same similarity space. Let $\Sigma_\sim := \Sigma/_{\sim_\mu}$.

To generate a binary-valued control function w from a set $U \in \mathscr{B}(X)$, we simply map U to its indicator function

$$w(U) := \chi_U \in L^1(\Sigma, \mu) \quad \forall U \in \Sigma_\sim.$$

Let $w \mapsto x(w)$ be the solution mapping of the initial value problem constraining Problem (4.1). Then our objective functional takes the form

$$F(U) := \int_{t_0}^{t_f} \left\| x\big(w(U)\big) - (1,\ 1) \right\|^2 \mathrm{d}t \quad \forall U \in \Sigma_\sim.$$

In Section 2.4.3, we had developed a standardized way to demonstrate the differentiability of such functionals. The conditions for this differentiability argument are summarized in Assumption 2.4.17 on page 173.

4.1.1.1 STATE BOUNDS

Let $x = x(w)$ for some $w \in L^1(\Sigma, \mu)$ with $w(t) \in [0,\ 1]$ almost everywhere in X. We first note that, assuming that for all $t \in X$ such that both $x_i(w)$ are differentiable in t and satisfy $x_i(w)(t) > 0$, the quantity

$$E \coloneqq \sum_{i=1}^{2} \big(x_i(w) - \ln(x_i(w))\big)$$

is differentiable in t and satisfies

$$\begin{aligned}
\frac{\mathrm{d}E}{\mathrm{d}t} &= \sum_{i=1}^{2} \frac{\mathrm{d}x_i(w)}{\mathrm{d}t} \cdot \left(1 - \frac{1}{x_i(w)}\right) \\
&= x_1(w) \cdot \big(1 - x_2(w) - c_1 \cdot w\big) \cdot \frac{x_1(w) - 1}{x_1(w)} \\
&\quad + x_2(w) \cdot \big(-1 + x_1(w) - c_2 \cdot w\big) \cdot \frac{x_2(w) - 1}{x_2(w)} \\
&= \big(x_1(w) - 1\big) \cdot \big(1 - x_2(w) - c_1 \cdot w\big) \\
&\quad + \big(x_2(w) - 1\big) \cdot \big(-1 + x_1(w) - c_2 \cdot w\big) \\
&= x_1(w) - 1 - x_1(w)x_2(w) + x_2(w) - c_1 x_1(w) \cdot w + c_1 \cdot w \\
&\quad - x_2(w) + 1 + x_1(w)x_2(w) - x_1(w) - c_2 x_2(w) \cdot w + c_2 \cdot w \\
&= c_1 \cdot \big(1 - x_1(w)\big) \cdot w + c_2 \cdot \big(1 - x_2(w)\big) \cdot w.
\end{aligned}$$

The quantity E is a well-known conserved quantity of the Lotka-Volterra system. It can be thought of as the "energy" of the system. To account for the impact of our control, we have to introduce an additional fictional state x_3 with initial value $x_3(t_0) = 0$ and temporal derivative

$$\dot{x_3}(w) = -c_1 \cdot \big(1 - x_1(w)\big) \cdot w - c_2 \cdot \big(1 - x_2(w)\big) \cdot w.$$

The state x_3 acts as an external "energy pool" with which energy is exchanged through the effect of our control. By adding x_3 to E, we obtain

$$\overline{E} \coloneqq x_3(w) + \sum_{i=1}^{2} \Big(x_i(w) - \ln\big(x_i(w)\big)\Big),$$

which is a conserved quantity of the Lotka-Volterra fishing ODE system. Since both upper and lower bounds on the states x_1 and x_2 derive from upper bounds on the species energy term $x_i \mapsto x_i - \ln x_i$, it is necessary to establish an a priori lower bound on x_3. We have $x_3(w)(t_0) = 0$ and

$$\dot{x_3}(w) = -c_1 \cdot \underbrace{\big(1 - x_1(w)\big)}_{<1} \cdot \underbrace{w}_{\leq 1} - c_2 \cdot \underbrace{\big(1 - x_2(w)\big)}_{<1} \cdot \underbrace{w}_{\leq 1} \geq -(c_1 + c_2)$$

which means that, as long as $x_1(w) > 0$ and $x_2(w) > 0$, we have

$$x_3(w)(t) \geq -(c_1 + c_2) \cdot (t - t_0).$$

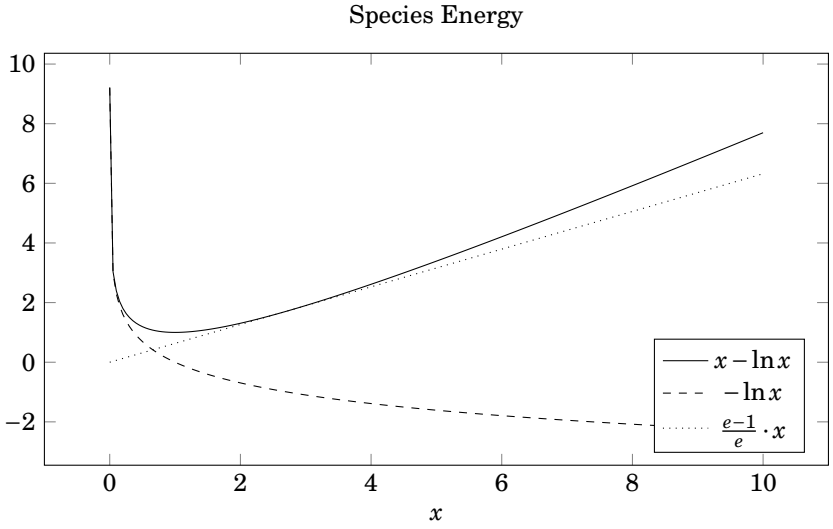


Figure 4.1: Plot of the single-species energy term $x \mapsto x - \ln x$ with the underestimators $x \mapsto -\ln x$ and $x \mapsto \frac{e-1}{e} \cdot x$.

The single-species energy term is convex and bounded below. Figure 4.1 depicts its graph. However, there is no straightforward closed-form expression for solutions of equations of the form $x - \ln x = y$ with fixed y. Therefore, we use underestimators. The energy term $x \mapsto x - \ln x$ assumes its minimum at $x = 1$, where we have $x - \ln x = 1 - 0 = 1$. For $x \in (0,\ 1)$, the function is strictly decreasing. Here, the logarithm dominates the behavior of the function and we have the underestimator

$$-\ln x < x - \ln x.$$

This means that, for a given energy budget $\overline{E}$, we obtain the estimate

$$\begin{aligned} \ln x_i &> -(x_i - \ln x_i) \\ &= -\big(\overline{E} - x_3 - \underbrace{(x_j - \ln x_j)}_{\geq 1}\big) \\ &\geq -\overline{E} - x_3 - 1 \end{aligned}$$

where j is the population state index that is not i (i.e., $j = 2$ if $i = 1$ and vice versa). Because the natural logarithm is strictly increasing, we can infer that

$$x_i > e^{-\overline{E} - x_3 - 1} \geq \underbrace{e^{-\overline{E} - (c_1 + c_2)\cdot(t_f - t_0) - 1}}_{=:\underline{x}} \quad \forall i \in \{1,\ 2\}.$$

For $x > 1$, the linear term begins to dominate the energy term. However, there is no constant shift $c \in \mathbb{R}$ for which $x \mapsto x + c$ is an underestimator of $x \mapsto x - \ln x$. Instead, we make use of the fact that $x \mapsto x - \ln x$ is convex and underestimate with a tangent. Describing the tangent in $x_0 > 0$ as an affine linear function $x \mapsto mx + c$, the parameters $m, c \in \mathbb{R}$ must satisfy the equation

$$\frac{\mathrm{d}}{\mathrm{d}x}(mx + c) = m = 1 - \frac{1}{x_0} = \frac{x_0 - 1}{x_0}$$

and

$$\underbrace{mx_0}_{=x_0-1} + c = x_0 - \ln x_0,$$

which yields $c = 1 - \ln x_0$. The general tangent function has the form

$$x \mapsto \frac{x_0 - 1}{x_0} \cdot x + 1 - \ln x_0$$

Aside from the requirement that $x_0 > 1$, the choice of x_0 is arbitrary. We set $x_0 := e > 1$ and obtain the underestimator $x \mapsto \frac{e-1}{e} \cdot x$. Indeed, we can verify that

$$x - \ln x - \frac{e-1}{e} \cdot x = \frac{1}{e}x - \ln x = 0 \quad \text{for } x = e$$

and

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{1}{e}x - \ln x\right) = \frac{1}{e} - \frac{1}{x}\begin{cases} < 0 & \text{for } x < e, \\ > 0 & \text{for } x > e. \end{cases}$$

From this underestimator, we can derive the upper bound

$$x_i \leq \frac{e}{e-1} \cdot (x_i - \ln x_i) \leq \frac{e}{e-1} \cdot (\overline{E} + x_3 + 1) \leq \underbrace{\frac{e}{e-1} \cdot \left(\overline{E} + (c_1 + c_2) \cdot (t_f - t_0) + 1\right)}_{=:\overline{x}}.$$

Both upper and lower bound depend on the energy budget $\overline{E}$, which is time-invariant and can be calculated from the initial state:

$$\overline{E} = x_{0,1} - \ln x_{0,1} + x_{0,2} - \ln x_{0,2}.$$

This gives us relatively tight state bounds for the entire time horizon. We can further relax them by $\varepsilon := \frac{1}{2} \cdot \underline{x}$ to preserve a lower bound strictly greater than zero while simultaneously establishing the margin between x and the boundary of the state domain required by Assumption 2.4.17. Our state domain is

$$D := (\underline{x} - \varepsilon,\ \overline{x} + \varepsilon)^2,$$

which is an open and convex subset of $\mathbb{R}^2$. Because the energy budget $\overline{E}$ is derived from the initial state, the lower and upper bounds also apply to x_0. We therefore have $x_0 \in D$. Any absolutely continuous function with $x(0) = x_0$ that satisfies the ODE system almost everywhere up to a point $\tau \in [t_0,\ t_f]$ then satisfies $B_\varepsilon(x(t)) \subseteq D$.

4.1.1.2 DIFFERENTIABILITY OF F

The right hand side of the ODE system has the form

$$f_0(x) + f_1(x) \cdot w$$

with

$$f_0(x) = \begin{pmatrix} x_1 \cdot (1 - x_2) \\ x_2 \cdot (-1 + x_1) \end{pmatrix}, \qquad f_1(x) = \begin{pmatrix} -c_1 \cdot x_1 \\ -c_2 \cdot x_2 \end{pmatrix},$$

$$\nabla f_0(x) = \begin{pmatrix} 1 - x_2 & -x_1 \\ x_2 & -1 + x_1 \end{pmatrix}, \qquad \nabla f_1(x) = \begin{pmatrix} -c_1 & 0 \\ 0 & -c_2 \end{pmatrix}.$$

Both f_0 and f_1 are quadratic polynomials, which guarantees that they are continuously differentiable and that their derivatives are Lipschitz continuous on $\mathbb{R}^2$. Since f_1 is linear, f_1 is trivially Lipschitz continuous on D. To demonstrate that f_0 is Lipschitz continuous on D, we have to make use of the fact that D is bounded. For the individual components of f_0, we can estimate that

$$\begin{aligned}
\left|f_{0,1}(x) - f_{0,1}(y)\right| &= \left|x_1 \cdot (1 - x_2) - y_1 \cdot (1 - y_2)\right| \\
&= |x_1 - y_1 + y_1 y_2 - x_1 x_2| \\
&\leq \left(1 + \max\{|y_2|, |x_2|\}\right) \cdot |x_1 - y_1| \\
&\leq (1 + \overline{x}) \cdot |x_1 - y_1|,
\end{aligned}$$

$$\begin{aligned}
\left|f_{0,2}(x) - f_{0,2}(y)\right| &= \left|x_2 \cdot (-1 + x_1) - y_2 \cdot (-1 + y_1)\right| \\
&= |y_2 - x_2 + x_1 x_2 - y_1 y_2| \\
&\leq \left(1 + \max\{|x_1|, |y_1|\}\right) \cdot |x_2 - y_2| \\
&\leq (1 + \overline{x}) \cdot |x_2 - y_2|.
\end{aligned}$$

This then implies that

$$\begin{aligned}
\left\|f_0(x) - f_0(y)\right\| &= \sqrt{\left|f_{0,1}(x) - f_{0,1}(y)\right|^2 + \left|f_{0,2}(x) - f_{0,2}(y)\right|^2} \\
&\leq \sqrt{(1 + \overline{x})^2 \cdot \left(|x_1 - y_1|^2 + |x_2 - y_2|^2\right)} \\
&= (1 + \overline{x}) \cdot \|x - y\|
\end{aligned}$$

which demonstrates the Lipschitz continuity of f_0 with Lipschitz constant $1 + \overline{x}$.

The Lagrange term has the form $l(x, w) = l_0(x)$ with

$$\begin{aligned}
l_0(x) &= (x_1 - 1)^2 + (x_2 - 1)^2, \\
\nabla l_0(x) &= \begin{pmatrix} 2x_1 - 2 & 0 \\ 0 & 2x_2 - 2 \end{pmatrix}.
\end{aligned}$$

This is a quadratic polynomial and therefore straightforwardly continuously differentiable with a Lipschitz continuous derivative.

With this, we have demonstrated that our setting satisfies all requirements enumerated in Assumption 2.4.17. We do not apply any scaling because doing so complicates error control. By satisfying these assumptions, the entirety of Section 2.4.3 becomes applicable to the objective functional F, proving that F is differentiable in the sense of Definition 2.4.1 and that the gradient density function g_U of F in $U \in \Sigma_\sim$ has the form

$$g_U = (1 - 2\chi_U) \cdot \xi_U^T \cdot f_1(x_U)$$

where x_U denotes the solution of the initial value problem, which exists and is confined to D according to Theorem 2.4.19, and ξ_U denotes the adjoint state function, which exists and is bounded according to Proposition 2.4.21. Because D is bounded and f_1 is Lipschitz continuous on D, g_U is also essentially bounded, meaning that F is benignly differentiable.

4.1.1.3 LIPSCHITZ CONTINUITY OF THE SET DERIVATIVE

In order to apply the limited curvature suboptimality estimator (see Proposition 2.4.8 on page 157), we have to show that the derivative of F is Lipschitz continuous. This is relatively simple because we have an explicit expression for the gradient density function. Let $U, V \in \Sigma_{\sim}$. For almost all $t \in U \triangle V$, we have

$$\begin{aligned}\left|g_U(t) + g_V(t)\right| &= \Big|\underbrace{\big(1-2\chi_U(t)\big)}_{\in\{-1,1\}} \cdot \xi_U^{\intercal}(t) \cdot f_1\big(x_U(t)\big) + \underbrace{\big(1-2\chi_V(t)\big)}_{=-\left(1-2\chi_U(t)\right)} \cdot \xi_V^{\intercal}(t) \cdot f_1\big(x_V(t)\big)\Big| \\ &= \left|\xi_U^{\intercal}(t) \cdot f_1\big(x_U(t)\big) - \xi_V^{\intercal}(t) \cdot f_1\big(x_V(t)\big)\right|.\end{aligned}$$

For $t \in (U \triangle V)^{\complement}$, we have

$$\begin{aligned}\left|g_U(t) - g_V(t)\right| &= \Big|\underbrace{\big(1-2\chi_U(t)\big)}_{\in\{-1,1\}} \cdot \xi_U^{\intercal}(t) \cdot f_1\big(x_U(t)\big) - \underbrace{\big(1-2\chi_V(t)\big)}_{=1-2\chi_U(t)} \cdot \xi_V^{\intercal}(t) \cdot f_1\big(x_V(t)\big)\Big| \\ &= \left|\xi_U^{\intercal}(t) \cdot f_1\big(x_U(t)\big) - \xi_V^{\intercal}(t) \cdot f_1\big(x_V(t)\big)\right|.\end{aligned}$$

This case distinction is to account for the local inversion of the gradient density. The remainder of the estimate is common to both cases:

$$\begin{aligned}&\left|\xi_U^{\intercal}(t) \cdot f_1\big(x_U(t)\big) - \xi_V^{\intercal}(t) \cdot f_1\big(x_V(t)\big)\right| \\ &\quad= \left|\xi_U^{\intercal}(t) \cdot \Big(f_1\big(x_U(t)\big) - f_1\big(x_V(t)\big)\Big) - \big(\xi_V(t) - \xi_U(t)\big)^{\intercal} \cdot f_1\big(x_V(t)\big)\right| \\ &\quad\leq \big\|\xi_U(t)\big\| \cdot \Big\|f_1\big(x_U(t)\big) - f_1\big(x_V(t)\big)\Big\| + \Big\|f_1\big(x_V(t)\big)\Big\| \cdot \big\|\xi_U(t) - \xi_V(t)\big\| \\ &\quad\leq \big\|\xi_U(t)\big\| \cdot L_1 \cdot \big\|x_U(t) - x_V(t)\big\| + \Big\|f_1\big(x_V(t)\big)\Big\| \cdot \big\|\xi_U(t) - \xi_V(t)\big\|\end{aligned}$$

where $L_1 \geq 0$ is a Lipschitz constant of f_1. As we have demonstrated in Proposition 2.4.21, there exists a uniform constant M_ξ such that $\big\|\xi_U(t)\big\| \leq M_\xi$ regardless of the choice of U. Similarly, because D is bounded and f_1 is continuous, $\|f_1\|$ assumes its maximum $M_f \geq 0$ over the closure $\overline{D}$. Because $x_V(t) \in D$ for all $t \in I$, we have

$$\begin{aligned}&\left|\xi_U^{\intercal}(t) \cdot f_1\big(x_U(t)\big) - \xi_V^{\intercal}(t) \cdot f_1\big(x_V(t)\big)\right| \\ &\quad\leq \big\|\xi_U(t)\big\| \cdot L \cdot \big\|x_U(t) - x_V(t)\big\| + \Big\|f_1\big(x_V(t)\big)\Big\| \cdot \big\|\xi_U(t) - \xi_V(t)\big\| \\ &\quad\leq M_\xi L_1 \cdot \big\|x_U(t) - x_V(t)\big\| + M_f \cdot \big\|\xi_U(t) - \xi_V(t)\big\|.\end{aligned}$$

The difference $x_U - x_V$ is an absolutely continuous function with $x_U(t_0) - x_V(t_0) = 0$ and

$$\frac{\mathrm{d}}{\mathrm{d}t}(x_U - x_V) = f_0(x_U) - f_0(x_V) + \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \cdot (\chi_V - \chi_U)$$

and therefore

$$\begin{aligned}\left\|\frac{\mathrm{d}}{\mathrm{d}t}(x_U - x_V)\right\| &\leq \big\|f_0(x_U) - f_0(x_V)\big\| + \left\|\begin{pmatrix} c_1 \\ c_2 \end{pmatrix}\right\| \cdot \underbrace{|\chi_V - \chi_U|}_{=\chi_{U \triangle V}} \\ &\leq L_0 \cdot \big\|x_U - x_V\big\| + \sqrt{c_1^2 + c_2^2} \cdot \chi_{U \triangle V}\end{aligned}$$

almost everywhere. Here, $L_0 \geq 0$ is a Lipschitz constant of f_0. By using a standard argument based on a variant of Gronwall's inequality, we obtain the estimate

$$\begin{aligned}\|x_U(t) - x_V(t)\| &\leq \sqrt{c_1^2 + c_2^2} \cdot \underbrace{\left(\int_{t_0}^{t} \chi_{U \triangle V} \, \mathrm{d}t \right)}_{=\mu(U \triangle V)} \cdot e^{L_0 \cdot (t - t_0)} \\ &= \underbrace{\sqrt{c_1^2 + c_2^2} \cdot e^{L_0 \cdot (t_f - t_0)}}_{=:C_1} \cdot \mu(U \triangle V).\end{aligned}$$

This estimate is very generous and likely exceeds the a priori bounds on the distance between two points in D. It is therefore almost certainly unsuitable for error estimation. However, it is theoretically relevant because it is proportional to $\mu(U \triangle V)$. For the adjoint state, the situation is somewhat simplified by the fact that ∇f_1 is constant. We have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(\xi_U - \xi_V) &= -\big(\nabla l_0(x_U) - \nabla l_0(x_V)\big) \\ &\quad - \big(\nabla f_0(x_U) + \nabla f_1(x_U) \cdot \chi_U\big)^{\mathsf{T}} \cdot \xi_U \\ &\quad + \big(\nabla f_0(x_V) + \nabla f_1(x_V) \cdot \chi_V\big)^{\mathsf{T}} \cdot \xi_V \\ &= -2(x_U - x_V) - \big(\nabla f_0(x_U)\big)^{\mathsf{T}} \cdot (\xi_U - \xi_V) - \big(\nabla f_0(x_U) - \nabla f_0(x_V)\big)^{\mathsf{T}} \cdot \xi_V \\ &\quad - \big(\nabla f_1(x_U)\big)^{\mathsf{T}} \cdot (\chi_U \cdot \xi_U - \chi_V \cdot \xi_V) \\ &= -2(x_U - x_V) - \big(\nabla f_0(x_U)\big)^{\mathsf{T}} \cdot (\xi_U - \xi_V) - \big(\nabla f_0(x_U) - \nabla f_0(x_V)\big)^{\mathsf{T}} \cdot \xi_V \\ &\quad - \big(\nabla f_1(x_U)\big)^{\mathsf{T}} \cdot \chi_U \cdot (\xi_U - \xi_V) - \big(\nabla f_1(x_U)\big)^{\mathsf{T}} \cdot (\chi_U - \chi_V) \cdot \xi_V.\end{aligned}$$

From this, we can infer that

$$\begin{aligned}\left\| \frac{\mathrm{d}}{\mathrm{d}t}(\xi_U - \xi_V) \right\| &\leq 2\|x_U - x_V\| + \big\|\nabla f_0(x_U)\big\| \cdot \|\xi_U - \xi_V\| + \big\|\nabla f_0(x_U) - \nabla f_0(x_V)\big\| \cdot \|\xi_V\| \\ &\quad + \big\|\nabla f_1(x_U)\big\| \cdot |\chi_U| \cdot \|\xi_U - \xi_V\| + \big\|\nabla f_1(x_U)\big\| \cdot |\chi_U - \chi_V| \cdot \|\xi_V\| \\ &\leq 2C_1 \cdot \mu(U \triangle V) + L_0 \cdot \|\xi_U - \xi_V\| + M_\xi \cdot L_0' \cdot \|x_U - x_V\| \\ &\quad + L_1 \cdot \|\xi_U - \xi_V\| + M_\xi \cdot L_1 \cdot \chi_{U \triangle V} \\ &\leq (2 + M_\xi \cdot L_0') \cdot C_1 \cdot \mu(U \triangle V) + (L_0 + L_1) \cdot \|\xi_U - \xi_V\| \\ &\quad + M_\xi \cdot L_1 \cdot \chi_{U \triangle V}\end{aligned}$$

where $L_0' \geq 0$ is a Lipschitz constant for ∇f_0. We can once more invoke a variant of Gronwall's inequality to obtain the estimate

$$\begin{aligned}\|\xi_U(t) - \xi_V(t)\| &\leq (2C_1 + M_\xi \cdot L_0' \cdot C_1 + M_\xi \cdot L_1) \cdot \mu(U \triangle V) \cdot e^{(L_0 + L_1) \cdot (t_f - t)} \\ &\leq \underbrace{(2C_1 + M_\xi \cdot L_0' \cdot C_1 + M_\xi \cdot L_1) \cdot e^{(L_0 + L_1) \cdot (t_f - t_0)}}_{=:C_2} \cdot \mu(U \triangle V).\end{aligned}$$

We note that the exponent is $t_f - t$ because we need to integrate backwards with the "initial" datum $\xi_U(t_f) - \xi_V(t_f) = 0$. By applying both estimates to our initial

inequality, we obtain

$$\begin{aligned}
&\left|\xi_U^\intercal(t)\cdot f_1\big(x_U(t)\big) - \xi_V^\intercal(t)\cdot f_1\big(x_V(t)\big)\right| \\
&\quad\le \underbrace{\left\|\xi_U^\intercal(t)\right\|}_{\le M_\xi}\cdot \underbrace{\left\|f_1\big(x_U(t)\big) - f_1\big(x_V(t)\big)\right\|}_{\le L_1\cdot\|x_U(t)-x_V(t)\|} + \underbrace{\left\|f_1\big(x_V(t)\big)\right\|}_{\le M_f}\cdot\left\|\xi_U^\intercal(t) - \xi_V^\intercal(t)\right\| \\
&\quad\le M_\xi L_1\cdot\underbrace{\left\|x_U(t)-x_V(t)\right\|}_{\le C_1\cdot\mu(U\triangle V)} + M_f\cdot\underbrace{\left\|\xi_U(t)-\xi_V(t)\right\|}_{\le C_2\cdot\mu(U\triangle V)} \\
&\quad\le (M_\xi L_1 C_1 + M_f C_2)\cdot\mu(U\triangle V).
\end{aligned}$$

For every $W\in\Sigma_\sim$, we have

$$\begin{aligned}
\big(\nabla F(U) \underset{U\triangle V}{\ominus} \nabla F(V)\big)(W) &= \int_{W\cap(U\triangle V)} |g_U + g_V|\,\mathrm{d}\mu + \int_{W\setminus(U\triangle V)} |g_U - g_V|\,\mathrm{d}\mu \\
&\le \int_W \left|\xi_U^\intercal\cdot f_1(x_U) - \xi_V^\intercal\cdot f_1(x_V)\right|\mathrm{d}\mu \\
&\le \int_W (M_\xi L_1 C_1 + M_f C_2)\cdot\mu(U\triangle V)\,\mathrm{d}\mu \\
&\le (M_\xi L_1 C_1 + M_f C_2)\cdot\mu(U\triangle V)\cdot\mu(W).
\end{aligned}$$

This demonstrates that the derivative of F is Lipschitz continuous in the sense of Definition 2.4.4.

4.1.1.4 NUMERICAL INTEGRATION AND ERROR CONTROL

We solve both the forward and adjoint initial value problems as well as all initial value problems required for error estimation using the Dormand-Prince method as implemented in version 1.15.0 of the SCIPY Python package [Vir+20]. The Dormand-Prince method is a Runge-Kutta method with seven stages that was originally put forward in [DP80] under the name "`RK5(4)7M`". For the purpose of sublevel set determination, we also require dense interpolants of the solution trajectory.

The Dormand-Prince method is known to have a quintic polynomial interpolant with an interpolation error of fourth order that requires no additional function evaluations, as well as a quintic interpolant with fifth-order interpolation error that requires additional function evaluations (see, e.g., [HWN93, Sec. II.6]). Additionally, there exists an alternative quartic polynomial interpolant proposed by Shampine [Sha86] that achieves a local interpolation error of fourth order. SCIPY uses the latter interpolant.

For error estimation we rely on residuals. Given an interpolated trajectory $\widetilde{x}\to I\to\mathbb{R}^n$, the residual in $t\in I$ is

$$r(t) = \dot{\widetilde{x}}(t) - f\big(\widetilde{x}(t), w(t)\big).$$

Here, the time derivative of $\widetilde{x}$ can be calculated explicitly because $\widetilde{x}$ is a continuous piecewise polynomial. In principle, because f is polynomial, the integral of r up to a given point $t\in I$ can be calculated exactly. However, this requires integration of eighth-degree polynomials, i.e., evaluation of ninth-degree polynomials, so we perform a numerical integration instead. Let

$$e_x \coloneqq \widetilde{x} - x$$

be the difference between the known approximate trajectory $\widetilde{x}$ and the exact trajectory x. Then the time derivative of the error satisfies

$$\dot{e}_x = \dot{\widetilde{x}} - f(x,w) \approx \dot{\widetilde{x}} - f(\widetilde{x},w) - \nabla_x f(\widetilde{x},w)\cdot e_x = r - \nabla_x f(\widetilde{x},w)\cdot e_x.$$

We obtain a pointwise error estimator for the solution state by solving the initial value problem

$$\begin{aligned} \dot{e}_x(t) &= \dot{\widetilde{x}}(t) - f\big(\widetilde{x}(t),w(t)\big) - \nabla_x f\big(\widetilde{x}(t),w(t)\big)\cdot e_x(t) \quad \text{for a.a. } t\in I,\\ e_x(t_0) &= (0,\ 0) \end{aligned}$$

in a generalized sense. We evaluate our objective error estimator at the same time by treating the integral as a third state with

$$\begin{aligned} f_3(x,w) &= \big\|x-(1,\ 1)\big\|^2,\\ x_{0,3} &= 0. \end{aligned}$$

The objective error estimator is then

$$e_F \coloneqq \big|e_{x,3}(t_f)\big|.$$

For the gradient error we obtain a similar pointwise error estimator. We have to additionally take into account the impact of the error in x on the right hand side of the adjoint ODE. For the approximate adjoint trajectory $\widetilde{\xi}$ and the exact adjoint trajectory ξ, the error $e_\xi \coloneqq \widetilde{\xi} - \xi$ is absolutely continuous and satisfies

$$\begin{aligned} \dot{e}_\xi &= \dot{\widetilde{\xi}} - \dot{\xi}\\ &= \dot{\widetilde{\xi}} + \big(\nabla_x l(x,w)\big)^\intercal + \big(\nabla_x f(x,w)\big)^\intercal \xi\\ &\approx \dot{\widetilde{\xi}} + \big(\nabla_x l(\widetilde{x},w) - e_x^\intercal\cdot\nabla^2_{xx} l(\widetilde{x},w)\big)^\intercal + \big(\nabla_x f(\widetilde{x},w) - e_x^\intercal\cdot\nabla^2_{xx} f(\widetilde{x},w)\big)^\intercal(\widetilde{\xi}-e_\xi) \end{aligned}$$

almost everywhere. Here, we integrate the pointwise error estimator backward in time starting at $e_\xi(t_f) = (0,\ 0)$ using dense interpolants for the forward error estimator e_x and the approximate adjoint state $\widetilde{\xi}$. This gives us an approximate trajectory of the pointwise error of the approximate adjoint solution $\widetilde{\xi}$.

As a surrogate for the gradient density function, we approximate the trajectory of

$$\phi \coloneqq \big(\nabla_w f(x,w)\big)^\intercal\cdot\xi \approx \underbrace{\big(\nabla_w f(\widetilde{x},w)\big)^\intercal\cdot\widetilde{\xi}}_{=:\widetilde{\phi}}.$$

This is essentially the gradient density function up to sign inversions. The advantage of working with ϕ is that ϕ is an absolutely continuous function. Rather than approximating ϕ using numerical integration, we interpolate ϕ using piecewise quartic polynomials on the coarsest joint refinement of the time grids used to approximate x and ξ. On each interval of the time grid, the quartic interpolant is fitted using function values $\widetilde{\phi}(t)$ on either end of the interval, the function value $\widetilde{\phi}(t)$ at the midpoint of the interval, and the time derivative $\dot{\widetilde{\phi}}(t)$ on either end of the interval. This is similar to the interpolation approach used by [Sha86] for the quartic interpolant of the Dormand-Prince solution.

We make use of the fact that f is a quadratic polynomial and therefore, the second derivatives of f are constant. The time derivative of ϕ is therefore

$$\begin{aligned}
\dot{\phi} &= \underbrace{\left(\nabla^2_{wx} f(x,w)\cdot\dot{x}\right)^\intercal}_{=\nabla^2_{wx} f(\widetilde{x},w)}\cdot\xi + \left(\nabla_w f(x,w)\right)^\intercal\cdot\dot{\xi} \\
&= \left(\nabla^2_{wx} f(\widetilde{x},w)\cdot f(x,w)\right)^\intercal\cdot\xi + \left(\nabla_w f(x,w)\right)^\intercal\cdot\left(-\nabla_x l(x,w) - \nabla_x f(x,w)\cdot\xi\right) \\
&\approx \left(\nabla^2_{wx} f(\widetilde{x},w)\cdot\left(f(\widetilde{x},w) - \nabla_x f(\widetilde{x},w)\cdot e_x\right)\right)^\intercal\cdot(\widetilde{\xi} - e_\xi) \\
&\quad - \left(\nabla_w f(\widetilde{x},w) - \nabla^2_{wx} f(\widetilde{x},w)\cdot e_x\right)^\intercal\cdot\left(\nabla_x l(\widetilde{x},w) - \nabla^2_{xx} l(\widetilde{x},w)\cdot e_x\right) \\
&\quad - \left(\nabla_w f(\widetilde{x},w) - \nabla^2_{wx} f(\widetilde{x},w)\cdot e_x\right)^\intercal\cdot\left(\nabla_x f(\widetilde{x},w) - \nabla^2_{xx} f(\widetilde{x},w)\cdot e_x\right)\cdot(\widetilde{\xi} - e_\xi).
\end{aligned}$$

With this approximation of the exact derivative, we can calculate a residual for the derivative approximation of the interpolant $\widetilde{\phi}$. We approximate the pointwise error of $\widetilde{\phi}$ by integrating this residual backward in time starting at $e_\phi(t_f) = 0$.

In order to calculate an L^∞ error estimator for the approximate gradient density, we interpolate the approximate trajectory of $e_\phi := \widetilde{\phi} - \phi$ with a quartic polynomial on each integration step interval, solve for the roots of the derivative of each such polynomial, filter out the roots that lie outside of the interpolation interval, and evaluate the polynomial at each derivative root as well as at the interpolation interval bounds. We then collect the resulting values for all interpolation intervals and select the maximal absolute value of that collection. This is guaranteed to be the maximal absolute value of our interpolant of the approximated pointwise error, which is approximately equal to the pointwise error of the gradient density function because the sign inversion does not impact the absolute value of the error. We apply a safety factor of 2 to compensate for the numerous inaccuracies in this process.

4.1.2 Implementation Notes

Rather than fixing a control mesh and refining mesh cells, we encode similarity classes as strictly increasing lists of switching times. At each switching time, numerical integrators are interrupted and restarted to avoid numerical problems that could arise from the discontinuity in the right-hand side of the ODE system.

All numerical integration is performed using the Dormand-Prince IVP solver provided by the SCIPY module `scipy.integrate`. SCIPY provides dense interpolants through the `OdeSolution` interface. Because we interrupt integration at each switching time, we obtain multiple such interpolants. We collect these into larger "trajectories" that consist of one `OdeSolution` per switching interval.

By default, SCIPY does not provide facilities to calculate derivatives or integrals of its interpolants. Rootfinding is also not supported by default. To work around this problem, we inject additional code into the relevant SCIPY classes that adds this functionality. This is implemented in the `lotka_volterra.ext.scipy` module of the Lotka-Volterra fishing code inside of the PYCOIMSET codebase. The practice of injecting patches into code at runtime is generally ill-advised. This code is expected to break if a future release of SCIPY modifies the internal data layout. For this reason, the code is developed for a specific version of SCIPY. We choose this method because it is less time-intensive than attempting to submit the code for proper inclusion in SCIPY.

We implement gradient measures using the aforementioned trajectory collections. Projected descent is calculated by evaluating an exact integral of the

piecewise approximation polynomials. Sublevel sets are found by performing rootfinding on the relevant piecewise polynomial. We find polynomial roots by finding the eigenvalues of the companion matrix, which is the rootfinding method used by the underlying NUMPY [Har+20] library. One of the advantages of this approach is that it works for all polynomial degrees and that NUMPY can vectorize this procedure, which means that roots of many polynomials can be found without the need for loops in Python. This can result in faster execution times than "smarter" methods.

At the end of the step-finding process, we have to pick an arbitrary subset of given measure from a similarity class. We always pick the earliest possible time points and the largest allowed subset.

We regulate error using absolute and relative tolerances that are imposed on the local truncation error by SCIPY's integrator. Relative tolerance is always set to the lowest allowed value, which is 100 times machine precision. Absolute tolerance is initialized to the minimum of 10^{-6}, $\frac{\beta_f}{t_f - t_0}$, and $\frac{\beta_g}{2(t_f - t_0)}$, where β_f and β_g are the error bounds for objective and gradient, respectively, that are given by the optimization loop. If the error estimate exceeds the requested bound, absolute tolerance is halved as many times as is necessary to satisfy the requested bounds. The local truncation error tolerance is shared between all integrator invocations.

4.1.3 Experiment

We solve the Lotka-Volterra fishing problem using the unconstrained optimization loop (see Algorithm 4 on page 232) with the following parameters:

$$\begin{aligned} \varepsilon &:= 0.002, & \sigma_0 &:= 0.2, & \sigma_1 &:= 0.4, \\ \sigma_2 &:= 0.6, & \xi_\delta &:= 0.01, & \xi_g &:= 0.1, \\ \xi_\tau &:= 0.1, & U_0 &:= \emptyset, & \Delta_0 &:= 12. \end{aligned}$$

Our implementation uses a single main execution thread on an Intel Core i5-10210U laptop CPU with 8 GB of total random-access memory. We note that the underlying libraries, particularly NUMPY and SCIPY, may use multiple threads. The test system is a desktop system and therefore should not be relied upon for accurate benchmarking.

The solver terminates after 202 iterations with an execution time of 350.46 CPU seconds, a final objective of 1.344426, and an instationarity of $1.887439 \cdot 10^{-3}$. Assuming convexity, instationarity is an upper bound on the optimality gap, meaning that the real objective should be no greater than 1.342538, which aligns with [Sag+06], where the authors give an optimal relaxed solution value of 1.34408 or 1.34466, depending on the solution method chosen. Even though [Sag+06] no longer reflects the state of the art in terms of approximate solution methods for binary optimal control, we note that the authors obtain binary solutions with objective function values no better than 1.34541. A rounded solution with an objective function value of 1.34416 is noted in [Sag08].

Table 4.1 on the next page is a shortened solver log from the run. It shows the log lines for every tenth iteration as well as for the final solution. The column "Step Size" shows the step size for the immediately preceding step, while "Rejected" shows the total number of rejections since the last iteration shown in the table. For a more accurate account of the progression of step size over time, we refer to Figure 4.2 on the following page, which juxtaposes instationarity

Table 4.1: Abbreviated solver log for the Lotka-Volterra Fishing Problem. Step sizes represent single iteration, number of rejected steps is accumulated over multiple iterations.

#	Objective	Instationarity	Step Size	Rejected	CPU Time
0	6.062259	7.917×10^{0}	0.000×10^{0}	0	0:00:00.15
10	1.622398	5.278×10^{-1}	1.875×10^{-1}	9	0:00:03.23
20	1.378280	5.525×10^{-2}	9.375×10^{-2}	9	0:00:07.90
30	1.364264	5.272×10^{-2}	1.172×10^{-2}	10	0:00:13.93
40	1.357410	2.951×10^{-2}	2.344×10^{-2}	7	0:00:22.76
50	1.351796	1.575×10^{-2}	2.930×10^{-3}	9	0:00:35.63
60	1.350527	1.403×10^{-2}	1.465×10^{-3}	5	0:00:49.09
70	1.349387	1.104×10^{-2}	1.172×10^{-2}	3	0:01:04.60
80	1.347689	7.550×10^{-3}	2.344×10^{-2}	6	0:01:21.02
90	1.347212	1.971×10^{-2}	1.172×10^{-2}	6	0:01:37.39
100	1.346539	5.321×10^{-3}	1.172×10^{-2}	3	0:01:54.83
110	1.345980	7.538×10^{-3}	2.930×10^{-3}	9	0:02:17.17
120	1.345557	6.701×10^{-3}	2.930×10^{-3}	5	0:02:34.95
130	1.345238	4.920×10^{-3}	2.930×10^{-3}	5	0:02:53.84
140	1.345071	4.113×10^{-3}	5.859×10^{-3}	2	0:03:11.85
150	1.344843	5.242×10^{-3}	2.930×10^{-3}	5	0:03:31.61
160	1.344740	4.946×10^{-3}	2.930×10^{-3}	2	0:03:50.69
170	1.344637	4.534×10^{-3}	2.930×10^{-3}	3	0:04:14.41
180	1.344566	4.977×10^{-3}	1.465×10^{-3}	7	0:04:47.03
190	1.344515	3.029×10^{-3}	1.465×10^{-3}	2	0:05:12.86
200	1.344432	2.948×10^{-3}	1.465×10^{-3}	7	0:05:45.23
202	1.344426	1.887×10^{-3}	1.465×10^{-3}	0	0:05:50.46

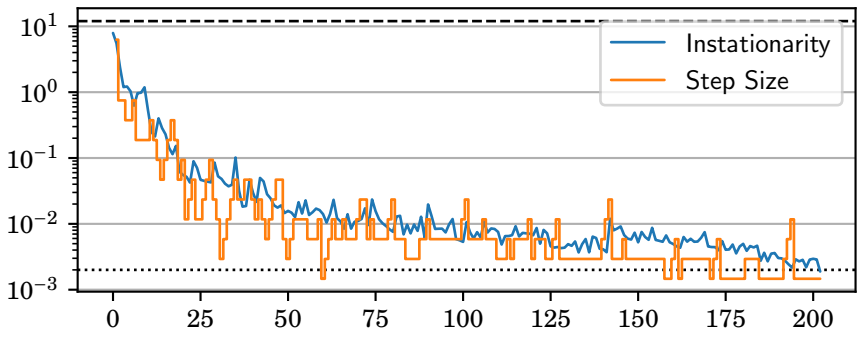


Figure 4.2: Progression of step size and instationarity over time for the Lotka-Volterra Fishing problem. Dashed line indicates upper bound for step size. Dotted line indicates instationarity tolerance.

and step size over time. The vertical axis of the plot is logarithmic to make it easier to recognize step rejections. Because every step rejection causes the trust region radius to be halved, as long as steps are approximately as large as the trust region will allow, a step rejection appears as a vertical jump by a constant amount. Multiple rejections appear as a vertical jump by an integer multiple of that amount.

Figure 4.2 indicates that, for most of the time, step size and instationarity track very closely. This is not entirely unexpected. For an objective function with constant curvature, the difference between the actual function value and the projected function value calculated using the first-order approximation is a fixed quadratic function. Projected descent is roughly proportional to the product of instationarity and step size. For a function in $\mathbb{R}^n$, we would roughly have

$$\begin{aligned}\frac{f(y)-f(x)}{\nabla f(x)(y-x)} &= 1+\frac{1}{2}\cdot\frac{(y-x)^\intercal\nabla^2 f(x)(y-x)}{\nabla f(x)(y-x)}\\ &\geq 1-\frac{1}{2}\cdot\frac{\|\nabla^2 f(x)\|\cdot\|y-x\|^2}{\|\nabla f(x)\|\cdot\|y-x\|}\\ &= 1-\frac{1}{2}\cdot\frac{\|\nabla^2 f(x)\|}{\|\nabla f(x)\|}\cdot\|y-x\|.\end{aligned}$$

If we presume that the curvature in the step direction is always roughly the same and $\|\nabla f(x)\|\to 0$, then it follows that $\frac{\|y-x\|}{\|\nabla f(x)\|}$ must roughly stay the same to ensure that the step quality stays above the acceptance threshold. Therefore, we would expect the step size to be roughly proportional to instationarity. This is somewhat more difficult to formalize for set functionals, because we do not have a good understanding of how second derivatives work, but it appears that a similar principle is at work here.

Figure 4.3 on pages Pages 350 and 351 shows solution trajectories and gradient densities for every fiftieth iterate as well as the final solution. Because this aligns with our excerpt from the solver log, objective function values and instationarities for these iterates can be found in Table 4.1. The shaded background area in the plots indicates the times for which the control is active. We note that, although it may appear as if the gradient density function assumes multiple values at once at some points, this is merely a plotting artifact that is caused by very fast switching.

We observe that the solver generates a very high number of switches. This is not a failure of the algorithm because the Lotka-Volterra fishing problem is designed to cause this type of behavior in nearly optimal solutions. Figure 4.4 on page 352 displays the number of distinct control intervals (both “on” and “off”) prescribed by the control function $w=\chi_U$ in each iteration. We sort these intervals into “bins” based on the logarithm of their size to get a more accurate idea of the fragmentation of the control set. The graph suggests that, up to around iteration 25, newly generated intervals are predominantly between 10^{-2} and 10^{0} in size. Afterwards, they appear to be mostly between 10^{-4} and 10^{-2}. This aligns roughly with the time at which step size drops below 10^{-2} according to Figure 4.2. This suggests that the algorithm tends to generate switching intervals of its current step size. From this, we could infer that the algorithm prefers creating isolated switching intervals over modifying existing ones. The final number of switching intervals is slightly above 200, which aligns with

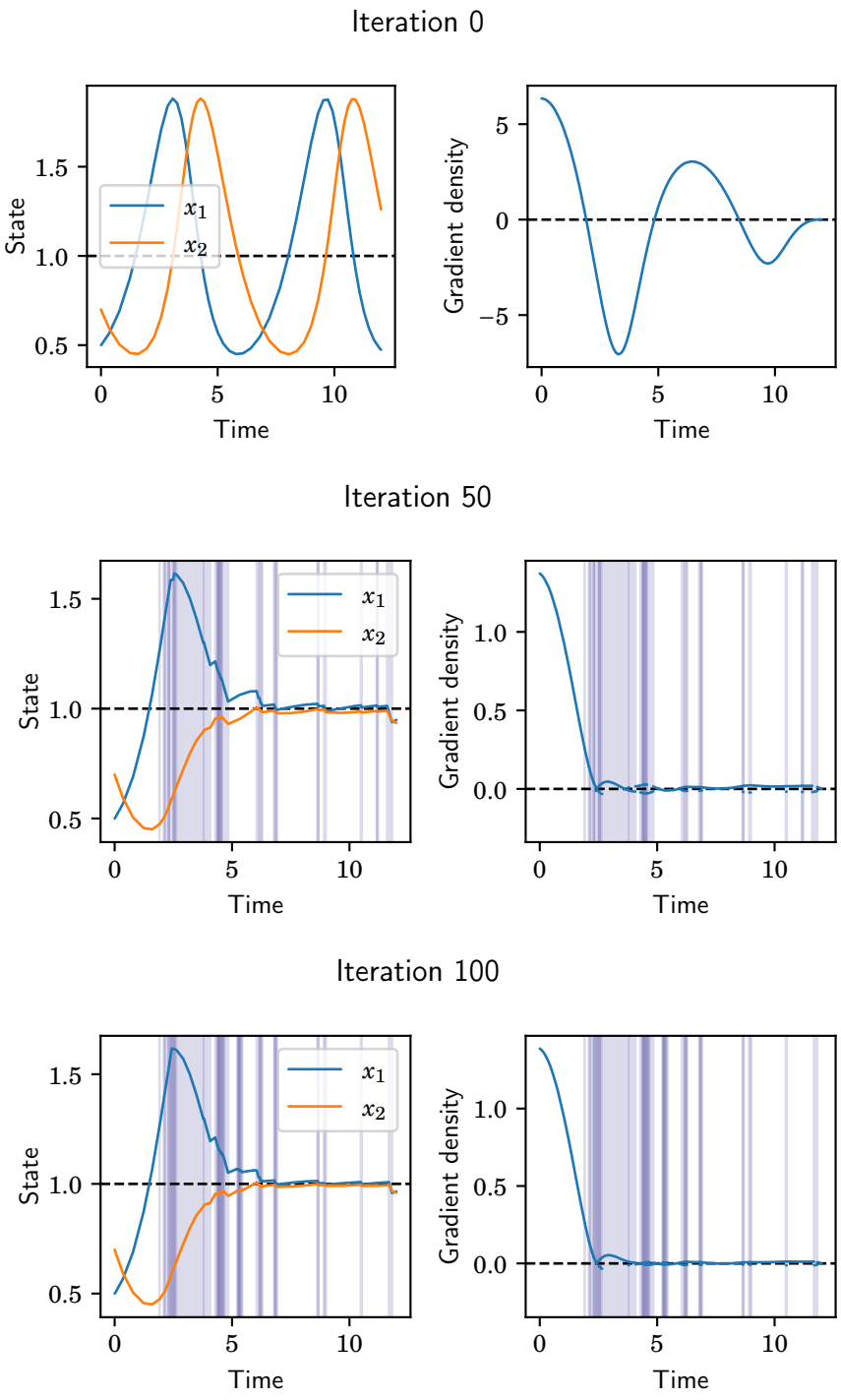


Figure 4.3: Trajectory and gradient density plots at selected iterations.

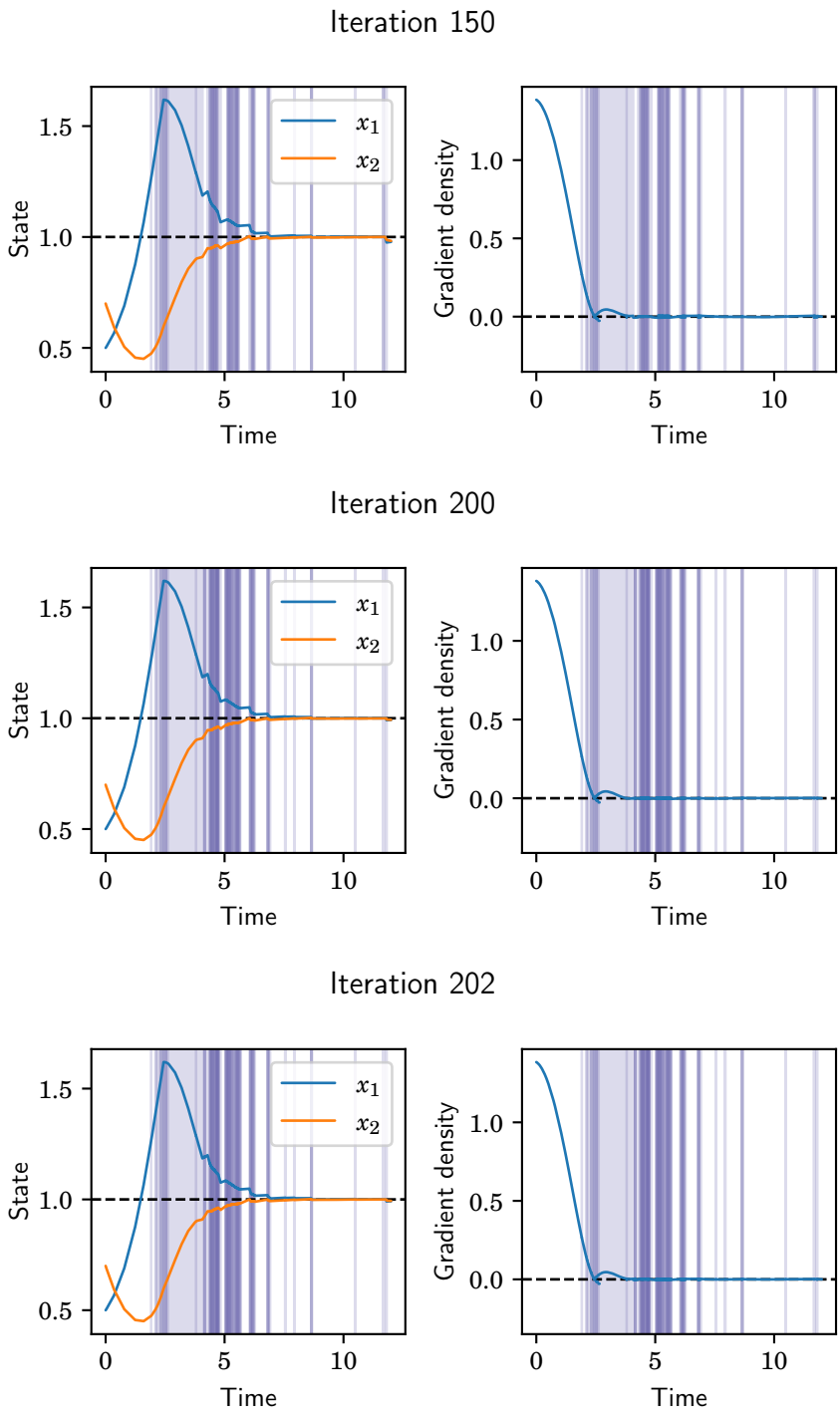


Figure 4.3: Trajectory and gradient density plots at selected iterations (cont.).

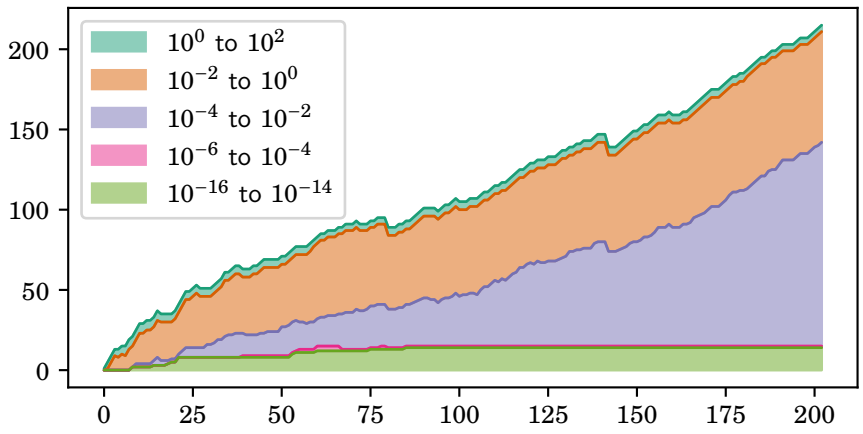


Figure 4.4: Distribution of contiguous interval sizes (both on and off) over all solver iterations as a stacked area graph. Length bins are logarithmically sized. Empty bins have been removed.

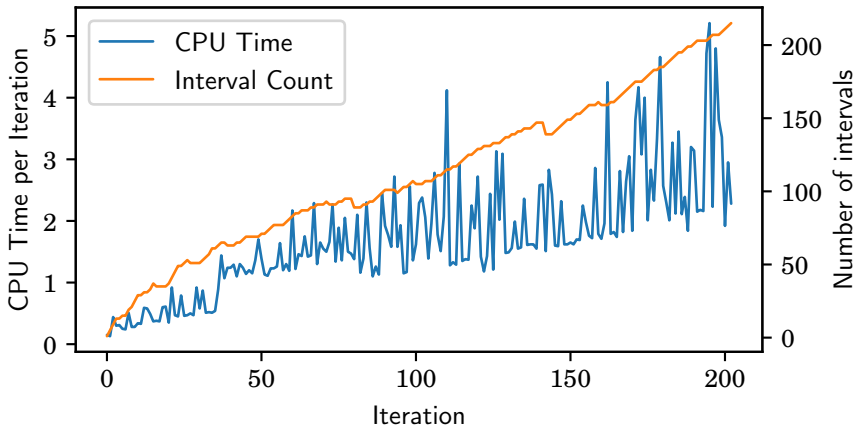


Figure 4.5: CPU time per iteration and number of switching intervals for the Lotka-Volterra fishing problem.

the number of iterations. On average, every iteration generates roughly one additional switching interval.

This behavior is not necessarily ideal because large numbers of switches can be detrimental to evaluation performance. Figure 4.5 suggests that there is an upward trend in per-iteration runtimes that appears to correlate with the number of switching intervals, though it is not clear whether the increase in CPU time per iteration stems from the increasing number of control intervals or from the decreasing error tolerances. Runtime measurements can also be affected by a multitude of environmental factors, which makes runtime measurements a highly unreliable performance metric unless the algorithm is executed in a highly controlled environment.

In [HLS22], the authors (including the author of this thesis) suggest a scaled measure for the Lotka-Volterra problem. This mitigates the fact that gradient densities naturally become larger in magnitude with increasing distance from the end of the time horizon. We can observe this in Figure 4.3. We forego this option because it further complicates our error control scheme, which is already very complex as it is. However, an error controlled problem implementation with a scaling measure could outperform the problem implementation we present here.

4.2 POISSON DESIGN PROBLEM

The second test problem is a classic topological design problem which dates back at least as far as [GBS06], from where we draw most of our knowledge about the problem. We originally adopt this problem from the example problems of the DOLFIN-ADJOINT project [Far+13] [3]. Our numerical solution approach is also based on that example.

The problem itself as described in [GBS06] is based around the partial differential equation

$$-\operatorname{div}(k\nabla y) = f$$

where $y \in C^2(\Omega) \cap C^1(\overline{\Omega})$ is a twice continuously differentiable temperature function on a bounded domain $\Omega \subset \mathbb{R}^n$ that has a continuously differentiable extension to $\overline{\Omega}$, $k : \Omega \to \mathbb{R}$ is the heat conduction coefficient, and $f : \Omega \to \mathbb{R}$ is a volumetric heat source term.

The boundary of Ω is $\Gamma \coloneqq \partial\Omega$ and is essentially partitioned into the Dirichlet boundary Γ_D and the Neumann boundary Γ_N in the sense that $\Gamma_D \cap \Gamma_N = \emptyset$ and $\overline{\Gamma_D \cup \Gamma_N} = \Gamma$. On Γ_D, we have a Dirichlet boundary condition

$$y = 0 \quad \text{on } \Gamma_D.$$

The Dirichlet boundary simulates contact with a medium that has a fixed temperature. On the Neumann boundary, we have a Neumann boundary condition

$$\langle k\nabla y,\, \vec{n}\rangle = 0 \quad \text{on } \Gamma_N$$

where $\vec{n}$ is the outer unit normal of Ω. The homogeneous Neumann boundary condition simulates contact with an insulator because it does not allow heat to diffuse across the boundary. The objective is to minimize a compliance term

$$\int_\Omega f y \,\mathrm{d}\mu.$$

Intuitively, compliance quantifies the degree to which the "temperature" y changes with the application of a volumetric heat source f. The variable through which we "control" the system is the heat conduction coefficient k. We have two materials at our disposal: one has relatively high heat conduction ($k = 1$) and one has very low heat conduction ($k = \varepsilon \in (0,\ 1)$). The choice between both materials is binary, i.e., we cannot place arbitrary mixtures of both materials. As

[3] The project has recently been renamed PYADJOINT and appears to be under new maintenance. We cite an older publication because the example upon which we base our methodology was authored by Patrick E. Farrell, who is not a co-author of the newer papers. We do not use PYADJOINT itself in our code.

an additional constraint, we have a limited amount of the better conductor at our disposal, i.e.,

$$\lambda(\{k = 1\}) \leq M$$

for some constant $M > 0$.

We now want to rewrite this problem as a set-valued optimization problem. We work within the similarity space associated with $(X,\ \Sigma,\ \mu) \coloneqq (\Omega,\ \mathscr{B}(\Omega),\ \lambda)$ where $\mathscr{B}(\Omega)$ is the σ-algebra of Borel measurable subsets of Ω and λ is the Lebesgue measure in $\mathbb{R}^2$. Let $\Sigma_{\sim} \coloneqq \Sigma/_{\sim_\mu}$. We define

$$k(w) \coloneqq \varepsilon + (1 - \varepsilon) \cdot w \quad \forall w \in L^\infty(\Omega) \colon w > 0 \text{ a.e.}$$

where the input function w is the indicator function of a set variable U. As the authors of [GBS06] point out, it is common practice in topology optimization to attach an arbitrary exponent $p > 1$ to the optimization variable. This type of "penalization" is not valid for our method because it violates Assumption 2.4.25 (6) due to the fact that the derivative of $k(w)$ with respect to the value of w would become zero in points outside of U. The specific problem instance under discussion here has the following parameters:

$$\begin{aligned}
\Omega &\coloneqq (0,\ 1)^2, & \Gamma_D &\coloneqq \big(\{0\} \times (0,\ 1]\big) \cup \big([0,\ 1) \times \{1\}\big),\\
\Gamma_N &\coloneqq \big((0,\ 1] \times \{0\}\big) \cup \big(\{1\} \times [0,\ 1)\big), & \varepsilon &\coloneqq 0.1,\\
M &\coloneqq 0.4.
\end{aligned}$$

Figure 4.6 on the next page illustrates the layout of the problem domain. Intuitively, we expect the algorithm to lay out its limited amount of the better heat conductor in a pattern that best conducts heat away from the Neumann boundary and towards the Dirichlet boundary. Figure 4.7 on the facing page shows an example configuration returned by the penalty solver. We note that ε is relatively high compared to the value suggested in [GBS06]. This is to mitigate numerical challenges during evaluation.

4.2.1 Theoretical Discussion

Because our heat conduction coefficient k is discontinuous, we have no reasonable expectation of there being a twice differentiable temperature function y that satisfies the PDE. However, we can solve the boundary value problem in a weak sense.

4.2.1.1 WEAK FORMULATION AND OPTIMIZATION PROBLEM

An in-depth derivation of the variational formulation of the underlying boundary value problem is beyond the scope of this thesis. We refer to [KA21, Sec. 3.2] and particularly Theorem 3.16 from that work for an excellent discussion of the underlying theory.

The variational formulation is obtained by applying the divergence theorem to the continuously differentiable vector field $k\nabla y \cdot v$ where v is an infinitely smooth test function that is equal to zero on Γ_D. This yields

$$\int_\Omega \operatorname{div}(k\nabla y \cdot v)\,\mathrm{d}\mu = \int_{\partial\Omega} \langle k\nabla y \cdot v,\ \vec{n}\rangle\,\mathrm{d}\sigma$$

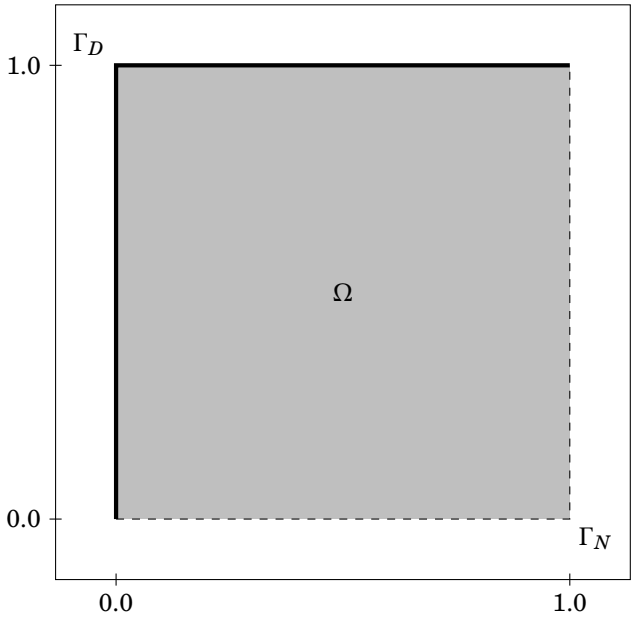


Figure 4.6: Illustration of the problem domain Ω.

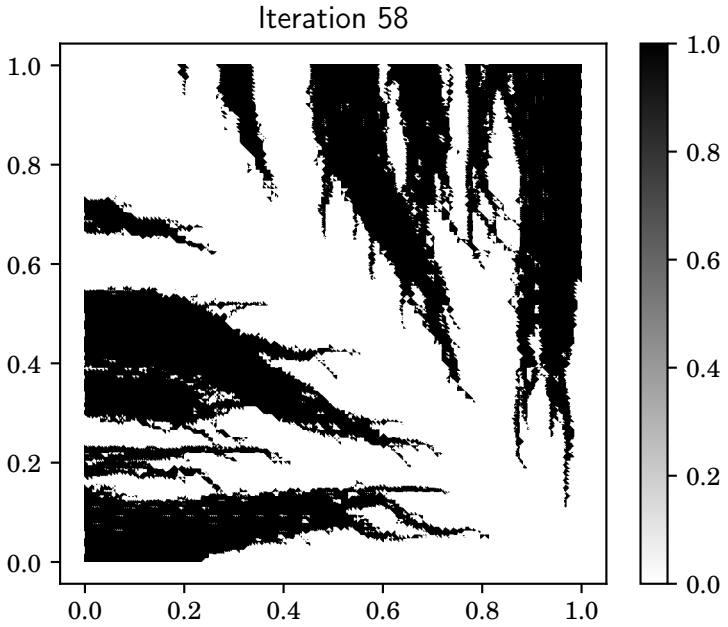


Figure 4.7: Solution obtained by the Poisson design problem obtained by the Algorithm 8 with the approximate steepest descent step. Black color indicates placement of better conductor.

where σ is the standard surface measure on $\partial\Omega$. We can then use the fact that

$$\begin{aligned}\operatorname{div}(k\nabla y \cdot v) &= \langle k\nabla y,\ \nabla v\rangle + \overbrace{\operatorname{div}(k\nabla y)}^{=-f \text{ a.e.}}\cdot v \\ &= \langle k\nabla y,\ \nabla v\rangle - fv\end{aligned}$$

almost everywhere in Ω, and that

$$\begin{aligned}\langle k\nabla y \cdot v,\ \vec{n}\rangle &= \langle k\nabla y,\ \vec{n}\rangle \cdot v \\ &= \begin{cases} 0\cdot v & \text{on } \Gamma_N, \\ \langle k\nabla y,\ \vec{n}\rangle \cdot 0 & \text{on } \Gamma_D \end{cases} \\ &= 0.\end{aligned}$$

By applying this to the integral equation, we obtain

$$\int_\Omega \langle k\nabla y,\ \nabla v\rangle \,\mathrm{d}\mu = \int_\Omega fv \,\mathrm{d}\mu. \tag{4.2}$$

In general, this equation does not have a twice continuously differentiable solution. We can extend the search space to the Sobolev space $H^1(\Omega) = W^{1,2}(\Omega)$. More precisely, to satisfy the Dirichlet boundary condition, we choose the search space

$$V \coloneqq \left\{ y \in H^1(\Omega) \,\middle|\, Ty|_{\Gamma_D} = 0 \right\}$$

where $T\colon H^1(\Omega) \to L^2(\partial\Omega)$ is the trace operator in the sense of [RR04, Sec. 7.2.5]. Because $y \mapsto Ty|_{\Gamma_D}$ is the composition of the trace operator and a restriction, both of which are continuous linear operators, V is the kernel of a continuous linear operator and therefore closed. As a closed linear subspace of the real Hilbert space $H^1(V)$, V is itself a real Hilbert space with the familiar $W^{1,2}(\Omega)$ inner product

$$\langle y,\ z\rangle_V \coloneqq \langle y,\ z\rangle_{L^2(\Omega)} + \langle \nabla y,\ \nabla z\rangle_{L^2(\Omega)} \quad \forall y, z \in V.$$

This is not a substantial search space extension. There exist results proving that $C^\infty(\overline{\Omega})$ is dense within $H^1(\Omega)$ if Ω is a bounded Lipschitz domain (see, e.g., [RR04, Lemma 7.49]). This proves, of course, that $C^\infty(\overline{\Omega})$ is also dense in V. We will not discuss the impact of a restriction to $y \in C^\infty(\overline{\Omega})$ with $y = 0$ on Γ_D here. However, this does not appear to be a major concern in most literature discussing the Poisson problem with Dirichlet boundary conditions on parts of the boundary.

Thus, our final optimization problem takes the form

$$\begin{aligned}\min_{U,y}\ & \int_\Omega fy \,\mathrm{d}\mu \\ \text{s.t.}\ & \int_\Omega \langle k(\chi_U)\nabla y,\ \nabla v\rangle \,\mathrm{d}\mu = \int_\Omega fv \,\mathrm{d}\mu \quad \forall v \in V, \\ & \lambda(U) \le M, \\ & y \in V, \\ & U \in \Sigma_\sim.\end{aligned} \tag{4.3}$$

Next, we discuss whether we can transform this into a reduced-space optimization problem.

4.2.1.2 EXISTENCE, UNIQUENESS, AND DIFFERENTIABILITY

To prove that the objective functional of Problem (4.3) is differentiable as a set functional, we invoke the argument from Section 2.4.4, which requires that we satisfy Assumptions 2.4.25 (1) to 2.4.25 (9). Again, we work without a scaling density function and the number of individual control functions is $n_w = 1$.

We choose $\delta \coloneqq \frac{\varepsilon}{2(1-\varepsilon)}$ as the radius for our differentiation environment. For $w \in \mathscr{W}_\delta^{n_w}(\Omega)$, we have $w \geq -\delta$ almost everywhere and therefore

$$k(w) = \varepsilon + (1-\varepsilon)\cdot w \geq \varepsilon - \frac{\varepsilon}{2} = \frac{\varepsilon}{2}$$

almost everywhere, which is sufficient to guarantee the existence and uniqueness of the weak solution of the boundary value problem. The solution space is $Y \coloneqq V$. The space Z is $\mathscr{L}(V,\, \mathbb{R})$. Both are Banach spaces.

To prove the existence and uniqueness of the solution, we can simply invoke [KA21, Thm. 3.16], which is applicable because $k \in L^\infty(\mathscr{B}(\Omega),\, \lambda)$ with $k \geq \frac{\varepsilon}{2} > 0$ almost everywhere, and because $\sigma(\Gamma_D) > 0$. However, to prove differentiability, we have to dive a little deeper into the solution theory of the boundary value problem. We start with the fact that [KA21, Cor. 3.15] states that there exists a constant $C_F > 0$ such that

$$\|y\|_{L^2(\Omega)} \leq C_F \cdot \|\nabla y\|_{L^2(\Omega)} \quad \forall y \in V.$$

The weak equation system has the form

$$a(y,v,w) = L(v) \quad \forall v \in V$$

where $a\colon V \times V \times \mathscr{W}_\delta^{n_w}(\Omega) \to \mathbb{R}$ with

$$a(y,v,w) \coloneqq \int_\Omega \langle k(w)\nabla y,\, \nabla v\rangle \,\mathrm{d}\mu \quad \forall (y,\, v,\, w) \in V \times V \times \mathscr{W}_\delta^{n_w}(\Omega)$$

is a bilinear form with respect to $(y,\, v)$ that satisfies

$$\begin{aligned}
\big|a(y,v,w)\big| &\leq k(1+\delta)\cdot |\langle \nabla y,\, \nabla v\rangle_{L^2}| \\
&\leq k(1+\delta)\cdot \|\nabla y\|_{L^2}\cdot \|\nabla v\|_{L^2} \\
&\leq k(1+\delta)\cdot \|y\|_V \cdot \|v\|_V,
\end{aligned}$$

$$\begin{aligned}
a(v,v,w) &= \int_\Omega \langle \underbrace{k(w)}_{\geq \frac{\varepsilon}{2}} \nabla v,\, \nabla v\rangle \,\mathrm{d}\mu \\
&\geq \frac{\varepsilon}{2}\cdot \|\nabla v\|_{L^2}^2 \\
&= \frac{\varepsilon}{2\cdot(1+C_F)^2}\cdot \big(\|\nabla v\|_{L^2} + C_F\cdot\|\nabla v\|_{L^2}\big)^2 \\
&\geq \frac{\varepsilon}{2\cdot(1+C_F)^2}\cdot \big(\|\nabla v\|_{L^2} + \|v\|_{L^2}\big)^2 \\
&\geq \frac{\varepsilon}{2\cdot(1+C_F)^2}\cdot \|v\|_V^2
\end{aligned}$$

for all $(y,\, v,\, w) \in V \times V \times \mathscr{W}_\delta^{n_w}(\Omega)$. Thus, a is bounded and V-coercive. The linear form L satisfies

$$\begin{aligned}
\big|L(v)\big| &= |\langle f,\, v\rangle_{L^2}| \\
&\leq \|f\|_{L^2}\cdot\|v\|_{L^2} \\
&\leq \|f\|_{L^2}\cdot\|v\|_V.
\end{aligned}$$

The Lax-Milgram theorem (see, e.g., [RR04, Thm. 9.14] or [KA21, Thm. 3.13]), then guarantees that for every $w \in W_\delta^{n_w}(\Omega)$, the equation

$$a(y,v,w) = L(v) \quad \forall v \in V$$

has exactly one solution $y_w \in V$ and that

$$\|y_w\|_V \leq \frac{2 \cdot (1+C_F)^2 \cdot \|f\|_{L^2}}{\varepsilon}.$$

We note that this upper bound on $\|y_w\|_V$ is independent of w, but can become arbitrarily large for $\varepsilon \to 0$, which is why we choose a relatively high value for ε to avoid numerical issues.

To align notation with Section 2.4.4, the map $f\colon V \times \mathcal{W}_\delta^{n_w}(\Omega) \to \mathcal{L}(V, \mathbb{R})$ is given by

$$f(y,w) \coloneqq \big(v \mapsto a(y,v,w) - L(v)\big) \quad \forall y \in V,\ w \in \mathcal{W}_\delta^{n_w}(\Omega).$$

Because a is a bounded, V-coercive linear form in $(y,\ v)$, f is F-differentiable with respect to y and the derivative has a bounded inverse. The derivative of a with respect to w has the form

$$a_w(y,v,w)d_w = \int_\Omega \langle (1-\varepsilon) d_w \nabla y,\ \nabla v\rangle \,\mathrm{d}\mu \quad \forall (y,\ v,\ w) \in V \times V \times \mathcal{W}_\delta^{n_w}(\Omega)$$

and therefore satisfies

$$\begin{aligned}
\big|a_w(y,v,w)d_w\big| &= (1-\varepsilon) \cdot \left| \int_\Omega d_w \cdot \langle \nabla y,\ \nabla v\rangle \,\mathrm{d}\mu \right| \\
&\leq (1-\varepsilon) \cdot \|\nabla y\|_{L^2} \cdot \|\nabla v\|_{L^2} \cdot \|d_w\|_{L^\infty} \\
&\leq (1-\varepsilon) \cdot \|y\|_V \cdot \|v\|_V \cdot \|d_w\|_{L^\infty}.
\end{aligned}$$

The objective j is simply given by

$$j(y,w) \coloneqq \langle f,\ y\rangle_{L^2} = L(y).$$

It is independent of w and a bounded linear form in y. It is therefore F-differentiable and its own derivative in both arguments.

With this, we have proven Assumptions 2.4.25 (1) to 2.4.25 (7). Assumption 2.4.25 (8) presents some problems.

Remark 4.2.1 (Notes on Assumption 2.4.25 (8)).
In order to apply our set-valued optimization methods to Problem (4.2), we have to show that the derivative of $w \mapsto y_w$, which exists only as a Fréchet derivative from $\mathcal{W}_\delta^{n_w}(\Omega) \subset L^\infty(\Omega)$ to V, behaves like a derivative from $L^1(\Omega)$ to V for steps between indicator functions.

According to the implicit function theorem, the Fréchet derivative of $w \mapsto y_w$ has the form

$$D_w y_w = -\big(D_y f(y,w)\big)^{-1} \circ D_w f(y,w).$$

Reading the application of $\big(D_y f(y,w)\big)^{-1}$ as solving an equation system, the linearized change in solution for a perturbation d_w in control space is the solution of the equation system

$$D_y f(y,w) d_y = -D_w f(y,w) d_w,$$

or equivalently, the image $d_y \in V$ of $D_w(w \mapsto y_w)(w)d_w$ is the unique solution of

$$\int_\Omega \langle k(w)\nabla d_y,\, \nabla v\rangle \,\mathrm{d}\mu = \int_\Omega \langle (1-\varepsilon)d_w \nabla y_w,\, \nabla v\rangle \,\mathrm{d}\mu \quad \forall v \in V.$$

If we have no additional information except that ∇y_w and ∇v are in $L^2(\Omega, \mathbb{R}^n)$, then the right hand side of this equation system would only be bounded with respect to the $L^\infty(\Omega)$ norm of d_w, which would not be sufficient for our purposes because the the L^∞ norm does not gradually decrease to zero as the measure of the step set decreases.

As a first step, we could exploit properties of the solution y_w. If the solution could be shown to be twice weakly differentiable, i.e., if it were in $H^2(\Omega)$ rather than $H^1(\Omega)$ (see, e.g., [Bre11, Cor. 9.13] and [RR04, Sec. 7.4.3]), then it would be equal almost everywhere to a Lipschitz continuous function. It would therefore be differentiable almost everywhere and its gradient would satisfy $\|\nabla y_w\|_2 \leq L < \infty$ almost everywhere for a suitably chosen Lipschitz constant L. This would mean that $\|\nabla y_w\|_2 \in L^\infty(\Omega)$, which would make the right hand side of the equation bounded with respect to the L^2 norm of d_w.

Such regularity results are generally difficult to prove and depend on both the properties of the coefficient function $k(w)$, the shape of the domain boundary, and the precise boundary conditions imposed thereon. There are numerous such regularity results, most of which rely on smooth parameters or a smooth boundary. For instance, [RR04, Sec. 9.5 and 9.6] permits coefficients in $W^{1,\infty}(\Omega)$, but assumes a C^2 boundary and only deals with Dirichlet boundary conditions, while [Gri85, Ch. 3] deals with more general convex and polygonal domains, but only permits smooth coefficients and only deals with the Dirichlet problem. There exist results for pure Dirichlet and pure Neumann boundary conditions. However, there are few results for mixed boundary conditions and almost none for mixed conditions on general convex domains. Notably, [MPS00] deals with general L^∞ coefficients, but presumes a C^2 boundary. We take this as an indication that general L^∞ coefficients are not a fundamental barrier to H^2 regularity results.

The second and arguably more severe problem is getting from an estimate with respect to the L^2 norm to an estimate with respect to the L^1 norm of d_w. This would technically require a restriction of the test function space from which the test function v is drawn. Whether and under which circumstances this might be valid is significantly beyond the scope of this work.

For the remainder of this section, we will make a technical assumption that the unique solution d_y of

$$\int_\Omega \langle k(w)\nabla d_y,\, \nabla v\rangle \,\mathrm{d}\mu = \int_\Omega \langle (1-\varepsilon)d_w \nabla y_w,\, \nabla v\rangle \,\mathrm{d}\mu \quad \forall v \in V.$$

satisfies $\|d_y\|_{L^1(\Omega)} \leq C \cdot \|d_w\|_{L^1}$ for some constant $C > 0$. For strong solutions of the Poisson equation, estimates with respect to the L^1 norm of the right hand side are discussed, for instance, in [RS19], though whether such results could be transfered to this case is questionable.

It is important to understand the manner in which an L^2 estimate could "break" our algorithms. Because all of our control functions are binary-valued, differences between them satisfy

$$|d_w|^p = |d_w|$$

almost everywhere for $0 < p < \infty$. Therefore, we have

$$\|d_w\|_{L^p} = \left(\int_\Omega |d_w|^p \,\mathrm{d}\mu\right)^{\frac{1}{p}} = \left(\int_\Omega |d_w| \,\mathrm{d}\mu\right)^{\frac{1}{p}} = \|d_w\|_{L^1}^{\frac{1}{p}}$$

for all $1 \leq p < \infty$. The problem here lies in the fact that for $p > 1$, the mapping $x \mapsto x^{\frac{1}{p}}$ becomes infinitely steep for x near zero. This means that, for infinitesimally small steps, the L^p norm is not equivalent to the L^1 norm. If, however, there is a fixed lower bound on step sizes, then the slope of the L^p norm can be bounded and the L^1 and L^p norms become functionally equivalent as long as steps remain above that fixed threshold.

Because the step acceptance criterion examines the actual descent based on point evaluations of the objective, accepted steps would still be true descent steps. We would, however, lose the guarantee that there can never be an infinite sequence of step rejections. We can prepare for this by setting a minimal trust region radius and aborting the main loop if the trust region radius falls below that threshold.

It is not entirely clear how this affects the ε-stationarity criterion. However, we would expect the linearized model function to still be an accurate model for all but the smallest of steps.

We note that this is not entirely beneficial. It implies that we may be able to continue using our algorithms as local improvement heuristics for problems where the derivative is not bounded with respect to the L^1 norm. However, it also means that we may not be able to detect mistakes in our theoretical line of argumentation by observing whether the optimization algorithm works as expected. It may still operate as expected, even if the set functional in question is technically not differentiable as a set functional. ◁

Under the given assumptions, the set differentiability of the objective functional $J\colon \Sigma/_{\sim_\mu} \to \mathbb{R}$ follows from Proposition 2.4.26. The derivative of the objective functional has the form

$$\begin{aligned}\nabla J(U)(D) &= D_y j(y_{\chi_U}, \chi_U)\big(D_y f(y_{\chi_U}, \chi_U)\big)^{-1} D_w f(y_{\chi_U}, \chi_U)(\chi_{U \triangle D} - \chi_U) \\ &= D_y j(y_{\chi_U}, \chi_U)\big(D_y f(y_{\chi_U}, \chi_U)\big)^{-1} D_w f(y_{\chi_U}, \chi_U)(\chi_{D \setminus U} - \chi_{D \cap U})\end{aligned}$$

Instead of solving a variational equation for each possible perturbation, such derivatives are generally evaluated using the adjoint method. The adjoint method makes use of the fact that there exists a test function $z \in V^*$ (where V^* is the dual space of V) such that

$$D_y f(y_{\chi_U}, \chi_U)(d_y)(z) = a(d_y, z, \chi_U) = L(z) = D_y j(y_{\chi_U}, \chi_U)(z) \quad \forall d_y \in V.$$

This test function is generally known as the "adjoint" solution because it solves a variational equation involving the adjoint of the bilinear form a. In our case, because the objective j is exactly the same as the linear form on the right hand side of the variational equation (4.2), and because a is symmetric, the adjoint problem is exactly the same as the original variational problem and the adjoint solution $z \in V^*$ is the element of the dual space that corresponds to the primal solution $y_{\chi_U} \in V$. We can make use of this to substantially reduce solution times.

By applying the adjoint of $-D_w f(y_{\chi_U}, \chi_U)$ to z, we obtain a linear form that maps a control space perturbation d_w to its corresponding linearized effect on the objective function.

The measure constraint has the form

$$G(U) \leq 0$$

with $G : \Sigma/_{\sim_\mu} \to \mathbb{R}$ given by $G(U) \coloneqq \lambda(U) - M$. The set derivative of G has the form

$$\nabla G(U)(D) = \lambda(D \setminus U) - \lambda(D \cap U) = G(U \triangle D) - G(U) \quad \forall U, D \in \Sigma/_{\sim_\mu}.$$

There is no truncation error and the functional is plainly Lipschitz-continuously differentiable with a Lipschitz constant of zero. This constraint is a prototypical example of a constraint that behaves like a linear constraint in conventional optimization. The density function of the derivative $\nabla G(U)$ is a difference of indicator functions on disjoint sets and is therefore in L^∞, making G a benignly and Lipschitz-continuously differentiable set functional.

The benign differentiability of the objective functional J can be proven using the Lebesgue differentiation theorem (see, e.g., [BC09, Ch. 8]), where Assumption 2.4.25 (8) ensures that the ratio between the gradient measure and the underlying measure μ is uniformly bounded over all sets.

We omit discussions of continuity of the derivative here. Normally, these would follow from the implicit function theorem. However, the implicit function theorem only yields continuity guarantees in $\mathscr{L}(L^\infty, \mathbb{R})$, not in $\mathscr{L}(L^1, \mathbb{R})$. We make a technical assumption of appropriate continuity in the sense of Remark 4.2.1.

In order to approximate the derivative, we have to discretize the problem. We partition $\overline{\Omega}$ into successively refined triangle meshes. If these meshes satisfy Assumptions 2.4.25 (10) to 2.4.25 (14), then the value of the density function of $\nabla J(U)$ can be approximated by the mean density of $\nabla J(U)$ over the surrounding mesh cell for almost all points in Ω. We need not consider Assumption 2.4.25 (9) because we do not apply any scaling.

4.2.1.3 CONSTRAINT QUALIFICATION AND KKT CONDITIONS

In order to satisfy Assumption 3.2.26, the problem must satisfy a constraint qualification in all feasible points. Indeed, it is fairly trivial to show that the measure constraint satisfies the Mangasarian-Fromovitz Constraint Qualification in the sense of Definition 3.2.15.

Let $U \in \Sigma/_{\sim_\lambda}$ be such that $\lambda(U) \leq M$. If $\lambda(U) = 0$, then there are no active constraints, because $M > 0$. Therefore, we can choose $N \coloneqq [\Omega]_{\sim_\lambda}$ to prove that the problem satisfies an MFCQ in U. If $\lambda(U) > 0$, then we choose $N \coloneqq U$.

Evidently, this choice satisfies $\lambda(N) > 0$. For $\lambda(U) < M$, the single constraint is inactive. Therefore, there is nothing further to prove. For $\lambda(U) = M$, every $D \subseteq_\lambda N$ satisfies

$$\nabla G(U)(D) = \lambda(\underbrace{D \setminus U}_{=\emptyset}) - \lambda(\underbrace{D \cap U}_{=D}) = -\lambda(D).$$

This proves that the problem also satisfies an MFCQ in points where $\lambda(U) \leq M$ is active.

Because there is only one constraint, there is at most one active constraint. We can therefore invoke Theorem 3.2.21 to show that the problem satisfies a GCQ in all feasible points.

4.2.1.4 NUMERICAL METHODS AND ERROR CONTROL

We numerically approximate solutions of Problem (4.2) using finite element methods (FEM). For a general introduction to FEM, we refer to [Bra13; LB13]. For details on the practical implementation of FEM, we refer to [LMW12]. In particular, we want to point out the small catalogue of commonly used finite elements provided in [Kir+12], in which all elements that we will use are explained.

The basic idea behind FEM is to approximate the solution space V with a finite-dimensional approximation. To do this, the domain is partitioned into a mesh of bounded, relatively simple "cells." On each cell, the function is approximated with a relatively simple function drawn from a finite-dimensional local function space. Elements of the local function space are described by their image under a basis of the local function space's dual space that is commonly referred to as the "degrees of freedom" of the local function (see [KL12]).

This yields a finite-dimensional subspace of the actual search space to which both trial and test functions can be restricted to obtain a finite-dimensional equation system that can be solved using conventional equation solvers.

In FEM, there exist various approaches to error estimation and adaptivity. A priori estimates are usually very inaccurate. They are mostly used for general convergence proofs and can be found in most textbooks on the subject (see, e.g., [Bra13, Ch. 7]). There exist some a posteriori norm-based error estimates based on residuals of the variational equation (see, e.g., [BR01, Sec. 4]). However, we will focus primarily on the so-called "dual-weighted residual" (DWR) method for objective error estimates. More precisely, we adhere closely to the methodology descibed in [BR01, Sec. 3 and 5]. The DWR method is a so-called "goal-oriented" error estimation and control method because it controls the error of a linear "goal functional." In our case, the goal functional is the objective functional, which is linear. For nonlinear objective functionals, the goal functional could be the linearization of the objective.

The DWR method derives error estimates from the residual of the variational equation, but weighs them according to their impact on the goal functional. Because the residual is accumulated from residuals on individual mesh cells and their interfaces with one another, the estimated error can be attributed to individual mesh cells, thus giving an indication of which cells most need to be refined in order to lower the error. This allows for targeted local refinement.

We rely mostly on Remark 3.5 in [BR01] as both justification for the application of DWR and as a source for residual terms. The error estimator put forward there is

$$\left|J(U) - \widetilde{J}(U)\right| \leq \eta_\omega(y_h) \coloneqq \sum_{K \in T} \rho_K \omega_K$$

where y_h is the discretized FEM solution of the variational equation system, T is the set of all mesh cells, ρ_K is a cell-wise term that derives from the residual of the actual variational equation on the cell K, and ω_K is a weight term that derives from the error of the dual solution.

According to [BR01, Prop. 3.1], the residual and weight terms have the form

$$\begin{aligned} \rho_K &\coloneqq \|R(y_h)\|_{L^2(K)} + h_K^{-1/2} \cdot \|r(y_h)\|_{\partial K} \quad \forall K \in T, \\ \omega_K &\coloneqq \|z - \varphi_h\|_{L^2(K)} + h_K^{1/2} \cdot \|z - \varphi_h\|_{\partial K} \quad \forall K \in T, \end{aligned}$$

where

$$R(y_h)|_K \coloneqq f + \operatorname{div}(k(w) \cdot \nabla y_h)$$

is the residual of the equation in the interior of a cell K, and

$$r(y_h)|_\Gamma \coloneqq \begin{cases} \frac{1}{2}\langle [k(w)\cdot\nabla y_h],\, \vec{n}\rangle & \text{if } \Gamma \subseteq \partial K \setminus \partial\Omega, \\ 0 & \text{if } \Gamma \subseteq \Gamma_D, \\ \langle k(w)\cdot\nabla y_h,\, \vec{n}\rangle & \text{if } \Gamma \subseteq \Gamma_N \end{cases}$$

is the residual associated with non-smoothess on a subset Γ of the boundary ∂K of a cell K. Here, $[k(w)\cdot\nabla y_h]$ is the jump term of the vector field $k(w)\cdot\nabla y_h$ across the boundary Γ. Because r only appears as an absolute value, the orientation of $\vec{n}$ and the sign of the jump term do not affect the error estimator.

The constant h_K represents the diameter of K, and z and φ_h are the exact adjoint solution and a mesh-dependent approximation thereof. Of course, the exact adjoint solution z is not known. In Section 5.1, [BR01] proposes three approaches for calculating the weights ω_K:

(i) *Global higher-order approximation* replaces z with a higher order adjoint solution and calculates the weights from the difference of two approximate adjoint solutions in a manner reminiscent of an embedded Runge-Kutta method;

(ii) *Local higher-order approximation*, which performs a higher-order interpolation of the approximate solution in each cell separately;

(iii) *Approximation by difference quotients*, which uses a local finite-difference approximation of the second derivative of z to calculate the weights and only works for piecewise linear approximations.

Due to ease of implementation, we choose global higher-order approximation. More precisely, we approximate

- Control functions w using discontinuous cell-wise constant functions;
- Lower order solutions y_h and ϕ_h using continuous cell-wise linear functions;
- Higher order solutions y and z using continuous cell-wise quadratic functions.

Our mesh is a conforming triangle mesh. The finite elements that we use are the Discontinuous Lagrange element of 0-th order, and the Continuous Lagrange elements of first and second order, respectively (see [Kir+12]). Here, the word "continuous" indicates that degrees of freedom are shared between adjacent cells to ensure continuity of the overall function. We refer to these elements as DG_0, CG_1, and CG_2, respectively.

With this, we have an error estimator for the objective function value that allows for errors to be attributed to individual mesh cells. For the gradient error estimator, we deviate from the proven methodology of DWR and improvise slightly. We remember that, for a given control perturbation d_w, the linearized change in objective for a given reference point $(y,\, w)$ with adjoint solution z is

$$-D_w f(y,w)(d_w)(z) = -\int_\Omega \langle (1-\varepsilon)\cdot d_w\cdot\nabla y,\, \nabla z\rangle \,\mathrm{d}\mu.$$

For approximations y_h of y and z_h of z, we find that

$$\begin{aligned}
&\left|\int_\Omega \langle (1-\varepsilon)\cdot d_w\cdot\nabla y,\ \nabla z\rangle \,\mathrm{d}\mu - \int_\Omega \langle (1-\varepsilon)\cdot d_w\cdot\nabla y_h,\ \nabla z_h\rangle \,\mathrm{d}\mu\right| \\
&\quad = (1-\varepsilon)\cdot\left|\int_\Omega d_w\cdot\big(\langle\nabla y,\ \nabla z\rangle - \langle\nabla y_h,\ \nabla z_h\rangle\big)\,\mathrm{d}\mu\right| \\
&\quad = (1-\varepsilon)\cdot\left|\int_\Omega d_w\cdot\Big(\Big\langle\nabla(y-y_h),\ \nabla\Big(\frac{z+z_h}{2}\Big)\Big\rangle + \Big\langle\nabla\Big(\frac{y+y_h}{2}\Big),\ \nabla(z-z_h)\Big\rangle\Big)\,\mathrm{d}\mu\right|.
\end{aligned}$$

Here, we make use of the fact that

$$\begin{aligned}
&\Big\langle\nabla(y-y_h),\ \nabla\Big(\frac{z+z_h}{2}\Big)\Big\rangle + \Big\langle\nabla\Big(\frac{y+y_h}{2}\Big),\ \nabla(z-z_h)\Big\rangle \\
&\quad = \frac{1}{2}\cdot\langle\nabla y,\ \nabla z\rangle + \frac{1}{2}\cdot\langle\nabla y,\ \nabla z_h\rangle - \frac{1}{2}\cdot\langle\nabla y_h,\ \nabla z\rangle - \frac{1}{2}\cdot\langle\nabla y_h,\ \nabla z_h\rangle \\
&\quad + \frac{1}{2}\cdot\langle\nabla y,\ \nabla z\rangle - \frac{1}{2}\cdot\langle\nabla y,\ \nabla z_h\rangle + \frac{1}{2}\cdot\langle\nabla y_h,\ \nabla z\rangle - \frac{1}{2}\cdot\langle\nabla y_h,\ \nabla z_h\rangle \\
&\quad = \langle\nabla y,\ \nabla z\rangle - \langle\nabla y_h,\ \nabla z_h\rangle.
\end{aligned}$$

We further simplify this expression by replacing the arithmetic mean $\frac{y+y_h}{2}$ and $\frac{z+z_h}{2}$ with y_h and z_h, respectively. The reasoning behind this is that, for sufficiently precise approximations, we expect the difference between the gradients of the two to be of negligible magnitude compared to either value, which means that the arithmetic mean will be close to both functions. For the difference between them, such a substitution is not possible. Here, we approximate the difference between the actual and approximate solution with the difference between the higher and lower order approximations that we already need to calculate for the DWR estimator of the objective error. This means that the gradient error estimator can be assembled at relatively low additional cost.

For a perturbation between two indicator functions, the pointwise value of d_w is either 0 or ± 1 almost everywhere and we find that

$$\begin{aligned}
&\left|\int_\Omega \langle (1-\varepsilon)\cdot d_w\cdot\nabla y,\ \nabla z\rangle \,\mathrm{d}\mu - \int_\Omega \langle (1-\varepsilon)\cdot d_w\cdot\nabla y_h,\ \nabla z_h\rangle \,\mathrm{d}\mu\right| \\
&\quad \le (1-\varepsilon)\cdot\int_{\{d_w\neq 0\}}\left|\Big\langle\nabla(y-y_h),\ \nabla\Big(\frac{z+z_h}{2}\Big)\Big\rangle + \Big\langle\nabla\Big(\frac{y+y_h}{2}\Big),\ \nabla(z-z_h)\Big\rangle\right|\mathrm{d}\mu \\
&\quad \approx (1-\varepsilon)\cdot\int_{\{d_w\neq 0\}}\big|\big\langle\nabla(y-y_h),\ \nabla z_h\big\rangle + \big\langle\nabla y_h,\ \nabla(z-z_h)\big\rangle\big|\,\mathrm{d}\mu
\end{aligned}$$

is actually a tight bound because we can adapt the sign of d_w to make the integrand non-negative almost everywhere. By accumulating the error estimator over all of Ω, we obtain an overall L^1 error estimator for the gradient density function.

In practice, the error estimator is difficult to compute because the integrand is not itself within the finite element space. We use

$$\eta_{g,K} \coloneqq (1-\varepsilon)\cdot\left|\int_K\big(\big\langle\nabla(y-y_h),\ \nabla z_h\big\rangle + \big\langle\nabla y_h,\ \nabla(z-z_h)\big\rangle\big)\,\mathrm{d}\mu\right|$$

as an approximate error contribution for each cell K and

$$\eta_g \coloneqq \sum_{K\in T}\eta_{g,K}$$

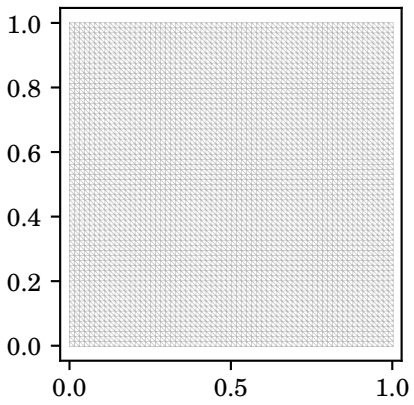


Figure 4.8: Initial mesh for the Poisson problem.

as an overall error estimator. While this is not very precise, we expect that it will yield sufficiently good approximations.

We start initially with a uniform triangle mesh. The initial mesh is displayed in Figure 4.8. This initial choice certainly satisfies the bounded eccentricity assumption (Assumption 2.4.25 (14)). For refinement, we select those mesh cells whose error indicator exceeds $\frac{\beta}{\#T}$ where β is the requested error tolerance and $\#T$ is the number of cells. This is a refinement-only variant of the "error-balancing strategy" used in [BR01]. We additionally impose a restriction that each round of mesh refinement may at most refine one tenth of all mesh cells. To avoid over-refinement across iterations, refined meshes are only built upon once the step for which they were generated is accepted. If a step is rejected or a gradient is re-calculated, all refinements generated by evaluations that do not impact the solver output are discarded.

Once the mesh cells to be refined are selected, the mesh is refined using a red-green-blue refinement scheme as, for instance, described in [FS20] and as implemented in the underlying finite element assembly library SCIKIT-FEM [GM20]. Here, the selected cells are refined using the "red" scheme and surrounding cells are refined using other schemes as needed to create a conforming mesh. We do not perform coarsening, although coarsening schemes for red-green-blue refinements do exist (see, e.g., [FS21]). Conservation of the shape regularity condition (Assumption 2.4.25 (14)) is demonstrated in [FS20].

4.2.2 Implementation Details

For the assembly of the discretized finite element equation systems, we use the SCIKIT-FEM[4] Python library [GM20]. In contrast to other, more feature-complete libraries, SCIKIT-FEM is a pure Python library with a relatively small dependency footprint and negligible setup cost. It does not support distributed solving across multiple machines, but makes extensive use of vectorization through the use of NUMPY and SCIPY for most internal computations.

[4] SCIKIT-FEM is licensed under an open-source 3-clause BSD license and available for download on `https://github.com/kinnala/scikit-fem`.

During refinement, control functions must be interpolated from the coarse mesh to its refinement. This operation incurs a large amount of unnecessary computational overhead because SCIKIT-FEM does not disclose information about parent-child relations between cells in the coarse and fine mesh. We have to reconstruct this information after the fact. This requires locating the coarse cell containing the midpoint of each cell of the fine mesh once per refinement, making refinement very costly and making runtimes almost certainly unrepresentative of what is actually practically achievable.

Aside from the implementation details that we have already discussed in the previous section, we apply a factor of 0.05 to all error estimates for the objective functional. This is an empirically determined value that is meant to mitigate over-refinement issues. We consider the use of a factor below 1 justifiable due to inherent safety margins within the optimization algorithm as well as the self-correcting nature of iterative optimization methods. For the gradient error, the fact that norm-based error estimators are very likely to significantly overestimate the error in the actual quantities of interest also plays a role. A higher factor is desirable, but requires better estimators and more sophisticated mesh refinement schemes.

During error bound apportionment for the penalty functional, we assign none of the error bound to the constraint functional. We have discussed this as an edge case in penalty error control in Section 3.2.1.3. It is valid because the constraint functional is evaluated without truncation error and because no division by zero is caused by this choice.

4.2.3 Experiment

Because the optimization problem (4.3) is a constrained problem, we perform two different experiments. In the first experiment, we optimize the value of the weighted sum

$$\widetilde{J}(U) \coloneqq J(U) + m \cdot G(U)$$

for a fixed weight $m = 8.75 \cdot 10^{-5}$. The choice of m is somewhat arbitrary and informed by the difference between the gradient terms of this regularization term and the quadratic penalty term. We have

$$\nabla\big(m \cdot G\big) = m \cdot \nabla G,$$
$$\nabla\Big(\frac{m}{2} \cdot \max\big\{0, G\big\}^2\Big) = m \cdot \max\big\{0, G\big\} \cdot \nabla G.$$

In order for the two terms to yield similar stationarity conditions, we must choose m such that

$$m \approx m^* \cdot \max\big\{0, G(U^*)\big\},$$

where m^* is the final penalty parameter chosen by the penalty method and $\max\{0, G(U^*)\}$ is the constraint violation at the penalty method's final iterate. The feasibility threshold for the penalty method is $\varepsilon_v \coloneqq 10^{-2}$. Therefore, if the final iterate is precisely at the feasibility threshold, $m = 8.75 \cdot 10^{-5}$ corresponds to an anticipated final penalty parameter of $m^* = 8.75 \cdot 10^{-3}$. If the final feasibility violation is only $8.75 \cdot 10^{-3}$, then the value corresponds to an anticipated penalty parameter of $m^* = 10^{-2}$. Conversely, this choice can be seen as minimizing the difference between the termination conditions at an anticipated penalty

parameter of $m^* = 10^{-2}$ and a final constraint violation between $7.5 \cdot 10^{-3}$ and 10^{-2}.

We solve the regularized problem with the unconstrained steepest descent method described in Section 3.1 with the following parameters:

$$\begin{aligned} \varepsilon &= 10^{-4}, & \sigma_0 &= 10^{-5}, & \sigma_1 &= 0.05, \\ \sigma_2 &= 0.9, & \xi_\delta &= 0.01, & \xi_g &= 0.9, \\ \xi_\tau &= 0.9, & \Delta_0 &= 0.01. \end{aligned}$$

We subsequently refer to this experiment as the "unconstrained run" because it is performed using an unconstrained optimization algorithm.

In the second experiment, we use the quadratic penalty method described in Section 3.2.1.3 to solve the constrained problem with the measure bound $M = 0.4$ and the algorithmic parameters

$$\begin{aligned} \varepsilon_\tau &= 10^{-4}, & \overline{\varepsilon}_\tau &= 10^{-4}, & \varepsilon_v &= 10^{-2}, \\ \sigma_0 &= 10^{-5}, & \sigma_1 &= 0.05, & \sigma_2 &= 0.9, \\ \xi_\delta &= 0.1, & \xi_g &= 0.9, & \xi_\tau &= 0.9, \\ \xi_v &= 0.1, & \Delta_0 &= 0.1, & m_{\text{init}} &= 10^{-4}, \\ \overline{m} &= 1. \end{aligned}$$

All controlled evaluation loops (see Algorithm 1 on page 215) are run with initial bound $\beta_0 = \infty$ and decay rate $\xi = 0.9$ unless explicitly otherwise stated by the calling algorithm. We fix the error apportionment parameters ζ and ϵ to 0.1. We subsequently refer to this experiment as the "penalty run" because it uses the penalty method. Both runs start with the initial solution $U_0 \coloneqq \emptyset$, which is a feasible solution.

In both experiments, step acceptance thresholds are very low and the tuning parameters ξ_g and ξ_τ are very high. Both of these are measures to reduce the risk of over-refinement due to high gradient error estimates. Because the gradient error estimate is a norm estimate, it cannot be expected to be very tight. We raise ξ_g and ξ_τ to increase the gradient error tolerance. On the other hand, we keep σ_0 low to keep the trust region radius high. Because the gradient error is controlled with respect to the L^1 norm, gradient error tolerances scale with the trust region radius, which means that maintaining a high trust region radius throughout the iteration also lowers the risk of over-refinement.

As before, we perform these experiments on an Intel Core i5-10210U laptop CPU with 8 GB of total random-access memory. The CPU has four physical cores and eight logical cores via Hyperthreading. While the Python code is still single-threading, the PDE solver uses NumPy and SciPy for significant workloads. These underlying libraries are capable of distributing over multiple cores and do so more effectively for the Poisson Design problem than they did for the Lotka Volterra Fishing problem.

The unconstrained run terminates after 43 iterations with a total execution time of 424.03 CPU seconds, a final objective of $1.424346 \cdot 10^{-4}$, and a final instationarity of $9.548901 \cdot 10^{-5}$. Throughout the optimization process, there are no step rejections. The number of triangle cells increases from an initial cell count of 8192 to a final cell count of 41595. The development of the cell count over the iterations of the main loop is shown in Figure 4.9 on the next page. An excerpt of the solver log is shown in Table 4.2 on the following page.

Table 4.2: Abbreviated solver log for the unconstrained run. Step sizes represent single iteration, number of rejected steps is accumulated over multiple iterations.

#	Objective	Instationarity	Step Size	Rejected	CPU Time
0	5.622×10^{-4}	4.973×10^{-3}	0.000×10^{0}	0	0:00:02.70
5	4.290×10^{-4}	3.393×10^{-3}	9.987×10^{-3}	0	0:00:16.64
10	3.115×10^{-4}	1.820×10^{-3}	9.995×10^{-3}	0	0:00:32.24
15	2.580×10^{-4}	1.106×10^{-3}	9.995×10^{-3}	0	0:00:47.82
20	2.097×10^{-4}	6.069×10^{-4}	9.995×10^{-3}	0	0:01:06.02
25	1.859×10^{-4}	3.953×10^{-4}	9.989×10^{-3}	0	0:01:37.04
30	1.680×10^{-4}	2.449×10^{-4}	9.995×10^{-3}	0	0:02:14.39
35	1.569×10^{-4}	1.729×10^{-4}	9.998×10^{-3}	0	0:03:19.65
40	1.463×10^{-4}	1.139×10^{-4}	9.995×10^{-3}	0	0:04:57.27
43	1.424×10^{-4}	9.549×10^{-5}	9.989×10^{-3}	0	0:07:04.03

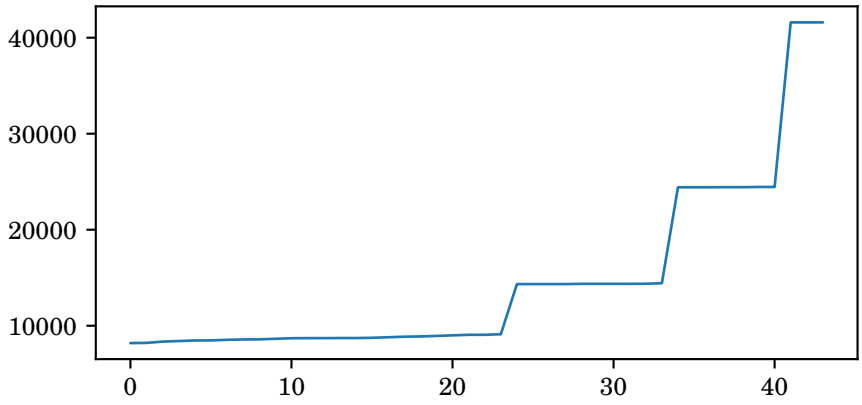


Figure 4.9: Cell count over iterations of the unconstrained run.

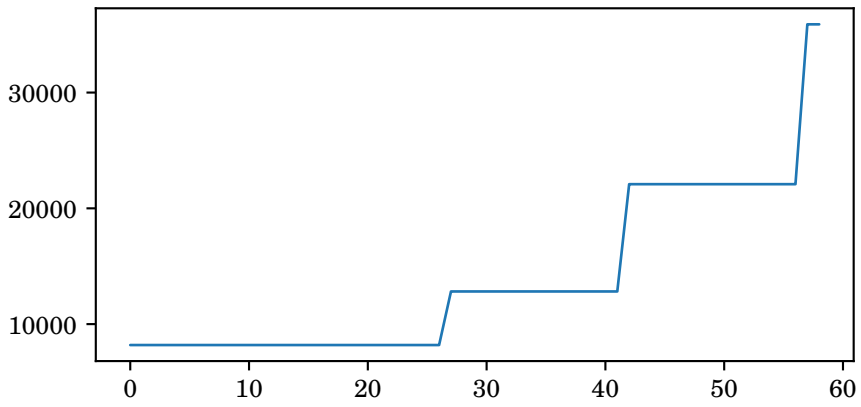


Figure 4.10: Cell count over iterations of the penalty run.

Table 4.3: Abbreviated solver log for the penalty run. Step sizes represent single iteration, number of rejected steps is accumulated over multiple iterations.

#	Objective	Infeasibility	Instationarity	Penalty	Step Size	Rejected	CPU Time
0	5.622×10^{-4}	0.000×10^{0}	5.060×10^{-3}	1.000×10^{-4}	0.000×10^{0}	0	0:00:02.72
5	4.253×10^{-4}	0.000×10^{0}	3.476×10^{-3}	1.000×10^{-4}	9.888×10^{-3}	0	0:00:15.33
10	3.041×10^{-4}	0.000×10^{0}	1.892×10^{-3}	1.000×10^{-4}	9.888×10^{-3}	0	0:00:31.41
15	2.477×10^{-4}	0.000×10^{0}	1.205×10^{-3}	1.000×10^{-4}	9.888×10^{-3}	0	0:00:46.54
20	1.994×10^{-4}	0.000×10^{0}	7.779×10^{-4}	1.000×10^{-4}	9.888×10^{-3}	0	0:01:01.78
25	1.678×10^{-4}	0.000×10^{0}	5.085×10^{-4}	1.000×10^{-4}	9.888×10^{-3}	0	0:01:18.19
30	1.451×10^{-4}	0.000×10^{0}	3.520×10^{-4}	1.000×10^{-4}	9.979×10^{-3}	0	0:01:49.57
35	1.280×10^{-4}	0.000×10^{0}	2.369×10^{-4}	1.000×10^{-4}	9.796×10^{-3}	0	0:02:20.31
40	1.141×10^{-4}	0.000×10^{0}	1.747×10^{-4}	1.000×10^{-4}	9.979×10^{-3}	0	0:02:49.58
45	1.017×10^{-4}	4.586×10^{-2}	1.278×10^{-4}	1.000×10^{-4}	9.995×10^{-3}	0	0:04:02.57
50	1.124×10^{-4}	7.644×10^{-2}	1.597×10^{-4}	6.400×10^{-3}	9.995×10^{-3}	0	0:05:18.40
55	1.040×10^{-4}	2.961×10^{-2}	2.490×10^{-4}	2.560×10^{-2}	9.552×10^{-3}	0	0:06:31.97
58	1.044×10^{-4}	5.300×10^{-3}	5.141×10^{-5}	5.120×10^{-2}	9.998×10^{-3}	0	0:08:00.60

The penalty run terminates after 58 iterations with a total execution time of 480.60 CPU seconds, a final objective of $1.043635 \cdot 10^{-4}$, a final penalty function instationarity of $5.140746 \cdot 10^{-5}$, and a final infeasibility of $5.300140 \cdot 10^{-3}$ with an accompanying penalty parameter of $5.12 \cdot 10^{-2}$. Again, there are no step rejections. The cell count increases from an initial value of 8192 to 35893. The development over the iterations of the main loop can be seen in Figure 4.10 on page 368. An excerpt of the solver log can be seen in Table 4.3 on the preceding page.

We note that in the unconstrained run, individual iterations take longer on average than they do in the penalty run. There is also more refinement. Because the measure bound is 0.4, the initial solution is the empty set, the trust region radius is 0.01 throughout, and because increasing the amount of the better heat conductor is generally beneficial to reduce compliance, we may assume that it takes approximately 40 iterations until the constraint becomes active. Indeed, a look at the detailed solver log, which is available as part of the run data set [Hah25b], shows that constraint violation becomes nonzero for the first time after 41 iterations with an initial constraint violation of $6.219482 \cdot 10^{-3}$, which is only slightly less than the step size of $9.918213 \cdot 10^{-3}$. Prior to this point, the penalty run takes 175.49 CPU seconds for 41 steps. After this point, it takes 305.11 CPU seconds for the remaining 17 steps, which means that computational effort per step rises dramatically as the constraint becomes active.

For the quadratic penalty method, this is to be expected because the penalty term is quadratic and the model function is a "linearization" in the sense of Definition 2.4.1 on page 152. This is, in part, the motivation for choosing a low initial penalty factor m_{init}. The low penalty factor reduces the initial curvature of the penalty term, thereby increasing the accuracy of the projected descent. Once the penalty term becomes active, the curvature of the penalty term should reduce the actual step quality and therefore likely also decrease the gap between $\widetilde{\rho}$ and the acceptance threshold σ_0, which enters into the formula determining the gradient error bound.

For the unconstrained run, the reason why steps require more effort is somewhat mysterious because the penalty term is not quadratic. It enters into the objective linearly. Because the constraint functional is linear, it should be accounted for perfectly by the linearization. As a result, it should shift both the projected and the actual descent by the same value. If the decrease in the compliance objective is small compared to its projected descent, then such a shift could, if its sign was positive (i.e., if the measure of U would increase), move the numerator of the step quality estimate towards zero faster than the denominator, leading to a noticable decay in step quality. However, this is speculative and we are unsure as to what causes this effect in the unconstrained run.

The solver logs show that no steps are ever rejected in either run. This means that we have not observed the behavior of the algorithm under step rejections. During trials, it was observed that step rejections often result in long chains of gradient error control refinements, though whether this is caused by inadequacies of the gradient error estimator or whether it might be an early indication of problems that arise when one treats an L^2 derivative as if it was an L^1 derivative is difficult to discern.

Figure 4.11 on the facing page shows the gradient density function from a selected iteration during the penalty run. The boundary of the control set is marked in black. We can see that the highest absolute values occur around

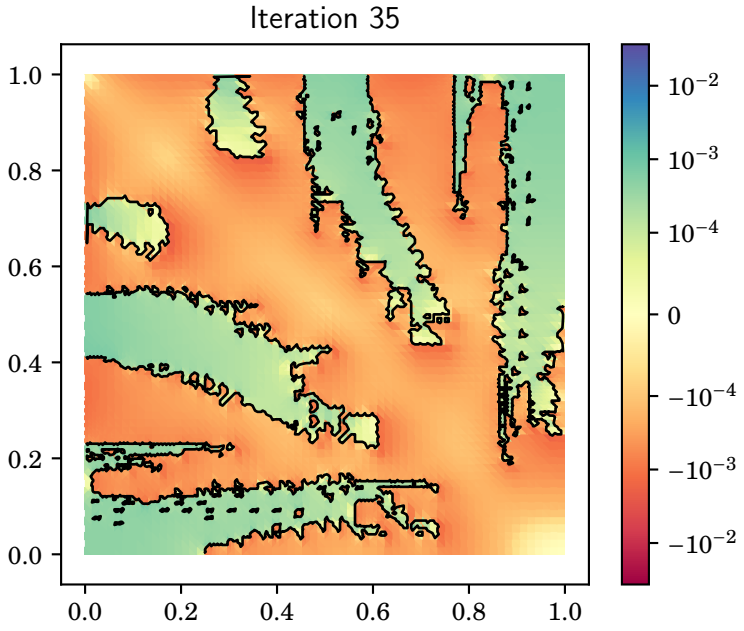


Figure 4.11: Gradient density function at iteration 35 of the penalty run. The bold black line marks the boundary of the control set.

the boundary. While the approximate values do not appear very large, it is conceivable that they could become infinitely concentrated around the boundary of the control set.

Finally, we examine the development of the control set over the course of the optimization. Figure 4.12 on the next page shows selected iterates from the unconstrained run. Figure 4.13 on page 373 shows selected iterates from the penalty run. Structurally, both solutions appear similar.

We observe that for both runs, the conductor "grows" gradually from the Dirichlet boundary into the domain. This is in line with our prior observation that low gradient densities are concentrated around the boundaries of the control set and may indicate that shape optimization methods starting from the Dirichlet boundary may be a more suitable optimization method for this problem.

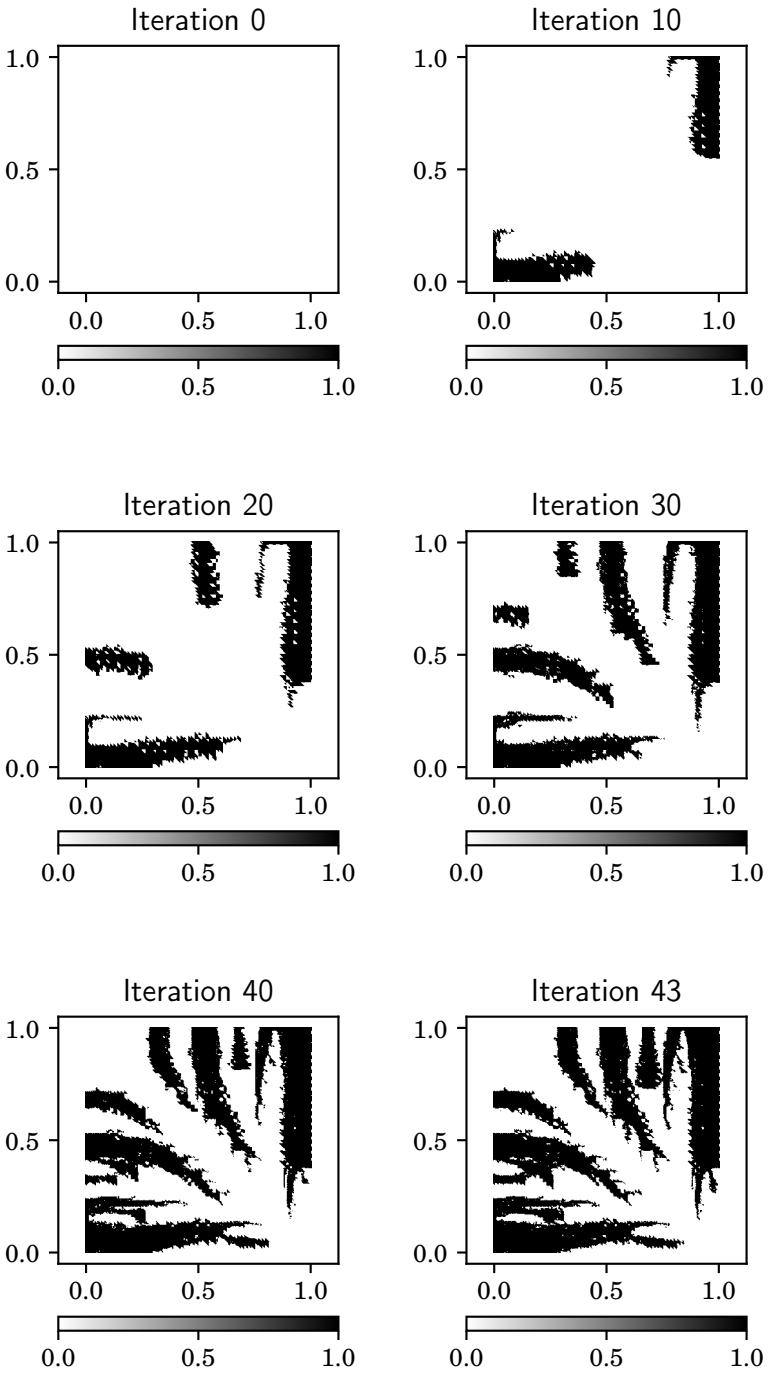


Figure 4.12: Control plots at selected iterations of the unconstrained run.

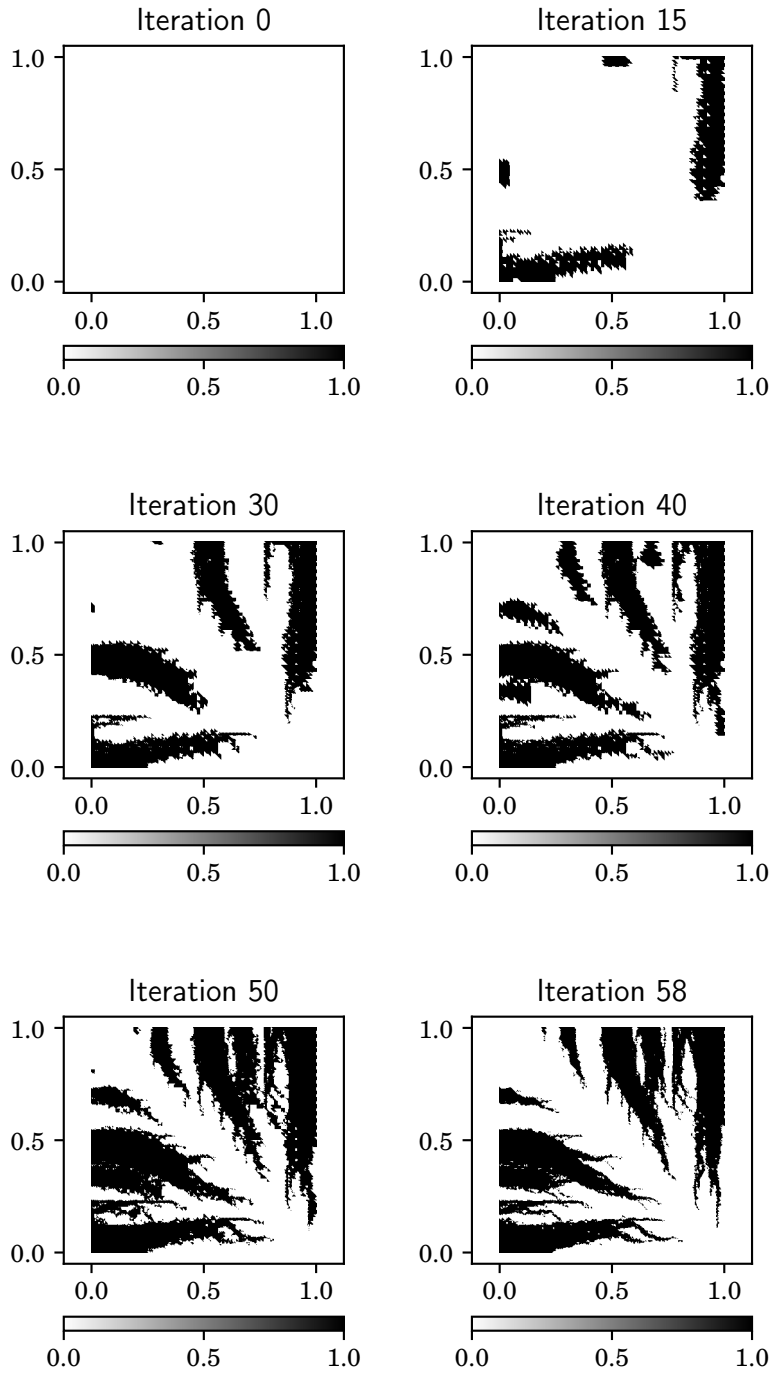


Figure 4.13: Control plots at selected iterations of the penalty run.

Chapter 5

Discussion And Outlook

In this chapter, we look back on the results of the previous chapters. We discuss some possible criticisms of these results. We also discuss how this work could be expanded upon in the future. Finally, we give some conclusions and closing words.

5.1 DISCUSSION AND CRITICISMS

The stated goal of this thesis is to develop iterative optimization schemes for problems with spatially distributed binary variables. In order to do so, we have developed theoretical tools to examine gradual improvement directions in the continuum of distributed binary-valued controls. We were then able to transfer traditional NLP methods to this setting with relative ease.

In summary, we consider this endeavor to be quite successful. Even when they are used with inexact error estimators, the methods derived in this way appear to be sufficiently resilient to still consistently produce iterative improvement. With the theoretical toolkit developed in Chapter 2 and parts of Chapter 3, it is very likely possible to go significantly beyond the two optimization methods presented here. However, there are also some plausible criticisms of our work that we will address here.

Lack of Comparison to Other Methods

In Chapter 4, we have not compared our methods with any other solution method. As we have noted there, this is primarily because of the implementation effort necessary to compare solution methods and because we do not believe that such comparisons can be made at this point in a manner that is not grossly misleading.

The main reason why such comparisons are necessarily misleading is that our methods require error control and mesh refinement, while most alternative methods work with fixed meshes and do not necessarily control for truncation errors. This means that other methods have an inherent benefit in performance comparisons because they do not incur the performance cost for error estimation and control. In many cases, this is not because other methods do not *need* error control. In principle, all discretizations of infinite-dimensional optimization problems require error control. However, this is often neglected in discretize-then-optimize methods.

In order to mitigate the cost of error control, we could attempt to start optimization with a sufficiently fine mesh such that error control becomes a non-issue. However, this could unfairly bias a comparison in favor of our methods. Discretize-then-optimize methods often use subproblem solvers whose performance deteriorates for large numbers of variables. This is especially true for methods that use Branch and Bound solvers, because the worst case performance of such solvers is well-known to increase exponentially with problem dimension. On the other hand, the performance of our optimization methods should only change in terms of the processing time required for a single iteration. The total number of iterations should not change by much as the degree of refinement is changed. Therefore, a higher initial degree of refinement would likely bias the comparison in our favor.

In summary, it is very difficult to make performance comparisons. In order to do so, we would need very well-implemented and performant error control methods for the test problems and we would have to be very careful in balancing the solver parameters. As we have noted in Chapter 4, this is significantly beyond the scope of what we can do in this thesis.

Combinatorial Constraints

At first glance, continuous optimization in measure spaces appears very powerful. The requirement of atomlessness gives the space a structure sufficiently reminiscent of vector spaces to allow for iterative nonlinear optimization. However, there are some constraints that are of significant practical interest that are fundamentally incompatible with our methodology because they violate atomlessness.

This applies to so-called "combinatorial constraints" (see, e.g., [SJK11]) such as variation bounds and dwell time constraints. For dwell time constraints or, more generally, "contiguous measure bounds," it is possible that these could be turned into non-convex constraints that could be solved with suitable relaxation methods. However, for variation bounds, this is a more fundamental problem.

A variation bound is a limitation on the surface area of a control set. We can evidently design a sequence of step sets in $\mathbb{R}^n$ whose Lebesgue measure approaches zero while its surface measure stays constant or even grows to infinity. This shows that variation bounds are plainly not differentiable constraints using our concept of differentiability. There exists some work on methods to solve problems with variation regularizations (see, e.g., [LM22]). However, solution methods appear limited to solving discretized problems with finite-dimensional discrete solvers.

Using finite-dimensional discrete solvers negates some of the benefits of looking at the problem as an optimization problem on a continuum. It is also not clear whether the resulting optimization method is transferrable to a solution method for the infinite-dimensional problem. While it may be practically useful to investigate such methods, we follow an optimize-then-discretize approach in which optimization methods are first developed for an undiscretized problem and their steps are then approximated in a discretized setting for practical purposes. We consider this approach superior because it yields more theoretical insight into the underlying problem and because the resulting methods tend to be less dependent on the specifics of the discretization that is being used. For instance, PYCOIMSET makes very few assumptions about underlying data structures and other problem implementation details. If our algorithms were formulated with a

specific discretization in mind, then PyCoimset would have to force the user to use that specific type of discretization.

It remains to be seen whether variation constraints can be integrated into our approach and how much of a change of setting would be necessary to do so. It is possible that the following notes on generality factor into this question.

Generality

We chose measure spaces as a setting for binary-valued optimization because it is, at first glance, a very natural setting. Binary-valued functions are indicator functions. Furthermore, the fact that measurable functions can be used to generate sublevel sets means that it is relatively easy to formulate a steepest descent step finding routine.

All of this means that measure spaces are very appealing as a setting. However, we do not make much use of the underlying measure space structure in Chapter 3. We only use this structure to prove geodesicity of the search space and to demonstrate properties of those geodesics.

Upon closer inspection, much of Chapter 3 could be transferred to any geodesic metric space with a commutative group structure in which there exists an efficient way to solve trust-region subproblems. Such a generalization could potentially expand the scope of our results and could allow for the inclusion of non-differentiable constraints such as variation bounds. Due to time constraints, we can, unfortunately, not attempt this generalization here. However, we leave this as a promising research question for anyone who wants to investigate it.

5.2 FUTURE RESEARCH

This thesis only scratches the surface of a large amount of potential future research. In this section, we want to highlight some research questions concerning this subject matter that could be examined in the future. This is in addition to the future research questions discussed in the appendices of this work.

We will proceed through these suggestions in the same order in which we had previously proceeded through the subjects in the main part of the thesis.

5.2.1 Measure Space Geodesics

Measure space geodesics are a powerful theoretical tool. We have only made sparing use of them in this thesis. Ignoring, for the moment, the theoretical foundation of optimization with scalar inequality constraints, the optimization algorithms that we propose are stated as trust-region algorithms and do not strictly require measure space geodesics as either an algorithmic guideline or a theoretical justification.

The amount of work that we have invested into their exploration may therefore appear exaggerated. However, this is out of a conviction that measure space geodesics are a powerful and potentially indispensible tool for future steps in algorithmic development.

Iterative Construction

The first promising area of future research is the *iterative construction* of measure space geodesics. In Section 2.3 we develop an extensive toolkit to simplify the construction of geodesics. At the foundation of this toolkit are these main pillars:

- geodesic interpolation (Theorems 2.3.18 and 2.3.22);
- geodesic level set functions (Theorem 2.3.28);
- geodesic composition (Theorem 2.3.34 and Definition 2.3.35);
- and pushforward and pullback operators (Definition 2.3.43).

These pillars were originally devised to answer two open questions that we will discuss in more detail later:

1. Can we use the secant inequality to define "convexity"?
2. Can we exploit second derivatives to obtain better search directions?

For either of these questions, an affirmative proof could potentially be based on the construction of special geodesics for which we cannot give an explicit "closed-form" expression.

We might therefore ask whether we could derive geodesics from implicit descriptions by some manner of fixed point iteration. The toolkit that we have developed allows iterative refinement of geodesics by the following scheme:

- using the pushforward operator, we can "unravel" an integrable function (e.g., a gradient density function) into a function defined on the parameter interval of the geodesic;
- using sparse interpolation, we can use the "unravelled" function as a quasi-GLSF for a minimal mean "parameter geodesic" connecting the empty set with the parameter interval of the original geodesic;
- using geodesic composition, we can use that parameter geodesic to reorder the original geodesic and repeat the process.

More generally, we can use any "criterion function" on the original geodesic's parameter interval to reorder the geodesic. Pushforward allows us to derive such criterion functions from density functions on the geodesic's total variation. By using generator geodesics as tie-breakers for the minimal mean geodesic, we may even be able to overcome the unavoidable decay of the generated σ-algebra that would arise whenever the criterion function is not essentially bijective.

What is missing from this generic scheme is a convergence proof. The procedure appears elegant and intuitive, but we do not yet know whether a procedure such as this could be made to converge in any meaningful sense. Therefore, the first major research question that we pose is:

> Under which conditions, if any, can a limit geodesic with desirable properties be derived from an iterative construction scheme such as the one proposed in this section? What other iterative construction schemes could be formulated?

An extensible and adaptible framework for such convergence arguments could open up the path to further research into the optimization of non-differentiable functionals, the exploitation of second-order derivatives, and other fields.

A simple example of such an iterative scheme would be the "simultaneous alignment procedure" proposed in Section A.2, though simultaneous alignment does not require a convergence proof because it terminates after a finite number of refinements.

5.2.2 Set Functionals

Our discussion of the theory of set functionals in Sections 2.4 and 2.5 leaves some open questions, particularly with regard to the proper definition of certain terms.

Cauchy Guarantees Under Strict Convexity

In Proposition 2.5.6, we have shown that for strongly convex set functionals, any sequence of arguments whose instationarity approaches zero is a Cauchy sequence. In "conventional" optimization in $\mathbb{R}^n$, the Cauchy property (and thereby convergence) already follows from *strict* convexity and does not require strong convexity. We do not yet know whether this argument is possible in our infinite-dimensional setting.

It is possible that the argument breaks down due to the infinite-dimensional nature of the space. However, we have not constructed a counterexample. Instead, the absence of a proof for strictly convex functionals stems from the fact that most straightforward proof methods use the optimum as a reference point. The fact that we generally do not have an optimum prohibits arguments like this in our setting.

Our question is therefore:

> Is there a way to prove that instationarity approaching zero implies a Cauchy sequence for strictly convex set functionals? If not, why does the argument break down and what properties of the setting can be exploited to construct counterexamples?

This question is mostly of theoretical interest. It hints at a significant gap in our understanding of the setting and the answer could improve our ability to reason about it.

Secant-Type Convexity

In Section 2.5, we transfer the concept of convexity from functions to set functionals. We base our definition there on the tangent inequality. This choice is attractive because it is more easily transferrable than the secant inequality, which is the traditional defining property for convexity.

The secant inequality states that a function $f : X \to \mathbb{R}$ is convex if and only if

$$f\big(\lambda y + (1-\lambda)x\big) \leq \lambda \cdot f(y) + (1-\lambda) \cdot f(x) \quad \forall x \in X, y \in X, \lambda \in [0,\ 1].$$

With geodesics, we have the ability to generate an analogue of the convex combination $\lambda \cdot y + (1-\lambda) \cdot x$. However, rather than having a single analogue of the convex combination, there are many suitable geodesics that can replicate the behavior of convex combinations that are all equivalent. This is a problem.

We could naïvely demand that a convex functional should appear convex along every geodesic. However, this breaks down for even the simplest of convex maps: the signed measure. Let (X, Σ, μ) be the space of Lebesgue measurable sets on the bounded interval $X = [0,\ 10]$ with the Lebesgue measure. We consider the set functional F with

$$F(U) \coloneqq \int_U x - 5 \,\mathrm{d}x \qquad \forall U \in \Sigma.$$

This is clearly a signed measure that is absolutely continuous with respect to μ and has the density function $x \mapsto x - 5$. It is therefore the closest equivalent to a bounded linear map that we have in our setting. It is differentiable (being its own derivative up to sign changes) and satisfies the tangent inequality. Therefore F is convex and should also satisfy the equivalent of the secant inequality.

Because we have access to the density function of F, we can formulate a maximum-mean geodesic of the density function by taking a minimum-mean geodesic of $x \mapsto 5 - x$. The map $\gamma\colon I \to \Sigma/_{\sim_\mu}$ with $I \coloneqq [0,\ 10]$ and $\gamma(t) \coloneqq [10 - t,\ 10]$ is evidently one such maximum mean geodesic. By taking the composition of F and γ, we obtain

$$\begin{aligned}
(F \circ \gamma)(t) &= \int_{10-t}^{t} x - 5 \,\mathrm{d}x \\
&= \left[\frac{1}{2}x^2 - 5x\right]_{10-t}^{10} \\
&= 50 - 50 - \left(\frac{(10-t)^2}{2} - 50 + 5t\right) \\
&= -\left(50 - 10t + \frac{1}{2}t^2 - 50 + 5t\right) \\
&= 5t - \frac{1}{2}t^2
\end{aligned}$$

This function is not convex. It is strictly concave. We can do the same locally with every continuously differentiable function in any point where its derivative measure is not a scaled version of the underlying measure μ. This means that our naïve definition yields a very small class of convex set functionals.

The next best approach is to demand that for any two arguments, there must exist some geodesic connecting the two arguments along which the set functional is a convex function. For signed measures, the constant mean geodesic (see Definition 2.3.56) is specifically designed to fulfill this criterion. For other set functionals, the question becomes what such a special geodesic would look like.

For convex differentiable set functionals, it is conceivable that we could design an iterative process by repeatedly finding constant mean geodesics of the set functional's directional derivative along the preceding geodesic. We would then have to prove that this process meaningfully converges to a limit geodesic.

This is not a viable approach for proving the secant inequality for non-differentiable set functionals because it requires the existence of a directional derivative.

We pose the following open research questions:

> Is it always possible to construct (iteratively or not) a geodesic along which differentiable set functionals that satisfy the tangent inequality appear convex? Can the same or a similar procedure be applied to

> certain classes of non-differentiable set functionals? If so, what additional assumptions would have to be made about these functionals? Is there a better way to formulate the secant inequality in a measure space setting?

The secant inequality is a fairly theoretical issue from our perspective because we are primarily interested in differentiable set functionals. However, it closely ties into the issue of iterative geodesic construction and directional derivatives along geodesics. It may also factor into the finding of useful generalizations of differentiability, which may be relevant to implementing common combinatorial constraints. This group of research questions is therefore suitable as a starting point for future research of significant value.

Generalized Concepts of Differentiability

In "conventional" NLP theory, there are generalizations of the derivative, such as the subderivative. This question is a very open one because we have not done much prior work on it:

> Is there a viable analogue of subderivatives in measure spaces? Could it be used to formulate optimization methods for convex non-differentiable optimization problems? Are there other generalizations of the derivative that could be transferred and what benefits could they bring?

This is potentially of great value. It could allow us to optimize with non-differentiable functionals or functionals whose derivative is not continuous.

5.2.3 Optimization Methods

The development of practical optimization methods is, of course, the central goal of this thesis. The following are suggestions for future expansion upon the optimization methods described in this thesis.

Step Finding Methods

We describe the main optimization loop in Algorithm 4 as a "framework." The idea behind this term is that we could obtain different optimization methods by using the framework with different step-finding routines. However, we have so far always used the steepest descent step. We could design other step-finding routines. As we have already pointed out, the barycentric constant mean geodesic is a straightforward example of a geodesic that could be turned into a step-finding routine.

Barycentric constant mean step-finding was briefly considered for inclusion in this thesis. It could potentially avoid issues that arise when numerical errors produce outlier values for the gradient density function. However, there are some algorithmic difficulties with a step finding routine based on barycentric constant mean geodesics. As opposed to the steepest descent step, which can be found using a single bisection procedure, the constant mean descent step requires that we simultaneously perform two bisections to determine two corresponding levels. It is not entirely clear how one would best execute these two simultaneous, interdependent bisections in practice. Particularly in the case of the final "fill out" step, where we pick an arbitrary subset within a given size range from within

the remaining candidate set, it is not clear how we could ensure that balance is maintained between the upper and lower layers such that the mean does not deviate too much from the desired mean.

This suggests that implementing such a step finding routine could be difficult. Meanwhile, whether alternative step-finding methods would provide any practical benefit is questionable. Thus, we have to ask the following three questions:

> Can we implement a step-finding method based on barycentric constant mean geodesics? Can we find other viable step-finding methods? Could such step-finding methods provide any benefit over the steepest descent step?

Norm-Free Gradient Error Control

In Chapter 4, we discovered that error control has a problematically large impact on our algorithms' performance. We suspect that this is because we control gradient errors, which are generally influenced and compounded by errors in objective evaluation, according to norm-based error estimates.

In general, the norm greatly overestimates the impact of errors on our algorithm. This is because our algorithm is only concerned with the projected descent associated with some steps, but norm-based error estimates always estimate the maximal error over all possible steps. It would be much better to estimate the error for steps of interest using methods like the dual-weighted residual error estimator that we use for the objective error in our second test problem.

This would require an architectural overhaul of PYCOIMSET where error estimates are evaluated and taken into consideration during the step finding process. This, in turn, would pose a fundamental problem: if the step set is derived from an approximate gradient density function, how does one estimate the aggregate of both the error in step determination and the error in the calculation of the projected descent over that step. Unless the combined error of both can be controlled with reasonable effort, we do not gain much.

We condense this into the following questions:

> Can we replace the norm-based error control for gradient density functions with a system that only requires error estimates for individual linear forms? How would this impact the complexity of the step finding routine?

Linear Optimization and Trust-Region Sequential Linear Programming

We have repeatedly leaned on the analogous behavior of signed measures and linear functionals. The next research topic we want to present is a compound of two issues where the first flows almost directly into the second.

If we analogize signed measures to linear functionals, then a linear optimization problem in a similarity space would take the form

$$\begin{aligned} &\min_{U \in \Sigma/_{\sim_\mu}} \varphi_0(U) \\ &\quad \text{s.t. } \varphi_i \leq b_i \qquad \forall i \in [n] \end{aligned}$$

where $n \in \mathbb{N}$, $b \in \mathbb{R}^n$, and φ_i are finite signed measures with $\varphi_i \ll \mu$ for all $i \in [n]_0$. The question is:

> Are there efficient algorithms to solve linear optimization problems in similarity spaces? Can they be made significantly more efficient than their nonlinear counterparts?

The latter question is of particular interest because our trust region constraint

$$\mu(U) \leq \Delta$$

is itself a "linear" constraint and can therefore be simply added to any linear optimization problem as a constraint.

This then opens up the possibility of using a linear solver as a step-finding subroutine within a trust region framework. Doing so would allow us to add linearizations of differentiable constraints to the step-finding routine, giving rise to what is essentially a trust-region sequential linear programming method (see, e.g. [BSS06, Sec. 10.3]). Therefore, we ask the following question as an extension of the previous one:

> Is there a viable trust-region sequential linear programming method for optimization in measure spaces with differentiable scalar inequality constraints?

Such a method would obviously greatly expand the scope of our ability to optimize with differentiable scalar inequality constraints.

Second Derivatives and Trust-Region Sequantial Quadratic Programming

In Section A.1, we briefly sketch how higher-order derivatives of set functionals could be shown to exist and have density functions. Leaving aside whether it is ever computationally feasible to calculate these higher-order derivatives, we do not yet know how one would exploit the information contained within them in an optimization algorithm.

In "conventional" NLP theory, the Hessian is used as part of the coefficient matrix of a linear equation system. This approach is not valid in the measure space setting because the optimality criterion is an inequality rather than an equation. We would therefore have to solve an infinite-dimensional inequality system.

It is not clear how we would solve such a system. We could attempt to base our approach on algorithms that are used to solve linear equations with very large Hessians. Matrix-free methods such as Krylov subspace methods seem appealing, but their convergence proofs are often based on the finite dimension of the underlying vector space. It may be possible to make arguments based on discretization that such methods eventually converge in all "relevant" spatial directions. However, it is not clear whether this approach is viable.

The questions are:

> Can second order derivatives be incorporated into step finding? How would one solve such step finding subproblems? Does their use provide any significant benefit over performing multiple first-order steps?

If we could find an approach to incorporate second derivatives into step-finding, this could also open up the possibility of merging such a method with measure space linear optimization to solve linear-quadratic optimization problems. These could then potentially be used to solve step finding subproblems in a sequential

quadratic programming (SQP) method (see, e.g., [NW06, Ch. 18] or [BSS06, Sec. 10.4]).

This field of research presents significant risks and promises substantial rewards. It is possible that second derivatives cannot be incorporated without accepting a performance penalty that makes them practically useless. This would mirror a phenomenon observable in PDE-constrained optimization where algorithms that use second derivatives are often ignored because Hessians would become too great a resource burden in the high-dimensional spaces under discussion.

Limited-Memory Quasi-Newton Methods

Among optimization methods that exploit second derivatives, we want to particularly highlight limited-memory quasi-Newton methods. This is a remarkable category of optimization methods where a low-rank approximation of the Hessians is derived from the observed behavior of the objective function during the preceding steps.

We particularly highlight the limited-memory BFGS method (L-BFGS) which can be implemented in such a way that it maintains an implicit representation of an approximation of the *inverse* of the Hessian (see, e.g., [NW06, Sec. 7.2]). We can then multiply a vector with that approximate inverse using nothing but vector-vector operations, which would completely eliminate the necessity of finding an analogue of matrix operations.

The efficacy of L-BFGS is mostly argued on account of it approximating the inverse of the Hessian. However, if we break down this variant of L-BFGS to its algorithmic steps, the fact that they are beneficial to step-finding may be arguable without the underlying approximation of the inverse of the Hessian.

If this is the case, then we could apply an analogue of L-BFGS in the measure space setting without first having to find analogues to the inverse of a matrix, which is potentially a problem in this setting.

We therefore pose the following questions:

> Can limited-memory quasi-Newton methods such as L-BFGS be transferred to the measure space setting? Can their efficacy be argued without the extensive theory of linear operators and their inverse that we have access to in more conventional vector space settings?

Taking into account the standing of L-BFGS in the field of PDE-constrained optimization, this may well turn out to be the only viable method by which second derivative information can practically be incorporated.

Penalty Methods, Barrier Methods, and Augmented Lagrangian

We have already addressed the possibility of transferring SLP and SQP to the measure space setting. These are constrained optimization methods that are based on incorporating linearized constraints into the step finding problem. However, there are also constrained optimization methods where the constraints are incorporated into the objective and are not directly visible as constraints to the step finding routine.

One such method is the penalty method, which we chose for this thesis because it is very easy to implement a naïve penalty method on top of our unconstrained optimization method. However, penalty methods are usually not the best choice.

There are several ways to expand upon this approach. We had to square our penalty term in order to make the penalty function continuously differentiable. This is regrettable because it discards the σ-additivity of the measure. We have essentially transformed a linear constraint into a quadratic objective term. If we could learn to work with subgradients in our setting, then we could apply an ℓ^1-like penalty term, which may lead to more desirable behavior. For a discussion of nonsmooth penalty methods and their potential benefits, we refer to [NW06, Sec. 17.2].

Secondly, there is the question of interior point methods. A naïve barrier method in the spirit of [NW06, Ch. 19] was briefly considered as the constrained optimization method for this thesis. However, the algorithm exhibited very undesirable behavior near the boundary of the feasible region, likely due to the large curvature of the barrier function in those regions. The benefit of barrier methods over penalty methods is that they produce interior points. While the solution returned by a penalty method will always slightly violate some constraints, barrier methods yield interior points which always (within numerical reason) satisfy the inequality constraints.

Thirdly, the fact that we can already use penalty methods suggests that we may be able to augment these methods with a Lagrangian term to create an analogue of the Augmented Lagrangian method (see, e.g., [NW06, Ch. 17]). We have already pointed out that the concept of the Lagrangian relaxation is transferrable to our setting. However, some care must be taken because Lagrange multipliers are generally not unique in our setting and their values may become very large unless properly controlled.

In summary, we pose the following questions:

> How can we expand on the penalty method presented in this thesis?
> Can we transfer other related optimization methods to our setting?

Hybrid Optimization

We develop a class of algorithms for infinite-dimensional binary-valued optimization that replicate the behavior of "conventional" iterative NLP solvers. An obvious next step could be to handle optimization problems with both set variables and conventional vector space variables. We will refer to such problems as "hybrid problems."

We could simply attempt to solve hybrid problems by defining an algorithm that applies both steps (a set step for the set variables and a vector step for the vector variables) simultaneously or alternates between both in a manner similar to coordinate descent or alternating direction methods. There is no real obstacle to doing this. Our algorithmic correctness arguments would have to be adjusted, but there is very little doubt that it is possible to do this.

The main obstacle that prevents us from dealing with hybrid optimization is what kinds of constraints we would need to formulate "interesting" optimization problems. It is quite possible that we would need to consider a form of vanishing constraint where constraints to vector values are only applied when those values are associated with points that are located within one of the set variables. Such vanishing constraints would likely cause significant theoretical complications.

We therefore pose the following questions:

> Can convergence (or a reasonable approximation thereof) be proven for hybrid optimization algorithms? In what contexts are vanishing constraints required and how do they complicate the theory of hybrid optimization?

Hybrid optimization is probably a requirement for a wide variety of practical optimization problems. It is not evident that it would pose a significant theoretical problem. Therefore, this is most likely one of the most straightforward extensions to the theory presented in this thesis.

Combinatorial Constraints

One of the advantages that combinatorial integral approximation methods have over optimization in measure spaces is the ease with which combinatorial constraints such as limits on the number of control switches or dwell time constraints can be enforced in the rounding problem. This advantage diminuishes somewhat in a context where CIA is applied on multi-dimensional domains, because it relies heavily on "unravelling" the multi-dimensional domain onto a one-dimensional interval. This process is likely equivalent to a pushforward through a generator geodesic. When pulling the result back to the original domain, some combinatorial constraints, such as limits on the number of switches, can easily lose their meaning.

There are some constraints, such as limits on total dwell time, that we can already enforce because they are simply upper and lower bounds on the measure of control sets.

Arguably the most interesting type of combinatorial constraint that we could meaningfully enforce in our setting are dwell time constraints. Dwell time constraints prescribe that, if a certain control is switched either on or off, then it should remain in that state for either a certain minimal or maximal time. The most direct translation of such constraints to our setting would be constraints that apply an upper or lower bound to the measure of contiguous components of sets.

The main problem with such constraints is that there are situations in which an infinitesimally small step can cause these quantities to jump, e.g., by connecting two contiguous sets together or splitting one contiguous set in two. Therefore, these constraints are likely not continuous and certainly not continuously differentiable. Even the introduction of subgradients may not help with such constraints, because subgradients are only meaningfully defined for functions that do not have jumps.

It is conceivable that we could relax these constraints into constraints that have approximately the same effect while being continuously differentiable. Such constraints could, for instance, be inspired by PDEs that model heat conduction, electric potential, or the transmission of mechanical strain. If, for instance, we modeled the heat exchange between two conductors with an insulator between them, then that exchange is almost non-existent if the conductors are clearly separated. As they grow closer to one another, heat would slowly start seeping between them. Finally, when they are connected, the exchange would become much stronger. Assessing contiguity through such a mechanism would effectively treat sets that are *almost connected* as if they had "fractional connectedness." It would sacrifice accuracy but could turn the constraint into a differentiable one.

Such constraints, even if they were differentiable, would likely never be convex. There is the remote possibility that there may be an infinite-dimensional Branch and Bound algorithm similar to the one that we will propose in the next section that could rigorously solve problems with dwell time constraints by branching on steps where sets are either connected or separated and treating both subproblems in separate subtrees. This would very likely be difficult to theoretically justify or practically implement.

We summarize this issue as follows:

> What properties do combinatorial constraints such as dwell time constraints have within the measure space setting? Are there combinatorial constraints that could benefit from subgradient methods? Can some combinatorial constraints be weakened to approximate their effect while turning them into differentiable constraints? Is it possible to solve problems with such constraints with infinite-dimensional Branch and Bound methods?

The benefit of such constraints is questionable compared to the amount of effort that would have to be spent to implement them. However, they could provide a good testbed to learn more about infinite-dimensional optimization. The potential use case for infinite-dimensional Branch and Bound appears particularly interesting considering the other uses of such an algorithm that we will discuss next.

Infinite-Dimensional Spatial Branch and Bound

Finally, there is the issue of nonconvex optimization. In "conventional" NLP theory, there exist various approaches to solving nonconvex optimization problems. Many of them are stochastic in nature and do not provide any sort of rigorous quality guarantee for their solutions. One method that stands in contrast to this trend is *spatial Branch and Bound*.

As the name suggests, spatial Branch and Bound is a Branch and Bound method. In spatial Branch and Bound, the lower bound for each node of the search tree is derived by minimizing a convex functional that underestimates the actual objective under convex relaxations of the constraints. Upper bounds are derived by evaluating the actual objective function. If the difference between the underestimators and the actual functionals is deemed too great, then the algorithm branches and calculates tighter underestimators for the child nodes.

The difference between spatial Branch and Bound and regular Branch and Bound is that branching is not limited to integer variables. The algorithm can branch on any variable and there is no guaranteed upper bound on the number of times that it can branch on any given variable. This makes spatial Branch and Bound much more complex than regular Branch and Bound.

One aspect in which most spatial and regular Branch and Bound methods are similar is that they usually branch on single scalar variables. In infinite-dimensional settings such as ours, this is not an option because our optimization variables do not consist of distinct scalar components.

Although we do not have scalar components to branch on, our settings tend to be "essentially finite-dimensional" in the sense that they can be discretized. A discretization considers a finite-dimensional space that is very close to objects of interest in the original space. The fact that we can use discretizations suggests

that, at least situationally, we have a finite number of "variables" that are much more significant to the quality of a solution than others.

If we could dynamically detect these significant branching directions as linear forms (or signed measures for similarity spaces), we could still branch on them even though we reside in an infinite-dimensional space where there are no distinct scalar components. Such a branching could still yield good results if the selection of branching directions is good enough.

Our questions are:

> Can we dynamically identify good branching directions for spatial Branch and Bound in infinite-dimensional vector spaces or similarity spaces? Is this a viable path for nonconvex optimization in similarity spaces?

This is not the most pressing research question that we present here. However, it is certainly very interesting from a theoretical standpoint. Infinite-dimensional Branch and Bound could be an exciting research topic. Even if it is not a viable method for infinite-dimensional optimization, the branching methods that we might discover during research could also feed back into finite-dimensional Branch and Bound solvers and thereby improve our ability to solve conventional nonconvex and mixed-integer problems.

5.3 CONCLUSIONS AND CLOSING WORDS

At this point, we are at the end of this expedition into the realm of iterative optimization in measure spaces. We have seen that, by slightly relaxing the assumptions made about the search space, very conventional and well-established optimization methods can become powerful tools to efficiently tackle the complexity of infinite-dimensional optimization with discretely valued variables.

We have seen that this does not require the use of discrete optimization methods, but can instead be seen as an extension of conventional optimization on continua. We have found an appropriate search space for this endeavor and have developed an extensive theoretical toolkit for arguing within it. This toolkit is the theory of measure space geodesics. We have then used this toolkit to transfer two simple optimization methods, one unconstrained and one constrained, to the measure space setting. Finally, we have tested them on two proof-of-concept test problems.

As a by-product of this thesis, we have developed PYCOIMSET, a Python library that is now available as open-source software that anyone can use to apply the algorithms to their own problems.

As we had stated in the opening chapter, the goal of this thesis is not to create entirely new algorithms, but rather to show the potential that can still be unlocked in old algorithms. We have demonstrated that simple iterative optimization schemes can be brought to bear on what is often spoken of as "integer optimization" by applying them in a metric space instead of the usual vector space setting.

This thesis document is hopefully not the end of that journey. I hope that it can serve others as a pathway through the initial theoretical barrier into a wide-open field of unexplored methods and new insights. After all, if the steepest descent method and the penalty method can be transferred to this setting, then why not all the other methods?

I hope that this thesis will inspire others to continue where it ends. If you, the reader, should feel inspired to continue beyond this point, then I wish you good success on your journey beyond, and I hope that my work makes your path a little easier than it would otherwise have been.

Appendix A

Additional Theory

In this chapter, we develop additional theoretical ideas that are beyond the scope of this thesis. The purpose of this appendix is to preserve unfinished or unused ideas for future research. Therefore, the ideas presented in here are not as rigorously argued and may depend on unproven assumptions or hypotheses. Where this is true, we note such assumptions explicitly.

A.1 HIGHER ORDER DERIVATIVES

In Section 2.4, we introduce the concept of differentiable set functionals. In Definition 2.4.1, we use the Taylor criterion as the defining property of a differentiable set functional. We stopped the Taylor expansion at the first derivative. However, it is reasonable to ask whether this definition can be expanded to higher orders of differentiability.

In Definition 2.4.1, we had defined derivatives as finite signed measures. The set of all finite signed measures on a measurable space (X, Σ) along with the total variation (see Definition 2.1.7) is a real Banach space. Because we are generally only interested in gradient measures that have density functions with respect to a measure $\mu\colon \Sigma \to \mathbb{R}_{\geq 0}$, i.e., that are absolutely continuous with respect to μ, for the remainder of this section we define

$$\begin{aligned}\mathbb{S}(\Sigma) &:= \{\phi\colon \Sigma \to \mathbb{R} \,|\, \phi \text{ is a finite signed measure}\},\\ \mathbb{S}_\mu(\Sigma) &:= \big\{\phi \in \mathbb{S}(\Sigma) \,\big|\, \phi \ll \mu\big\}.\end{aligned}$$

It is known that, when paired with the canonical addition and scaling operations and the total variation as a norm, $\mathbb{S}(\Sigma)$ becomes a real Banach space. It is easy to demonstrate that $\mathbb{S}_\mu(\Sigma)$ is closed under addition and scaling. Therefore, $\mathbb{S}_\mu(\Sigma)$ is a subspace of $\mathbb{S}(\Sigma)$. In order to show that $\mathbb{S}_\mu(\Sigma)$ is a Banach space, we have to demonstrate that it is sequentially closed.

Because sequential and topological closedness are equivalent in metric spaces, we can demonstrate the former by proving the latter. Let $\phi \in \mathbb{S}_\mu(\Sigma)$ and let $\nu \in \mathbb{S}(\Sigma) \setminus \mathbb{S}_\mu(\Sigma)$. We can perform a Lebesgue decomposition (see, e.g., [BS20, Thm. 3.9.5], [BK10, Prop. 9.3], [Kub15, Thm. 7.10]) of ϕ and ν with respect to μ. This gives us unique signed measures $\phi_1, \phi_2, \nu_1, \nu_2$ such that

$$\begin{aligned}\phi &= \phi_1 + \phi_2, & \phi_1 &\ll \mu, & \phi_2 &\perp \mu,\\ \nu &= \nu_1 + \nu_2, & \nu_1 &\ll \mu, & \nu_2 &\perp \mu,\end{aligned}$$

where "$\alpha \perp \beta$" signifies that α and β are "mutually singular." This means that there exists $A \in \Sigma$ such that, $|\alpha|(A) = 0$ and $|\beta|(A^\complement) = 0$. Because $\phi \in \mathbb{S}_\mu(\Sigma)$ and because the decomposition is unique, we know that $\phi_1 = \phi$ and $\phi_2 = 0$. However, because $\nu \notin \mathbb{S}_\mu(\Sigma)$, we know that $\nu_2 \neq 0$. Let $A \in \Sigma$ be such that $|\mu|(A) = 0$ and $|\nu_2|(A^\complement) = 0$. We have

$$\begin{aligned}
\|\phi - \nu\| &= \|\phi_1 - \nu_1 + \underbrace{\phi_2}_{=0} - \nu_2\| \\
&= \|\phi_1 - \nu_1 - \nu_2\| \\
&= |\phi_1 - \nu_1 - \nu_2|(A) + \underbrace{|\phi_1 - \nu_1 - \nu_2|(A^\complement)}_{\geq 0} \\
&\geq |\phi_1 - \nu_1 - \nu_2|(A).
\end{aligned}$$

Let $\Pi \subseteq \Sigma^{\mathbb{N}}$ be the set of all countable measurable partitions of A. We have

$$\begin{aligned}
\|\phi - \nu\| &\geq |\phi_1 - \nu_1 - \nu_2|(A) \\
&= \sup_{B \in \Pi} \sum_{i=1}^{\infty} \Big| \underbrace{\phi_1(B_i) - \nu_1(B_i)}_{=0 \text{ because } \mu(B_i)=0} - \nu_2(B_i) \Big| \\
&= \sup_{B \in \Pi} \sum_{i=1}^{\infty} \big| \nu_2(B_i) \big| \\
&= |\nu_2|(A) \\
&= |\nu_2|(A) + \underbrace{|\nu_2|(A^\complement)}_{=0} \\
&= \|\nu_2\| \\
&> 0
\end{aligned}$$

This means that ν is strictly separated from $\mathbb{S}_\mu(\Sigma)$. Since this holds for all $\nu \in \mathbb{S}(\Sigma) \setminus \mathbb{S}_\mu(\Sigma)$, $\mathbb{S}_\mu(\Sigma)$ is a closed subspace of $\mathbb{S}(\Sigma)$ and therefore a Banach space.

Because $\mathbb{S}_\mu(\Sigma)$ is a real Banach space, we can have vector measures that map into $\mathbb{S}_\mu(\Sigma)$, which is the space in which all gradient measures reside. In the spirit of Fréchet derivatives, where the second derivative is a bounded linear operator that maps into the space of bounded linear maps, we can define the second set derivative as a vector measure of bounded variation that maps to $\mathbb{S}_\mu(\Sigma)$.

Definition A.1.1 (Spaces of Vector Measures).
Let $(X,\ \Sigma,\ \mu)$ be a measure space, for every real Banach space V, we define

$$\begin{aligned}
\mathbb{V}(\Sigma, V) &:= \{\nu \colon \Sigma \to V \mid \nu \text{ vector measure},\ \|\nu\| < \infty\}, \\
\mathbb{V}_\mu(\Sigma, V) &:= \big\{\nu \in \mathbb{V}(\Sigma, V) \,\big|\, \nu \ll \mu\big\}.
\end{aligned}$$ ◁

We evidently have $\mathbb{V}(\Sigma, \mathbb{R}) = \mathbb{S}(\Sigma)$ and $\mathbb{V}_\mu(\Sigma, \mathbb{R}) = \mathbb{S}_\mu(\Sigma)$. To iterate to higher orders, we require a result of the following kind.

Hypothesis A.1.2 (Completeness of Spaces of Vector Measures).
Let $(X,\ \Sigma,\ \mu)$ be a measure space, and let V be a real Banach space, then $\mathbb{V}_\mu(\Sigma, V)$ is a real Banach space when equipped with the total variation norm. ◁

If this hypothesis is true, then the following definition becomes viable.

Definition A.1.3 (Recursive Spaces of Vector Measures).
Let $(X,\ \Sigma,\ \mu)$ be a measure space, and let V be a real Banach space. We define

$$\begin{aligned}\mathbb{V}_\mu^1(\Sigma,V) &:= \mathbb{V}_\mu(\Sigma,V),\\ \mathbb{V}_\mu^{i+1}(\Sigma,V) &:= \mathbb{V}_\mu\big(\Sigma,\mathbb{V}_\mu^i(\Sigma,V)\big) \quad \forall i\in\mathbb{N}\colon i>1.\end{aligned}$$ ◁

Under this definition, the $(i+1)$-st derivative of a set functional can be conceived of as a vector measure in $\mathbb{V}^{i+1}(\Sigma,\mathbb{R})$ whose output locally approximates the change of the i-th derivative. To formalize this definition, we introduce the concept of a differentiable vector-valued mapping on a similarity space.

Definition A.1.4 (Vector-Valued Differentiability).
Let $(X,\ \Sigma,\ \mu)$ be a measure space, let V be a real Banach space. We refer to a mapping $F\colon \Sigma/_{\sim_\mu} \to V$ as *differentiable in* $U\in\Sigma/_{\sim_\mu}$ if there exists an absolutely continuous vector measure $\nabla F(U)\in\mathbb{V}_\mu(\Sigma,V)$ of bounded variation such that

$$F(V) = F(U) + \nabla F(U)(U\mathbin{\triangle} V) + o\big(\mu(U\mathbin{\triangle} V)\big) \quad \forall V\in\Sigma/_{\sim_\mu}.$$

We refer to F as *differentiable on* $N\subseteq\Sigma/_{\sim_\mu}$ if F is differentiable in every $U\in N$. We refer to F as *differentiable* if F is differentiable on $\Sigma/_{\sim_\mu}$. ◁

Definition A.1.5 (Higher Order Differentiability).
Let $(X,\ \Sigma,\ \mu)$ be a measure space, let V be a real Banach space, let $U\in\Sigma/_{\sim_\mu}$, and let $F\colon\Sigma/_{\sim_\mu}\to V$. If F is differentiable in U, then we refer to $\nabla^1F(U) := \nabla F(U)$ as the *first derivative of* F *in* U.

Let $i\in\mathbb{N}$, and let $N\subseteq\Sigma/_{\sim_\mu}$ be a neighborhood of U such that $\nabla^iF(U')$ exists for all $U'\in N$ and such that the mapping $\nabla^iF\colon N\to\mathbb{V}_\mu^i(\Sigma,V)$ is differentiable in U. Then we refer to

$$\nabla^{i+1}F(U) := \nabla(\nabla^iF)(U)\in\mathbb{V}_\mu^{i+1}(\Sigma,V)$$

as the $(i+1)$*-st derivative of* F *in* U. We refer to F as i *times differentiable in* U if $\nabla^iF(U)$ exists, as i *times differentiable on* $N\subseteq\Sigma/_{\sim_\mu}$ if $\nabla^iF(U')$ exists for all $U'\in N$, and as i *times differentiable* if F is i times differentiable on $\Sigma/_{\sim_\mu}$. ◁

It is evident that Definition 2.4.1 is a special case of Definition A.1.4 with $V=\mathbb{R}$. Accordingly, we can use Definition A.1.5 to define higher-order derivatives of set functionals. When transferring the concept of continuous differentiability, we have to take into account the local inversion behavior of set derivatives. We had defined the locally inverted difference variation (LIDV, see Definition 2.4.3) with signed measures in mind. However, because the variation measure is an equally valid concept for vector measures, the LIDV can straightforwardly be applied to vector measures as well.

Definition A.1.6 (Locally Inverted Difference Variation).
Let $(X,\ \Sigma,\ \mu)$ be a measure space, let V be a real Banach space, and let $R\in\Sigma/_{\sim_\mu}$. For $\varphi,\nu\in\mathbb{V}(\Sigma,V)$, we refer to the measure $(\varphi \underset{R}{\ominus} \nu)$ with

$$(\varphi \underset{R}{\ominus} \nu)(D) := |\varphi-\nu|(D\setminus R) + |\varphi+\nu|(D\cap R) \quad \forall D\in\Sigma$$

as the R*-locally inverted difference variation* between φ and ν. ◁

Definition A.1.7 (Continuous Differentiability).
Let (X, Σ, μ) be a measure space, let V be a real Banach space, let $N \subseteq \Sigma/_{\sim_\mu}$ be open, and let $F\colon \Sigma/_{\sim_\mu} \to V$ be differentiable on N. We refer to F as *continuously differentiable in* $U \in N$ if for every $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\big(\nabla F(U) \underset{U \triangle U'}{\ominus} \nabla F(U')\big)(W) \leq \varepsilon \cdot \mu(W) \quad \forall U', W \in \Sigma/_{\sim_\mu} \colon \mu(U \triangle U') \leq \delta. \qquad \triangleleft$$

For continuously differentiable mappings, a result along the lines of Taylor's theorem may then be possible.

Hypothesis A.1.8 (Taylor's Theorem).
Let (X, Σ, μ) be a finite atomless measure space, let $n \in \mathbb{N}$, let V be a real Banach space, let $N \subseteq \Sigma/_{\sim_\mu}$ be open, let $U \in N$, and let $F\colon \Sigma/_{\sim_\mu} \to V$ be n times continuously differentiable on N. Then we have

$$F(U') = F(U) + \sum_{i=1}^{n} \frac{1}{i!} \nabla^i F(U) \underbrace{(U \triangle U') \ldots (U \triangle U')}_{i \text{ times}} + o\big(\mu(U \triangle U')^n\big) \quad \forall U' \in N. \qquad \triangleleft$$

Proving Hypothesis A.1.8 is essential to demonstrating the merit of using higher order derivatives in model functions for optimization. However, the hypothesis is likely not straightforward to prove. By requiring N to be open, we ensure that an ε-ball around U is within N. This ensures that, for U' within that ε-ball, the entirety of a connecting geodesic between U and U' is within the ball. This lays the groundwork for an error estimate based on a polygon chain argument, which could be used to prove that the residual scales with the n-th power of the distance. We do not perform this proof here.

Instead, we turn our attention to the second question that must be answered before further investigation into higher order derivatives is warranted. That question is whether such higher order derivatives, if shown to exist, are actually useful for optimization.

At first, this may appear to be a trivial question. Of course, incorporating second derivatives into our methods *should* improve the accuracy of our model functions and, by extension, the quality of the steps made by iterative optimization algorithms. However, rephrasing the question purely in terms of model accuracy misses a crucial point: Improving the accuracy of the model function is of no benefit unless we can formulate an algorithm for step determination that can benefit from higher order derivatives.

The first step towards finding a way to benefit from knowledge about higher order derivatives is to prove that they have a tractable form. For first order derivatives, we chose the setting such that the derivative measure always has a density function. Due to the way in which we have defined higher order derivatives, it should be possible to prove that higher order derivatives also have density functions.

Hypothesis A.1.9 (Higher Order Derivative Measures and Densities).
Let (X, Σ, μ) be a measure space, let $n \in \mathbb{N}$, let $U \in \Sigma/_{\sim_\mu}$, and let $F\colon \Sigma/_{\sim_\mu} \to \mathbb{R}$ be n times differentiable in U. Let

$$\big(X^n, \Sigma^{(n)}, \mu^{(n)}\big)$$

be the product measure space of n instances of (X, Σ, μ). Then there exists a unique signed measure $\varphi\colon \Sigma^{(n)} \to \mathbb{R}$ with $\varphi \ll \mu^{(n)}$ such that

$$\nabla^n F(U)\underbrace{(D_n)\dots(D_1)}_{n \text{ evaluations}} = \varphi\Big(\bigtimes_{i=1}^{n} D_i\Big) \quad \forall (D_i)_{i\in[n]} \in \Sigma^n. \qquad \triangleleft$$

A proof for this hypothesis would likely be based on an extension argument where φ is first defined on an algebra of sets that consists only of the finite unions of measurable rectangles:

$$\mathscr{R} \coloneqq \Big\{ \bigcup_{i=1}^{m} \bigtimes_{j=1}^{n} A_{i,j} \,\Big|\, m \in \mathbb{N}, (A_{i,j})_{i\in[m], j\in[n]} \in \Sigma^{m\times n} \Big\}.$$

For a measurable rectangle, we would define

$$\varphi\Big(\bigtimes_{i=1}^{n} D_i\Big) \coloneqq \nabla^n F(U)(D_n)\dots(D_1) \quad \forall (D_i)_{i\in[n]} \in \Sigma^n.$$

We would then extend this to countable disjoint unions of rectangles. For all $i \in \mathbb{N}$, let $A_i \in \Sigma^n$ and let $R_i \coloneqq \bigtimes_{j=1}^{n} A_{i,j}$ such that $R_i \cap R_j = \emptyset$ for $i \neq j$. Then we would define

$$\varphi\Big(\bigcup_{i=1}^{\infty} R_i\Big) \coloneqq \sum_{i=1}^{\infty} \varphi(R_i).$$

The first challenge is proving that this is well-defined. To establish that, we would have to show that a decomposition of a rectangle into disjoint rectangles would yield the same value as the original rectangle. Expanding upon this, by showing that every subdivision of a union of disjoint rectangles yields the same value as the original union, we could show well-definedness over all of $\mathscr{R}$, because any two unions yielding the same result would have the same value as their joint refinement. For non-disjoint unions, we would have to rewrite them as disjoint unions by subdividing the rectangles into their overlapping and non-overlapping parts.

Once the definition is properly extended to arbitrary countable unions of elements in $\mathscr{R}$, an extension argument along the lines of Carathéodory's extension theorem would have to be used to extend φ to a signed measure on $\Sigma^{(n)}$. Absolute continuity with respect to $\mu^{(n)}$ would likely follow from the definition of the outer measure as the infimum over the measures of all encompassing sets in $\mathscr{R}$ and the fact that disjoint unions of measurable rectangles can only have measure zero if each participating rectangle has measure zero along at least one axis.

A.2 MFCQ IMPLIES GCQ

In Section 3.2.1.1, we transfer the Karush-Kuhn-Tucker (KKT) conditions to our setting. An important ingredient for KKT conditions are so-called constraint qualifications. In our case, we primarily focused on the Mangasarian-Fromovitz Constraint Qualification (MFCQ, see Definition 3.2.15).

We were able to prove that the MFCQ always implies equality between the tangent cone and the linearized tangent cone. However, we were only able to prove that this implies equality of the normal and linearized normal cone in the

case in which at most one constraint is active. This is because we were only able to demonstrate the inclusion

$$\big(\widetilde{T}_G(U)\big)^\circ \subseteq \widetilde{N}_G(U)$$

for points $U \in \Sigma_\sim$ where at most one constraint satisfies $G_j(U) = 0$. This is a substantial restriction. In this section, we outline how a proof of the more general Hypothesis 3.2.20 could proceed. Before we begin, we have to prove some preliminary results.

A.2.1 Preliminaries: Constant Mean Property

In Section 2.3.5, we introduced constant mean geodesics. Constant mean geodesics are specifically designed such that a specific signed measure will behave like a linear functional along their path (see Definition 2.3.56). As far as we have seen so far, this is a specific property that only holds for some signed measures along specifically constructed geodesics.

However, this is not quite true. The constant mean property is not a property of the constant mean geodesic. Rather, it is a property that applies for all elements of its generated similarity space (see Definition 2.3.38).

Lemma A.2.1.
Let (X, Σ, μ) be a finite atomless measure space, let $f \in L^1(\Sigma, \mu)$, let $I \subseteq \mathbb{R}$ be an interval, let $\gamma\colon I \to {}^{\Sigma}\!/_{\sim_\mu}$ be a constant mean geodesic of f, and let $M \in \mathbb{R}$ be such that

$$\int_{\gamma(s)\triangle\gamma(t)} f \,\mathrm{d}\mu = M \cdot \mu\big(\gamma(s) \triangle \gamma(t)\big) \quad \forall s, t \in I.$$

Then we have

$$\int_A f \,\mathrm{d}\mu = M \cdot \mu(A) \quad \forall A \in \sigma(\gamma).$$

◁

PROOF. Let $g \in \mathscr{G}(X, \Sigma, \mu, I)$ be a GLSF such that

$$\{g \leq t\} \in \gamma(t) \quad \forall t \in I.$$

Let $C_\gamma \geq 0$ be a geodesic constant of γ. It is sufficient to show that for every $B \in \mathscr{B}(I)$, we have

$$\int_{g^{-1}(B)} f \,\mathrm{d}\mu = M \cdot \mu\big(g^{-1}(B)\big).$$

For every interval $J \subseteq I$,

$$\int_{g^{-1}(J)} f \,\mathrm{d}\mu = M \cdot \mu\big(g^{-1}(J)\big)$$

follows from the fact that γ is a constant mean geodesic and the definition of M. Let $U \subseteq \mathbb{R}$ be an open set. Then U is a countable disjoint union of open intervals:

$$U = \bigcup_{i=1}^{\infty} J_i$$

where $J_i \subseteq \mathbb{R}$ is an open interval for each $i \in \mathbb{N}$. For each $i \in \mathbb{N}$ the intersection $I \cap J_i$ is an interval that is a subset of I. Therefore, we have

$$\begin{aligned}\int_{g^{-1}(I\cap U)} f \,\mathrm{d}\mu &= \sum_{i=1}^{\infty} \int_{g^{-1}(I\cap J_i)} f \,\mathrm{d}\mu \\ &= M \cdot \sum_{i=1}^{\infty} \mu\big(g^{-1}(I \cap J_i)\big) \\ &= M \cdot \mu\big(g^{-1}(I \cap U)\big).\end{aligned}$$

Let $F \subseteq \mathbb{R}$ be a closed set. Then $F^{\complement}$ is open and we have

$$\begin{aligned}\int_{g^{-1}(I\cap F)} f \,\mathrm{d}\mu &= \int_{g^{-1}(I)} f \,\mathrm{d}\mu - \int_{g^{-1}(I\cap F^{\complement})} f \,\mathrm{d}\mu \\ &= M \cdot \Big(\mu\big(g^{-1}(I)\big) - \mu\big(g^{-1}(I \cap F^{\complement})\big)\Big) \\ &= M \cdot \mu\big(g^{-1}(I) \setminus g^{-1}(I \setminus F)\big) \\ &= M \cdot \mu\big(g^{-1}(I \cap F)\big).\end{aligned}$$

Let $B \in \mathscr{B}(I)$ and let $B^* \in \mathscr{B}(\mathbb{R})$ be such that $B = B^* \cap I$. Let $\varepsilon > 0$. Due to the absolute continuity of the Lebesgue integral, there exists $\delta > 0$ such that

$$\int_A |f| \,\mathrm{d}\mu \leq \frac{\varepsilon}{2} \quad \forall A \in \Sigma \colon \mu(A) \leq \delta.$$

Let $\delta^* \coloneqq \min\{\delta, \frac{\varepsilon}{2 \cdot C_\gamma \cdot |M|}\}$, where $\frac{\varepsilon}{2 \cdot C_\gamma \cdot |M|} = \infty$ for $C_\gamma \cdot |M| = 0$. We know that $\delta^* < \infty$ because $\delta < \infty$. There exist an open set $U_\varepsilon \subseteq \mathbb{R}$ and a closed set $F_\varepsilon \subseteq \mathbb{R}$ such that $F_\varepsilon \subseteq B^* \subseteq U_\varepsilon$ and

$$\lambda(U_\varepsilon \setminus F_\varepsilon) \leq \delta^*.$$

This implies that

$$\begin{aligned}\lambda(\underbrace{U_\varepsilon \setminus B^*}_{\subseteq U_\varepsilon \setminus F_\varepsilon}) &\leq \delta^*, \\ \lambda(\underbrace{B^* \setminus F_\varepsilon}_{\subseteq U_\varepsilon \setminus F_\varepsilon}) &\leq \delta^*.\end{aligned}$$

We have

$$\begin{aligned}&\left|\int_{g^{-1}(B)} f \,\mathrm{d}\mu - M \cdot \mu\big(g^{-1}(B)\big)\right| \\ &\quad= \left|\int_{g^{-1}(B^*\cap I)} f \,\mathrm{d}\mu - M \cdot \mu\big(g^{-1}(B^* \cap I)\big)\right| \\ &\quad= \left|\underbrace{\int_{g^{-1}(U_\varepsilon \cap I)} f \,\mathrm{d}\mu}_{= M \cdot \mu(g^{-1}(U_\varepsilon \cap I))} - \int_{g^{-1}((U_\varepsilon \setminus B^*)\cap I)} f \,\mathrm{d}\mu - M \cdot \mu\big(g^{-1}(B^* \cap I)\big)\right| \\ &\quad= \left|\int_{g^{-1}((U_\varepsilon \setminus B^*)\cap I)} f \,\mathrm{d}\mu - M \cdot \mu\Big(g^{-1}\big((U_\varepsilon \setminus B^*) \cap I\big)\Big)\right| \\ &\quad\leq \int_{g^{-1}((U_\varepsilon \setminus B^*)\cap I)} |f| \,\mathrm{d}\mu + |M| \cdot \underbrace{\mu\Big(g^{-1}\big((U_\varepsilon \setminus B^*) \cap I\big)\Big)}_{= C_\gamma \cdot \lambda((U_\varepsilon \setminus B^*)\cap I)}\end{aligned}$$

$$
\begin{aligned}
&= \int_{g^{-1}((U_\varepsilon \setminus B^*)\cap I)} |f| \,\mathrm{d}\mu + C_\gamma \cdot |M| \cdot \underbrace{\lambda\big((U_\varepsilon \setminus B^*) \cap I\big)}_{\leq \lambda(U_\varepsilon \setminus B^*) \leq \delta^*} \\
&\leq \int_{g^{-1}((U_\varepsilon \setminus B^*)\cap I)} |f| \,\mathrm{d}\mu + \underbrace{C_\gamma \cdot |M| \cdot \delta^*}_{\leq \frac{\varepsilon}{2}} \\
&\leq \int_{g^{-1}((U_\varepsilon \setminus B^*)\cap I)} |f| \,\mathrm{d}\mu + \frac{\varepsilon}{2}.
\end{aligned}
$$

At this point, we make use of the fact that $\mu(g^{-1}((U_\varepsilon \setminus B^*) \cap I)) \leq \lambda(U_\varepsilon \setminus B^*) \leq \delta^* \leq \delta$ because this implies that

$$
\begin{aligned}
&\left| \int_{g^{-1}(B)} f \,\mathrm{d}\mu - M \cdot \mu\big(g^{-1}(B)\big) \right| \\
&\quad \leq \underbrace{\int_{g^{-1}((U_\varepsilon \setminus B^*)\cap I)} |f| \,\mathrm{d}\mu}_{\leq \frac{\varepsilon}{2}} + \frac{\varepsilon}{2} \\
&\quad \leq \varepsilon.
\end{aligned}
$$

Because this holds for every $\varepsilon > 0$, we have

$$
\int_{g^{-1}(B)} f \,\mathrm{d}\mu = M \cdot \mu\big(g^{-1}(B)\big) \quad \forall B \in \mathscr{B}(I).
$$

For each $A \in \sigma(g)$, there exists $B \in \mathscr{B}(I)$ such that $A = g^{-1}(B)$. Therefore, we have

$$
\int_A f \,\mathrm{d}\mu = M \cdot \mu(A) \quad \forall A \in \sigma(g).
$$

The claim then follows because $\sigma(\gamma) = {}^{\sigma(g)}\!/_{\sim_\mu}$ and because the integral is well-defined on similarity spaces. □

The significance of Lemma A.2.1 is easily underestimated. It decouples the constant mean property from a specific geodesic and transfers it to an entire similarity space. This is especially significant in light of Theorem 2.3.47, because that theorem implies that the generated similarity space of any rearrangement of γ generates a subspace of $\sigma(\gamma)$. This means that the constant mean property is preserved under geodesic rearrangement.

A.2.2 Preliminaries: Simultaneous Alignment

Because the constant mean property is preserved under rearrangement, if we can rearrange an existing geodesic such that it becomes constant mean, then we could iterate that process to obtain geodesics that are constant mean with respect to several functions at once. We will refer to this process as "simultaneous alignment" because it realigns a geodesic such that it acts as a constant mean geodesic for multiple functions simultaneously.

The alignment process for a single function is remarkably simple. We first note that the integral of an integrable function along a geodesic is always absolutely continuous. This is a simple consequence of the Lipschitz continuity of geodesics in conjunction with the absolute continuity of the Lebesgue integral. For a definition of absolute continuity, we refer to [BS20, Def. 4.3.1].

Lemma A.2.2.
Let (X, Σ, μ) be a measure space, let $f \in L^1(\Sigma, \mu)$, let $I \subseteq \mathbb{R}$ be an interval, and let $\gamma\colon I \to \Sigma/_{\sim_\mu}$ be a geodesic. Then the mapping $F\colon I \to \mathbb{R}$ with

$$F(t) \coloneqq \int_{\gamma(t)} f \,\mathrm{d}\mu \quad \forall t \in I$$

is absolutely continuous in the sense that for every $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\sum_{i=1}^{\infty} \big|F(b_i) - F(a_i)\big| \leq \varepsilon$$

for every sequence of pairwise disjoint intervals $(a_i,\ b_i) \subseteq I$ with $a_i \leq b_i$ for all $i \in \mathbb{N}$ and

$$\sum_{i=1}^{\infty} |b_i - a_i| \leq \delta.$$

◁

Proof. Let $g \in \mathscr{G}(X, \Sigma, \mu, I)$ be a GLSF representation of γ, and let $C_\gamma \geq 0$ be a geodesic constant of γ. Let $\varepsilon > 0$. Due to the absolute continuity of the integral, there exists $\delta^* > 0$ such that

$$\int_A |f| \,\mathrm{d}\mu \leq \varepsilon \quad \forall A \in \Sigma\colon \mu(A) \leq \delta^*.$$

If $C_\gamma = 0$, then we set $\delta \coloneqq \infty$. Otherwise, we set $\delta \coloneqq \frac{\delta^*}{C_\gamma}$. In either case, this ensures that

$$C_\gamma \cdot \lambda(A) \leq \delta^* \quad \forall A \in \mathscr{B}(I)\colon \lambda(A) \leq \delta.$$

Let $\big((a_i,\ b_i)\big)_{i\in\mathbb{N}}$ be a sequence of pairwise disjoint intervals such that $a_i \leq b_i$ and $(a_i,\ b_i) \subseteq I$ for all $i \in \mathbb{N}$, as well as

$$\sum_{i=1}^{\infty} |b_i - a_i| \leq \delta.$$

We have

$$\lambda\Big(\underbrace{\bigcup_{i=1}^{n} (a_i,\ b_i)}_{=:\widetilde{A}\in\mathscr{B}(I)}\Big) = \sum_{i=1}^{\infty} |b_i - a_i| \leq \delta$$

and therefore

$$\mu\big(g^{-1}(A)\big) = C_\gamma \cdot \lambda(A) \leq \delta^*.$$

Because γ is canonical and because the Lebesgue integral is countably additive, we have

$$\begin{aligned}
\sum_{i=1}^{\infty} \big|F(b_i) - F(a_i)\big| &= \sum_{i=1}^{\infty} \left| \int_{\gamma(b_i)} f \,\mathrm{d}\mu - \int_{\gamma(a_i)} f \,\mathrm{d}\mu \right| \\
&= \sum_{i=1}^{\infty} \left| \int_{\gamma(a_i) \triangle \gamma(b_i)} f \,\mathrm{d}\mu \right| \\
&\leq \sum_{i=1}^{\infty} \int_{\gamma(a_i) \triangle \gamma(b_i)} |f| \,\mathrm{d}\mu \\
&= \sum_{i=1}^{\infty} \int_{g^{-1}((a_i,\ b_i))} |f| \,\mathrm{d}\mu
\end{aligned}$$

$$= \int_{g^{-1}(A)} |f| \, \mathrm{d}\mu$$
$$\leq \varepsilon. \qquad \square$$

We note that the statement of Lemma A.2.2 is much stronger than that required by the cited definition of absolute continuity. It does not require the parameter interval to be bounded and works for countably infinite collections of intervals. However, if the parameter interval is compact, then this is sufficient to show that F is absolutely continuous.

This is relevant because absolutely continuous functions are differentiable almost everywhere in the sense that they have an integrable derivative function whose integral can be used to calculate differences between function values at different points in the parameter interval.

Theorem A.2.3 (Lebesgue's Theorem [BS20, Thm. 4.3.7]).
A function f is absolutely continuous on $[a,\ b]$ if and only if there exists a function g integrable on $[a,\ b]$ such that

$$f(t) = f(a) + \int_a^t g(s) \, \mathrm{d}s \quad \forall t \in [a,\ b]. \qquad \triangleleft$$

If we have such a derivative g, then we can use it to generate a constant mean geodesic and use that to rearrange the original geodesic.

Lemma A.2.4 (Single Function Alignment).
Let (X, Σ, μ) be a measure space, let $I := [a,\ b] \subseteq \mathbb{R}$ be a compact interval, let $\gamma \colon I \to \Sigma/_{\sim_\mu}$ be a canonical geodesic, and let $f \in L^1(\Sigma,\mu) \cap L^\infty(\Sigma,\mu)$. Then there exists a parameter geodesic $\rho \colon [0,\ b-a] \to \mathscr{B}(I)/_{\sim_\lambda}$ such that the parameterization $\gamma \odot \rho$ of γ by ρ is a canonical geodesic that satisfies $\mathrm{TV}(\gamma \odot \rho) = \mathrm{TV}(\gamma)$ and

$$\int_{(\gamma\odot\rho)(s)\triangle(\gamma\odot\rho)(t)} f \, \mathrm{d}\mu = M \cdot \mu\big((\gamma \odot \rho)(s) \triangle (\gamma \odot \rho)(t)\big) \quad \forall s,t \in [0,\ b-a]$$

for some constant $M \in \mathbb{R}$. $\triangleleft$

PROOF. Let $g \in \mathscr{G}(X,\Sigma,\mu,I)$ be a GLSF representation of γ, let $C_\gamma \geq 0$ be a geodesic constant of γ. To simplify notation, let $I' := [0,\ b-a]$. According to Lemma A.2.2, the function $F \colon I \to \mathbb{R}$ with

$$F(t) := \int_{\gamma(t)} f \, \mathrm{d}\mu \quad \forall t \in I$$

is absolutely continuous. According to Theorem A.2.3, there exists a function $G \in L^1(\mathscr{B}(I),\lambda)$ such that

$$F(t) = \underbrace{F(a)}_{=0} + \int_a^t G \, \mathrm{d}\lambda \quad \forall t \in I.$$

Let $\rho \colon I' \to \mathscr{B}(I)/_{\sim_\lambda}$ be a constant mean geodesic of G with geodesic constant $C_\rho = 1$ and $\mathrm{TV}(\rho) = I$. Let $M_\rho \in \mathbb{R}$ be constant such that

$$\int_{\rho(s)\triangle\rho(t)} G \, \mathrm{d}\lambda = M_\rho \cdot \lambda\big(\rho(s) \triangle \rho(t)\big) \quad \forall s,t \in I'.$$

Let $\phi := \gamma \odot \rho$ be the parameterization of γ by ρ. Because $C_\rho = 1$, we know that C_γ is a geodesic constant of ϕ. Because constant mean geodesics are always canonical by definition, we have

$$\phi(0) = g^{-1}\big(\rho(0)\big) = \big[g^{-1}(\varnothing)\big]_{\sim_\mu} = [\varnothing]_{\sim_\mu},$$

which means that ϕ is canonical. Because ϕ is canonical, we have

$$\mathrm{TV}(\phi) = \phi(b-a) = g^{-1}\big(\rho(b-a)\big) = g^{-1}\big(\mathrm{TV}(\rho)\big) = g^{-1}(I) = \mathrm{TV}(\gamma).$$

We still have to prove that ϕ has the constant mean property with respect to f. We do so by an outer approximation argument. Let $s,t \in I'$. We have

$$\begin{aligned}\int_{\phi(s)\triangle\phi(t)} f \,\mathrm{d}\mu &= \int_{g^{-1}(\rho(s))\triangle g^{-1}(\rho(t))} f \,\mathrm{d}\mu \\ &= \int_{g^{-1}(\rho(s)\triangle\rho(t))} f \,\mathrm{d}\mu.\end{aligned}$$

Let $B \in \mathscr{B}(I)$ be a representative of $\rho(s) \triangle \rho(t)$. Let $\varepsilon > 0$. Because of the absolute continuity of the Lebesgue integral, there exists $\delta^* > 0$ such that

$$\int_A |f| \,\mathrm{d}\mu \leq \frac{\varepsilon}{2} \quad \forall A \in \Sigma\colon \mu(A) \leq \delta^*.$$

If $C_\gamma = 0$, then we set $\delta_1 := \infty$. Otherwise, we set $\delta_1 := \frac{\delta^*}{C_\gamma}$. In either case, this ensures that

$$C_\gamma \cdot \lambda(A) \leq \delta^* \quad \forall A \in \mathscr{B}(I)\colon \lambda(A) \leq \delta_1.$$

There also exists $\delta_2 > 0$ such that

$$\int_A |G| \,\mathrm{d}\lambda \leq \frac{\varepsilon}{2} \quad \forall A \in \mathscr{B}(I)\colon \lambda(A) \leq \delta_2.$$

Let $\delta := \min\{\delta_1, \delta_2\}$. Let $U_\varepsilon \subseteq \mathbb{R}$ be an open set with $B \subseteq U_\varepsilon$ such that $\lambda(U_\varepsilon \setminus B) \leq \delta$. Because U_ε is an open subset of the real numbers, it is a countable union of pairwise disjoint open intervals $\big((a_i,\ b_i)\big)_{i\in\mathbb{N}}$. Without loss of generality, let $a_i \leq b_i$ for all $i \in \mathbb{N}$. For each $i \in \mathbb{N}$, let $J_i := (a_i,\ b_i) \cap I$ and let $\widetilde{a}_i$ and $\widetilde{b}_i$ be maximally tight lower and upper bounds on J_i, respectively, such that $\widetilde{a}_i \leq \widetilde{b}_i$ for all $i \in \mathbb{N}$. We note that, because nullsets are irrelevant to integrals, it does not matter whether these bounds are included in J_i. Let subsequently

$$B' := \bigcup_{i=1}^{\infty} J_i \in \mathscr{B}(I).$$

We now have

$$\begin{aligned}\int_{\phi(s)\triangle\phi(t)} f \,\mathrm{d}\mu &= \int_{g^{-1}(\rho(s)\triangle\rho(t))} f \,\mathrm{d}\mu \\ &= \int_{g^{-1}(B)} f \,\mathrm{d}\mu \\ &= \int_{g^{-1}(B')} f \,\mathrm{d}\mu - \int_{g^{-1}(B'\setminus B)} f \,\mathrm{d}\mu \\ &= \left(\sum_{i=1}^{\infty} \int_{g^{-1}(J_i)} f \,\mathrm{d}\mu\right) - \int_{g^{-1}(B'\setminus B)} f \,\mathrm{d}\mu\end{aligned}$$

$$
\begin{aligned}
&= \left(\sum_{i=1}^{\infty} \int_{\gamma(\widetilde{a}_i)\triangle\gamma(\widetilde{b}_i)} f \,\mathrm{d}\mu\right) - \int_{g^{-1}(B'\setminus B)} f \,\mathrm{d}\mu \\
&= \sum_{i=1}^{\infty} \big(F(\widetilde{b}_i) - F(\widetilde{a}_i)\big) - \int_{g^{-1}(B'\setminus B)} f \,\mathrm{d}\mu \\
&= \left(\sum_{i=1}^{\infty} \int_{J_i} G \,\mathrm{d}\lambda\right) - \int_{g^{-1}(B'\setminus B)} f \,\mathrm{d}\mu \\
&= \int_{\bigcup_{i=1}^{\infty} J_i} G \,\mathrm{d}\lambda - \int_{g^{-1}(B'\setminus B)} f \,\mathrm{d}\mu \\
&= \int_{B'} G \,\mathrm{d}\lambda - \int_{g^{-1}(B'\setminus B)} f \,\mathrm{d}\mu \\
&= \int_{B} G \,\mathrm{d}\lambda + \int_{B'\setminus B} G \,\mathrm{d}\lambda - \int_{g^{-1}(B'\setminus B)} f \,\mathrm{d}\mu \\
&= \int_{\rho(s)\triangle\rho(t)} G \,\mathrm{d}\lambda + \int_{B'\setminus B} G \,\mathrm{d}\lambda - \int_{g^{-1}(B'\setminus B)} f \,\mathrm{d}\mu \\
&= M_\rho \cdot \underbrace{\lambda\big(\rho(s)\triangle\rho(t)\big)}_{=|s-t|} + \int_{B'\setminus B} G \,\mathrm{d}\lambda - \int_{g^{-1}(B'\setminus B)} f \,\mathrm{d}\mu \\
&= M_\rho \cdot |s-t| + \int_{B'\setminus B} G \,\mathrm{d}\lambda - \int_{g^{-1}(B'\setminus B)} f \,\mathrm{d}\mu.
\end{aligned}
$$

Because $\lambda\big(B'\setminus B\big) \le \lambda(U_\varepsilon \setminus B) \le \delta \le \delta_2$, we have

$$\left|\int_{B'\setminus B} G \,\mathrm{d}\lambda\right| \le \int_{B'\setminus B} |G| \,\mathrm{d}\lambda \le \frac{\varepsilon}{2}.$$

Furthermore, because

$$\mu\big(g^{-1}(B'\setminus B)\big) = C_\gamma \cdot \lambda(B'\setminus B) \le C_\gamma \cdot \lambda(U_\varepsilon \setminus B) \le \delta^*,$$

we have

$$\left|\int_{g^{-1}(B'\setminus B)} f \,\mathrm{d}\mu\right| \le \int_{g^{-1}(B'\setminus B)} |f| \,\mathrm{d}\mu \le \frac{\varepsilon}{2}.$$

Together, this yields the estimate

$$\left|\int_{\phi(s)\triangle\phi(t)} f \,\mathrm{d}\mu - M_\rho \cdot |s-t|\right| \le \left|\int_{B'\setminus B} G \,\mathrm{d}\lambda\right| + \left|\int_{g^{-1}(B'\setminus B)} f \,\mathrm{d}\mu\right| \le \varepsilon.$$

Because this holds for all $\varepsilon > 0$, we have

$$\int_{\phi(s)\triangle\phi(t)} f \,\mathrm{d}\mu = M_\rho \cdot |s-t|.$$

This is not quite the intended identity. For $C_\gamma > 0$, we can simply set $M \coloneqq \frac{M_\rho}{C_\gamma}$ and obtain

$$\int_{\phi(s)\triangle\phi(t)} f \,\mathrm{d}\mu = M_\rho \cdot |s-t| = M \cdot \mu\big(\phi(s)\triangle\phi(t)\big).$$

For $C_\gamma = 0$, we have to consider that F, γ, and ϕ are all constant functions. Because F is constant, we have $G \equiv 0$ and therefore $M_\rho = 0$. We can therefore set $M \coloneqq 0$ and obtain the same identity. □

Simultaneous alignment is then just an iterative application of Lemma A.2.4.

Theorem A.2.5 (Simultaneous Alignment).
Let (X, Σ, μ) be a finite measure space, let $I \coloneqq [a,\ b] \subseteq \mathbb{R}$ be a compact interval, let $\gamma\colon I \to \Sigma/_{\sim_\mu}$ be a canonical geodesic, let $n \in \mathbb{N}$, and let $f_i \in L^1(\Sigma,\mu) \cap L^\infty(\Sigma,\mu)$ for every $i \in [n]$. Then there exists a canonical geodesic $\gamma'\colon [0,\ b-a] \to \Sigma/_{\sim_\mu}$ with $\mathrm{TV}(\gamma') = \mathrm{TV}(\gamma)$ such that for every $i \in [n]$, there exists a constant $M_i \in \mathbb{R}$ such that

$$\int_{\gamma'(s)\triangle\gamma'(t)} f_i \,\mathrm{d}\mu = M_i \cdot \mu\big(\gamma'(s) \triangle \gamma'(t)\big) \quad \forall s,t \in [0,\ b-a]. \qquad \lhd$$

Proof. We inductively construct a tuple $(\gamma_i)_{i\in[n]}$ of canonical rearrangements $\gamma_i\colon [0,\ b-a] \to \Sigma/_{\sim_\mu}$ of γ with $\mathrm{TV}(\gamma_i) = \mathrm{TV}(\gamma)$ and of constants $(M_i)_{i\in[n]} \in \mathbb{R}^n$ such that for each $i \in [n]$, we have

$$\int_{\gamma_i(s)\triangle\gamma_i(t)} f_j \,\mathrm{d}\mu = M_j \cdot \mu\big(\gamma_i(s) \triangle \gamma_i(t)\big) \quad \forall s,t \in [0,\ b-a], j \le i.$$

To prove the latter, we demonstrate that

$$\sigma(\gamma_i) \subseteq \sigma(\gamma_j) \subseteq \sigma(\gamma) \quad \forall i,j \in [n]\colon i \ge j.$$

The claim then simply follows by choosing $\gamma' \coloneqq \gamma_n$. To simplify notation, let subsequently $I' \coloneqq [0,\ b-a]$.

Part 1 (Induction Start). According to Lemma A.2.4, there exists a parameter geodesic $\rho_1\colon I' \to \mathscr{B}(I)/_{\sim_\lambda}$ such that the parameterization $\gamma_1 \coloneqq \gamma \odot \rho_1$ of γ by ρ_1 is a canonical geodesic that satisfies $\mathrm{TV}(\gamma_1) = \mathrm{TV}(\gamma)$ and

$$\int_{\gamma_1(s)\triangle\gamma_1(t)} f_1 \,\mathrm{d}\mu = M_1 \cdot \mu\big(\gamma_1(s) \triangle \gamma_1(t)\big) \quad \forall s,t \in I'$$

for some constant $M_1 \in \mathbb{R}$. Because γ_1 is a rearrangement of γ, we have $\sigma(\gamma_1) \subseteq \sigma(\gamma)$.

Part 2 (Induction Step). Let $i \in [n-1]$ be such that $\gamma_i\colon I' \to \Sigma/_{\sim_\mu}$ is a canonical geodesic with $\mathrm{TV}(\gamma_i) = \mathrm{TV}(\gamma)$ such that $\sigma(\gamma_i) \subseteq \sigma(\gamma_j) \subseteq \sigma(\gamma)$ for all $j \in [n-2]$ as well as

$$\int_{\gamma_i(s)\triangle\gamma_i(t)} f_j \,\mathrm{d}\mu = M_j \cdot \mu\big(\gamma_i(s) \triangle \gamma_i(t)\big) \quad \forall s,t \in I', j \le i.$$

According to Lemma A.2.4, there exists a parameter geodesic $\rho_{i+1}\colon I' \to \mathscr{B}(I')/_{\sim_\lambda}$ such that the parameterization $\gamma_{i+1} \coloneqq \gamma_i \odot \rho_{i+1}$ of γ_i by ρ_{i+1} is a canonical geodesic with $\mathrm{TV}(\gamma_{i+1}) = \mathrm{TV}(\gamma_i) = \mathrm{TV}(\gamma)$ and

$$\int_{\gamma_{i+1}(s)\triangle\gamma_{i+1}(t)} f_{i+1} \,\mathrm{d}\mu = M_{i+1} \cdot \mu\big(\gamma_{i+1}(s) \triangle \gamma_{i+1}(t)\big) \quad \forall s,t \in I'$$

for some constant $M_{i+1} \in \mathbb{R}$. Because γ_{i+1} is a reparameterization of γ_i, we have

$$\sigma(\gamma_{i+1}) \subseteq \sigma(\gamma_i) \subseteq \sigma(\gamma_j) \subseteq \sigma(\gamma) \quad \forall j \in [i-1].$$

According to Lemma A.2.1,

$$\int_{\gamma_i(s)\triangle\gamma_i(t)} f_j \,\mathrm{d}\mu = M_j \cdot \mu\big(\gamma_i(s)\triangle\gamma_i(t)\big) \quad \forall s,t \in I', j \leq i$$

implies that

$$\int_A f_j \,\mathrm{d}\mu = M_j \cdot \mu(A) \quad \forall A \in \sigma(\gamma_i), j \leq i.$$

Because $\gamma_{i+1}(s)\triangle\gamma_{i+1}(t) \in \sigma(\gamma_{i+1}) \subseteq \sigma(\gamma_i)$ for all $s,t \in I'$, this means that

$$\int_{\gamma_{i+1}(s)\triangle\gamma_{i+1}(t)} f_j \,\mathrm{d}\mu = M_j \cdot \mu\big(\gamma_{i+1}(s)\triangle\gamma_{i+1}(t)\big) \quad \forall s,t \in I', j \leq i+1.$$

□

As we can see, the proof of Theorem A.2.5 is rather simple if Lemma A.2.4 is available.

A.2.3 Proof Sketch for Hypothesis 3.2.20

As we had stated initially, the central hypothesis that we would like to prove is Hypothesis 3.2.20, which states that

$$\big(\widetilde{T}_G(U)\big)^\circ \subseteq \widetilde{N}_G(U)$$

in any feasible point $U \in \mathscr{F}_G$ where an MFCQ is satisfied. We would likely prove this indirectly, i.e., we would prove that for any signed measure ϕ that is not in $\widetilde{N}_G(U)$, there exists a direction $v \in \widetilde{T}_G(U)$ such that

$$\limsup_{\substack{t\to 0\\ t>0}} \frac{(\phi \circ v)(t)}{t} > 0.$$

This is the negation of the defining property of the polar cone of a pseudo-cone (see Definition 3.2.6). By definition of the linearized tangent cone, this means that $v \in \mathrm{Dir}(\Sigma/_{\sim_\mu})$ with

$$\begin{aligned} \limsup_{\substack{t\to 0\\ t>0}} \frac{(\phi \circ v)(t)}{t} &> 0, \\ \limsup_{\substack{t\to 0\\ t>0}} \frac{\big(\nabla G_i(U) \circ v\big)(t)}{t} &\leq 0 \quad \forall i \in [n] \colon G_i(U) = 0, \end{aligned}$$

where $n \in \mathbb{N}$ is the number of scalar inequality constraints. Simultaneous alignment greatly simplifies this process. If we were to identify a similarity class $N \in \Sigma/_{\sim_\mu}$ such that $\phi(N) > 0$ and $\nabla G_i(U)(N) \leq 0$ for all $i \in [n]$ with $G_i(U) = 0$, then we could take an arbitrary geodesic connecting $[\varnothing]_{\sim_\mu}$ with N and align it for the density functions of ϕ and the relevant $\nabla G_i(U)$ to obtain the direction v. It is therefore sufficient to find such a similarity class N.

We begin by clarifying the theoretical implications of ϕ "not being in $\widetilde{N}_G(U)$." For the purposes of this discussion, we work in the real vector space of benign measures, which is a linear subspace of the Banach space of finite signed measures. Therefore, ϕ is a benign measure on the measure space (X, Σ, μ). The

linearized normal cone $\widetilde{N}_G(U)$ consists of all benign measures that can be written in the form

$$\phi_- + \sum_{i=1}^{n} q_i \cdot \nabla G_i(U)$$

where ϕ_- is a benign, non-positive measure and q is a vector of non-negative real coefficients such that $q_i = 0$ for all i with $G_i(U) < 0$. We can also write $\phi \notin \widetilde{N}_G(U)$ as

$$\phi - \sum_{i=1}^{n} q_i \cdot \nabla G_i(U) \notin M^-(\Sigma/_{\sim_\mu}) \quad \forall q \in \mathbb{R}^n_{\geq 0} \colon G_i(U) < 0 \implies q_i = 0 \; \forall i \in [n]$$

where $M^-(\Sigma/_{\sim_\mu})$ is the convex cone of non-positive benign measures. Let subsequently $\overline{M}^-$ be the closed convex cone of non-positive finite signed measures. According to Lemma 3.2.5, the closure of $M^-(\Sigma/_{\sim_\mu})$ within the space of finite signed measures is a subset of $\overline{M}^-$.

We want to separate the set of all relevant

$$\phi - \sum_{i=1}^{n} q_i \cdot \nabla G_i(U)$$

from $\overline{M}^-$. Because $M^-(\Sigma/_{\sim_\mu})$ is a cone, we can equivalently rewrite non-membership in $M^-(\Sigma/_{\sim_\mu})$ as

$$q_0 \cdot \phi - \sum_{i=1}^{n} q_i \cdot \nabla G_i(U) \notin M^-(\Sigma/_{\sim_\mu}) \; \forall q \in \mathbb{R}^{[n]_0}_{\geq 0} \colon q_0 > 0 \wedge \big(G_i(U) < 0 \implies q_i = 0 \; \forall i \in [n]\big)$$

and restrict ourselves to a normalized subset of coefficients. Let

$$Q \coloneqq \left\{ q_0 \cdot \phi - \sum_{i=1}^{n} q_i \cdot \nabla G_i(U) \,\middle|\, q \in \mathbb{R}^{[n]_0}_{\geq 0}, \|q\|_1 = 1, q_0 > 0, q_i = 0 \; \forall i \in [n] \colon G_i(U) < 0 \right\}.$$

Because the problem satisfies an MFCQ in U, we can invoke Lemma 3.2.16, which demonstrates that every convex combination of active constraint gradient measures assumes strictly negative values for some similarity classes. Accordingly, the combination

$$q_0 \cdot \phi - \sum_{i=1}^{n} q_i \cdot \nabla G_i(U)$$

can never yield a non-positive measure if $q_0 = 0$. Therefore, we can relax the strict positivity of q_0 to non-strict positivity without losing separation from $M^-(\Sigma/_{\sim_\mu})$. Let

$$\overline{Q} \coloneqq \left\{ q_0 \cdot \phi - \sum_{i=1}^{n} q_i \cdot \nabla G_i(U) \,\middle|\, q \in \mathbb{R}^{[n]_0}_{\geq 0}, \|q\|_1 = 1, q_i = 0 \; \forall i \in [n] \colon G_i(U) < 0 \right\}.$$

Because $\overline{Q}$ consists entirely of benign measures and does not intersect $M^-(\Sigma/_{\sim_\mu})$, $\overline{Q}$ consists entirely of benign measures that are not non-positive. $\overline{Q}$ is the image of the unit simplex, which is the convex hull of all unit vectors and a compact set in $\mathbb{R}^{n+1}$, under a finite-rank linear operator. Therefore, $\overline{Q}$ is compact. Because $\overline{Q}$ is compact, the continuous mapping $\varphi \mapsto \operatorname{dist}(\varphi, \overline{M}^-)$ assumes its minimum on $\overline{Q}$. Because all elements of $\overline{Q}$ are not non-positive, that minimum is strictly positive, i.e.,

$$\operatorname{dist}(\overline{Q}, \overline{M}^-) > 0.$$

$\overline{Q}$ is also convex. Because both $\overline{\overline{M}^{-}}$ and $\overline{Q}$ are non-empty and convex, and because their intersection is empty, we can invoke a variant of the Hahn-Banach separation theorem (see, e.g., [Cla20, Thm. 8.10]) to show that there exists an element of the dual space of the Banach space of finite signed measures that separates $\overline{\overline{M}^{-}}$ and $\overline{Q}$ in the sense that it maps all elements of the former to non-negative numbers and all elements of the latter to strictly negative numbers.

The next step is to investigate this separating "hyperplane" in the space of finite signed measures and attempt to derive an offending similarity class N from it. This requires some discussion of the nature of the space of finite signed measures. If we can appropriately restrict ourselves to the space of finite signed measures with density functions, then the space is isomorphic to L^1. Because we are operating exclusively in finite measure spaces, the dual of that space is isometrically isomorphic to L^∞, which means that we would have a function $f \in L^\infty(\Sigma, \mu)$ such that

$$\int_X f \,\mathrm{d}\varphi \geq 0 \quad \forall \varphi \in \overline{\overline{M}^{-}},$$
$$\int_X f \,\mathrm{d}\varphi < 0 \quad \forall \varphi \in \overline{Q}.$$

The former proves that f is non-positive almost everywhere. We would now have to identify a way to derive the similarity class N from these properties of the function f to prove the claim.

A.3 SET DERIVATIVE AND TOPOLOGICAL DERIVATIVE

The results in this section were developed in the time between submission and defense. They were not part of the original submission version and are therefore subject to a lesser standard of review. They are provided as a theoretical addendum for the interested reader.

The concept of using iterative optimization methods to optimize sets is not novel. Such optimization problems have long been the subject of research under the umbrella of shape and topology optimization. In Section 1.2, we briefly note the remarkable proximity between our approach and optimization methods that utilize the so-called "topological derivative". In this section, we will briefly investigate this relationship.

The topological derivative is a concept that is commonly attributed to Eschenauer et al. [EKS94], though the authors of that work refer to the topological derivative as a "characteristic function" for the positioning of "bubbles". For a more fleshed out definition of the topological derivative, we turn to more recent work by Novotny and Sokołowsky [NS13]. There, the topological derivative is defined by a so-called "topological asymptotic expansion" (see [NS13, Sec. 1.1]) of the form

$$F\big(A \setminus \omega_\varepsilon(\hat{x})\big) = F(A) + f(\varepsilon) \cdot T(\hat{x}) + o\big(f(\varepsilon)\big).$$

Here, $F\colon \Sigma \to \mathbb{R}$ is what we have previously referred to as a set functional. In shape and topology optimization, this is often referred to as a "shape functional". The parameterized measurable set $\omega_\varepsilon(\hat{x}) \coloneqq \hat{x} + \varepsilon \cdot \omega$ represents a small hole based on a measurable set ω that is made within the set A. The function $T\colon A \to \mathbb{R}$ that

maps an arbitrary point in A to a real number and can be written as

$$T(\hat{x}) = \lim_{\varepsilon \to 0} \frac{F\big(A \setminus \omega_\varepsilon(\hat{x})\big) - F(A)}{f(\varepsilon)} \quad \forall \hat{x} \in A$$

is then referred to as the "topological derivative" of F with respect to the insertion of a hole of the given shape within A.

The relationship between the topological derivative and our set derivative appears evident. Let F be set differentiable in A. Because the definition above applies only for insertions of holes within the current iterate A, the set difference is equal to the symmetric difference. If we also define f such that $f(\varepsilon) = \mu\big(\omega_\varepsilon(\hat{x})\big)$ for all $\hat{x}$, then it is easy to see that under very mild additional assumptions about $\omega_\varepsilon(\hat{x})$, T is the gradient density function of F in A. In this sense, the existence of a set derivative implies the existence of a topological derivative.

A converse of this implication is much harder to establish. This is primarily because of the very loose restrictions on $f(\varepsilon)$ and ω. The restrictions that the authors impose (see [NS13, Condition 1.1]) are tailored to allow for the use of so-called "nucleation methods". Nucleation methods are methods that are similar to the original bubble method proposed in [EKS94]. They use the topological derivative to place individual "holes" and modify the shape of those holes using the shape derivative. We evidently have to impose additional restrictions to turn the topological derivative into a set derivative.

A.3.1 Theoretical Prerequisites

At the foundation of the relationship between the set and topological derivative, there are two theorems: the Lebesgue differentiation theorem and the Vitali covering theorem. For both theorems, we need to begin by introducing the concept of "bounded eccentricity," which is required for both theorems

Definition A.3.1 (Sets of Bounded Eccentricity [Leb10, Par. 25]).
A collection $\mathscr{F}$ of Lebesgue measurable subsets of $\mathbb{R}^n$ is said to be *of bounded eccentricity* if and only if there exists a uniform strictly positive constant $C > 0$ such that for each $A \in \mathscr{F}$, there exists a Euclidean ball B with $A \subseteq B$ and

$$\lambda(A) > C \cdot \lambda(B)$$

where λ is the n-dimensional Lebesgue measure. ◁

We note that, although we cite Lebesgue in this definition, Lebesgue refers to these collections as "regular families" rather than "collections of bounded eccentricity". We stress that the constant C has to be uniform across all members of the collection $\mathscr{F}$.

We first turn our attention to the Vitali covering theorem, which is discussed by many textbooks as a prerequisite to proving the Lebesgue differentiation theorem. One of the limiting factors in most treatments of the Vitali covering theorem is that they only discuss cases where the covering consists entirely of closed balls. This is insufficient for our purposes because the topological derivative does not restrict the hole shape in this way. Therefore, we require a more general variant of the theorem that can be challenging to find in modern literature.

Theorem A.3.2 (Vitali Covering Theorem[1]).
Let $A \subseteq \mathbb{R}^n$ be a Lebesgue measurable set of finite measure, let $\mathcal{V} \subseteq \mathscr{L}(\mathbb{R}^n)$ be a collection of closed sets of bounded eccentricity such that for every $\widehat{x} \in A$ and every $\delta > 0$, there exists a set $U \in \mathcal{V}$ with $\widehat{x} \in U$ and $0 < \operatorname{diam}(U) < \delta$. Let futher $B \supseteq A$ be a Lebesgue measurable superset of A that is also of finite measure and satisfies $U \subseteq B$ $\forall U \in \mathcal{V}$.

Then there exists a finite or countably infinite pairwise disjoint sub-collection $\{U_j\}_j \subseteq \mathcal{V}$ such that

$$\lambda\Big(A \setminus \bigcup_j U_j\Big) = 0.$$

◁

The collection $\mathcal{V}$ in Theorem A.3.2 is commonly referred to as a "Vitali covering" of the set A. Later on, we will use finite sub-collections of Vitali coverings to approximate arbitrary set changes with a finite number of disjoint "holes" in the sense of the topological derivative. This will allow us to reduce the question of whether a topological derivative also functions as a set derivative in our sense to the much simpler question whether it is additive for finite disjoint unions of holes.

The final theorem that we need in order to establish the relationship between the set derivative and the topological derivative is the well-known Lebesgue differentiation theorem. We cannot cite this theorem directly from Lebesgue, although [Leb10] is commonly held to be the work from which the multi-dimensional variant of this theorem originates. This is because Lebesgue's variant of the theorem only extends to functions of bounded variation. Instead, we cite the generalized variant of the theorem from Stein and Shakarchi.

Theorem A.3.3 (Lebesgue Differentiation Theorem [SS05, Sec. 3.1.2]).
Let $\Omega \subseteq \mathbb{R}^n$ be open, and let $f : \Omega \to \mathbb{R}$ be locally Lebesgue integrable. Let $\mathscr{F}$ be a collection of Lebesgue measurable subsets of Ω that is of bounded eccentricity such that for every $\overline{x} \in \Omega$, there exists a sub-collection $\mathscr{F}_{\overline{x}} \subseteq \mathscr{F}$ with $\overline{x} \in B$ $\forall B \in \mathscr{F}_{\overline{x}}$ that contains sets of arbitrarily small diameter. Then for almost all $\overline{x} \in \Omega$, we have

$$\lim_{\substack{\lambda(B)\to 0 \\ B\in\mathscr{F}_{\overline{x}}}} \frac{1}{\lambda(B)} \int_B \big|f(x) - f(\overline{x})\big| \,\mathrm{d}x = 0.$$

◁

The reason why we have to cite these specific versions of the two theorems is because they allow for the use of arbitrary measurable sets of bounded eccentricity. This is important because we will be using sets of whatever shape the topological derivative has been developed for. That shape is arbitrary and largely unrestricted by the definition of the topological derivative. If we were to use more common variants of Theorems A.3.2 and A.3.3 that only use balls, then this would generate the false impression that the relationship that we are about to prove is restricted to cases where the topological derivative is developed for ball-shaped holes. Such a restriction is not necessary.

[1] This is a minor correction of a variant of the theorem that can – to the author's best knowledge – only be found on [Wik25]. It is derived from a variant for the Hausdorff measure (see [Fal85, Thm. 1.10]). By making use of the fact that the Lebesgue and Hausdorff measures on $\mathbb{R}^n$ are equal up to a constant scaling factor, this variant can be transferred to the Lebesgue measure. An additional edge case can be ignored because A is of finite measure and because the collection $\{U_j\}_j$ is of bounded eccentricity. The existence of the shared superset B of finite measure then ensures that the edge case in which $\sum_j \operatorname{diam}(U_j)^n = \infty$ cannot occur. The original version on [Wik25] treats the existence of B as implied, but does not properly justify that implication.

A.3.2 Integrable Topological Derivatives by Measure (ITDMs)

As we had initially noted, it is necessary to restrict the definition of the topological derivative in order to establish a relationship with the set derivative. Now that we know the requirements of the two central theorems, we can formulate the precise restrictions needed. To simplify this restricted definition, we first formulate our restrictions on the holes with respect to which the topological derivative must be generated.

Definition A.3.4 (Closed Holes of Bounded Eccentricity).
Let $\Omega \subseteq \mathbb{R}^n$ be Lebesgue measurable, and let $\Sigma := \mathscr{L}(\Omega)$. We refer to a mapping $\omega\colon \Omega \times (0,\ \infty) \to \Sigma$ as a *closed hole generator of bounded eccentricity* if and only if

1. $\omega(\widehat{x}, \varepsilon)$ is closed $\forall \widehat{x} \in \Omega, \varepsilon > 0$;
2. $\widehat{x} \in \omega(\widehat{x}, \varepsilon)$ $\forall \widehat{x} \in \Omega, \varepsilon > 0$; and
3. there exist constants $C_1 > 0$ and $C_2 > 0$ and a mapping $\widehat{\varepsilon}_\omega\colon \Omega \to (0,\ \infty)$ such that

$$\begin{aligned} \operatorname{diam}\big(\omega(\widehat{x}, \varepsilon)\big) &< C_1 \cdot \varepsilon \quad && \forall \widehat{x} \in \Omega, \varepsilon > 0, \\ \lambda\big(\omega(\widehat{x}, \varepsilon)\big) &> C_2 \cdot \varepsilon^n \quad && \forall \widehat{x} \in \Omega, \varepsilon \in \big(0,\ \widehat{\varepsilon}_\omega(\widehat{x})\big). \end{aligned}$$

◁

The primary purpose of Definition A.3.4 is to ensure that the holes generated by ω are always suitable for use with both the Lebesgue differentiation theorem and the Vitali covering theorem.

Definition A.3.5 (Integrable Topological Derivatives by Measure).
Let $\Omega \subseteq \mathbb{R}^n$ be a Lebesgue measurable set, let $\Sigma := \mathscr{L}(\Omega)$, let $\Sigma_\sim := \Sigma/_{\sim_\lambda}$, and let $\omega\colon \Omega \times (0,\ \infty) \to \Sigma$ be a closed hole generator of bounded eccentricity.

We say that a continuous set functional $F\colon \Sigma_\sim \to \mathbb{R}$ admits an *integrable topological derivative by measure* (ITDM) with respect to ω in $U \in \Sigma$ if there exists an integrable function $T_U \in L^1(\Sigma, \lambda)$ such that

$$T_U(\widehat{x}) = \lim_{\varepsilon \to 0} \frac{F\big(U \mathbin{\triangle} \omega(\widehat{x}, \varepsilon)\big) - F(U)}{\lambda\big(\omega(\widehat{x}, \varepsilon)\big)} \quad \text{for a.a. } \widehat{x} \in \Omega.$$

◁

Definition A.3.5 is a restriction of the generalized definition of the topological derivative in the sense that:

1. it requires that $f(\varepsilon) = \lambda\big(\omega(\widehat{x}, \varepsilon)\big)$ (it is "by measure");
2. it requires that T_U be integrable;
3. it restricts the choice of the hole.

The restrictions on the hole ω are the least straightforward part of Definition A.3.5. They exist largely to ensure the applicability of the Vitali covering theorem and the Lebesgue differentiation theorem to the hole. Aside from the additional restrictions, we also slightly relax the definition put forward by [NS13] by allowing for hole shapes that vary based on location and on the scaling parameter ε. This is in line with more recent work in the field of topology optimization (see, e.g., [LOS24]), where derivatives for the subtraction of holes at the boundary are calculated with modified hole shapes.

Finally, we specify a Lipschitz-like continuity criterion that implies additivity. This is to demonstrate that the additivity condition can also be understood as a continuity condition on the topological derivative.

Definition A.3.6 (Lipschitz Continuous ITDMs).
Let $\Omega \subseteq \mathbb{R}^n$ be a Lebesgue measurable set, let $\Sigma := \mathscr{L}(\Omega)$, let $\Sigma_\sim := \Sigma/_{\sim_\lambda}$, let $\mathscr{N} \subseteq \Sigma_\sim$ be open with respect to the topology induced by the metric of $\Sigma_\sim$, and let $\omega\colon \Omega \times (0,\ \infty) \to \Sigma$ be a closed hole generator of bounded eccentricity. Let $F\colon \Sigma_\sim \to \mathbb{R}$ be a continuous set functional that admits an ITDM with respect to ω in every $U \in \mathscr{N}$.

We say that F admits a *Lipschitz continuous* ITDM with respect to ω on $\mathscr{N}$ if and only if there exists a constant $L > 0$ such that

$$\Big|F\big(V \triangle \omega(\widehat{x},\varepsilon)\big) - F(V) - F\big(U \triangle \omega(\widehat{x},\varepsilon)\big) - F(U)\Big| \leq L \cdot \lambda(U \triangle V)$$

for all $U, V \in \mathscr{N}$, all $\widehat{x}$ for which the Taylor expansion in Definition A.3.5 holds in U, and all $\varepsilon > 0$ such that $\omega(\widehat{x},\varepsilon)$ and $U \triangle V$ are essentially disjoint. We refer to L as a *Lipschitz constant* of the ITDM. ◁

We will use the Lipschitz continuity property to prove that a topological derivative is "additive" in the sense that the linearized change in objective can be added up over any finite number of disjoint holes without a qualitative loss of precision beyond what would be expected due to the accumulated measure. As we will see shortly, additivity is the central requirement that turns an ITDM into a set derivative.

It is important to note that this kind of additivity is not an automatic property of topological derivatives. For set derivatives, additivity follows directly from the additivity of measures and integrals. However, the definition of the topological derivative does not imply it. As we will note later on, this is both an advantage and a major weakness of the topological derivative concept.

A.3.3 Set Derivatives are ITDMs

We begin by proving the simpler implication: if a set functional admits a set derivative, then it also admits an ITDM and the topological derivative function is equal to the gradient density function of the set functional. This implication can simply be demonstrated by applying the Lebesgue differentiation theorem to the gradient density function.

Theorem A.3.7 (Set Derivatives are ITDMs).
Let $\Omega \subseteq \mathbb{R}^n$ be a Lebesgue measurable set, let $\Sigma := \mathscr{L}(\Omega)$, let $\Sigma_\sim := \Sigma/_{\sim_\lambda}$, and let $\omega\colon \Omega \times (0,\ \infty) \to \Sigma$ be a closed hole generator of bounded eccentricity. Let $F\colon \Sigma_\sim \to \mathbb{R}$ be a continuous set functional that is set differentiable according to Definition 2.4.1 on page 152 in $U \in \Sigma_\sim$, and let $g_U \in L^1(\Sigma,\lambda)$ be the gradient density function of F in U.

Then F admits an ITDM with respect to ω in U and the topological derivative function $T_U \in L^1(\Sigma,\mu)$ satisfies $T_U = g_U$. ◁

PROOF. According to Definition A.3.4, there exists a constant $C_1 > 0$ such that for each $\widehat{x} \in \Omega$, we have

$$\operatorname{diam}\big(\omega(\widehat{x},\varepsilon)\big) < C_1 \cdot \varepsilon \quad \forall \varepsilon > 0.$$

This ensures that for every $\varepsilon > 0$, there exists a ball $B_{\hat{x},\varepsilon} \subseteq \mathbb{R}^n$ of diameter $C_1 \cdot \varepsilon$ such that $\omega(\hat{x},\varepsilon) \subseteq B_{\hat{x},\varepsilon}$. Because the measure of an n-dimensional ball goes to zero as its diameter goes to zero, we have

$$\lambda\big(\omega(\hat{x},\varepsilon)\big) \leq \lambda\big(B_{\hat{x},\varepsilon}\big) \xrightarrow{\varepsilon\to 0} 0.$$

Similarly, according to Definition A.3.4, there exists a constant $C_2 > 0$ such that for every $\hat{x} \in \Omega$, there exists $\hat{\varepsilon}_\omega(\hat{x}) > 0$ such that

$$\lambda\big(\omega(\hat{x},\varepsilon)\big) > C_2 \cdot \varepsilon^n \quad \forall \varepsilon \in \big(0,\ \hat{\varepsilon}(\hat{x})\big).$$

This has two effects. It ensures that $\lambda\big(\omega(\hat{x},\varepsilon)\big) \to 0$ always implies $\varepsilon \to 0$. However, it also ensures bounded eccentricity in the sense that for each $\hat{x} \in \Omega$ and every $\varepsilon \in \big(0,\ \hat{\varepsilon}(\hat{x})\big)$, we have

$$\begin{aligned}
\lambda\big(\omega(\hat{x},\varepsilon)\big) &> C_2 \cdot \varepsilon^n \\
&= \frac{2^n \cdot C_2}{C_1^n} \cdot \left(\frac{C_1 \cdot \varepsilon}{2}\right)^n \\
&= \underbrace{\frac{2^n \cdot C_2 \cdot \Gamma\big(\frac{n}{2}+1\big)}{C_1^n \cdot \pi^{n/2}}}_{=:C>0} \cdot \underbrace{\frac{\pi^{n/2}}{\Gamma\big(\frac{n}{2}+1\big)} \cdot \left(\frac{C_1 \cdot \varepsilon}{2}\right)^n}_{=\lambda(B_{\hat{x},\varepsilon})} \\
&= C \cdot \lambda(B_{\hat{x},\varepsilon}).
\end{aligned}$$

This demonstrates that the collection

$$\mathscr{F} := \left\{\omega(\hat{x},\varepsilon) \,\middle|\, \hat{x} \in \Omega, \varepsilon \in \big(0,\ \hat{\varepsilon}_\omega(\hat{x})\big)\right\} \subseteq \Sigma$$

is of bounded eccentricity. Furthermore, for every $\hat{x} \in \Omega$, the sub-collection

$$\mathscr{F}_{\hat{x}} := \left\{\omega(\hat{x},\varepsilon) \,\middle|\, \varepsilon \in \big(0,\ \hat{\varepsilon}_\omega(\hat{x})\big)\right\} \subseteq \mathscr{F}$$

satisfies $\hat{x} \in B\ \forall B \in \mathscr{F}_{\hat{x}}$ and contains sets of arbitrarily small diameter. Therefore, $\mathscr{F}$ satisfies all requirements of the Lebesgue differentiation theorem (see Theorem A.3.3). We define $T_U := g_U \in L^1(\Sigma,\lambda)$. Bearing in mind that $\lambda\big(\omega(\hat{x},\varepsilon)\big) \to 0$ implies $\varepsilon \to 0$ and vice versa, we have

$$\begin{aligned}
T_U(\hat{x}) &= g_U(\hat{x}) \\
&= \lim_{\substack{\lambda(B)\to 0 \\ B\in\mathscr{F}_{\hat{x}}}} \frac{\int_B g_U \,\mathrm{d}x}{\lambda(B)} \\
&\overset{2.4.1}{=} \lim_{\substack{\lambda(B)\to 0 \\ B\in\mathscr{F}_{\hat{x}}}} \frac{F(U \triangle B) - F(U) - o\big(\lambda(B)\big)}{\lambda(B)} \\
&= \lim_{\substack{\lambda(B)\to 0 \\ B\in\mathscr{F}_{\hat{x}}}} \frac{F(U \triangle B) - F(U)}{\lambda B} \\
&= \lim_{\varepsilon\to 0} \frac{F\big(U \triangle \omega(\hat{x},\varepsilon)\big) - F(U)}{\lambda\big(\omega(\hat{x},\varepsilon)\big)}
\end{aligned}$$

for almost all $\hat{x}$. This demonstrates that F admits an ITDM with respect to ω in U and that $T_U = g_U$ is the topological derivative function. □

Theorem A.3.7 demonstrates that set differentiable set functionals form a subset of those set functionals to which the topological derivative concept can be applied. This is relatively evident from the definitions of both derivatives. However, it is useful because it allows us to quickly construct test functionals for topology optimization algorithms. We will use this capability to construct an edge case in Section A.3.5.

We note that, because the closedness of the holes generated by ω is only needed for the Vitali covering theorem, Theorem A.3.7 does not require it. We do not make this theoretical distinction here because it would require an additional definition.

A.3.4 Lipschitz Continuous ITDMs are Set Derivatives

Next, we come to the more interesting implication. As we have noted before, the general statement that all topological derivatives are also set derivatives is not true. Topological derivatives, in general, need not be integrable functions. They need not even be measurable functions by definition. The function f can incorporate boundary integrals and the topological derivative need not be additive in the sense that we had discussed earlier. It is therefore almost certain that the set of all set functionals that admit a set derivative is a strict subset of those that admit a topological derivative. However, it is interesting to ask how large this subset is.

In this section, we demonstrate that every set functional that admits a Lipschitz continuous ITDM also admits a set derivative and that the gradient density function is equal to the topological derivative function in that case. For brevity, we do not repeat the same argument about the bounded eccentricity of holes generated by ω again.

Theorem A.3.8 (Lipschitz Continuous ITDMs are Set Derivatives). *Let $\Omega \subseteq \mathbb{R}^n$ be a set of finite Lebesgue measure, let $\Sigma := \mathscr{L}(\Omega)$, let $\Sigma_\sim := \Sigma/_{\sim_\lambda}$, let $\mathscr{N} \subseteq \Sigma_\sim$ be open with respect to the metric topology of $\Sigma_\sim$, and let $\omega \colon \Omega \times (0, \infty) \to \Sigma$ be a closed hole generator of bounded eccentricity. Let $F \colon \Sigma_\sim \to \mathbb{R}$ be a continuous set functional that admits a Lipschitz continuous ITDM with respect to ω on $\mathscr{N}$.*

Then F is set differentiable in all $U \in \mathscr{N}$ according to Definition 2.4.1 and in every $U \in \mathscr{N}$, the gradient density function of F in U is equal to the topological derivative function T_U. ◁

PROOF. Let $U \in \mathscr{N}$. Because $\mathscr{N}$ is open, there exists a constant $R > 0$ such that $B_R(U) \subseteq \mathscr{N}$. Our goal is to prove that

$$0 = \lim_{\lambda(D) \to 0} \frac{F(U \triangle D) - F(U) - \int_D T_U \,\mathrm{d}x}{\lambda(D)}.$$

Equivalently, we can show that for every ratio bound $\rho > 0$, there exists a radius $\delta > 0$ such that

$$\left| F(U \triangle D) - F(U) - \int_D T_U \,\mathrm{d}x \right| \leq \rho \cdot \lambda(D) \quad \forall D \in \Sigma_\sim \colon \lambda(D) \leq \delta.$$

Let subsequently $U^* \in U$ be a representative of U, let $D \in \Sigma_\sim$ be a step with an arbitrary representative $D^* \in D$, and let $\rho > 0$ be an arbitrary ratio bound. We will divide this bound into smaller fractions to compensate for five errors that occur when we replace the step with an approximation by a finite union of holes:

1. the change in $F(U \triangle D)$, which is bounded by continuity;
2. the change in $\int_D T_U \,\mathrm{d}x$, which is bounded by the absolute continuity of the integral;
3. the error incurred by replacing the integral with a weighted sum, which is bounded by the Lebesgue differentiation theorem;
4. the error incurred due to the change in topological derivative;
5. the linearization error of the asymptotic topological expansion.

Without loss of generality, let $\lambda(D) > 0$. If $\lambda(D) = 0$, then the claim evidently holds for all $\delta > 0$.

Part 1 (Step Approximation Margins). We begin by constructing an approximation $\widetilde{D}$ of the step D^* that consists of finitely many pairwise disjoint holes. In order to use the Lipschitz continuity of the ITDM of F, we have to ensure that neither the original nor the approximate step move outside of the neighborhood $\mathcal{N}$. We can achieve this by choosing $\delta \leq \delta_1 \coloneqq \frac{R}{2}$ and demanding that the step approximation error not exceed $\widehat{\delta}_1 \coloneqq \frac{\mu(D)}{2} \leq \frac{R}{4}$.

Next, we have to bound $\left|F(U^* \triangle D^*) - F(U^* \triangle \widetilde{D})\right|$. We can do this by using the continuity of F. There exists $\widehat{\delta}_2 > 0$ such that

$$\left|F(U^* \triangle D^*) - F(U^* \triangle \widetilde{D})\right| \leq \frac{\rho}{6} \cdot \lambda(D) \quad \forall \widetilde{D} \colon \lambda(D^* \triangle \widetilde{D}) \leq \widehat{\delta}_2.$$

In order to bound $\left|\int_{D^*} T_U \,\mathrm{d}x - \int_{\widetilde{D}} T_U \,\mathrm{d}x\right|$, we invoke the absolute continuity of the Lebesgue integral, which guarantees that there exists $\widehat{\delta}_3 > 0$ such that

$$\int_{D^* \triangle \widetilde{D}} |T_U| \,\mathrm{d}x \leq \frac{\rho}{6} \cdot \lambda(D) \quad \forall \widetilde{D} \colon \lambda(D^* \triangle \widetilde{D}) \leq \widehat{\delta}_3.$$

This yields the desired estimate because

$$\left|\int_{D^*} T_U \,\mathrm{d}x - \int_{\widetilde{D}} T_U \,\mathrm{d}x\right| = \left|\int_{D^* \setminus \widetilde{D}} T_U \,\mathrm{d}x - \int_{\widetilde{D} \setminus D^*} T_U \,\mathrm{d}x\right| \leq \int_{D^* \triangle \widetilde{D}} |T_U| \,\mathrm{d}x.$$

Therefore, we have to find a step approximation $\widetilde{D}$ that satisfies

$$\lambda(D^* \triangle \widetilde{D}) \leq \underbrace{\min\{\widetilde{\delta}_1, \widetilde{\delta}_2, \widetilde{\delta}_3\}}_{=:\widehat{\delta} > 0}.$$

According to the Lebesgue differentiation theorem (see Theorem A.3.3), there exists a nullset N_1 such that

$$0 = \lim_{\varepsilon \to 0} \frac{1}{\lambda\big(\omega(x,\varepsilon)\big)} \cdot \int_{\omega(x,\varepsilon)} \left|T_U(y) - T_U(x)\right| \mathrm{d}y \quad \forall x \in D^* \setminus N_1.$$

We note the equivalence between $\varepsilon \to 0$ and $\lambda\big(\omega(x,\varepsilon)\big) \to 0$ that we have already demonstrated in the previous section. According to Definition A.3.5, there exists a nullset N_2 such that

$$T_U(x) = \lim_{\varepsilon \to 0} \frac{F\big(U \triangle \omega(x,\varepsilon)\big) - F(U)}{\lambda\big(\omega(x,\varepsilon)\big)} \quad \forall x \in D^* \setminus N_2.$$

The set $D_0^* := D^* \setminus (N_1 \cup N_2)$ is Lebesgue measurable and satisfies $\lambda(D^* \triangle D_0^*) = 0$. Because D_0^* is a subset of Ω and Ω is of finite measure, we can find an open set $G \supseteq D_0^*$ and a compact set $F \subseteq D_0^*$ such that $\lambda(G \setminus F) \leq \frac{\widehat{\delta}}{3}$.

PART 2 (VITALI COVERING). By subtracting N_1 and N_2 from D^*, we have ensured that we can apply the Lebesgue differentiation theorem and the asymptotic topological expansion around every point in F. For each $x \in F$, let $\varepsilon_1(x,\rho) > 0$ be such that

$$\frac{1}{\lambda\big(\omega(x,\varepsilon)\big)} \cdot \int_{\omega(x,\varepsilon)} \big|T_U(y) - T_U(x)\big| \,\mathrm{d}y \leq \frac{\rho}{6} \quad \forall \varepsilon \in \big(0,\ \varepsilon_1(x,\rho)\big].$$

For each $x \in F$, let further $\varepsilon_2(x,\rho) > 0$ be such that

$$F\big(U \triangle \omega(x,\varepsilon)\big) - F(U) - T_U(x) \cdot \lambda\big(\omega(x,\varepsilon)\big) \leq \frac{\rho}{6} \cdot \lambda\big(\omega(x,\varepsilon)\big) \quad \forall \varepsilon \in \big(0,\ \varepsilon_2(x,\rho)\big],$$

and let $\varepsilon_3(x) > 0$ be such that $\omega(x,\varepsilon) \subseteq G$ for all $\varepsilon \in \big(0,\ \varepsilon_3(x)\big]$. An appropriate $\varepsilon_3(x)$ can always be chosen because $x \in F \subseteq G$, because G is open, and because $\operatorname{diam}\big(\omega(x,\varepsilon)\big) \xrightarrow{\varepsilon \to 0} 0$. We define

$$\overline{\varepsilon}(x,\rho) := \min\big\{\widehat{\varepsilon}_\omega(x),\, \varepsilon_1(x,\rho),\, \varepsilon_2(x,\rho),\, \varepsilon_3(x)\big\} > 0 \quad \forall x \in F.$$

The collection

$$\mathcal{V} := \big\{\omega(x,\varepsilon) \,\big|\, x \in F,\, \varepsilon \leq \overline{\varepsilon}(x,\rho)\big\}$$

consists of closed sets of bounded eccentricity such that every $x \in F$ is contained within member sets of arbitrarily small measure. Furthermore, all members of $\mathcal{V}$ are subsets of G, which has finite measure. Therefore, $\mathcal{V}$ is a Vitali covering of F. According to the Vitali covering theorem (see Theorem A.3.2), there exists a finite or countably finite pairwise disjoint sub-collection $\big\{\omega(x_j,\varepsilon_j)\big\}_j \subseteq \mathcal{V}$ such that

$$\lambda\Big(F \setminus \bigcup_j \omega(x_j,\varepsilon_j)\Big) = 0.$$

If the sub-collection is countably infinite, then we can truncate it after finitely many elements to obtain a finite pairwise disjoint sub-collection with

$$\lambda\Big(F \setminus \bigcup_j \omega(x_j,\varepsilon_j)\Big) \leq \frac{\widehat{\delta}}{3}.$$

We subsequently define

$$\widetilde{D} := \bigcup_j \omega(x_j,\varepsilon_j).$$

PART 3 (STEP APPROXIMATION ERROR). We first examine the distance between D^* and $\widetilde{D}$. We have

$$\begin{aligned}\lambda(D^* \triangle \widetilde{D}) &\leq \underbrace{\lambda(D^* \triangle D_0^*)}_{=0} + \lambda(D_0^* \triangle \widetilde{D}) \\ &= \lambda(\underbrace{D_0^* \setminus \widetilde{D}}_{\subseteq G \setminus \widetilde{D}}) + \lambda(\underbrace{\widetilde{D} \setminus D_0^*}_{\subseteq \widetilde{D} \setminus F})\end{aligned}$$

$$\begin{aligned}
&\leq \lambda(G \setminus \widetilde{D}) + \underbrace{\lambda(\widetilde{D} \setminus F)}_{\leq \lambda(G \setminus F)} \\
&\leq 2 \cdot \lambda(G \setminus F) + \lambda(F \setminus \widetilde{D}) \\
&\leq 3 \cdot \frac{\widehat{\delta}}{3} \\
&= \widehat{\delta}.
\end{aligned}$$

As we had discussed earlier, this ensures that

$$\left|F(U^* \triangle D^*) - F(U^* \triangle \widetilde{D})\right| \leq \frac{\rho}{6} \cdot \lambda(D),$$

and that

$$\left|\int_{D^*} T_U \, \mathrm{d}x - \int_{\widetilde{D}} T_U \, \mathrm{d}x\right| \leq \frac{\rho}{6} \cdot \lambda(D).$$

PART 4 (TELESCOPE SUM ESTIMATE). Having appropriately bounded the error incurred by substituting the actual step with a finite disjoint union of holes, we can now prove the main estimate. We have

$$\begin{aligned}
&\left|F(U \triangle D) - F(U) - \int_D T_U \, \mathrm{d}x\right| \\
&\quad = \left|F(U^* \triangle D^*) - F(U^*) - \int_{D^*} T_U \, \mathrm{d}x\right| \\
&\quad \leq \left|F(U^* \triangle D^*) - F(U^* \triangle \widetilde{D})\right| \\
&\qquad + \left|F(U^* \triangle \widetilde{D}) - F(U^*) - \int_{\widetilde{D}} T_U \, \mathrm{d}x\right| \\
&\qquad + \left|\int_{\widetilde{D}} T_U \, \mathrm{d}x - \int_{D^*} T_U \, \mathrm{d}x\right| \\
&\quad \leq \frac{2 \cdot \rho}{6} \cdot \lambda(D) + \left|F(U^* \triangle \widetilde{D}) - F(U^*) - \int_{\widetilde{D}} T_U \, \mathrm{d}x\right|
\end{aligned}$$

For the final component estimate, we make use of the particular form of the approximate step $\widetilde{D}$. To simplify notation, let $N \in \mathcal{N}$ be the cardinality of the finite sub-collection selected by the Vitali covering theorem. We define

$$U_k^* := U^* \triangle \bigcup_{j=1}^{k} \omega(x_j, \varepsilon_j) \quad \forall k \in [N]_0.$$

These are the intermediate sets generated by nucleating the selected holes sequentially. We note that for all $k \in [N]_0$, we have

$$\begin{aligned}
\lambda(U^* \triangle U_k^*) &= \lambda\Big(\underbrace{\bigcup_{j=1}^{k} \omega(x_j, \varepsilon_j)}_{\subseteq \widetilde{D}}\Big) \\
&\leq \lambda(\widetilde{D}) \\
&\leq \lambda(D) + \lambda(D^* \triangle \widetilde{D}) \\
&\leq \delta + \widehat{\delta}
\end{aligned}$$

$$\begin{aligned}&\leq \frac{R}{2} + \frac{R}{4}\\ &< R.\end{aligned}$$

Therefore, we have $U_k^* \in \mathcal{N}\ \forall k \in [N]_0$. Let $L > 0$ be a Lipschitz constant for the ITDM of F with respect to ω on $\mathcal{N}$. We can now make the following estimate:

$$\begin{aligned}&\left|F(U^* \triangle \widetilde{D}) - F(U^*) - \int_{\widetilde{D}} T_U \,\mathrm{d}x\right|\\
&\quad\leq \sum_{j=1}^{N} \left|F(U_j^*) - F(U_{j-1}^*) - \int_{U_j^* \triangle U_{j-1}^*} T_U \,\mathrm{d}x\right|\\
&\quad= \sum_{j=1}^{N} \left|F\big(U_{j-1}^* \triangle \omega(x_j, \varepsilon_j)\big) - F(U_{j-1}^*) - \int_{\omega(x_j,\varepsilon_j)} T_U \,\mathrm{d}x\right|\\
&\quad= \sum_{j=1}^{N} \Bigg(\left|F\big(U^* \triangle \omega(x_j, \varepsilon_j)\big) - F(U^*) - \int_{\omega(x_j,\varepsilon_j)} T_U \,\mathrm{d}x\right|\\
&\qquad\quad + \underbrace{\left|F\big(U_{j-1}^* \triangle \omega(x_j, \varepsilon_j)\big) - F(U_{j-1}^*) - F\big(U^* \triangle \omega(x_j, \varepsilon_j)\big) + F(U^*)\right|}_{\leq L \cdot \lambda(U_{j-1}^* \triangle U^*)} \Bigg)\\
&\quad\overset{A.3.6}{\leq} \sum_{j=1}^{N} \Bigg(L \cdot \lambda(U_{j-1}^* \triangle U^*)\\
&\qquad\quad + \left|F\big(U^* \triangle \omega(x_j, \varepsilon_j)\big) - F(U^*) - T_U(x_j) \cdot \lambda\big(\omega(x_j, \varepsilon_j)\big)\right|\\
&\qquad\quad + \left|T_U(x_j) \cdot \lambda\big(\omega(x_j, \varepsilon_j)\big) - \int_{\omega(x_j,\varepsilon_j)} T_U \,\mathrm{d}x\right| \Bigg)\\
&\quad\leq \sum_{j=1}^{N} \Bigg(L \cdot \lambda(U_{j-1}^* \triangle U^*)\\
&\qquad\quad + \underbrace{\left|F\big(U^* \triangle \omega(x_j, \varepsilon_j)\big) - F(U^*) - T_U(x_j) \cdot \lambda\big(\omega(x_j, \varepsilon_j)\big)\right|}_{\leq \frac{\rho}{6} \cdot \lambda(\omega(x_j,\varepsilon_j))}\\
&\qquad\quad + \underbrace{\int_{\omega(x_j,\varepsilon_j)} \big|T_U(y) - T_U(x_j)\big| \,\mathrm{d}y}_{\leq \frac{\rho}{6} \cdot \lambda(\omega(x_j,\varepsilon_j))} \Bigg)\\
&\quad\leq \sum_{j=1}^{N} \left(L \cdot \lambda\Big(\bigcup_{k=1}^{j} \omega(x_k, \varepsilon_k)\Big) + \frac{2 \cdot \rho}{6} \cdot \lambda\big(\omega(x_j, \varepsilon_j)\big) \right)\\
&\quad\leq \frac{L}{2} \cdot \lambda(\widetilde{D})^2 + \frac{2 \cdot \rho}{6} \cdot \lambda(\widetilde{D}).\end{aligned}$$

At this point, we make use of the fact that $\lambda(D^* \triangle \widetilde{D}) \leq \frac{1}{2} \cdot \lambda(D)$, which yields the estimate

$$\lambda(\widetilde{D}) \leq \lambda(D^*) + \lambda(D^* \triangle \widetilde{D}) \leq \frac{3}{2} \cdot \lambda(D).$$

By making an appropriate substitution in the above estimate, we obtain

$$\left|F(U^* \triangle \widetilde{D}) - F(U^*) - \int_{\widetilde{D}} T_U \,\mathrm{d}x\right| \leq \frac{9 \cdot L}{8} \cdot \lambda(D)^2 + \frac{3 \cdot \rho}{6} \cdot \lambda(D).$$

We can then introduce an additional bound on the step length in the form of

$$\delta \leq \delta_2 \coloneqq \frac{8 \cdot \rho}{54 \cdot L},$$

which gives us the following overall estimate for the telescope sum:

$$\left| F(U^* \triangle \widetilde{D}) - F(U^*) - \int_{\widetilde{D}} T_U \, \mathrm{d}x \right| \leq \frac{\rho}{6} \cdot \lambda(D) + \frac{3 \cdot \rho}{6} \cdot \lambda(D) = \frac{4 \cdot \rho}{6} \cdot \lambda(D).$$

By substituting this into the main estimate, we obtain

$$\begin{aligned}
&\left| F(U \triangle D) - F(U) - \int_D T_U \, \mathrm{d}x \right| \\
&\quad \leq \frac{2 \cdot \rho}{6} \cdot \lambda(D) + \left| F(U^* \triangle \widetilde{D}) - F(U^*) - \int_{\widetilde{D}} T_U \, \mathrm{d}x \right| \\
&\quad \leq \frac{2 \cdot \rho}{6} \cdot \lambda(D) + \frac{4 \cdot \rho}{6} \cdot \lambda(D) \\
&\quad = \rho \cdot \lambda(D).
\end{aligned}$$

Part 5 (Summary). In conclusion, we have demonstrated, that for every $\rho > 0$, there exists

$$\delta \coloneqq \min\{\delta_1, \delta_2\} = \min\left\{ \frac{R}{2}, \frac{8 \cdot \rho}{54 \cdot L} \right\} > 0$$

such that for every $D \in \Sigma_\sim$ with $\lambda(D) \leq \delta$, we have

$$\left| F(U \triangle D) - F(U) - \int_D T_U \, \mathrm{d}x \right| \leq \rho \cdot \lambda(D).$$

This shows that F is set differentiable according to Definition 2.4.1 in U with a derivative measure of

$$\nabla F(U)(D) = \int_D T_U \, \mathrm{d}x \quad \forall D \in \Sigma.$$

Evidently, the gradient density function of F in U is T_U. □

We stress that Theorem A.3.8 only requires the Lipschitz continuity of the ITDM to justify that the holes making up the approximate step $\widetilde{D}$ can be nucleated sequentially without rendering the topological derivative invalid. It is possible that this kind of additivity can be established under milder assumptions. We also note that, because the individual holes in the approximate step are closed and disjoint, any additivity argument need not consider cases where the holes touch, as they are always separated by some strictly positive distance.

The fact that additivity of an ITDM implies set differentiability is also interesting because it implies a converse implication: any ITDM that is not a set derivative is also not additive for finite disjoint unions of closed holes. This means that if an ITDM is not also a set derivative, then there is some finite disjoint union of holes for which it does not accurately predict the change of the functional value. This substantially reduces the usefulness of the topological derivative for optimality conditions, because a solution that looks optimal according to the topological derivative may not be optimal with respect to the simultaneous nucleation of even as few as two holes.

A.3.5 Notes on Methods of Topology Optimization

We close this section with some general notes on existing methods used in the field of topology optimization. Topology optimization is a well-established field of research with a history of several decades. Given the evident link between the set derivative and the topological derivative, it is legitimate to question why it is nearly absent from the discussions in the main part of this thesis.

One reason for this absence is the difficulty involved in rigorously establishing the link even under substantial restrictions to the generalized concept of the topological derivative. Theorems A.3.7 and A.3.8 clearly demonstrate that this argument is not straightforward. It could be argued that the modern topological derivative concept as presented in [NS13] is over-generalized, which makes it very difficult to make general statements about topological derivatives. Here, we have addressed this issue by making our own definition of a kind of topological derivative – the ITDM. By restricting the most nebulous aspects of the definition of the topological derivative, we create a scenario where it is possible to make such general statements at the cost of losing some of the concept's generality.

The other reason why topology optimization is mostly absent from our discussion is that the methods of topology optimization are sometimes justified empirically. This means that some optimization methods in the field are not rigorously demonstrated to work at all. One notable example of this is what the authors of [NSŻ19] present as Algorithm 1. This is a first-order method with a level-set representation of the solution based on the topological derivative and is similar to earlier methods proposed by [AA06; Ams11]. The idea behind this algorithm is very simple:

- the current iterate is encoded as the sublevel set $\{\Psi_i < 0\}$ of a function Ψ that is iteratively improved;
- the sign of the topological derivative is inverted inside of $\{\Psi_i < 0\}$ to compensate for the natural sign flip of the topological derivative;
- improvements are made by adding a multiple of the sign-corrected topological derivative function to Ψ_i.

There are many pitfalls in this algorithm as it is presented in [NSŻ19]. First, we note that level-set function adjustments of the sort that is used here do not create individual holes in the precise shape for which the topological derivative is developed. This means that there is no theoretical guarantee that the topological derivative has any predictive capability for the kinds of changes made by the algorithm. In Section 3 of their paper, the authors of [NSŻ19] state that a parameter determining the step size "is a step size determined by a line search performed in order to decrease the value of the objective function". However, they never argue that such a step size can actually be found.

This is a general problem with level-set algorithms that use multiples of the topological derivative as steps to adjust the level-set function. Because the topological derivative cannot be relied upon to predict objective changes for arbitrarily shaped set adjustments, there is also no guarantee that a descent step can be found. The authors acknowledge shortfalls in the theoretical justification of their algorithm in Section 5 of [NSŻ19].

In cases where the topological derivative is also a set derivative, the possible non-existence of a descent step "in the direction" of the topological derivative

disappears. However, even in this case, the first-order algorithm presented in [NSŻ19] falls short. To demonstrate this, we consider the set functional

$$J(U) \coloneqq \frac{1}{2} \cdot \big(\lambda(U) - 1\big)^2$$

on $\big([-1,\ 1], \mathscr{L}([-1,\ 1]), \lambda\big)$. This functional is evidently a composition of the Lebesgue measure λ, which is benignly differentiable according to Definition 2.4.1, with the polynomial $h\colon x \mapsto \frac{1}{2} \cdot (x-1)^2$. According to Theorem 2.4.6, J is benignly set differentiable in all $U \in \mathscr{L}([-1,\ 1])$ and the derivative satisfies

$$\nabla J(U) = h'\big(\lambda(U)\big) \cdot \nabla\lambda(U) = \big(\lambda(U) - 1\big) \cdot \nabla\lambda(U).$$

For each step D, we have

$$\nabla J(U)(D) = \big(\lambda(U) - 1\big) \cdot \big(\lambda(D \setminus U) - \lambda(U \setminus D)\big) = \int_D \underbrace{(1 - 2 \cdot \chi_U) \cdot \big(\lambda(U) - 1\big)}_{=: g_U} \,\mathrm{d}x.$$

Evidently, J is minimized by any subset of $[-1,\ 1]$ that has Lebesgue measure 1. An example of this would be $[0,\ 1]$. Next, we make use of Theorem A.3.7, which states that J admits an ITDM in every U with respect to any closed hole generator ω of bounded eccentricity and that the topological derivative function of J in U is equal to the gradient density function g_U.

After applying the sign correction proposed by the authors of [NSŻ19], we obtain the constant function $\widetilde{g}_U \equiv \lambda(U) - 1$. The fact that this sign-corrected topological derivative is a constant function presents a substantial problem to the first-order algorithm proposed by the authors of [NSŻ19]. If we were, for instance, to start with any constant level-set function Ψ_0 (representing either the empty set or the set $[-1,\ 1]$), then there would be no step size that did not either make no changes at all, or replace the current iterate with its complement. The algorithm would therefore keep oscillating between the starting set and its complement with no hope of ever progressing towards the optimum.

The issue is that the algorithm put forward in [NSŻ19] does not provide for a tie-breaking step such as the one performed in Line 18 of Algorithm 5 on page 242. The need for such a procedure is readily apparent in an in-depth theoretical discussion of the algorithm. However, it is quite easy to miss such edge cases when the algorithm is only justified empirically, because topological derivative functions with actual plateaus are rare in practical applications.

Another major class of topology optimization methods are nucleation methods such as the bubble method [EKS94] or its more modern descendants. For both the original paper proposing the modern form of this algorithm and a more recent example, we refer to [AJ05; LOS24]. Nucleation methods only use the topological derivative to place holes of the particular given shape and use the shape derivative for all other changes. They are therefore on a much more solid theoretical foundation with respect to their use of the topological derivative than the previous method. However, as we had discussed at the end of Section A.3.4, in cases where the topological derivative is not additive, the sense in which such methods can achieve optimality is somewhat questionable.

In conclusion, while the field of topology optimization provides veritable riches of theory concerning the nature and calculation of topological derivatives, due to its empirical approach, the optimization methods derived from this concept are

difficult to incorporate into rigorous theoretical frameworks. Many publications in the field make unstated and unjustified assumptions about the topological derivative that can easily trick the unprepared reader into drawing unfounded conclusions. For this reason, we make a conscious choice to not discuss results from topology optimization in the main part of this thesis.

A.4 FRACTIONAL PERIMETER AND CONDITIONALLY DIFFERENTIABLE FUNCTIONALS

The results in this section were developed in the time between submission and defense. They were not part of the original submission version and are therefore subject to a lesser standard of review. They are provided as a theoretical addendum for the interested reader.

One of the primary weaknesses of our approach is that the set derivative cannot capture changes that are happening on the boundary of a set. For instance, the perimeter or boundary integrals of a set are not set differentiable. This is a substantial weakness because boundary integrals are of great interest in PDE-constrained optimization. However, as we will see in this section, it is possible to approximate boundary integrals with set differentiable expressions.

This is not entirely unexpected, because boundary integrals are defined using the trace operator, which implicitly averages functions over small volumes near the boundary. It is therefore quite evident that boundary integrals could be expressed as limits of volumetric integrals.

In this section, we will focus exclusively on the perimeter and an approximation known as the "fractional perimeter". Demonstrating that the fractional perimeter is, in principle, accessible to our theoretical framework opens up the possibility of designing solvers that work with the true perimeter through adaptive approximation.

A.4.1 On Regularization of the True Perimeter

Before we begin, we first recapitulate why incorporating the true perimeter into the set derivative is not possible. This is not a conceptual weakness of the set derivative, but rather a property of the perimeter. To simplify, we will assume that "perimeter" is equal to the total variation of the indicator function of a set and that, for sufficiently benign sets, it can be thought of as the "surface area" of a set.

Let $A \subseteq \mathbb{R}^n$ be a Lebesgue-measurable set of finite perimeter, and let $\varepsilon > 0$. By an adjustment of A of volume less than or equal to $\frac{\varepsilon}{2}$, we ensure that there exists an ℓ^∞ ball B of ℓ^∞ diameter $\sqrt[n]{\frac{\varepsilon}{2}}$ such that B is either fully contained or disjoint from the modified set A'. This modification can change the perimeter of the set. However, it cannot increase the perimeter by more than the perimeter of B, which is finite. Therefore, A' is still of finite perimeter.

Let $M > 0$ be an arbitrary constant, and let $N \coloneqq \left\lceil \frac{M}{2 \cdot (\varepsilon/2)^{\frac{n-1}{n}}} \right\rceil$. Depending on whether B is a part of A' or disjoint from A', we now either remove or add N truncated hyperplanes (obtained by intersecting hyperplanes with B), spaced equidistantly along one coordinate axis inside of B. We then thicken these

truncated hyperplanes such that the resulting construct has strictly positive measure not exceeding $\frac{\varepsilon}{2}$, and that the thickened hyperplanes do not touch.

The resulting combined adjustment has a volume of less than ε. However, the perimeter of the resulting set is at least as large as the sides of the N thickened truncated hyperplanes, which have an $n-1$ dimensional measure of

$$2N \cdot \left(\sqrt[n]{\varepsilon/2}\right)^{n-1} \geq 2 \cdot (\varepsilon/2)^{\frac{n-1}{n}} \cdot \frac{M}{2 \cdot (\varepsilon/2)^{\frac{n-1}{n}}} = M.$$

Because this holds for all $\varepsilon > 0$ and all $M > 0$, around every set A of finite perimeter, within any distance $\varepsilon > 0$, the perimeter exceeds every positive real number. Therefore, the perimeter is discontinuous in every point and cannot be accounted for via the set derivative. For this reason, it was not taken into consideration during the research for the thesis.

It may be possible to accommodate perimeter bounds through specialized step finding routines, though it is not clear whether this would cause issues with the geodesic connectedness of the feasible set or convexity of the objective. Here, problems may arise due to the more counterintuitive structural aspects of similarity spaces. We will further elaborate on these issues when discussing the fractional perimeter.

We note that perimeter regularization is not as simple as adding the perimeter to the objective function. Unless treated with the utmost care, regularizing the perimeter yields extremely mesh-dependent solutions. To illustrate this, consider the set that minimizes the ratio between perimeter and volume. In $\mathbb{R}^n$ with the Lebesgue measure and its $n-1$ dimensional counterpart, it is well known that this ratio is minimized by the Euclidean ball. However, if we approximate sets via unions of entire cells in a mesh made entirely of axis-aligned rectangles, then the ℓ^∞ ball of the same radius has the same perimeter and a larger volume. This means that if we optimize with a regularization of the true perimeter on such a mesh, then our solutions will tend to have large axis-aligned rectangles in them because this shape accommodates the largest possible volume with the smallest possible perimeter. Other mesh geometries favor other shapes. However, it is very difficult to account for and control the impact of these effects.

Therefore, perimeter regularization should only be attempted with very sophisticated mesh control. At minimum, a perimeter-regularized optimization method should employ some sort of combined mesh movement and refinement or at least a level set encoding for the solution. Many level set encodings are, unfortunately, not capable of straightforwardly modelling the symmetric difference. Mesh movement comes with many numerical pitfalls and is significantly beyond the capabilities of the thesis' author. We will further elaborate on mesh dependency when discussing the fractional perimeter.

A.4.2 On Regularization of the Fractional Perimeter

Perimeter being somewhat problematic as part of the objective function, the question of how perimeter regularization might otherwise be achieved arises. In some recent publications in the field of optimal control, regularization based on the so-called *fractional perimeter* has been put forward as an alternative (see, e.g., [AM24]).

The fractional parameter of a Lebesgue-measurable set $A \subseteq \mathbb{R}^n$ is given by

$$P_\alpha(A) \coloneqq \int_A \int_{A^\complement} \frac{1}{\|x-y\|^{n+\alpha}} \,\mathrm{d}y \,\mathrm{d}x$$

for $\alpha \in (0,1)$. In this section, our primary goal is to show that, for a given fixed $\alpha \in (0,1)$, a suitably chosen feasible set, and appropriately constrained steps, the difference of P_α over a given step D would likely admit a second-order Taylor expansion in a lifted search space.

First, we note that, for a given point $x \in A \setminus \partial A$, we can bound the inner integral from above by

$$\begin{aligned}
\int_{A^\complement} \frac{1}{\|x-y\|^{n+\alpha}} \,\mathrm{d}y &= \int_{A^\complement - x} \frac{1}{\|y\|^{n+\alpha}} \,\mathrm{d}y \\
&\leq \int_{\mathbb{R}^n \setminus B_{\operatorname{dist}(x,A^\complement)}(0)} \frac{1}{\|y\|^{n+\alpha}} \,\mathrm{d}y \\
&= \int_{\operatorname{dist}(x,A^\complement)}^{\infty} \int_{S_r^{n-1}} \frac{1}{r^{n+\alpha}} \,\mathrm{d}s(x) \,\mathrm{d}r \\
&= \int_{\operatorname{dist}(x,A^\complement)}^{\infty} \frac{\operatorname{vol}(S_r^{n-1})}{r^{n+\alpha}} \,\mathrm{d}r \\
&= \frac{2\pi^{n/2}}{\Gamma(n/2)} \cdot \int_{\operatorname{dist}(x,A^\complement)}^{\infty} \frac{1}{r^{1+\alpha}} \,\mathrm{d}r \\
&= \frac{2\pi^{n/2}}{\Gamma(n/2)} \cdot \left[\frac{-\alpha}{r^\alpha}\right]_{\operatorname{dist}(x,A^\complement)}^{\infty} \\
&= \frac{2\pi^{n/2}\alpha}{\Gamma(n/2) \cdot \operatorname{dist}(x, A^\complement)}.
\end{aligned}$$

We therefore must observe the following during our reformulation:

- the iterated integrals under discussion should be assumed to not be finite unless the inner and outer integration domain are essentially disjoint;
- regardless of whether it is set differentiable, the fractional perimeter is unlikely to be benignly set differentiable because its density goes to infinity near the boundary of the sets under discussion;
- because of this, if we exchange the order of integration, the prerequisites of Tonelli's theorem must first be ensured by showing that the iterated integral is dominated by the fractional perimeter of either input set (which we will assume to have finite fractional perimeter).

In light of the last point, we must assume that our feasible set is a subset of

$$\mathscr{F}_\alpha \coloneqq \left\{A \in \Sigma_\sim \,\middle|\, P_\alpha(A) < \infty\right\}.$$

To simplify the following reformulation, we introduce the shorthand notation

$$\text{“}\int_X \int_Y \text{”} \coloneqq \int_X \int_Y \frac{1}{\|x-y\|^{n+\alpha}} \,\mathrm{d}y \,\mathrm{d}x.$$

for disjoint X and Y. We also allow for sums of integrals on the inner level with an analogous notation. Let $A, B \in \mathscr{F}_\alpha$. The fractional perimeter of A can be

decomposed into

$$\begin{aligned}\int_A \int_{A^\complement} &= \int_A \left(\int_{B\setminus A} + \int_{A^\complement \cap B^\complement} \right) \\ &= \int_{A\setminus B} \left(\int_{B\setminus A} + \int_{A^\complement \cap B^\complement} \right) + \int_{A\cap B} \left(\int_{B\setminus A} + \int_{A^\complement \cap B^\complement} \right).\end{aligned}$$

We note that these individual integrals are all dominated by $\int_A \int_{A^\complement}$, which is finite. Therefore, Tonelli's theorem allows us to freely exchange the order of integration for all four integrals as need arises. We can similarly argue that

$$\int_B \int_{B^\complement} = \int_{B\setminus A} \left(\int_{A\setminus B} + \int_{A^\complement \cap B^\complement} \right) + \int_{A\cap B} \left(\int_{A\setminus B} + \int_{A^\complement \cap B^\complement} \right).$$

The previous remark on the applicability of Tonelli's theorem applies here as well with $\int_B \int_{B^\complement}$ as the dominating integral. To clarify the following reformulation, we will mark equalities based on Tonelli's theorem with an asterisk ($*$). We observe that

$$\begin{aligned}&P_\alpha(A) - P_\alpha(B) \\ &\quad= \int_A \int_{A^\complement} - \int_B \int_{B^\complement} \\ &\quad= \int_{A\setminus B} \left(\int_{B\setminus A} + \int_{A^\complement \cap B^\complement} \right) - \int_{B\setminus A} \left(\int_{A\setminus B} + \int_{A^\complement \cap B^\complement} \right) + \int_{A\cap B} \left(\int_{B\setminus A} - \int_{A\setminus B} \right) \\ &\quad\overset{*}{=} \int_{A\setminus B} \left(\int_{B\setminus A} + \int_{A^\complement \cap B^\complement} - \int_{A\cap B} \right) - \int_{B\setminus A} \left(\int_{A\setminus B} + \int_{A^\complement \cap B^\complement} - \int_{A\cap B} \right) \\ &\quad\overset{*}{=} \int_{A\setminus B} \left(\int_{B\setminus A} + \int_{A^\complement \cap B^\complement} - \int_{A\cap B} \right) - \int_{A\setminus B} \int_{B\setminus A} - \int_{B\setminus A} \left(\int_{A^\complement \cap B^\complement} - \int_{A\cap B} \right) \\ &\quad= \int_{A\setminus B} \left(\int_{A^\complement \cap B^\complement} - \int_{A\cap B} \right) - \int_{B\setminus A} \left(\int_{A^\complement \cap B^\complement} - \int_{A\cap B} \right).\end{aligned}$$

Next, we use the fact that $A \triangle B = (A\setminus B) \cup (B\setminus A)$ and $(A\triangle B)^\complement = (A^\complement \cap B^\complement) \cup (A\cap B)$ to aggregate these integrals into

$$P_\alpha(A) - P_\alpha(B) = \underbrace{\int_{A\triangle B} \int_{(A\triangle B)^\complement} \frac{(1-\chi_B(x))\cdot(1-\chi_B(y))}{\|x-y\|^{n+\alpha}} \,\mathrm{d}y \,\mathrm{d}x}_{=:\Delta P_{\alpha,B}(A\triangle B)}.$$

We note that the density function in the integral on the right hand side depends only on B. While the integral does not behave like the simple set derivatives discussed in the thesis, this means that we can, in principle, incorporate a pre-calculated approximation of this density function in the step finding process.

The iterated integral representation of $\Delta P_{\alpha,B}$ shows interesting parallels to the way one would intuitively expect a second derivative to work. Indeed, as the thesis' section on logical constraints demonstrates, it is possible to optimize multivariate problems whose variables are subject to "partition constraints". If we interpret $\Delta P_{\alpha,B}$ as a functional in two set-valued variables, then we obtain the functional

$$G_{\alpha,B}(X,Y) \coloneqq \int_X \int_Y \frac{(1-\chi_B(x))\cdot(1-\chi_B(y))}{\|x-y\|^{n+\alpha}} \,\mathrm{d}y \,\mathrm{d}x,$$

which is only defined for essentially disjoint X and Y. Then $\Delta P_{\alpha,B}$ can be seen as the restriction of $G_{\alpha,B}$ to the feasible set of a partition constraint. In this restricted search space, $\Delta P_{\alpha,B}$ would then likely be the equivalent of a quadratic function whose second derivative has the density function

$$X \times X \ni (x,y) \mapsto \frac{(1-\chi_B(x))\cdot(1-\chi_B(y))}{\|x-y\|^{n+\alpha}}.$$

To account for the fact that the inner integral has to extend to all of $\mathbb{R}^n$, irrespective of whether the problem domain Ω is a true subset of $\mathbb{R}^n$ or not, a first-order term with the density function

$$x \mapsto \int_{\mathbb{R}^n \setminus \Omega} \frac{1-\chi_B(x)}{\|x-y\|^{n+\alpha}} \,\mathrm{d}y$$

would have to be added to account for the integration mass outside of Ω. Because $B \subseteq X$, we can simplify $1-\chi_B(y) = 1$ here.

In conclusion, a mesh-agnostic step finding solver for problems with fractional perimeter regularization can possibly be constructed on the foundation of the work presented in the thesis. Such a step-finding solver would have to approximately infimize a quadratic set functional subject to simultaneous partition and measure constraints. This requires a combination of the theories of set differentiability, scalar-valued inequality constraints (for the measure constraint), and logical constraints (for the partition constraint). In addition, it would likely require elaboration on the theory of higher order set derivatives sketched in Section A.1.

An additional concern is the likely necessity to restrict step choice to sets D with $\Delta P_{\alpha,B}(D) < \infty$. Directions γ that satisfy

$$\limsup_{t\to 0} \Delta P_{\alpha,B}(\gamma(t)) < \infty$$

or the stricter

$$\Delta P_{\alpha,B}(\gamma(t)) < \infty \quad \forall t$$

form a pseudocone. To ensure that $\Delta P_{\alpha,B}$ decays to zero as the underlying measure μ (likely a weighted version of λ in this case) decays to zero, it may be desirable to further restrict admissible search directions to

$$\mathscr{D}^{\infty}_{\alpha,B} \coloneqq \left\{ \gamma \in \mathrm{Dir}(\Sigma_{\sim}) \,\middle|\, \limsup_{t\to 0} \frac{\Delta P_{\alpha,B}(\gamma(t))}{t} < \infty \right\},$$

or the more restrictive

$$\mathscr{D}^{M}_{\alpha,B} \coloneqq \left\{ \gamma \in \mathrm{Dir}(\Sigma_{\sim}) \,\middle|\, \limsup_{t\to 0} \frac{\Delta P_{\alpha,B}(\gamma(t))}{t} \leq M \right\}$$

for $M > 0$, where $t = \mu(\gamma(t))$ because directions are minimizing by definition. Because all of these sets are pseudocones, they can be integrated into the KKT framework developed in the thesis, though the results of this are uncertain. It would then have to be investigated how all of these cones relate to the radial pseudocone of $\mathscr{F}_\alpha$ in B to ensure that all points within the feasible set $\mathscr{F}_\alpha$ remain reachable under such a restriction of search direction. We note that there is likely some sort of "topology of pseudocones" under which $\mathscr{D}^{\infty}_{\alpha,B}$ is open and $\mathscr{D}^{M}_{\alpha,B}$ is a closed approximation that converges to $\mathscr{D}^{M}_{\alpha,B}$ for $M \to \infty$.

The fractional perimeter is likely the first practical example of a class of set functionals that we have not properly addressed in this thesis: functionals that are only set differentiable along certain search directions. In general, the derivatives of such functionals would come with a pseudocone of admissible search directions. KKT theory would have to account for this by using the polar cone of that pseudocone in place of the polar cone of the universal pseudocone.

A further, less obvious issue with such search direction restrictions is that weakly secant convex functionals may not appear convex along any of the admissible search directions. This may cause some weakly secant convex functions to always appear non-convex to the optimization algorithm. As the concept of secant convexity is not yet sufficiently explored, the precise impact of this is unknown.

Due to the breadth of different areas touched upon in the thesis that are combined in this problem, the thesis should be understood as only performing preliminary work for the development of a mesh-agnostic step finding solver for problems with fractional perimeter regularization. It also appears evident that the additional amount of work necessary to consider fractional perimeter regularization in this thesis would be substantial.

It is conceivable that perimeter regularization could be better adressed in a metric space with a metric that combines measure and perimeter into one metric. However, this may require discarding the immensely powerful tool of density functions. Therefore, it may be difficult to solve step finding problems in the resulting metric space. We will address this possibility in Section A.5.

A.5 GENERALIZED DERIVATIVE FOR GEODESIC METRIC SPACES

The results in this section were developed in the time between submission and defense. They were not part of the original submission version and are therefore subject to a lesser standard of review. They are provided as a theoretical addendum for the interested reader.

In Sections A.3 and A.4, we have briefly touched upon alternative derivative concepts that apply to functionals that are not necessarily set differentiable in the sense of Definition 2.4.1 on page 152. These concepts are

- the generalized topological derivative, which can incorporate perimeter and boundary integrals through the choice of the function f, but requires a specific hole shape to be prescribed around each point;
- the shape derivative, which can deform set boundaries, but cannot change set topology;
- conditional set derivatives, which apply only along a limited pseudocone of search directions, but can potentially be used to approximate the perimeter and other boundary integrals.

In this section, we briefly contemplate a generalized derivative concept that may be able to unify all three of these concepts. At the foundation of this concept is the geodesic metric space.

Definition A.5.1 (Geodesic Metric Space).
A metric space (X, d) is called *geodesic* if and only if for every $x, y \in X$, there exist an interval $I \subseteq \mathbb{R}$, a geodesic $\gamma\colon I \to X$, and parameters $a, b \in I$ such that $\gamma(a) = x$ and $\gamma(b) = y$. ◁

The definition of a geodesic measure space in Section 2.3.6 is based on this concept. We now have to generalize the pseudocone concept to arbitrary metric spaces.

Definition A.5.2 (Directions and Pseudocones).
Let (X, d) be a geodesic metric space, and let $x \in X$. We refer to a minimizing geodesic γ as a *direction anchored in* x if and only if $0 \in \operatorname{dom}(\gamma)$, if there exists at least one $\varepsilon > 0$ with $\varepsilon \in \operatorname{dom}(\gamma)$, and if $\gamma(0) = x$. We refer to a set C of directions anchored in x as a *pseudocone anchored in* x. ◁

By demanding that a directions' parameter interval include both 0 and some strictly positive number, we ensure that each geodesic's parameter interval includes not only 0, but also some relative neighborhood of 0 within the non-negative numbers, which allows us to make the types of limit arguments that are essential for KKT theory and the theory of iterative optimization.

Specifying the anchor point of a pseudocone avoids the need to demand that X must have a group structure. While this is true of measure spaces, it is not always possible to create such a structure in arbitrary metric spaces.

Definition A.5.3 (Geodesic Derivative).
Let (X, d) be a geodesic metric space, let $x \in X$, and let $f : X \to \mathbb{R}$. We refer to f as *geodesically differentiable in* x *along a pseudocone* C that is anchored in x if and only if there exists a mapping $\nabla f(x) \colon C \ni \gamma \mapsto \nabla f(x)(\gamma) \in L^1\big(\operatorname{dom}(\gamma)\big)$ such that

$$f\big(\gamma(t)\big) = f(x) + \int_{\operatorname{conv}\{0,t\}} \nabla f(x)(\gamma)\,\mathrm{d}s + o(t)$$

for $t \in \operatorname{dom}(\gamma)$. ◁

This definition is very general. It incorporates all conditional set derivatives, including the unconditional set derivative, which is essentially a conditional derivative that applies along all directions. Furthermore, it subsumes all topological derivatives that use hole shapes that can be expressed as geodesics. It can likely be argued to subsume shape derivatives, conventional vector space derivatives, and even vector space semi-derivatives.

As opposed to the set derivative, it can be applied with alternate metrics. This likely makes it capable of incorporating perimeter and boundary integrals. As opposed to the topological derivative, the geodesic derivative allows us to assign different derivatives to different geodesics, which obviates the need to specify a single hole shape around each point. It could be argued that Definition A.5.3 is close to being a minimal requirement for the definition of iterative optimization schemes with line search globalization strategies.

The power of Definition A.5.3 comes at a substantial cost. An optimization scheme using the geodesic derivative would have to use a two-stage step finding procedure. In the first step, it would have to identify a descent direction within the cone of allowed search directions C. In the second step, a traditional line search procedure would have to applied to find the step length. The first step is much more problematic than the second.

In the main part of this thesis, we were able to explicitly derive a descent direction: the min-mean geodesic of the gradient density function. We were able to do this because we had the gradient density function as a tool that allows for the construction of the geodesic. In topology optimization, nucleation methods

artificially create a similar situation by limiting themselves to nucleating holes around points where the topological derivative exists and has the correct sign. In general metric spaces, there is no unified way to find descent directions. Finding descent directions requires in-depth study of the structure of geodesics in each individual metric space, similar to what we have done in Section 2.3. In general, there is no guarantee that the construction of descent directions is possible or easy. If optimality conditions do not take a sufficiently large subset of all possible directions into account, then they may be meaningless from a practical standpoint.

In light of all of these problems, Definition A.5.3 may not be practically viable. However, it could be worthwhile to investigate how much of optimization theory can be transferred to this derivative concept. It is likely that any theoretical result proven within such a general framework would transfer to most optimization settings.

Appendix $\mathcal{B}$

Additional Algorithms

B.1 POINTWISE MERGESORT

In Section 3.2.2, we discuss partition constraints, which are a special form of logical constraints. There, we point out that all logical constraints can be transformed into partition constraints, though doing so requires enumerating all boolean vectors that satisfy the underlying boolean formula, which may not always be practical.

To devise a step-finding procedure that works for arbitrary boolean functions in DNF, we would have to identify those feasible configurations which offer the best ratio between transition benefit and transition cost. Then, if additional measure "budget" is available within the trust region, we can switch to configurations whose cost-benefit ratio is worse, as long as doing so increases the overall projected descent.

Because boolean formulae in DNF are disjunctions, each clause of the formula has its own set of feasible configurations. In these feasible configurations, those variables that appear in the clause are "fixed." We cannot choose the setting for these variables. All other variables are free. To transition to a feasible configuration for a given clause, we must first set all fixed variables to the correct value. We then have to choose which of the free variables to switch.

It would stand to reason that we would switch these free variables in an order that is determined by the cost-benefit ratio of the switch, starting with the variables that promise the steepest descent. The cost-benefit ratio depends on gradient measures and is therefore different across the domain. The order in which we would perform switches in the free variables is therefore also different depending on where in the domain we are.

This means that we would need a sorting algorithm that sorts a list of density functions pointwise almost everywhere based on their value. In this section, we design a variant of the well-known MERGESORT algorithm (see, e.g., [Knu98, Sec. 5.2.4]) that does this. For brevity, we will not perform detailed proofs here. However, the general idea of the algorithm is easily understood.

The reason why we can transfer the MERGESORT algorithm is because it is a comparison-based sorting algorithm and because measurable functions support pointwise comparison through the determination of sublevel sets. We now proceed to lay out the basic operations necessary for MERGESORT's merge operation. In

Figure B.1 on the following page, these operations are symbolically depicted to make them easier to understand. Without loss of generality, we sort all lists in ascending order.

We begin with the simplest operation: Given two measurable functions f and g return their pointwise minimum and maximum. To do this, we first determine $\{f \leq g\}$, which is the same as $\{f - g \leq 0\}$. We then create an output function which is equal to f on $\{f \leq g\}$ and equal to g everywhere else. This is the minimum of both functions. Similarly, the maximum is equal to g on $\{f \leq g\}$ and equal to f everywhere else. This is formalized in Algorithm 12 and depicted in Figure B.1a.

Algorithm 12 Pointwise Two-Way Merge

Require: $n \in \mathbb{N}$, tuple f of measurable functions $f_i \colon X \to \mathbb{R} \cup \{\infty\}$, and measurable $i, j \colon X \to [n]$.

Ensure: Yields measurable $k, l \colon X \to [n]$ such that $\{k(x), l(x)\} = \{i(x), j(x)\}$ and $f_{k(x)}(x) \leq f_{l(x)}(x)$ everywhere.

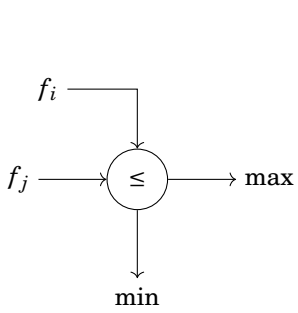
1: **function** MERGE-1-1(f, i, j)
2: $\quad g_1 = \sum_{k=1}^{n} \chi_{\{i=k\}} \cdot f_k$
3: $\quad g_2 = \sum_{k=1}^{n} \chi_{\{j=k\}} \cdot f_k$
4: $\quad A \leftarrow \{g_1 - g_2 \leq 0\}$
5: $\quad k \leftarrow \chi_A \cdot i + \chi_{A^\complement} \cdot j$
6: $\quad l \leftarrow \chi_A \cdot j + \chi_{A^\complement} \cdot i$
7: $\quad$ **return** $(k,\ l)$
8: **end function**

We note that we choose a slightly different notation. Instead of sorting two functions directly, we instead generate index functions that indicate how to combine functions from a pool of measurable functions to obtain the minimum and maximum. We do this because, in practice, we need to retain the index information to know which variables (i.e., which layers of the layered similarity space) the functions correspond to.

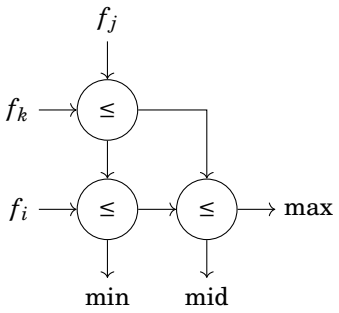
The second operation that we need to implement is the three-way merge. This is the basic operation of MERGESORT's overall merge operation. It takes three input functions. In practice, two of these input functions come from the two lists to be merged. The third is the maximum from the previous merge step, which is carried over to the next step. The three-way merge takes these three input functions and outputs the pointwise minimum, median, and maximum functions.

The three-way merge consists of three two-way merges. We first select any two inputs and perform a two-way merge between them. We then perform a two-way merge of the minimum with the remaining input. The minimum of that merge is the overall minimum. The third two-way merge is between the maxima of the two preceding two-way merges. Its minimum is the overall median, while its maximum is the overall maximum. This operation is formally described in Algorithm 13 on page 431 and depicted in Figure B.1b.

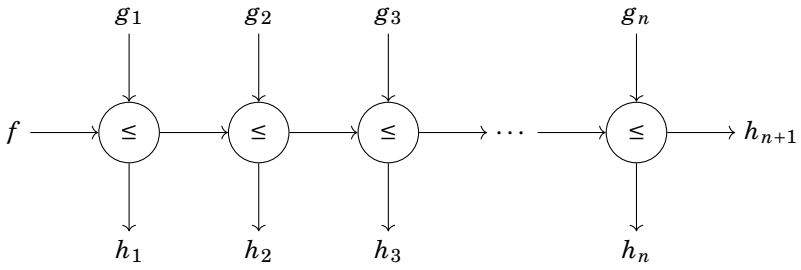
With these two elementary merge operations, we can proceed to merging larger lists of functions. The first larger merge that we discuss is the "1-n merge," which merges a single function f into a pointwise sorted list of measurable functions $(g_i)_{i \in [n]}$. This is essentially just a sequence of two-way merges. In the first merge, we merge f and g_1. The minimum of the merge becomes the



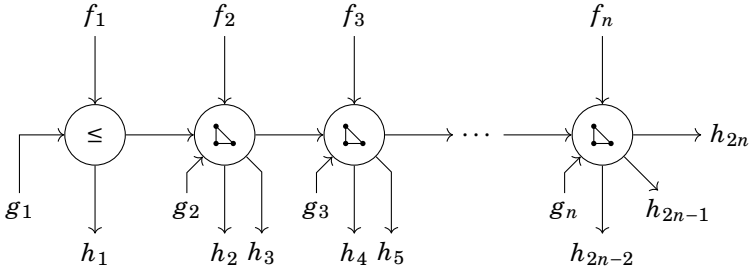
(a) Two-way merge



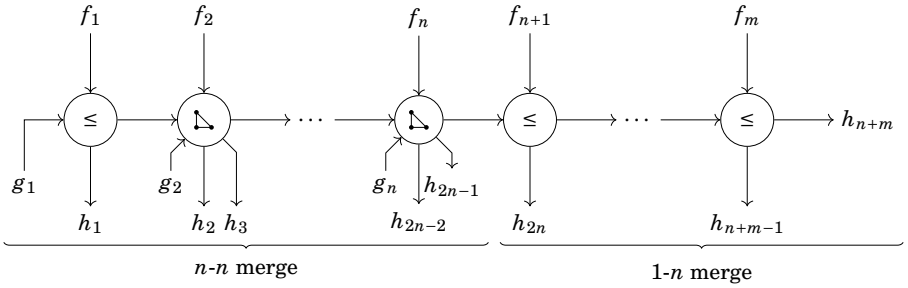
(b) Three-way merge (◺)



(c) 1-n merge of f into sorted $(g_i)_{i\in[n]}$ to obtain sorted $(h_i)_{i\in[n+1]}$



(d) n-n merge of sorted $(f_i)_{i\in[n]}$ and $(g_i)_{i\in[n]}$ to obtain sorted $(h_i)_{i\in[2n]}$



(e) m-n merge of sorted $(f_i)_{i\in[m]}$ and $(g_i)_{i\in[n]}$ with $m \geq n$ to obtain sorted $(h_i)_{i\in[n+m]}$

Figure B.1: Symbolic representation of the pointwise merge operation used by the MERGESORT algorithm to merge two pointwise sorted lists of functions into one pointwise sorted list. Location of arrows relative to operation nodes indicates role of function in operation.

Algorithm 13 Pointwise Three-Way Merge

Require: $n \in \mathbb{N}$, f n-tuple of measurable $f_i : X \to \mathbb{R} \cup \{\infty\}$, $i, j, k : X \to [n]$.
Ensure: Yields measurable $l_1, l_2, l_3 : X \to [n]$ such that

$$\{l_1(x),\ l_2(x),\ l_3(x)\} = \{i(x),\ j(x),\ k(x)\},$$
$$f_{l_m(x)}(x) \le f_{l_{m+1}(x)}(x) \qquad \forall m \in \{1,\ 2\}.$$

```
1: function MERGE-1-1-1(f, i, j, k)
2:     (m_1, m_2) ← MERGE-1-1(f, j, k)
3:     (l_1, m_3) ← MERGE-1-1(f, i, m_1)
4:     (l_2, l_3) ← MERGE-1-1(f, m_3, m_2)
5:     return (l_1, l_2, l_3)
6: end function
```

first element of the output list. The maximum gets carried over into a two-way merge with g_2, whose minimum becomes the second element of the output list. This continues until we obtain the maximum of the final merge with g_n, which becomes the $(n+1)$-st element of the output list. The transitivity of the order relation "$\le$" ensures that the maximum that is carried over to the next two-way merge is always pointwise greater than or equal to all prior elements of the list, which is essential to proving that the output list is pointwise sorted. The 1-n merge is formally described in Algorithm 14 and depicted in Figure B.1c.

Algorithm 14 Pointwise 1-n Merge

Require: $m \in \mathbb{N}$, f m-tuple of measurable $f_i : X \to \mathbb{R} \cup \{\infty\}$, $j : X \to [m]$ measurable, $n \in \mathbb{N}$, $i : X \to [m]^n$ componentwise measurable such that

$$f_{i_l(x)}(x) \le f_{i_{l+1}(x)}(x) \qquad \forall l \in [m-1]$$

almost everywhere.
Ensure: Yields componentwise measurable $k : X \to [m]^{n+1}$ such that

$$\{k_l(x) \,|\, l \in [m+1]\} = \{j(x)\} \cup \{i_l(x) \,|\, l \in [n]\},$$
$$f_{k_l(x)}(x) \le f_{k_{l+1}(x)}(x) \qquad \forall l \in [n]$$

almost everywhere.

```
1: function MERGE-1-N(f, i, j)
2:     l ← 1
3:     while l ≤ n do
4:         (k_l, j) ← MERGE-1-1(f, j, i_l)
5:         l ← l + 1
6:     end while
7:     k_{n+1} ← j
8:     return k
9: end function
```

The primary merge operation is the "n-n merge," which we address next.

It merges two lists of equal size. Like a 1-n merge, it begins with a two-way merge between the leading elements of both lists. The minimum becomes the first output element. The maximum is carried over. For all subsequent steps, however, we perform a three-way merge between the two next elements in the input lists and the carried maximum. The minimum and median are appended to the output list in the correct order, while the maximum is carried over to the next three-way merge. After both lists are exhausted, the final maximum becomes the final element of the output list. The n-n merge is formally described in Algorithm 15 and depicted in Figure B.1d.

Algorithm 15 Pointwise n-n Merge

Require: $m \in \mathbb{N}$, f m-tuple of measurable $f_i \colon X \to \mathbb{R} \cup \{\infty\}$, $n \in \mathbb{N}$, $j \colon X \to [m]^n$ and $i \colon X \to [m]^n$ componentwise measurable such that

$$\begin{aligned} f_{i_l(x)}(x) &\leq f_{i_{l+1}(x)}(x) && \forall l \in [n-1], \\ f_{j_l(x)}(x) &\leq f_{j_{l+1}(x)}(x) && \forall l \in [n-1] \end{aligned}$$

almost everywhere.

Ensure: Yields componentwise measurable $k \colon X \to [m]^{2n}$ such that

$$\begin{aligned} \{k_l(x) \,|\, l \in [2n]\} &= \{j_l(x) \,|\, l \in [n]\} \cup \{i_l(x) \,|\, l \in [n]\}, \\ f_{k_l(x)}(x) &\leq f_{k_{l+1}(x)}(x) && \forall l \in [2n] \end{aligned}$$

almost everywhere.

1: **function** MERGE-N-N(f, i, j)
2: $(k_1, r) \leftarrow$ MERGE-1-1(f, i_1, j_1)
3: $l \leftarrow 2$
4: **while** $l \leq n$ **do**
5: $(k_{2l-2}, k_{2l-1}, r) \leftarrow$ MERGE-1-1-1(f, r, i_l, j_l)
6: $l \leftarrow l + 1$
7: **end while**
8: $k_{2n} \leftarrow r$
9: **return** k
10: **end function**

Finally, we combine the n-n merge and the 1-n merge to obtain the more general m-n merge which is capable of merging sorted lists of differing length. This is quite simple. If $m \geq n$, then we perform an n-n merge on the first n elements of both lists. We then carry the maximum of that n-n merge into a 1-n merge with the remaining $m - n$ elements of the longer list. This is formally stated in Algorithm 16 on the next page and depicted in Figure B.1e.

Now that we have defined all necessary merge operations, we can state the full MERGESORT algorithm. We do so in Algorithm 17 on the facing page. The algorithm starts with a tuple $M^{(1)}$ of n single-element tuples of index functions, each of which is initialized with a constant corresponding to its index in $M^{(1)}$. In each step of the main loop, $M^{(j+1)}$ is generated by merging two adjacent index tuples with an m-n merge and the variable m_j is updated to keep track of the length of $M^{(j)}$. In each step, this length is roughly halved, until only one element

Algorithm 16 Pointwise m-n Merge

Require: $n_f \in \mathbb{N}$, f n_f-tuple of measurable $f_i : X \to \mathbb{R} \cup \{\infty\}$, $m \in \mathbb{N}$, $n \in \mathbb{N}$ with $n \leq m$, $i : X \to [n_f]^m$ and $j : X \to [n_f]^n$ componentwise measurable such that

$$\begin{aligned} f_{i_l(x)}(x) &\leq f_{i_{l+1}(x)}(x) \qquad \forall l \in [m-1], \\ f_{j_l(x)}(x) &\leq f_{j_{l+1}(x)}(x) \qquad \forall l \in [n-1] \end{aligned}$$

almost everywhere.

Ensure: Yields componentwise measurable $k : X \to [n_f]^{m+n}$ such that

$$\begin{aligned} \left\{k_l(x) \,\middle|\, l \in [m+n]\right\} &= \left\{j_l(x) \,\middle|\, l \in [n]\right\} \cup \left\{i_l(x) \,\middle|\, l \in [m]\right\}, \\ f_{k_l(x)}(x) &\leq f_{k_{l+1}(x)}(x) \qquad \forall l \in [m+n-1] \end{aligned}$$

almost everywhere.

1: **function** MERGE-M-N(f, i, j)
2: $k' \leftarrow$ MERGE-N-N(f, $i_{[n]}$, $j_{[n]}$)
3: $k'' \leftarrow$ MERGE-1-N(f, $i_{[m]\setminus[n]}$, k'_{2n})
4: Let k be the concatenation of $k'_{[2n-1]}$ and k'' in that order.
5: **return** k
6: **end function**

Algorithm 17 Pointwise MERGESORT

Require: $n \in \mathbb{N}$, f n-tuple of measurable $f_i : X \to \mathbb{R} \cup \{\infty\}$.
Ensure: Yields $k \in X \to [n]^n$ componentwise measurable such that

$$\begin{aligned} \left\{k_i(x) \,\middle|\, i \in [n]\right\} &= [n], \\ f_{k_i(x)}(x) &\leq f_{k_{i+1}(x)}(x) \qquad \forall i \in [n-1] \end{aligned}$$

almost everywhere.

function MERGESORT(f)
$M^{(1)} \leftarrow \left((X \ni x \mapsto i)\right)_{i \in [n]}$
$m_1 \leftarrow n$
$j \leftarrow 1$
while $m_j > 1$ **do**
for all $k \in \left[\lfloor m_j/2 \rfloor\right]$ **do**
$M_k^{(j+1)} \leftarrow$ MERGE-M-N$\left(f, M_{2k-1}^{(j)}, M_{2k}^{(j)}\right)$
end for
if $m_j \equiv 1 \pmod 2$ **then**
$M_{\lceil m_j/2 \rceil}^{(j+1)} \leftarrow M_{m_j}^{(j)}$
end if
$m_{j+1} \leftarrow \left\lceil \frac{m_j}{2} \right\rceil$
$j \leftarrow j+1$
end while
$k \leftarrow M_1^{(j)}$
return k
end function

remains. Due to the guarantees for the output of the m-n merge operation, the output is pointwise sorted.

Appendix C

PyCoimset Continuous Improvement of Sets in Python

As part of this thesis, we develop a reusable software package called PyCoimset, which aims to be a reusable and flexible framework that third parties can use to apply iterative set-valued optimization algorithms to their own problems.

At the time of writing, PyCoimset is at version 0.1.7. The package follows semantic versioning guidelines, which means that the interface is currently considered immature and subject to future change. The full source code for PyCoimset is available under the Apache 2.0 License[1]. Version 0.1.7 has also been archived under [Hah25a].

C.1 DESIGN CHOICES

The design of PyCoimset follows some overarching principles that are intended to maximize its flexibility for future extension and application.

Implementation Agnosticism

PyCoimset is designed to not force its user into any particular way of implementing the underlying set and density function encodings as well as functional evaluations. This is achieved by following the overarching paradigm of *implementation agnosticism*.

The software is roughly divided into two segments, which we will call the *problem implementation layer* and the *algorithmic layer*. The problem implementation layer is the lower level layer where, ideally, all code specific to the problem implementation resides. The algorithmic layer operates on top of the problem implementation layer and interacts with problem-specific code exclusively through a defined interface exposed by the problem implementation layer.

[1] It can be downloaded at `https://github.com/mirhahn/pycoimset`.

Python as Implementation Language

We choose Python as the implementation language because it allows for rapid prototyping and has a very low barrier to entry. The Python ecosystem already has a wide variety of mature numerical software, such as, e.g., NumPy, SciPy, or the PDE solver library FEniCS. Python is also known to have very mature tools to generate bindings to C/C++ code, such as Cython and PyBind11, which allows for Python code to interface with high-performance problem implementations written in nearly every other language that can interface with C/C++. Other languages, such as Julia, also provide facilities to bind directly to Python to make use of its rich ecosystem of pre-existing packages.

Python is a weakly typed language with strong support for so-called "duck typing." Duck typing is an object-oriented programming paradigm in which objects do not formally have to inherit from a base type in order to be used by an interface. Instead, they only have to implement the necessary functionality in order to be used. This makes problem experts much more flexible in how they design their type hierarchies. With version 3.8, Python has introduced so-called *protocols*, which provide a way of defining type requirements for an interface without the need for inheritance. Protocols are designed to be used in conjunction with static type checkers to ease error diagnosis and avoidance in a duck typing context. PyCoimset defines the interface of the problem implementation layer using protocols.

The choice of Python as an implementation language is not without drawbacks. Python is known to have irreconcilable difficulties with multithreading and to have very poor performance when processing large amounts of data in pure Python code. This is offset when much of that work can be outsourced to natively compiled extensions such as NumPy. However, it is not clear whether a pure Python implementation is a long-term viable path for PyCoimset. At current time, having a Python interface is almost inevitable for any piece of scientific software and it is always possible to move the library to C/C++ and provide the Python interface through a binding in the future.

C.2 PROBLEM IMPLEMENTATION LAYER

The problem implementation layer of PyCoimset consists of a system of interdependent duck typing protocols. These protocols define the behavior of

- similarity classes;
- signed measures;
- similarity spaces;
- and set functionals.

These protocols are declared in the `pycoimset.typing` module.

Similarity Classes

Similarity class types are used to encode solutions of the optimization problem as well as steps between such solutions. Their behavior is prescribed by the `SimilarityClass` protocol. This protocol requires the following functionality

- "`set.space`" yields the underlying similarity space, which must be an object implementing the `SimilaritySpace` protocol;
- "`set.measure`" yields the class' measure using the underlying space's default measure;
- "`set.subset(lb, ub, [hint])`" yields a similarity class with measure between `lb` and `ub` that is an essential subset of `set` with `hint` providing an optional signed measure that the algorithm should (but does not need to) try to minimize;
- basic set operations:
  - "`~set`" yields the complement of `set`,
  - "`set | other`" yields the union of `set` and `other`,
  - "`set & other`" yields the intersection of `set` and `other`,
  - "`set - other`" yields the difference between `set` and `other`,
  - "`set ^ other`" yields the symmetric difference of `set` and `other`.

Similarity classes are assumed to be immutable, i.e., a given similarity class cannot be modified after it has been created. This is meant to prevent inadvertent corruption of solutions and steps after they have been stored. It also simplifies caching logic. Many internal caches used by PyCoimset solvers are indexed using object references rather than the content of those objects.

Users are free to implement several different types of similarity classes for a given problem. This can, for instance, be used to simplify the encoding of specific classes such as a similiarity space's universal and empty classes. However, all similarity classes that share the same underlying similarity space must implement all basic set operations.

Implementers are encouraged to use copy-on-write and lazy evaluation to avoid unnecessary work. However, they should bear in mind that such methods, particularly lazy evaluation, can also slow down execution if they are implemented in pure Python.

Signed Measures

Signed measures are primarily used to encode gradient measures. All signed measures are assumed to be absolutely continuous with respect to the native measure of the underlying similarity space. The signed measure object must have access to a representation of the measure's density function and provide level set generation operations.

The `SignedMeasure` protocol requires the following functionality:

- "`meas.space`" yields the underlying similarity space, which is an object implementing `SimilaritySpace`;
- "`meas(set)`" yields the measure of a similarity class `set` with the same underlying similarity space as `meas`;
- "`meas.norm('L1')`" and "`meas.l1_norm`" yield the L^1 norm of the density function;

- "`meas.norm('Linfty')`" and "`meas.linfty_norm`" yield the L^∞ norm of the density function;
- basic linear arithmetic:
  - "`-meas`" yields the additive inverse of `meas`,
  - "`num * meas`" yields a real scalar multiple of `meas`,
  - "`meas / num`" yields the quotient of `meas` by a nonzero real scalar `num`,
  - "`meas + other`" yields the sum of signed measures `meas` and `other` that share an underlying similarity space,
  - "`meas - other`" yields the difference between signed measures `meas` and `other` that share an underlying similarity space;
- level set generation (`other` can be a scalar or another measure):
  - "`meas < other`" yields the strict sublevel set relative to `other`,
  - "`meas <= other`" yields the sublevel set relative to `other`,
  - "`meas > other`" yields the strict superlevel set relative to `other`,
  - "`meas >= other`" yields the superlevel set relative to `other`.

We note that all norm-related code is only needed as part of the penalty solver and can be omitted. The protocol offers a default implementation of the L^1 norm and a version of `meas.norm` that caches its outputs as default mixins.

As with similarity classes, signed measures are immutable and must be compatible with every other signed measure or similarity class type within the same similarity space.

Similarity Spaces

Similarity spaces are relatively lightweight representations of the underlying spaces from which similarity classes are drawn and on which signed measures are defined. Their primary function within PyCoimset is that their object ID marks similarity classes and signed measures that are compatible with one another.

As such, it is essential that similarity space objects persist throughout the optimization process and are never re-created or replaced. They are not immutable, but must satisfy certain invariance requirements.

The `SimilaritySpace` protocol requires the following functionality:

- "`spc.measure`" yields the measure of the universal similarity class under the space's native measure and must never change;
- "`spc.empty_class`" yields an encoding of the empty similarity class;
- "`spc.universal_class`" yields an encoding of the universal similiarity class.

The objects returned by `spc.empty_class` and `spc.universal_class` may change over time and should reflect global refinement decisions as they are made. However, `spc.measure` is not allowed to change and must always be an upper bound on the `measure` property of all similarity classes within the space.

Set Functionals

Finally, set functionals encode objective and constraint functionals. They accept similarity classes as input and produce signed measures as gradient outputs. Set functionals must implement the `Functional` protocol.

Set functionals are not immutable, because they have a central role in PyCoimset's caching strategy. Each functional retains a reference to its last evaluation argument as an attribute. This allows it to compare new arguments to the last argument and delete cached results only when the argument changes. This caching interface is very intrusive and may be abandoned in favor of a less intrusive one in a future release.

At the moment, the `Functional` prescribes the following interface for a given function `func`:

- "`func.input_space`" is the `SimilaritySpace` from which arguments must be drawn;
- "`func.grad_tol_type`" is a read-only enum value from the enumeration `pycoimset.typing.ErrorNorm` that is either `ErrorNorm.L1` or `ErrorNorm.Linfty` to indicate that gradient error is controlled using the L^1 or L^∞ norm, respectively;
- "`func.arg`" is the current input `SimilarityClass` (or `None`) and is set via `func.arg = set`;
- "`func.val_tol`" is the current functional evaluation tolerance, which is a settable floating-point property;
- "`func.grad_tol`" is the current gradient evaluation tolerance, which is a settable floating-point property;
- "`func.get_value()`" evaluates the functional value at the current argument and tolerance and returns a tuple `(val, err)` of value and error estimate;
- "`func.get_gradient()`" evaluates the gradient at the current argument and tolerance and returns a tuple `(grad, err)` of gradient and error estimate.

In addition to this interface, the following expressions must be defined for floating-point numbers `num` and are implemented as mixins for classes that inherit from the protocol:

- "`func <= num`" returns an object of type `pycoimset.typing.Constraint` that represents a scalar inequality constraint of the form $G(U) \leq b$;
- "`func >= num`" returns an object of type `pycoimset.typing.Constraint` that represents a scalar inequality constraint of the form $G(U) \geq b$.

The basic evaluation procedure for a functional is stated in Listing C.1 on the next page. Generally, both tolerances are set to some value before any evaluation is done. This is to increase the likelihood of intermediate results being reusable.

Listing C.1: Evaluation template for set functionals in PyCoimset.

```
# Set arguments and tolerances before evaluation
func.arg = my_set
func.val_tol = vtol
func.grad_tol = gtol

# Perform evaluation
fval, ferr = func.get_value()
gval, gerr = func.get_gradient()
```

C.3 ALGORITHMIC LAYER

The algorithmic layer is centered around the solver implementations in the `pycoimset.solver` module. At the time of writing, there are two solver implementations: `UnconstrainedSolver` and `PenaltySolver`. Both of these take the objective functional, constraints (if applicable), and initial solutions as constructor arguments. In addition, they accept a variety of algorithmic parameters and an optional implementation of `pycoimset.typing.UnconstrainedStepFinder`, if the user wants to specify an alternative step finder. Currently, the only implementation of that protocol is `pycoimset.step.SteepestDescentStepFinder`, which is used by default if no step finder is specified.

Once constructed, both solvers provide the method `solver.solve()`, which solves the optimization problem to the tolerances specified in the algorithmic parameters.

Retrieving Solution Information

Once the solver is finished, the termination status can be retrieved through the `solver.status` attribute, the precise meaning of which differs between solvers. The solver's current iterate is retrievable through `solver.x`. Some statistics about iteration counts, step sizes, and objective functionals can be retrieved through `solver.stats`, though the precise format differs between solvers.

Monitoring through Callbacks

Both solver variants have a callback interface that allows the user to inject a callable through the `solver.callback` attribute. This callable is called once per iteration and receives only the solver itself as an argument. This can be used to preserve intermediate solutions and progress information for later evaluation that would otherwise be discarded.

Functional Helpers

PyCoimset provides three simple proxies that adapt functionals in some way. These are located in `pycoimset.helpers.functionals`. The helper `transform` applies an affine linear transformation to the output of an existing functional, `with_safety_factor` applies a safety factor to error estimates and bounds, and `weighted_sum` combines multiple functionals into a single functional whose output is a weighted sum of the input functionals.

These helpers are somewhat non-trivial to implement because they change error bounds and estimates and perform arithmetic with multiple gradient measures.

Usage Example

To illustrate the usage of PyCoimset, we provide Listings C.2 and C.3 on the following page and on page 443. These are adapted from the main solver file of the test problem in Section 4.2. They are substantially simplified and include none of the code implementing the relevant functionals. They are merely intended to illustrate the usage of the solver classes in conjunction with functional helpers.

Listing C.2: Simplified evaluation code for the penalty run of the test problem from Section 4.2.

```python
# Import scikit-fem
import skfem

# Import types from problem implementation
from functionals import MeasureFunctional, ObjectiveFunctional
from space import BoolArrayClass, SimilaritySpace

# Import pycoimset types and functions
from pycoimset import PenaltySolver
from pycoimset.helpers.functionals import with_safety_factor


# Construct initial mesh using scikit-fem.
initial_mesh = skfem.MeshTri().refined(6)
space = SimilaritySpace(initial_mesh)
ctrl = BoolArrayClass(space, space.mesh)

# Set up solver.
sol_param = PenaltySolver.Parameters()
f = ObjectiveFunctional(space)
g = MeasureFunctional(space)

solver = PenaltySolver(
    # Objective
    with_safety_factor(f, 0.05),
    # Constraint
    g <= 0.4,
    # Parameters
    x0=ctrl,                # Initial solution
    mu=0.01,                # Initial penalty parameter
    err_wgt=[0.0, 1.0],     # Tolerance apportionment
    param=sol_param         # Additional parameters
)

# Solve the problem.
solver.solve()
```

Listing C.3: Simplified evaluation code for the unconstrained run of the test problem from Section 4.2.

```
# Import scikit-fem
import skfem

# Import types from problem implementation
from functionals import MeasureFunctional, ObjectiveFunctional
from space import BoolArrayClass, SimilaritySpace

# Import pycoimset types and functions
from pycoimset import UnconstrainedSolver
from pycoimset.helpers.functionals import (
    with_safety_factor,
    weighted_sum
)
from pycoimset.solver.unconstrained.solver import (
    SolverParameters
)


# Construct initial mesh.
initial_mesh = skfem.MeshTri().refined(6)
space = SimilaritySpace(initial_mesh)
ctrl = BoolArrayClass(space, space.mesh)

# Set up solver.
sol_param = SolverParameters()
f = ObjectiveFunctional(space)
g = MeasureFunctional(space)

solver = UnconstrainedSolver(
    weighted_sum(
        [
            with_safety_factor(f, 0.05),
            g
        ],
        [1.0, 8.75e-5],
        [1.0, 0.0],          # Tolerance apportionment (value)
        [1.0, 0.0]           # Tolerance apportionment (grad)
    ),
    initial_sol=ctrl,        # Initial solution
    param=sol_param          # Additional parameters
)

# Solve the problem.
solver.solve()
```

Bibliography

[AJ05] Grégoire Allaire and François Jouve. “Coupling the Level Set Method and the Topological Gradient in Structural Optimization.” In: *IUTAM Symposium on Topological Design Optimization of Structures, Machines and Materials*. Vol. 137. Rungstedgaard, Denmark: Springer Netherlands, Oct. 2005, pp. 3–12. DOI: 10.1007/1-4020-4752-5_1. URL: https://cnrs.hal.science/hal-04497790.

[Ams11] Samuel Amstutz. “Connections between topological sensitivity analysis and material interpolation schemes in topology optimization.” In: *Structural and Multidisciplinary Optimization* 43.6 (June 2011), pp. 755–765. ISSN: 1615-1488. DOI: 10.1007/s00158-010-0607-6.

[AA06] Samuel Amstutz and Heiko Andrä. “A new algorithm for topology optimization using a level-set method.” In: *Journal of Computational Physics* 216.2 (2006), pp. 573–588. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2005.12.015. URL: https://www.sciencedirect.com/science/article/pii/S0021999105005656.

[AM24] Harbir Antil and Paul Manns. “Integer Optimal Control with Fractional Perimeter Regularization.” In: *Applied Mathematics & Optimization* 90.1 (July 2024), p. 14. ISSN: 1432-0606. DOI: 10.1007/s00245-024-10157-y.

[BSS06] Mokhtar S. Bazaraa, Hanif D. Sherali, and C. M. Shetty. *Nonlinear Programming: Theory and Applications*. 3rd ed. Hoboken, New Jersey, USA: John Wiley & Sons, Inc., 2006. ISBN: 978-0-4717-8777-8. DOI: 10.1002/0471787779.

[BR01] Roland Becker and Rolf Rannacher. “An optimal control approach to a posteriori error estimation in finite element methods.” In: *Acta Numerica* (2001), pp. 1–102. DOI: 10.1017/S0962492901000010.

[BS04] Martin P. Bendsøe and Ole Sigmund. *Topology Optimization: Theory, Methods, and Applications*. 2nd ed. Springer-Verlag Berlin Heidelberg, 2004. ISBN: 978-3-662-05086-6. DOI: 10.1007/978-3-662-05086-6.

[BC09] John J. Benedetto and Wojciech Czaja. *Integration and Modern Analysis*. 1st ed. Birkhäuser Boston, 2009. ISBN: 978-0-8176-4656-1. DOI: 10.1007/978-0-8176-4656-1.

[Ber74] Leonard D. Berkovitz. *Optimal Control Theory*. New York, NY: Springer New York, 1974. ISBN: 978-1-4757-6097-2. DOI: 10.1007/978-1-4757-6097-2.

[Bet+21] Johanna Bethge et al. "Mathematical Optimization and Machine Learning for Efficient Urban Traffic." In: *German Success Stories in Industrial Mathematics*. Ed. by Hans Georg Bock et al. Cham: Springer International Publishing, 2021, pp. 113–120. ISBN: 978-3-030-81455-7. DOI: 10.1007/978-3-030-81455-7_19.

[BL85] Hans Georg Bock and Richard W. Longman. "Computation of optimal controls on disjoint control sets for minimum energy subway operation." In: *Advances in the Astronautical Sciences* 50 (1985), pp. 949–972.

[Boc+17] Hans Georg Bock et al. "Minimum Energy Time Tables for Subway Operation - And Hamiltonian Feedback to Return to Schedule." In: *Modeling, Simulation and Optimization of Complex Processes HPSC 2015*. Cham: Springer International Publishing, 2017, pp. 1–14. ISBN: 978-3-319-67168-0. DOI: 10.1007/978-3-319-67168-0_1.

[Bog07] Vladimir Bogachev. *Measure Theory*. Springer-Verlag Berlin Heidelberg, 2007. ISBN: 978-3-540-34514-5. DOI: 10.1007/978-3-540-34514-5.

[BS20] Vladimir I. Bogachev and Oleg G. Smolyanov. *Real and Functional Analysis*. Springer Cham, 2020. ISBN: 978-3-030-38219-3. DOI: 10.1007/978-3-030-38219-3.

[Bol13] Oskar Bolza. "Über den ‚Anormalen Fall' beim Lagrangeschen und Mayerschen Problem mit gemischten Bedingungen und variablen Endpunkten." In: *Mathematische Annalen* 74 (1913), pp. 430–446. URL: https://resolver.sub.uni-goettingen.de/purl?PPN235181684_0074.

[Bra13] Dietrich Braess. *Finite Elemente. Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie*. 5th ed. Springer-Verlag Berlin Heidelberg, 2013. ISBN: 978-3-642-34796-9. DOI: 10.1007/978-3-642-34796-9.

[Bre11] Haim Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. 1st ed. Universitext. Springer New York, 2011. ISBN: 978-0-387-70914-7. DOI: 10.1007/978-0-387-70914-7.

[BK10] Martin Brokate and Götz Kersting. *Measure and Integral*. 1st ed. Compact Textbooks in Mathematics. Birkhäuser Cham, 2010. ISBN: 978-3-319-15364-0. DOI: 10.1007/978-3-319-15364-0.

[BBI01] Dmitri Burago, Yuri Burago, and Sergei Ivanov. *A Course in Metric Geometry*. Vol. 33. Graduate Studies in Mathematics. Oxford University Press, 2001. ISBN: 978-1-4704-1794-9. DOI: 10.1090/gsm/033.

[Bür+20] Adrian Bürger et al. "pycombina: An Open-Source Tool for Solving Combinatorial Approximation Problems Arising in Mixed-Integer Optimal Control." In: *21st IFAC World Congress*. International Federation of Automatic Control. 2020. DOI: 10.1016/j.ifacol.2020.12.1799.

[CGM74] J. Céa, A. Gioan, and J. Michel. “Adaptation de la methode du gradient a un probleme d’identification de domaine.” In: *Computing Methods in Applied Sciences and Engineering Part 2*. Ed. by R. Glowinski and J. L. Lions. Berlin, Heidelberg: Springer Berlin Heidelberg, 1974, pp. 391–402. ISBN: 978-3-540-38380-2. DOI: 10.1007/3-540-06769-8_19.

[Céa+00] Jean Céa et al. “The shape and topological optimizations connection.” In: *Computer Methods in Applied Mechanics and Engineering* 188.4 (2000). IVth World Congress on Computational Mechanics. (II). Optimum, pp. 713–726. ISSN: 0045-7825. DOI: 10.1016/S0045-7825(99)00357-6. URL: https://www.sciencedirect.com/science/article/pii/S0045782599003576.

[Cla20] Christian Clason. *Introduction to Functional Analysis*. 1st ed. Compact Textbooks on Mathematics. Birkhäuser Cham, 2020. ISBN: 978-3-030-52784-6. DOI: 10.1007/978-3-030-52784-6.

[Coh13] Donald L. Cohn. *Measure Theory*. Birkhäuser Advanced Texts. Birkhäuser Basel, 2013. ISBN: 978-1-4614-6956-8. DOI: 10.1007/978-1-4614-6956-8.

[Coo71] Stephen A. Cook. “The Complexity of Theorem-Proving Procedures.” In: *Proceedings of the Third Annual ACM Symposium on Theory of Computing*. STOC ’71. Shaker Heights, Ohio, USA: Association for Computing Machinery, 1971, pp. 151–158. ISBN: 9781450374644. DOI: 10.1145/800157.805047. URL: https://doi.org/10.1145/800157.805047.

[Dij+13] N. P. van Dijk et al. “Level-set methods for structural topology optimization: a review.” In: *Structural and Multidisciplinary Optimization* 48.3 (Sept. 2013), pp. 437–472. ISSN: 1615-1488. DOI: 10.1007/s00158-013-0912-y.

[DP80] J. R. Dormand and P. J. Prince. “A family of embedded Runge-Kutta formulae.” In: *Journal of Computational and Applied Mathematics* 6.1 (1980), pp. 19–26. ISSN: 0377-0427. DOI: 10.1016/0771-050X(80)90013-3. URL: https://www.sciencedirect.com/science/article/pii/0771050X80900133.

[EKS94] H. A. Eschenauer, V. V. Kobelev, and A. Schumacher. “Bubble method for topology and shape optimization of structures.” In: *Structural optimization* 8.1 (Aug. 1994), pp. 42–51. ISSN: 1615-1488. DOI: 10.1007/BF01742933.

[Fal85] K. J. Falconer. *The Geometry of Fractal Sets*. Cambridge Tracts in Mathematics. Cambridge University Press, 1985.

[Far+13] P. E. Farrell et al. “Automated Derivation of the Adjoint of High-Level Transient Finite Element Programs.” In: *SIAM Journal on Scientific Computing* 35.4 (2013), pp. C369–C393. DOI: 10.1137/120873558.

[Fre09] David H. Fremlin. *Measure Theory. Broad Foundations*. Vol. 2. Lulu Press Inc., Dec. 7, 2009, pp. 38–42. ISBN: 978-0-9538129-7-4.

[FS20] Stefan A. Funken and Anja Schmidt. “Adaptive Mesh Refinement in 2D – An Efficient Implementation in Matlab.” In: *Computational Methods in Applied Mathematics* 20.3 (2020), pp. 459–479. DOI: 10.1515/cmam-2018-0220.

[FS21] Stefan A. Funken and Anja Schmidt. “A coarsening algorithm on adaptive red-green-blue refined meshes.” In: *Numerical Algorithms* 87.3 (July 2021), pp. 1147–1176. ISSN: 1572-9265. DOI: 10.1007/s11075-020-01003-7.

[Gar+22] David J Gardner et al. “Enabling new flexibility in the SUNDIALS suite of nonlinear and differential/algebraic equation solvers.” In: *ACM Transactions on Mathematical Software (TOMS)* (2022). DOI: 10.1145/3539801.

[Geb+23] Anna Gebhard et al. “Pharmacokinetic–pharmacodynamic modeling of maintenance therapy for childhood acute lymphoblastic leukemia.” In: *Scientific Reports* 13.1 (July 2023), p. 11749. ISSN: 2045-2322. DOI: 10.1038/s41598-023-38414-0.

[GBS06] A. Gersborg-Hansen, M. P. Bendsøe, and O. Sigmund. “Topology optimization of heat conduction problems using the finite volume method.” In: *Structural and Multidisciplinary Optimization* 31 (2006), pp. 251–259. ISSN: 1615-1488. DOI: 10.1007/s00158-005-0584-3.

[Gor+22] Eduard Gorbunov et al. “Recent Theoretical Advances in Decentralized Distributed Convex Optimization.” In: *High-Dimensional Optimization and Probability: With a View Towards Data Science*. Ed. by Ashkan Nikeghbali et al. Cham: Springer International Publishing, 2022, pp. 253–325. ISBN: 978-3-031-00832-0. DOI: 10.1007/978-3-031-00832-0_8. URL: https://doi.org/10.1007/978-3-031-00832-0_8.

[Gor16] Alexey L. Gorodentsev. *Algebra I*. 1st ed. Springer International Publishing, 2016. ISBN: 978-3-319-45285-2. DOI: 10.1007/978-3-319-45285-2.

[Gri85] Pierre Grisvard. *Elliptic Problems in Nonsmooth Domains*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 1985. DOI: 10.1137/1.9781611972030.

[GM20] Tom Gustafsson and G. D. McBain. “scikit-fem: A Python package for finite element assembly.” In: *Journal of Open Source Software* 5.52 (2020), p. 2369. DOI: 10.21105/joss.02369.

[Hah25a] Mirko Hahn. *mirhahn/pycoimset: 0.1.7*. Version v0.1.7. Jan. 2025. DOI: 10.5281/zenodo.14726149.

[Hah25b] Mirko Hahn. *mirhahn/pycoimset: 0.1.7 Example Run Data*. Zenodo, Jan. 2025. DOI: 10.5281/zenodo.14726459.

[HLS22] Mirko Hahn, Sven Leyffer, and Sebastian Sager. “Binary optimal control by trust-region steepest descent.” In: *Mathematical Programming* (2022). DOI: 10.1007/s10107-021-01733-z.

[HWN93] Ernst Hairer, Gerhard Wanner, and Syvert P. Nørsett. *Solving Ordinary Differential Equations I: Nonstiff Problems*. 2nd ed. Springer-Verlag Berlin Heidelberg, 1993. ISBN: 978-3-540-78862-1. DOI: 10.1007/978-3-540-78862-1.

[Hal69] Jack K. Hale. *Ordinary Differential Equations*. Wiley, 1969. ISBN: 978-0-486-47211-9.

[HS13] Falk M. Hante and Sebastian Sager. "Relaxation methods for mixed-integer optimal control of partial differential equations." In: *Computational Optimization and Applications* 55.1 (May 2013), pp. 197–225. ISSN: 1573-2894. DOI: 10.1007/s10589-012-9518-3.

[Har+20] Charles R. Harris et al. "Array programming with NumPy." In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2.

[Hin+05] Alan C Hindmarsh et al. "SUNDIALS: Suite of nonlinear and differential/algebraic equation solvers." In: *ACM Transactions on Mathematical Software (TOMS)* 31.3 (2005), pp. 363–396. DOI: 10.1145/1089014.1089020.

[Hin+09] Michael Hinze et al. *Optimization with PDE Constraints*. 1st ed. Springer, Dordrecht, 2009. ISBN: 978-1-4020-8839-1. DOI: 10.1007/978-1-4020-8839-1.

[Hof71] J. Hoffmann-Jørgensen. "Vector Measures." In: *Mathematica Scandinavica* 28 (Dec. 1971), pp. 5–32. DOI: 10.7146/math.scand.a-11003.

[Jos+20] Felix Jost et al. "Model-Based Simulation of Maintenance Therapy of Childhood Acute Lymphoblastic Leukemia." In: *Frontiers in Physiology* 11 (2020). ISSN: 1664-042X. DOI: 10.3389/fphys.2020.00217. URL: https://www.frontiersin.org/journals/physiology/articles/10.3389/fphys.2020.00217.

[Jun14] Michael Jung. "Relaxations and Approximations for Mixed-Integer Optimal Control." PhD thesis. Ruprecht-Karls-Universität Heidelberg, 2014. DOI: 10.11588/heidok.00016036.

[Kar72] Richard M. Karp. "Reducibility among Combinatorial Problems." In: *Complexity of Computer Computations: Proceedings of a symposium on the Complexity of Computer Computations, held March 20–22, 1972, at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York, and sponsored by the Office of Naval Research, Mathematics Program, IBM World Trade Corporation, and the IBM Research Mathematical Sciences Department*. Ed. by Raymond E. Miller, James W. Thatcher, and Jean D. Bohlinger. Boston, MA: Springer US, 1972, pp. 85–103. ISBN: 978-1-4684-2001-2. DOI: 10.1007/978-1-4684-2001-2_9.

[KL12] Robert C. Kirby and Anders Logg. "The finite element method." In: *Automated Solution of Differential Equations by the Finite Element Method: The FEniCS Book*. Ed. by Anders Logg, Kent-Andre Mardal, and Garth Wells. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 77–94. ISBN: 978-3-642-23099-8. DOI: 10.1007/978-3-

642-23099-8_2. URL: https://doi.org/10.1007/978-3-642-23099-8_2.

[Kir+12] Robert C. Kirby et al. “Common and unusual finite elements.” In: *Automated Solution of Differential Equations by the Finite Element Method: The FEniCS Book*. Ed. by Anders Logg, Kent-Andre Mardal, and Garth Wells. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 95–119. ISBN: 978-3-642-23099-8. DOI: 10.1007/978-3-642-23099-8_3.

[KMU21] Christian Kirches, Paul Manns, and Stefan Ulbrich. “Compactness and convergence rates in the combinatorial integral approximation decomposition.” In: *Mathematical Programming* 188.2 (Aug. 2021), pp. 569–598. ISSN: 1436-4646. DOI: 10.1007/s10107-020-01598-8.

[Kir+10] Christian Kirches et al. “Time-optimal control of automobile test drives with gear shifts.” In: *Optimal Control Applications and Methods* 31.2 (2010), pp. 137–153. DOI: 10.1002/oca.892. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/oca.892.

[KA21] Peter Knabner and Lutz Angermann. *Numerical Methods for Elliptic and Parabolic Partial Differential Equations*. 2nd ed. Vol. 44. Texts in Applied Mathematics. Springer Nature Switzerland AG, 2021. DOI: 10.1007/978-3-030-79385-2.

[Knu98] Donald Knuth. *The Art Of Computer Programming, Vol. 3: Sorting And Searching*. 2nd ed. Addison-Wesley, 1998. ISBN: 978-0-201-89685-5.

[Kub15] Carlos S. Kubrusly. *Essentials of Measure Theory*. Springer International Publishing Switzerland, 2015. ISBN: 978-3-319-22506-7. DOI: 10.1007/978-3-319-22506-7.

[LT03] Emmanuel Laporte and Patrick Tallec. *Numerical Methods in Sensitivity Analysis and Shape Optimization*. 1st ed. Boston, MA: Birkhäuser, 2003. ISBN: 978-1-4612-6598-6. DOI: 10.1007/978-1-4612-0069-7.

[LB13] Mats G. Larson and Fredrik Bengzon. *The Finite Element Method: Theory, Implementation, and Applications*. 1st ed. Vol. 10. Texts in Computational Science and Engineering. Springer-Verlag Berlin Heidelberg, 2013. ISBN: 978-3-642-33287-6. DOI: 10.1007/978-3-642-33287-6.

[Le+22] Do Duc Le et al. “Autonomous traffic at intersections: an optimization-based analysis of possible time, energy, and CO2 savings.” In: *Networks* 79.3 (2022), pp. 338–363. DOI: 10.1002/net.22078.

[Leb10] Henri Lebesgue. “Sur l’intégration des fonctions discontinues.” In: *Annales scientifiques de l’École Normale Supérieure*. 3rd ser. 27 (1910), pp. 361–450. DOI: 10.24033/asens.624.

[LM22] Sven Leyffer and Paul Manns. “Sequential linear integer programming for integer optimal control with total variation regularization.” In: *ESAIM: COCV* 28 (2022), p. 66. DOI: 10.1051/cocv/2022059.

[LMW21] Sven Leyffer, Paul Manns, and Malte Winckler. “Convergence of sum-up rounding schemes for cloaking problems governed by the Helmholtz equation.” In: *Computational Optimization and Applications* 79.1 (May 2021), pp. 193–221. ISSN: 1573–2894. DOI: 10.1007/s10589-020-00262-3.

[LOS24] Yang Liu, Yuuki Oda, and Kazuki Sasahara. “Shape and topology optimization method with generalized topological derivatives.” In: *International Journal of Mechanical Sciences* 284 (2024), p. 109735. ISSN: 0020-7403. DOI: 10.1016/j.ijmecsci.2024.109735.

[LMW12] Anders Logg, Kent-Andre Mardal, and Garth N. Wells, eds. *Automated Solution of Differential Equations by the Finite Element Method. The FEniCS Book*. Ed. by Timothy J. Barth et al. 1st ed. Vol. 84. Lecture Notes in Computational Science and Engineering. Springer-Verlag Berlin Heidelberg, 2012. ISBN: 978-3-642-23099-8. DOI: 10.1007/978-3-642-23099-8.

[Lot25] Alfred J. Lotka. *Elements of Physical Biology*. Williams and Wilkins Company, 1925.

[MZ21] Zhongjing Ma and Suli Zou. “Pontryagin’s Minimum Principle.” In: *Optimal Control Theory: The Variational Method*. Singapore: Springer Singapore, 2021, pp. 147–218. ISBN: 978-981-33-6292-5. DOI: 10.1007/978-981-33-6292-5_4.

[MK20] Paul Manns and Christian Kirches. “Multidimensional Sum-Up Rounding for Elliptic Control Systems.” In: *SIAM Journal on Numerical Analysis* 58.6 (2020), pp. 3427–3447. DOI: 10.1137/19M1260682.

[Man+23] Paul Manns et al. “On Convergence of Binary Trust-Region Steepest Descent.” In: *Journal of Nonsmooth Analysis and Optimization* Volume 4 (July 2023). DOI: 10.46298/jnsao-2023-10164. URL: https://jnsao.episciences.org/11643.

[MPS00] Antonino Maugeri, Dian K. Palagachev, and Lubomira G. Softova. *Elliptic and Parabolic Equations with Discontinuous Coefficients*. 1st ed. Vol. 109. Mathematical Research. WILEY-VCH Verlag Berlin GmbH, 2000. ISBN: 978-3-52760086-1. DOI: 10.1002/3527600868.

[MP04] Bijan Mohammadi and Olivier Pironneau. “Shape Optimization in Fluid Mechanics.” In: *Annual Review of Fluid Mechanics* 36 (2004), pp. 255–279. ISSN: 1545-4479. DOI: 10.1146/annurev.fluid.36.050802.121926. URL: https://www.annualreviews.org/content/journals/10.1146/annurev.fluid.36.050802.121926.

[NW06] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. 2nd ed. Springer Series in Operations Research and Financial Engineering. 2006. ISBN: 978-0-387-40065-5. DOI: 10.1007/978-0-387-40065-5.

[NS13] Antonio André Novotny and Jan Sokołowski. *Topological Derivatives in Shape Optimization*. 1st ed. Interaction of Mechanics and Mathematics. Springer Berlin, Heidelberg, 2013. ISBN: 978-3-642-35245-4. DOI: 10.1007/978-3-642-35245-4.

[NSŻ19] Antonio André Novotny, Jan Sokołowski, and Antoni Żochowski. "Topological Derivatives of Shape Functionals. Part II: First-Order Method and Applications." In: *Journal of Optimization Theory and Applications* 180.3 (Mar. 2019), pp. 683–710. ISSN: 1573-2878. DOI: 10.1007/s10957-018-1419-x.

[Pat18] Hemant Kumar Pathak. *An Introduction to Nonlinear Analysis and Fixed Point Theory*. 1st ed. Springer, Singapore, 2018. ISBN: 978-981-10-8866-7. DOI: 10.1007/978-981-10-8866-7.

[RS19] Bogdan Raita and Daniel Spector. "A note on estimates for elliptic systems with L^1 data." In: *Comptes Rendus. Mathématique* 357.11-12 (2019), pp. 851–857. DOI: 10.1016/j.crma.2019.11.007.

[RR04] Michael Renardy and Robert C. Rogers. *An Introduction to Partial Differential Equations*. Ed. by J. E. Marsden, L. Sirovich, and S. S. Antman. 2nd ed. Vol. 13. Texts in Applied Mathematics. Springer Verlag New York, Inc., 2004. ISBN: 978-0-387-21687-4. DOI: 10.1007/b97427.

[Rob+21] Nicolò Robuschi et al. "Multiphase mixed-integer nonlinear optimal control of hybrid electric vehicles." In: *Automatica* 123 (2021), p. 109325. ISSN: 0005-1098. DOI: 10.1016/j.automatica.2020.109325.

[Sag06] Sebastian Sager. "Numerical methods for mixed–integer optimal control problems." PhD thesis. Universität Heidelberg, 2006. URL: https://katalog.ub.uni-heidelberg.de/titel/66091702.

[Sag08] Sebastian Sager. *Lotka Volterra fishing problem – mintOC*. Nov. 5, 2008. URL: https://mintoc.de/index.php?title=Lotka_Volterra_fishing_problem (visited on 04/05/2024).

[SJK11] Sebastian Sager, Michael Jung, and Christian Kirches. "Combinatorial Integral Approximation." In: *Mathematical Methods of Operations Research* 73.3 (2011), pp. 363–380. DOI: 10.1007/s00186-011-0355-4.

[Sag+06] Sebastian Sager et al. "Numerical Methods for Optimal Control with Binary Control Functions Applied to a Lotka-Volterra Type Fishing Problem." In: *Recent Advances in Optimization*. Ed. by Alberto Seeger. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 269–289. ISBN: 978-3-540-28258-7. DOI: 10.1007/3-540-28258-0_17.

[Sha86] Lawrence F. Shampine. "Some Practical Runge-Kutta Formulas." In: *Math. Comput.* 46.173 (Jan. 1986), pp. 135–150. ISSN: 0025–5718. DOI: 10.2307/2008219.

[Sha+20] Meenarli Sharma et al. "Inversion of convection–diffusion equation with discrete sources." In: *Optimization and Engineering* 22.3 (July 2020), pp. 1419–1457. DOI: 10.1007/s11081-020-09536-5.

[Shi18] Satish Shirali. *A Concise Introduction to Measure Theory*. Springer Nature Switzerland, 2018. ISBN: 978-3-030-03241-8. DOI: 10.1007/978-3-030-03241-8.

[Sor82] D. C. Sorensen. "Newton's Method with a Model Trust Region Modification." In: *SIAM Journal on Numerical Analysis* 19.2 (1982), pp. 409–426. DOI: 10.1137/0719026.

[Sor16] Stephan Sorgatz. “Optimization of Vehicular Traffic at Traffic-Light Controlled Intersections.” PhD thesis. Otto-von-Guericke University Magdeburg, 2016. URL: http://nbn-resolving.de/urn:nbn:de:gbv:ma9:1-8902.

[SS05] Elias M. Stein and Rami Shakarchi. *Real Analysis: Measure Theory, Integration, and Hilbert Spaces*. Vol. 3. Princeton Lectures in Analysis. Princeton and Oxford: Princeton University Press, 2005. ISBN: 9780691113869. DOI: 10.1515/9781400835560.

[UU12] Michael Ulbrich and Stefan Ulbrich. *Nichtlineare Optimierung*. 1st ed. Mathematik Kompakt. Birkhäuser Basel, 2012. ISBN: 978-3-0346-0654-7. DOI: 10.1007/978-3-0346-0654-7.

[Vir+20] Pauli Virtanen et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python.” In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[Vol26a] Vito Volterra. “Fluctuations in the Abundance of a Species considered Mathematically.” In: *nature* 118 (1926), pp. 558–560. DOI: 10.1038/118558a0.

[Vol26b] Vito Volterra. “Variazioni e fluttuazioni del numero di individui in specie animali conviventi.” In: *Memorie della R. Accademia dei Lincei*. 6th ser. 2 (1926), 85 pp.

[Wik25] Wikipedia contributors. *Vitali covering lemma — Wikipedia, The Free Encyclopedia*. 2025. URL: https://en.wikipedia.org/wiki/Vitali_covering_lemma (visited on 10/24/2025).

[Yan+19] Tao Yang et al. “A survey of distributed optimization.” In: *Annual Reviews in Control* 47 (2019), pp. 278–305. ISSN: 1367-5788. DOI: 10.1016/j.arcontrol.2019.05.006. URL: https://www.sciencedirect.com/science/article/pii/S1367578819300082.

[Zei95] Eberhard Zeidler. *Applied Functional Analysis. Main Principles and Their Applications*. 1st ed. Vol. 2. Springer New York, 1995. ISBN: 978-1-4612-0821-1. DOI: 10.1007/978-1-4612-0821-1.

[ZA15] Feng Zhu and Panos J. Antsaklis. “Optimal control of hybrid switched systems: A brief survey.” In: *Discrete Event Dynamic Systems* 25.3 (Sept. 2015), pp. 345–364. ISSN: 1573-7594. DOI: 10.1007/s10626-014-0187-5.