Title
Geometry, Analysis, and Optimization in Probability Theory

Author
Jaffe, Adam Quinn

Publication Date
2024

Peer reviewed|Thesis/dissertation

Geometry, Analysis, and Optimization in Probability Theory

by

Adam Quinn Jaffe

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Distinguished Professor Steven N. Evans, Chair
Associate Professor Shirshendu Ganguly
Associate Professor Adityanand Guntuboyina
Assistant Professor Nikita Zhivotovskiy

Spring 2024

Geometry, Analysis, and Optimization in Probability Theory

Abstract

Geometry, Analysis, and Optimization in Probability Theory

by

Adam Quinn Jaffe

Doctor of Philosophy in Statistics

University of California, Berkeley

Distinguished Professor Steven N. Evans, Chair

The focus of this thesis is the use of techniques from geometry, analysis, and optimization to several concrete problems in probability theory, and also to some problems in statistics and machine learning. Growing interest in the intersections of these fields is motivated by several recent developments: the recognition of the utility of optimal transport in probability and statistics, the large number of modern statistical applications where one encounters non-Euclidean data, and more.

The first part of the thesis studies couplings. Its main results include a form of infinite-dimensional linear programming duality for a rich class of coupling problems involving equivalence relations, and consequences of this abstract theory for various coupling problems encountered in stochastic calculus.

The second part studies the canonical notion of central tendency for probability measures on metric spaces, the Fréchet mean (also called the barycenter or the center of mass). The several chapters in this part establish: a limit theory for Fréchet means in a general class of infinite-dimensional metric spaces; a development of large deviations theory for Fréchet means in the Bures-Wasserstein space; a statistical optimality result for estimating Fréchet mean sets in a general metric space; and, an optimal adaptive algorithm for Fréchet mean set estimation in the space of phylogenetic trees.

The third part studies clustering, and provides asymptotic guarantees for k-means clustering and variants thereof. In particular, it establishes consistency results for adaptive variants of k-means (k-means when k is chosen according to the elbow method, k-medoids, and more) and it proves further limit theorems for the classic k-means problem.

Contents

List of Figures

Acknowledgments

I will always look back on my time at Berkeley as one of the great periods of my life. But graduate school is also a difficult time, and I could not have succeeded without the help of many people along the way. To those who appear in this acknowledgement (and, also, to so many who do not): I hope my sincere gratitude is evident.

Above all, I must thank my family. From both of my parents, Marc and Lisa, I inherited a curious nature and a strong social awareness, which have served me well in academia. My sister, Alex, has been a role model throughout my life and in my career. And, my partner, Ellen, has been an essential source of diversions, commiseration, and programming lessons. From all of you I gained the ability and the opportunity to follow my passions.

Next I thank my academic advisors, both formal and informal. First is my thesis advisor, Steve, from whom I learned my most important lessons, be they mathematical, academic, or otherwise. From our first interactions during his special topics course "Probability and Convexity" to the last days of writing this thesis, I have come to admire Steve's deep analytic perspective on probability and statistics, his encyclopedic memory for all things he has read, and his great humility; these are all characteristics that I strive to emulate in my life and work. Second is Shirshendu who (despite having, frankly, very little research interest in common with me) has provided me invaluable career advice, research direction, and more. I also thank Aditya and Nikita for discussing various research projects and for serving on my thesis committee. Additionally, I want to acknowledge the generosity of Fraydoun Rezakhanlou, Jon Niles-Weed, Victor Panaretos, Stefan Sommer, and Bodhi Sen, for their insights, recommendations, and encouragement over several years.

The staff in the Department of Statistics also deserves huge thanks. I am sure that they are aware of the role they play in making the department run smoothly and efficiently, but perhaps they are not reminded frequently enough of their role in creating the welcoming environment that makes the Berkeley Statistics so special. I particularly want to thank La Shana, David, and Tanisha for their friendship, their advice, and for saving me, in one way or another, on too many occasions to count.

Lastly, I want to thank all of the peers who have shaped my academic experience. This of course includes the members of my PhD cohort (Alice, Corrine, Facu, Katherine, Nikhil, Yassine), but also other peers in the department, in other departments, at other universities, and beyond. I also want to thank my housemates from Cedar street (Greg, Sally, Nate, Antoine, Apollo, Ilina, Franck, Donna, Facu, Guillaume, Sophie, Tsiresy, Georgia, Kris, Aisha) with whom I formed a strong community for the last five years.

Chapter 1

Introduction

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space with expectation denoted $\mathbb{E}$, and suppose that Y is a real-valued random variable with $\mathbb{E}[|Y|] < \infty$ and that $\mathcal{G}$ is a sub-σ-algebra of $\mathcal{F}$. A surprisingly difficult question, inevitably encountered by every student of probability theory, is: How should one make sense of the conditional expectation $\mathbb{E}[Y \mid \mathcal{G}]$?

This setting gives rise to my favorite theorem: If we further have $\mathbb{E}[|Y|^2] < \infty$, then it turns out that the conditional expectation $\mathbb{E}[Y \mid \mathcal{G}]$ is the unique solution to the optimization problem

$$\begin{cases} \text{minimize} & \mathbb{E}\Big[|X - Y|^2\Big] \\ \text{over} & \mathcal{G}\text{-measurable } X \\ \text{with} & \mathbb{E}\Big[|X|^2\Big] < \infty. \end{cases} \tag{1.1}$$

I have come to realize that this theorem exemplifies several mathematical themes that appear throughout my research: In service of studying a concrete question in probability and statistics, this result (i) highlights a geometric principle which yields valuable insights and intuitions, (ii) leverages technical analytic machinery (from functional analysis, point-set topology, and measure theory) to make it rigorous, and (iii) relates it to an optimization problem that can actually be computed or approximated in practice. (In another direction, it also emphasizes that extending probabilistic results from L^2 to L^1 can be very difficult.)

This thesis surveys the research I have completed throughout my time at Berkeley, with a focus on the ways that the various projects are connected by the themes (i), (ii), and (iii). The material is derived from the papers [66, 37, 100, 99, 168, 97, 56, 101] which were completed with the help of very many collaborators along the way. It is divided into three parts, which I will now briefly introduce.

Part I: Coupling

Coupling is one of the fundamental techniques of modern probability theory, and some[1] go as far as to say that coupling is *the* technique which distinguishes probability from analysis. Here, for two random variables Y, Y', possibly defined on different probability spaces, a *coupling* of Y, Y' is a pair of random variables X, X' defined on the same probability space in such a way that X and Y have the same distribution, X' and Y' have the same distribution, and the joint distribution of X and X' has some desirable properties.

Equivalently stated in terms of the distributions, a coupling of probability measures μ_1, μ_2 on a measurable space $(\Omega, \mathcal{F})$ is a probability measure $\tilde{\mu}$ on the product space $(\Omega\times\Omega, \mathcal{F}\otimes\mathcal{F})$ satisfying $\tilde{\mu}(\,\cdot\,\times\Omega) = \mu_1$ and $\tilde{\mu}(\Omega\times\,\cdot\,) = \mu_2$ and where $\tilde{\mu}$ has some desirable properties. From this perspective, it is easy to see that there usually exist many couplings of two fixed probability measures; see Figure 1.1 for an illustration. In fact, a common paradigm is that many concrete problems in probability can be formulated as a certain optimization problem over the space of couplings.

This first part of the thesis is focused on a particular class of optimization problems over couplings which has, in special cases, been studied many times [192, 82, 85, 163, 181, 193, 78, 119]. That is, for random variables Y, Y' and an equivalence relation E, we aim to solve:

$$\begin{cases} \text{minimize} & 1-\mathbb{P}(X \text{ and } X' \text{ are } E\text{-equivalent}) \\ \text{over} & \text{couplings } X, X' \text{ of } Y, Y'. \end{cases} \tag{1.2}$$

The importance of (1.2) is that its optimal value is zero if and only if Y, Y' can be coupled to be E-equivalent almost surely and that its optimal value is one if and only Y, Y' cannot be coupled to be E-equivalent with any positive probability at all.

[1] I recall learning this opinion from Jim Pitman at La Val's Pizza on Euclid Avenue

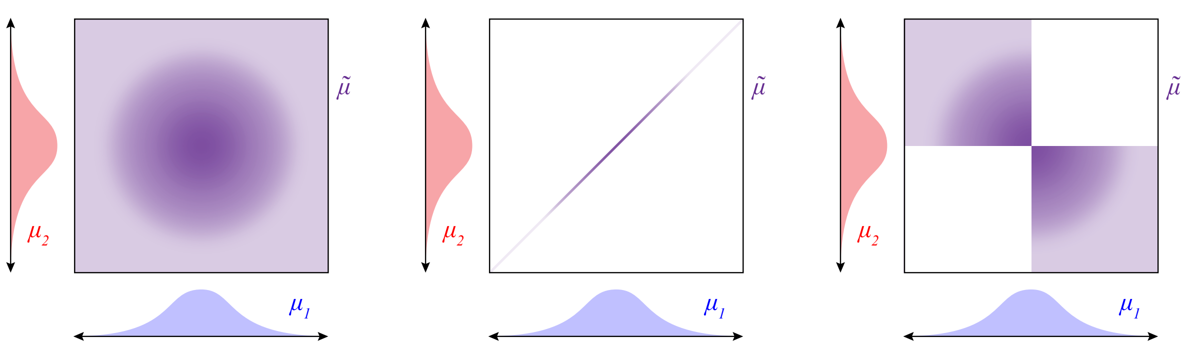

Figure 1.1: The space of all couplings of two probability measures on $\mathbb{R}$. Some elements include the independent coupling (left), the coupling concentrated on the diagonal (middle), and the independent coupling conditional on the coordinates having different signs (right).

The most interesting examples of (1.2) occur in the setting where Y and Y' take values in an infinite-dimensional space and where the equivalence relation E has very poor topological regularity. In fact, the prototypical example is when Y and Y' are Markov processes and E is the equivalence relation of eventual equality on path space; this version of the problem has many rich connections to Markov chain mixing times, potential theory, and Riemannian geometry (all due, roughly speaking, to its connection to the tail σ-algebra[2]), so it has been an object of much study [117, 125, 116, 42, 16, 115, 118, 25, 46, 41, 178, 134, 92].

This part is based on several works of mine which study coupling problems like those outlined above. Chapter 2 is based on my work [99] which establishes a general form of infinite-dimensional linear programming duality for (1.2). Chapter 3 is based on the works [99, 97, 56] and focuses on applications of these general results to the so-called *germ coupling problem* in stochastic calculus.

Part II: Centering

While classical statistical problems concern data or parameters that live in a Euclidean space (like $\mathbb{R}^m$ for some $m \in \mathbb{N}$, or an infinite-dimensional Hilbert space $\mathcal{H}$), many modern statistical problems concern data or parameters living in a space with more interesting geometry. Examples include

- network analysis [81, 121, 70, 149, 71], where data living on a graph can have an essentially arbitrary geometry, and where data consisting of graphs can inherit many different geometries depending on the choice of graph metric,
- mathematical imaging [61, 53, 54] and shape analysis [28, 127, 69], where the existing geometry of various spaces of matrices becomes further complicated by the presence of some interesting group actions, and
- computational phylogenetics [182, 34, 135, 17, 209, 106, 136], where the space of all phylogenetic trees can be represented as a union of a collection of orthants, glued together along sub-orthants of smaller dimension,

Even the most basic statistical questions become complicated in these settings. Indeed, one typically has to modify existing statistical methods to accomodate the non-Euclidean structure of the problem; in some cases, one has come up with entirely new methodologies

The difficulties posed by non-Euclidean geometry are already evident in the (surprisingly interesting) problem of estimating central tendency. In the Euclidean case of a random variable Y in a Hilbert space $(\mathcal{H}, \|\cdot\|)$, this is just estimating the expectation $\mathbb{E}[Y]$ on the basis of independent, identically distributed (IID) samples. In the non-Euclidean case of

[2]When I first became interested in (1.2) and related coupling problems, Steve warned me that "the tail σ-algebra is a notoriously slippery thing".

a random variable Y in a metric space $(\mathcal{X}, d)$, this is usually formulated as estimating a solution to the optimization problem

$$\begin{cases} \text{minimize} & \mathbb{E}\Big[d^2(x, Y)\Big] \\ \text{over} & x \in \mathcal{X} \end{cases} \tag{1.3}$$

on the basis of IID samples. Solutions to (1.3) are called *Fréchet means* (also called *barycenters* or *centers of mass*) and, if $Y_1, Y_2, \ldots$ are IID samples from the same distribution as Y, then one usually estimates this population object from the *empirical Fréchet mean*

$$, \begin{cases} \text{minimize} & \frac{1}{n}\sum_{i=1}^n d^2(x, Y_i) \\ \text{over} & x \in \mathcal{X}. \end{cases} \tag{1.4}$$

These are indeed canonical generalizations of Euclidean estimation, since, in the Euclidean setting, the unique solution to (1.3) is $x = \mathbb{E}[Y]$ and the unique solution to (1.4) is $x = \frac{1}{n}\sum_{i=1}^n Y_i$. However, we emphasize that, in the non-Euclidean setting when $(\mathcal{X}, d)$ is a general metric space, the problems (1.3) can have no solution, a unique solution, or multiple solutions.

This second part of the thesis is focused on probabilistic and statistical aspects of Fréchet means, which has been studied in very many works [213, 186, 31, 32, 95, 94]. The overal goal, simply stated, is to derive limit theorems establishing the way that empirical Fréchet means converge to population Fréchet means as $n \to \infty$; See Figure 1.2 for an illustration. There exists a particularly well-developed theory in the case that $\mathcal{X}$ is a Riemannian manifold and d is the metric induced by its metric tensor [31, 32, 28], but, for many of the application areas outlined above, we want to be able to say something about the case that $(\mathcal{X}, d)$ is either highly singular or infinite-dimensional.

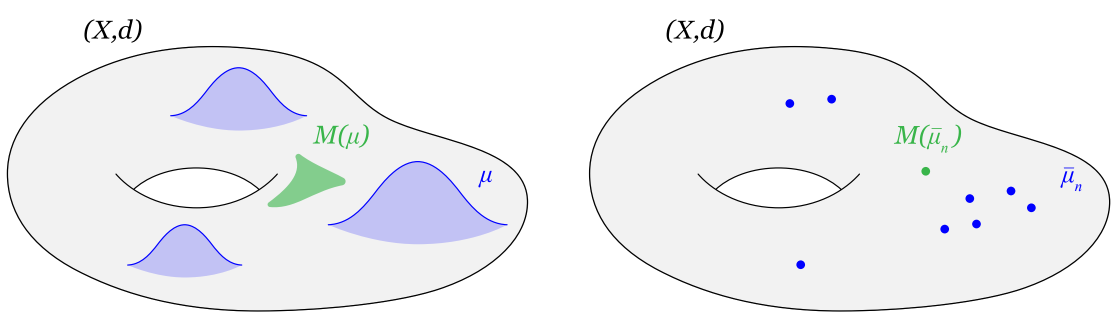

Figure 1.2: Fréchet means in a general metric space. There are many possible population Fréchet means (left), while, in this realization, there is a unique empirical Fréchet mean (right).

This part is based on several works of mine focused on Fréchet means and related objects. Chapter 4 is based on my works [66, 98] establishing the fundamental limit theorems under minimal conditions, and is closely related to concurrent work[3] in [180]. Chapter 5 is based on the project [101] studying large deviations theory for Wasserstein and Bures-Wasserstein barycenters. Chapters 6 and 7 are based on the work [37] which studies a particular set-valued Fréchet mean estimation problem and an application in computational phylogenetics.

Part III: Clustering

A fundamental task in unsupervised learning is that of *clustering*, namely, partitioning a set of data into a finite number of groups where elements within a group are similar (and, typically, elements between distinct groups are dissimilar). Among the most common clustering methods is k-means clustering: For data points $Y_1, \ldots Y_n$ in a (possibly infinite-dimensional) Hilbert space $(\mathcal{H}, \|\cdot\|)$ and any $k \in \mathbb{N} := \{1, 2, \ldots\}$, the set of *$k$-means cluster centers* is any solution to the optimization problem

$$\begin{cases} \text{minimize} & \frac{1}{n}\sum_{i=1}^{n} \min_{x\in S} \|x - Y_i\|^2 \\ \text{over} & S \subseteq \mathcal{H} \\ \text{with} & \#S = k. \end{cases} \tag{1.5}$$

Intuitively speaking, a set of k-means cluster centers for these data points is a set of points S_n in $\mathcal{H}$ to at least one of which all data are optimally close; the *k-means clusters* are then the sets

$$\{Y_i : \|x - Y_i\| \leq \|x' - Y_i\| \text{ for all } x' \in S_n\}$$

indexed by $x \in S_n$. See Figure 1.3 for an illustration of the k-means clusters for a toy data set, when $k \in \{2, 3, 4, 5\}$.

Over several decades, k-means has played an important role in both applied and theoretical statistics. On the applied side, it is among the simplest and most powerful clustering methods, and a chapter dedicated to its study appears in nearly every introductory machine learning textbook [79, 10, 87]. On the theoretical side, there is a rich body of literature studying its asymptotic theory [165, 1, 126, 133, 161, 160, 159, 194], computational feasibility [145, 142, 76, 9], and, more recently, concentration properties [33, 120, 164]. (It has also been given many different names throughout its history; see [39].)

This final part of the thesis (which consists only of Chapter 8) follows my work in [100] which studies some novel asymptotic aspects of k-means clustering. Particular attention is paid to adaptive variants of k-means that are often used in practice, yet for which no theoretical guarantees were previously known.

[3] The works [66] and [180] were coincidentally posted to arxiv on the same day (December 23, 2020), leading me to jokingly refer to Christof as my "nemesis" for several years. When I finally met him in January 2023 and he turned out to be very nice, I decided to drop the nickname.

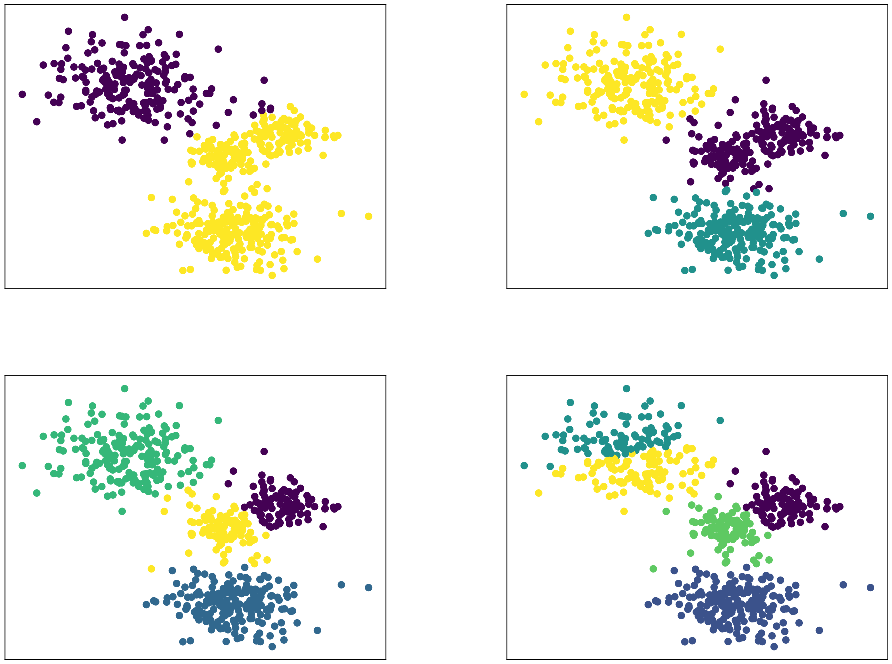

Figure 1.3: Applying k-means clustering to some data, for $k \in \{2, 3, 4, 5\}$. When $k = 2$, some true clusters are erroneously grouped together (top left). When $k = 3$ (top right) or $k = 4$ (bottom left), some meaningful clusters are recovered. When $k = 5$, some single clusters are erroneously split up (bottom right).

Notation

While all of the required terminology and notation will be outlined at the beginning of each chapter, a small amount of notation is shared by all parts of the thesis: We write $\mathbb{N} = \{1, 2, 3, \ldots\}$ for the set of natural numbers, which we take to exclude 0. We write $\mathcal{P}(S, \mathcal{S})$ for the space of all probability measures on a measurable space $(S, \mathcal{S})$. If (S, τ) is a topological space, we write $\mathcal{B}(S, \tau)$, $\mathcal{B}(S)$, or $\mathcal{B}(\tau)$ for the Borel σ-algebra, whichever is clearest from context. If (S, τ) is a Polish space, we write $\mathcal{P}(S)$ as shorthand for $\mathcal{P}(S, \mathcal{B}(S))$.

Part I

Coupling

Chapter 2

The Equivalence Coupling Problem

Let $(\Omega, \mathcal{F})$ be a measurable space and let E be an equivalence relation on Ω which satisfies $E \in \mathcal{F} \otimes \mathcal{F}$. For any probability measures $P, P' \in \mathcal{P}(\Omega, \mathcal{F})$, let $\Pi(P, P')$ denote the space of all couplings of P, P', and let us consider the optimization problem

$$\begin{cases} \text{minimize} & 1 - \tilde{P}(E) \\ \text{over} & \tilde{P} \in \Pi(P, P'), \end{cases} \tag{2.1}$$

which we refer to as the *E-equivalence coupling problem*. Special cases of the problem (2.1) have appeared in probabilty theory many times, but a general analysis is quite complicated since for the most interesting applications one needs to consider the case that Ω is infnite-dimensional and that E has very poor topological regularity.

The goal of this chapter is to show that, in sufficient generality, (2.1) is dual, in the sense of duality of infinite-dimensional linear programs, to the problem

$$\begin{cases} \text{maximize} & |P(A) - P'(A)| \\ \text{over} & A \in \mathcal{G}, \end{cases} \tag{2.2}$$

which we refer to as the $\mathcal{G}$-total variation problem. The upshot of this duality is that, from a probabilist's point of view, there are many classical tools (zero-one laws, expressions for Radon-Nikodým derivatives, continuity-singularity dichotomy theorems) that can be used to analyze (2.2); consequently they can be used to analyze (2.1).

There already exist a few particular instances duality for the equivalence coupling problem, which we now describe. First, if Ω is Polish space with $\mathcal{F}$ is its Borel σ-algebra, and if $\Delta = \{(x, x) \in \Omega \times \Omega : x \in \Omega\}$ denotes the diagonal in $\Omega \times \Omega$, we have

$$\max_{A \in \mathcal{F}} |P(A) - P'(A)| = \min_{\tilde{P} \in \Pi(P, P')} (1 - \tilde{P}(\Delta)) \tag{2.3}$$

for all Borel probability measures P, P' on Ω. This result has certainly been known for a long time (at least for countable sets Ω) so its exact source is difficult to track down [139, Chapter I.7].

A second known example concerns the space of binary sequences $\Omega := \{0,1\}^{\mathbb{N}}$ with $\mathcal{F}$ the Borel σ-algebra of its product topology. (Actually, $\{0,1\}$ can be replaced with any Polish space here.) If E_0 is the equivalence relation of eventual equality and if $\mathcal{T}$ is the tail σ-algebra on Ω, then it was shown in a series of works [163, 75, 192, 85, 82], that one has

$$\max_{A\in\mathcal{T}} |P(A) - P'(A)| = 0 \qquad \text{if and only if} \qquad \min_{\tilde{P}\in\Pi(P,P')} (1 - \tilde{P}(E_0)) = 0, \tag{2.4}$$

for all Borel probability measures P, P' on Ω.

A third example, from ergodic theory, shows that eventual equality and the tail σ-algebra in (2.4) can be replaced with the analogous objects for the notion of group-invariance. Indeed, suppose that Ω is a Polish space with $\mathcal{F}$ its Borel σ-algebra and that a locally compact Polish group G acts continuously on Ω. Then, writing E_G for its orbit equivalence relation and $\mathcal{I}_G$ for its invariant σ-algebra, we have [193]

$$\max_{A\in\mathcal{I}_G} |P(A) - P'(A)| = 0 \qquad \text{if and only if} \qquad \min_{\tilde{P}\in\Pi(P,P')} (1 - \tilde{P}(E_G)) = 0, \tag{2.5}$$

for P, P' any two Borel probability measures on Ω. This result is an extension of earlier work on the (two-sided) shift [8], and subsequent developments have provided many generalizations, primarily of an algebraic nature [78, 181, 119].

This chapter is based on the work [99] which provides some general results (Theorem 1 and Theorem 2) guaranteeing the existence of a duality between (2.1) and (2.2). These results can be seen as a form of Kantorovich duality for a suitable Monge-Kantorovich optimal transport problem, but the poor topological regularity of the cost function means that standard results do not apply; our main technical innovation is thus to replace topological approximation arguments with more delicate measure-theoretic approximation arguments. Together, these two main results are together powerful enough to recover all the duality statements given above, and they have some novel consequences of interest; in the next chapter, we will explore applications of these results to a few problems in stochastic calculus.

2.1 Preliminaries

In order to state our main results, we need to introduce our precise notion of duality for the equivalence coupling problem. Throughout, we assume that $(\Omega, \mathcal{F})$ is a standard Borel space and that E is a Borel equivalence relation on $(\Omega, \mathcal{F})$.

Definition 1. *A Borel equivalence relation E on $(\Omega, \mathcal{F})$ is called* strongly dualizable *if we have*

$$\max_{A\in E^*} |P(A) - P'(A)| = \min_{\tilde{P}\in\Pi(P,P')} (1 - \tilde{P}(E)) \tag{2.6}$$

for all $P, P' \in \mathcal{P}(\Omega, \mathcal{F})$, where the "max" and "min" assert that the supremum and infimum are both achieved and where

$$E^* := \{A \in \mathcal{F} : \text{for all } (x,x') \in E, \text{ we have } x \in A \text{ if and only if } x' \in A\}$$

is the E-invariant σ-algebra.

Let us give some remarks on this definition. First, the definition of the E-invariant σ-algebra generalizes the notion of the invariant σ-algebra for group actions, since, if E_G is the orbit equivalence relation for a group G acting measurably on Ω, then one can easily show the expected $E_G^* = \mathcal{I}_G$. Second, we note that E^* is a natural choice for a collection of events appearing on the left side of (2.6): If E^* is replaced with some other collection $\mathcal{G}$, then, by taking $P = \delta_x$ and $P' = \delta_{x'}$ for $x, x' \in \Omega$, we see that every $A \in \mathcal{G}$ has the property that $x \in A$ is equivalent to $x' \in A$ for all $(x, x') \in E$. This means $\mathcal{G} \subseteq E^*$, so E^* is, in some sense, the maximal $\mathcal{G}$ with which E can form a strongly dual pair.

The remainder of this section is spent proving some intermediate results that will be used in our main theorems. The first result identifies the left side of (2.6) with its convex relaxation, and shows that a maximizer always exists.

Lemma 1. *For any Borel equivalence relation E, we have*

$$\max_{A \in E^*} |P(A) - P'(A)| = \max_{\substack{f \in bE^* \\ 0 \leq f \leq 1}} \left| \int_\Omega f \, \mathrm{d}P - \int_\Omega f \, \mathrm{d}P' \right|$$

for all Borel probability measures P, P' on $(\Omega, \mathcal{F})$,

Proof. We consider the Hilbert space $L^2(\Omega, E^*, \frac{1}{2}(P + P'))$, and we define $K := \{f \in bE^* : 0 \leq f \leq 1\}$ which is clearly a convex set. In fact, by the Banach-Alaoglu theorem, K is compact in the weak topology. Now we claim that $\mathrm{ex}(K) = \{\mathbb{1}_A : A \in E^*\}$. Indeed, the "$\supseteq$" direction is obvious, and the "$\subseteq$" direction is shown as follows: If $f \notin \{\mathbb{1}_A : A \in E^*\}$ then there is some $x \in \Omega$ with $f(x) \in (0, 1)$. This means the event $\{0 < f < 1\}$ is non-empty, and, since it can also be written as

$$\{0 < f < 1\} = \bigcup_{0 < \varepsilon \leq \frac{1}{2}} \{\varepsilon \leq f \leq 1 - \varepsilon\},$$

there must exist a sufficiently small $\varepsilon > 0$ such that $A_\varepsilon := \{\varepsilon \leq f \leq 1 - \varepsilon\}$ is non-empty. We of course have $f = \frac{1}{2}(f + \varepsilon \mathbb{1}_{A_\varepsilon} + f - \varepsilon \mathbb{1}_{A_\varepsilon})$ and it follows from $f \in bE^*$ and our construction that $f \pm \varepsilon \mathbb{1}_{A_\varepsilon}$ are both in K. This shows that $f \notin \mathrm{ex}(K)$, whence $\mathrm{ex}(K) = \{\mathbb{1}_A : A \in E^*\}$. Finally, we note that the map $f \mapsto \int_\Omega f \, \mathrm{d}P - \int_\Omega f \, \mathrm{d}P'$ is linear and weakly continuous on K, so the result follows from Bauer's maximum principle. □

The second result shows that we always have a sort of "weak duality".

Lemma 2. *For any Borel equivalence relation E, we have*

$$\max_{A \in E^*} |P(A) - P'(A)| \leq \inf_{\tilde{P} \in \Pi(P, P')} (1 - \tilde{P}(E))$$

for all Borel probability measures P, P' on $(\Omega, \mathcal{F})$.

Proof. For any $A \in E^*$ we have $(A \times \Omega) \cap E = (\Omega \times A) \cap E$. Thus, for any Borel probability measures P, P' on $(\Omega, \mathcal{F})$ and any $\tilde{P} \in \Pi(P, P')$, we can bound:

$$\begin{aligned} P(A) - P'(A) = \tilde{P}(A \times \Omega) - \tilde{P}(\Omega \times A) &= \tilde{P}((A \times \Omega) \setminus E) \\ &\quad - \tilde{P}((\Omega \times A) \setminus E) \\ &\leq \tilde{P}((A \times \Omega) \setminus E). \end{aligned}$$

Now take the supremum over $A \in E^*$, use $\Omega \in E^*$ and apply Lemma 1. Finally, take the infimum over $\tilde{P} \in \Pi(P, P')$ to conclude. □

Third, we show that if a sub-probability measure has its marginals dominated by given marginals, then it is always possible to "complete" this to a probability measure.

Lemma 3. *Suppose P, P' are probability measures on $(\Omega, \mathcal{F})$ and that $\tilde{Q}$ is a sub-probability measure on $(\Omega, \mathcal{F})$ satisfying $\tilde{Q}(\,\cdot\, \times \Omega) \leq P$ and $\tilde{Q}(\Omega \times \,\cdot\,) \leq P'$. Then there exists sub-probability measures M, M' on $(\Omega, \mathcal{F})$ and a real number $0 \leq \gamma \leq 1$ such that $\tilde{P} := \tilde{Q} + \gamma\, M \otimes M' \in \Pi(P, P')$.*

Proof. We define $M := P - \tilde{Q}(\,\cdot\, \times \Omega)$ and $M' := P' - \tilde{Q}(\Omega \times \,\cdot\,)$, which are sub-probability measures on $(\Omega, \mathcal{F})$, and we write $\alpha \in [0, 1]$ for their common total mass. More explicitly, we have

$$\alpha = M(\Omega) = P(\Omega) - \tilde{Q}(\Omega \times \Omega) = 1 - \tilde{Q}(\Omega \times \Omega)$$

and similar for M' and P'. If $\alpha = 0$ then $\gamma = 0$ and $M = M' = 0$ are as desired. Otherwise $\alpha \in (0, 1]$, and we claim that $\gamma = 1/\alpha$ and M, M' are as desired. To see that $\tilde{P} := \tilde{Q} + \gamma\, M \otimes M'$ is a probability measure, compute

$$\tilde{P}(\Omega \times \Omega) = \tilde{Q}(\Omega \times \Omega) + \gamma\alpha^2 = \tilde{Q}(\Omega \times \Omega) + \alpha = 1.$$

To see that $\tilde{P}$ has the correct marginals, compute

$$\tilde{P}(\,\cdot\, \times \Omega) = \tilde{Q}(\,\cdot\, \times \Omega) + \gamma\alpha\, M(\Omega) = \tilde{Q}(\,\cdot\, \times \Omega) + M = P,$$

and likewise for $\tilde{P}(\Omega \times \,\cdot\,)$. □

Lastly, we give some information on the requisite notion of "smoothness" that will appear in our results. A Borel equivalence relation E on $(\Omega, \mathcal{F})$ is called *smooth* if there exists a standard Borel space $(S, \mathcal{S})$ and a measurable map $\phi : \Omega \to S$ such that $x, x' \in \Omega$ have $(x, x') \in E$ if and only if $\phi(x) = \phi(x')$. Our results will require the following novel characterization of smoothness, which we believe may be of independent interest; while the equivalence between (i) and (ii) is classical [183, Exercise 5.1.10], the equivalence with (iii) is novel and essential for our proof of Theorem 1.

Lemma 4. *The following are equivalent:*

(i) E is smooth.

(ii) E^* *is countably-generated.*

(iii) $E \in E^* \otimes E^*$.

Proof. By [183, Exercise 5.1.10], it suffices to prove that (i) and (iii) are equivalent. To show (i) implies (iii), suppose E is smooth so there exists a standard Borel space $(S, \mathcal{S})$ and a measurable map $\phi : \Omega \to S$ such that $x, x' \in \Omega$ have $(x, x') \in E$ if and only if $\phi(x) = \phi(x')$. It readily follows that $\phi : (\Omega, E^*) \to (S, \mathcal{S})$ is measurable, hence also that $f : (\Omega \times \Omega, E^* \otimes E^*) \to (S \times S, \mathcal{S} \otimes \mathcal{S})$ defined via $f(x, x') := (\phi(x), \phi(x'))$ is measurable. Finally, observe that we have $\Delta \in \mathcal{S} \otimes \mathcal{S}$ since $(S, \mathcal{S})$ is standard Borel, hence $E = f^{-1}(\Delta) \in E^* \otimes E^*$. Thus, (iii) holds.

It requires a bit more work to show (iii) implies (i), so suppose $E \in E^* \otimes E^*$. It is classical that $\Sigma_1 := \{B \in E^* \otimes E^* :$ there exist $A_1, A_2, , \ldots \in E^* \times E^*$ with $B \in \sigma(A_m \times A_n : m, n \in \mathbb{N})\}$ is a σ-algebra containing $E^* \times E^*$, hence $E^* \otimes E^* \subseteq \Sigma_1$. In particuar, there exist $A_1, A_2, \ldots \in E^* \times E^*$ with $E \in \sigma(A_m \times A_n : m, n \in \mathbb{N})$.

Next, we aim to show that the equivalence classes $[x]_E := \{x' \in \Omega : (x, x') \in E\}$ are in $\sigma(A_n : n \in \mathbb{N})$ for all $x \in \Omega$. To do this, take arbitrary $x \in \Omega$ and define $\Sigma_2(x) := \{B \in E^* \otimes E^* : ([x]_E \times [x]_E) \cap B \in \sigma(A_m \times A_n : m, n \in \mathbb{N})\}$, which is easily seen to be a σ-algebra. Moreover, observe that for all $m, n \in \mathbb{N}$ we have

$$([x]_E \times [x]_E) \cap (A_m \times A_n) = \begin{cases} A_m \times A_n, & \text{if } x \in A_m \cap A_n, \\ \varnothing, & \text{if } x \notin A_m \cap A_n, \end{cases}$$

since $A_m, A_n \in E^*$. This implies $\sigma(A_n \times A_m : m, n \in \mathbb{N}) \subseteq \Sigma_2(\omega)$, hence $E \in \Sigma_2(x)$. Consequently, $[x]_E \times [x]_E = ([x]_E \times [x]_E) \cap E \in \sigma(A_m \times A_n : m, n \in \mathbb{N}) \subseteq \sigma(A_n : n \in \mathbb{N}) \otimes \sigma(A_n : n \in \mathbb{N})$. Now Fubini's theorem gives $[x]_E \in \sigma(A_n : n \in \mathbb{N})$, as claimed.

Moving on, we claim that, for all $x, x' \in \Omega$ with $(x, x') \notin E$, there exists $n \in \mathbb{N}$ such that we have either $[x]_E \subseteq A_n$ and $[x']_E \cap A_n = \varnothing$ or $[x]_E \subseteq \Omega \setminus A_n$ and $[x']_E \subseteq A_n$. If this is not true, then, recalling $\{A_n\}_{n \in \mathbb{N}} \subseteq E^*$, there must exist $x, x' \in \Omega$ with $(x, x') \notin E$ such that for all $n \in \mathbb{N}$ we have $[x]_E, [x']_E \subseteq A_n$ or $[x]_E, [x']_E \subseteq \Omega \setminus A_n$. But $\Sigma_3(x, x') := \{A \in E^* : [x]_E, [x']_E \subseteq A$ or $[x]_E, [x']_E \subseteq \Omega \setminus A\}$ is a σ-algebra, so $\sigma(A_n : n \in \mathbb{N}) \subseteq \Sigma_3(x, x')$. This contradicts the conclusion of the previous paragraph.

Finally, we define the function $\phi : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ via the summation $\phi(x) := \sum_{n \in \mathbb{N}} 3^{-n} \mathbb{1}_{A_n}(x)$ for all $x \in \Omega$, which is clearly measurable. Also, the previous paragraph shows that $x, x' \in \Omega$ have $(x, x') \in E$ if and only if $\phi(x) = \phi(x')$. Therefore, E is smooth, so (i) holds. This finishes the proof. □

2.2 Two Duality Theorems

Now we can prove the main results of the chapter. The first result is based on a natural adaptation of classical proof of duality (2.3) for the total variation norm. Its proof is accompanied by the illustration in Figure 2.1.

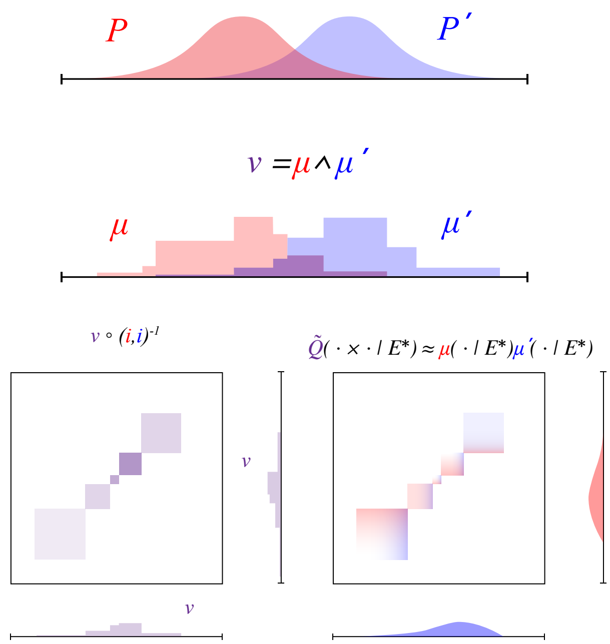

Figure 2.1: The proof of strong dualizability for smooth equivalence relations (Theorem 1). First, we coarsen P, P' on $(\Omega, \mathcal{F})$ to μ, μ' on (Ω, E^*). Second, we form their minimum $\nu = \mu \wedge \mu'$ and lift this to the "diagonal" in $(\Omega \times \Omega, E^* \otimes E^*)$. Then, we let the coordinates of the lifted measure be conditionally independent given their equivalence class, which we denote $\tilde{Q}$. Finally, we complete this to a bona fide coupling of P, P'.

Theorem 1. *Every smooth Borel equivalence relation is strongly dualizable.*

Proof. For arbitrary P, P', Lemma 2 gives

$$\max_{A \in E^*} |P(A) - P'(A)| \leq \inf_{\tilde{P} \in \Pi(P, P')} (1 - \tilde{P}(E)),$$

so it only remains to show that there exists $\tilde{P} \in \Pi(P, P')$ satisfying

$$1 - \tilde{P}(E) \leq \max_{A \in E^*} |P(A) - P'(A)|. \tag{2.7}$$

We perform this construction in several steps.

First, define $\mu := P|_{E^*}$ and $\mu' := P'|_{E^*}$ as probability measures on (Ω, E^*), and then set $\nu := \mu \wedge \mu'$. Now let $A_+, A_- \in E^*$ denote the positive part and negative part, respectively, of the Hahn-Jordan decomposition of the signed measure $\mu - \mu'$. Then we get

$$\nu(\Omega) = \mu(A_-) + \mu'(A_+) = 1 - \mu(A_+) + \mu'(A_+) = 1 - \sup_{A\in E^*} |P(A) - P'(A)|,$$

which will be useful in our later calculations.

Second, we use the fact that $(\Omega, \mathcal{F})$ is standard Borel to get [40, Corollary 10.4.6] a regular conditional distribution of P with respect to E^* denoted $K : \Omega \times \mathcal{F} \to [0,1]$ as well as a regular conditional distribution of P' with respect to E^* denoted $K' : \Omega \times \mathcal{F} \to [0,1]$. From this we define the set-function $\tilde{Q} : \mathcal{F} \times \mathcal{F} \to [0,1]$ via

$$\tilde{Q}(A \times A') := \int_\Omega K(y, A)K'(y, A') \,\mathrm{d}\nu(y) \tag{2.8}$$

for $A, A' \in \mathcal{F}$. We claim that $\tilde{Q}$ extends uniquely to a probability measure on $(\Omega\times\Omega, \mathcal{F}\otimes\mathcal{F})$, which (by a slight abuse of notation) we also denote by $\tilde{Q}$. This of course follows from Carathéodory's extension theorem [104, Theorem 2.5] if we can show that $\tilde{Q}$ is countably additive on the semi-ring $\mathcal{F} \times \mathcal{F}$, so suppose that we have $A \times A' = \bigcup_{n\in\mathbb{N}}(A_n \times A'_n)$ for $A, A', \{A_n\}_{n\in\mathbb{N}}$, and $\{A'_n\}_{n\in\mathbb{N}}$ in $\mathcal{F}$ such that $\{A_n \times A'_n\}_{n\in\mathbb{N}}$ are disjoint. This implies $\mathbb{1}_A \otimes \mathbb{1}_{A'} = \sum_{n\in\mathbb{N}}(\mathbb{1}_{A_n} \otimes \mathbb{1}_{A'_n})$, so for a fixed $y \in \Omega$ we can take the probability of both sides under the product measure $K(y, \cdot\,) \otimes K'(y, \cdot\,)$ and we get

$$K(y, A)K'(y, A') = \sum_{n\in\mathbb{N}} K(y, A_n)K'(y, A'_n).$$

Now integrate both sides with respect to ν, and use monotone convergence to get

$$\begin{aligned}
\tilde{Q}(A \times A') &= \int_\Omega K(y, A)K'(y, A') \,\mathrm{d}\nu(y) \\
&= \int_\Omega \sum_{n\in\mathbb{N}} K(y, A_n)K'(y, A'_n) \,\mathrm{d}\nu(y) \\
&= \sum_{n\in\mathbb{N}} \int_\Omega K(y, A_n)K'(y, A'_n) \,\mathrm{d}\nu(y) = \sum_{n\in\mathbb{N}} \tilde{Q}(A_n \times A'_n).
\end{aligned}$$

This is as desired, so we have a probability measure $\tilde{Q}$ on $(\Omega \times \Omega, \mathcal{F} \otimes \mathcal{F})$ whose values on rectangles are defined by (2.8).

Next we claim that the hypotheses of Lemma 3 are satisfied. To do this, observe that for any $A \in \mathcal{F}$ we can compute

$$\begin{aligned}
\tilde{Q}(A \times \Omega) &= \int_\Omega K(y, A)K'(y, \Omega) \,\mathrm{d}\nu(y) \\
&= \int_\Omega K(y, A) \,\mathrm{d}\nu(y) \\
&\le \int_\Omega K(x, A) \,\mathrm{d}\mu(x) \\
&= \int_\Omega K(x, A) \,\mathrm{d}P(x) = P(A),
\end{aligned}$$

and that an identical calculation shows $\tilde{Q}(\Omega \times A) \leq P'(A)$ for all $A \in \mathcal{F}$. We can get a coupling $\tilde{P} \in \Pi(P, P')$ with $\tilde{Q} \leq \tilde{P}$.

Finally, we will show that $\tilde{P}$ satisfies (2.7). To do this, write $f : (\Omega, E^*) \to (\Omega\times\Omega, E^*\otimes E^*)$ for the measurable map $f(y) := (y, y)$, and let us show that we have $\tilde{P}(S) = \nu(f^{-1}(S))$ for all $S \in E^* \otimes E^*$. Indeed, it is straightforward to show that for all $A, A' \in E^*$ we have

$$\begin{aligned}\tilde{P}(A \times A') &= \int_\Omega K(y, A)K'(y, A')\,\mathrm{d}\nu(y)\\ &= \int_\Omega \mathbb{1}_A(y)\mathbb{1}_{A'}(y)\,\mathrm{d}\nu(y) = \nu(A \cap A') = \nu(f^{-1}(A \times A')).\end{aligned}$$

In the second equality, we used that $K(y, A) = \mathbb{1}_A(y)$ holds P-almost surely hence μ-almost surely hence ν-almost surely, since $\nu \ll \mu$, and also the analogous result for K'. This shows that the probability measures $\tilde{P}$ and $\nu \circ f^{-1}$ agree on the π-system $E^* \times E^*$, so it follows that they agree on $E^* \otimes E^*$. Finally, we use Lemma 4 to get $E \in E^* \otimes E^*$, and we compute:

$$\tilde{P}(E) = \nu(f^{-1}(E)) \geq \nu(f^{-1}(\Delta)) = \nu(\Omega) = 1 - \sup_{A\in E^*} |P(A) - P'(A)|,$$

so rearranging this gives

$$1 - \tilde{P}(E) \leq \sup_{A\in E^*} |P(A) - P'(A)|,$$

as desired. □

One may be tempted to think that Theorem 1 is powerful enough to establish strong dualizability most examples of interest in probability theory. However, non-smooth Borel equivalence relations are common, for example the equivalence relation of eventual equality E_0. (The fact that E_0 is not smooth is exemplified by the Glimm-Effros dichotomy [111, Theorem 6.5], but it can also be shown directly via elementary considerations.) Thus, our goal is to widen our sufficient conditions for strong dualizability. Towards filling this gap, we establish a closure property for strong dualizability as our second main result. Its proof is accompanied by the illustration in Figure 2.2.

Theorem 2. *A countable increasing union of strongly dualizable Borel equivalence relations is strongly dualizable.*

Proof. Suppose that $E_1 \subseteq E_2 \subseteq \cdots$ are strongly dualizable Borel equivalence relations, and write $E := \bigcup_{n\in\mathbb{N}} E_n$. Since E_n is measurable for all $n \in \mathbb{N}$ it follows that E is measurable. As in the proof of Theorem 1, we begin by noting that for arbitrary P, P', Lemma 2 gives

$$\max_{A\in E^*} |P(A) - P'(A)| \leq \inf_{\tilde{P}\in\Pi(P,P')} (1 - \tilde{P}(E)),$$

so that we only need to construct some $\tilde{P} \in \Pi(P, P')$ satisfying

$$1 - \tilde{P}(E) \leq \max_{A\in E^*} |P(A) - P'(A)|. \tag{2.9}$$

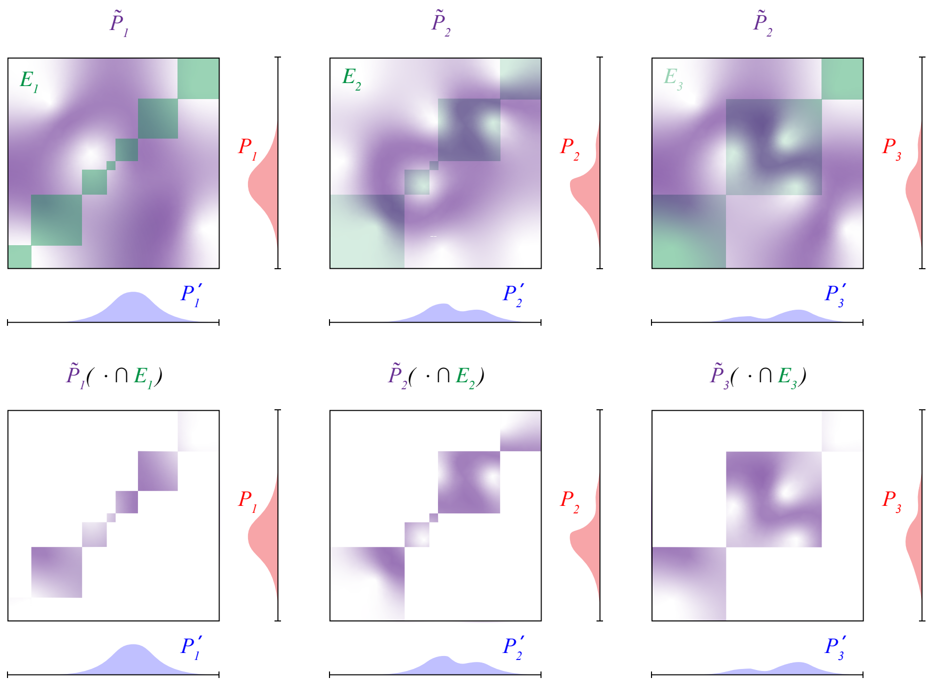

Figure 2.2: The proof of strong dualizability for countable increasing unions of strongly dualizable equivalence relations (Theorem 2). First, we set $P_1 = P$ and $P'_1 = P'$. Second, we let $\tilde{P}_1$ be an optimal coupling of P_1, P'_1 for the equivalence relation E_1. Third, we "set aside" $\tilde{P}_1(\cdot \cap E_1)$, and we subtract its marginals off of P_1, P'_1, yielding $P_2 = P_1 - \tilde{P}_1((\cdot \times \Omega) \cap E_1)$ and $P_2 = P_1 - \tilde{P}_1((\Omega \times \cdot) \cap E_1)$. Then we repeat the process to P_n, P'_n and E_n for $n \geq 2$. The optimal coupling is given by completing $\sum_{n \in \mathbb{N}} \tilde{P}_n(\cdot \cap E_n)$ to a bona fide coupling of P, P'.

Our construction is iterative and takes several steps.

To begin, set $P_1 := P$ and $P'_1 := P'$. Then, inductively for $n \in \mathbb{N}$, use the strong dualizability of E_n to get $\tilde{P}_n \in \Pi(P_n, P'_n)$ satisfying

$$\max_{A \in E_n^*} |P_n(A) - P'_n(A)| = 1 - \sum_{m<n} \tilde{P}_m(E_m) - \tilde{P}_n(E_n) = 1 - \sum_{m \leq n} \tilde{P}_m(E_m), \tag{2.10}$$

and then set $P_{n+1} := P_n - \tilde{P}_n((\cdot \times \Omega) \cap E_n)$ and $P'_{n+1} := P'_n - \tilde{P}_n((\Omega \times \cdot) \cap E_n)$. To ensure that this construction is well-defined, we must verify that P_n and P'_n are, for each $n \in \mathbb{N}$, sub-probability measures with the same total mass, $1 - \sum_{m<n} \tilde{P}_m(E_m)$. This follows by

induction where the base case $n = 1$ is trivial and the inductive step follows from combining

$$\begin{aligned} P_{n+1}(\Omega) &= P_n(\Omega) - \tilde{P}_n((\Omega \times \Omega) \cap E_n) \\ &= P_n(\Omega) - \tilde{P}_n(E_n) \\ &= 1 - \sum_{m<n} \tilde{P}_m(E_m) - \tilde{P}_n(E_n) = 1 - \sum_{m<n+1} \tilde{P}_m(E_m) \end{aligned}$$

with the analogous calculation for $P'_{n+1}(\Omega)$.

Next, fix $n \in \mathbb{N}$, and recall that by construction we have $P_n = P - \sum_{m<n} \tilde{P}_m((\,\cdot \times \Omega) \cap E_m)$ and $P'_n = P' - \sum_{m<n} \tilde{P}_m((\Omega \times \,\cdot\,) \cap E_m)$. Thus, combining (2.10) with the triangle inequality, we get

$$\begin{aligned} 1 - &\sum_{m \le n} \tilde{P}_m(E_m) \\ &= \sup_{A \in E_n^*} |P_n(A) - P'_n(A)| \\ &\le \sup_{A \in E_n^*} |P(A) - P'(A)| \\ &\qquad + \sum_{m<n} \sup_{A \in E_n^*} |\tilde{P}_m((A \times \Omega) \cap E_m) - \tilde{P}_m((\Omega \times A) \cap E_m)|. \end{aligned}$$

We claim that the sum in the last line above is equal to zero. In fact, we claim that all summands are equal to zero, in that for all $m < n$ we have

$$\sup_{A \in E_n^*} |\tilde{P}_m((A \times \Omega) \cap E_m) - \tilde{P}_m((\Omega \times A) \cap E_m)| = 0. \tag{2.11}$$

It follows by normalizing that this holds if and only if we have

$$\sup_{A \in E_n^*} |\tilde{P}_m(A \times \Omega \,|\, E_m) - \tilde{P}_m(\Omega \times A \,|\, E_m)| = 0. \tag{2.12}$$

By the strong dualizability of E_n we know that (2.12) holds if and only if there exists a coupling $\tilde{P} \in \Pi(\tilde{P}_m(\,\cdot \times \Omega \,|\, E_m), \tilde{P}_m(\Omega \times \,\cdot \,|\, E_m))$ satisfying $\tilde{P}(E_n) = 1$. Of course, the coupling $\tilde{P}_m(\cdot \,|\, E_m)$ is exactly what is needed, since $E_m \subseteq E_{n+1}$ implies $\tilde{P}_m(E_{n+1} \,|\, E_m) = 1$. Thus, we have shown

$$1 - \sum_{m \le n} \tilde{P}_m(E_m) \le \sup_{A \in E_{n+1}^*} |P(A) - P'(A)| \tag{2.13}$$

for all $n \in \mathbb{N}$.

Now let us get for each $n \in \mathbb{N}$ some $A_n \in E_n^*$ with

$$|P(A_n) - P'(A_n)| \ge \sup_{A \in E_n^*} |P(A) - P'(A)| - \frac{1}{2^n}$$

Now consider the Hilbert space $L^2(\Omega, \mathcal{F}\frac{1}{2}(P+P'))$, in which $\{\mathbb{1}_{A_n}\}_{n\in\mathbb{N}}$ form a norm-bounded sequence. By the Banach-Alaoglu theorem, there exists a subsequence $\{n_j\}_{j\in\mathbb{N}}$ and some $f \in L^2(\Omega, \mathcal{F}, \frac{1}{2}(P+P'))$ with $\mathbb{1}_{A_{n_j}} \to f$ weakly. Since $L^2(\Omega, E_n^*, \frac{1}{2}(P+P'))$ is a strongly closed subspace of $L^2(\Omega, \mathcal{F}, \frac{1}{2}(P+P'))$, it follows that it is also weakly closed. This implies that $f \in L^2(\Omega, E_n^*, \frac{1}{2}(P+P'))$ for all $n \in \mathbb{N}$, hence $f \in L^2(\Omega, E^*, \frac{1}{2}(P+P'))$. Also, we have

$$0 \le \lim_{j\to\infty} \int_\Omega \mathbb{1}_{A_{n_j}} \mathbb{1}_{\{f\le 0\}} \,\mathrm{d}\left(\frac{P+P'}{2}\right) = \int_\Omega f \mathbb{1}_{\{f\le 0\}} \,\mathrm{d}\left(\frac{P+P'}{2}\right) \le 0$$

which shows that $f \ge 0$ holds P- and P'-almost surely; a similar argument shows that $f \le 1$ holds P- and P'-almost surely. Putting this all together, we conclude that there exists a function $g : \Omega \to \mathbb{R}$ which is E^*-measurable and satisfies $0 \le g \le 1$ such that $f = g$ both P- and P'-almost surely. Consequently, Lemma 1 gives:

$$\begin{aligned}
\liminf_{n\to\infty} \sup_{A\in E_n^*} |P(A) - P'(A)| &\le \lim_{j\to\infty} |P(A_{n_j}) - P'(A_{n_j})| \\
&= \left| \int_\Omega f \,\mathrm{d}P - \int_\Omega f \,\mathrm{d}P' \right| \\
&= \left| \int_\Omega g \,\mathrm{d}P - \int_\Omega g \,\mathrm{d}P' \right| \\
&\le \sup_{A\in E^*} |P(A) - P'(A)|.
\end{aligned}$$

Therefore, we conclude

$$1 - \sum_{n\in\mathbb{N}} \tilde{P}_n(E_n) \le \sup_{A\in E^*} |P(A) - P'(A)|. \tag{2.14}$$

by taking $n \to \infty$ in (2.13).

Now we have all of the ingredients to construct our coupling. First, set

$$\tilde{Q} := \sum_{n\in\mathbb{N}} \tilde{P}_n(\cdot \cap E_n),$$

which is evidently a sub-probability measure on $(\Omega \times \Omega, \mathcal{F} \otimes \mathcal{F})$. Next we claim that the hypotheses of Lemma 3 are satisfied. Indeed, for any $A \in \mathcal{F}$ and $n \in \mathbb{N}$ we have

$$\begin{aligned}
\sum_{m\le n} \tilde{P}_m((A \times \Omega) \cap E_m) &= \sum_{m\le n} (P_{m+1}(A) - P_{m+1}(A)) \\
&= P(A) - P_{n+1}(A) \\
&\le P(A),
\end{aligned}$$

so taking $n \to \infty$ we get

$$\tilde{Q}(A \times \Omega) = \sum_{n\in\mathbb{N}} \tilde{Q}_n((A \times \Omega) \cap E_n) \le P(A).$$

We also get $\tilde{Q}(\Omega \times A) \leq P'(A)$ by the same calculation. Therefore, Lemma 3 gives us $\tilde{P} \in \Pi(P, P')$ with $\tilde{Q} \leq \tilde{P}$.

To finish the proof, we note that for all $n \in \mathbb{N}$ we have

$$\tilde{P}(E_n) \geq \tilde{Q}(E_n) \geq \sum_{m \leq n} \tilde{P}_m(E_m \cap E_n) = \sum_{m \leq n} \tilde{P}_m(E_m),$$

so taking $n \to \infty$ gives

$$\tilde{P}(E) \geq \sum_{n \in \mathbb{N}} \tilde{P}_n(E_n).$$

Finally, by applying (2.14), we find

$$1 - \tilde{P}(E) \leq 1 - \sum_{n \in \mathbb{N}} \tilde{P}_n(E_n) \leq \sup_{A \in E^*} |P(A) - P'(A)|,$$

and the result is proved. □

It appears that Theorem 1 and Theorem 2 are together powerful enough to establish strong dualizability for most concrete applications appearing in probability. However, we believe that an interesting question of independent interest is whether all Borel equivalence relations are strongly dualizable.

Chapter 3

Applications in Stochastic Calculus

Consider two Borel probability measures P, P' on $D := D_0([0,\infty);\mathbb{R})$, the Polish space of càdlàg ("right-continuous, left-limits") paths from $[0,\infty)$ to $\mathbb{R}$ vanishing at 0. Writing $E_{0+} := \bigcup_{t>0}\{(x,x') \in D \times D : x_s = x'_s \text{ for all } 0 \leq s < t\}$ for the equivalence relation of agreeing for some initial segment of time, we are interested in the E_{0+}-coupling problem:

$$\begin{cases} \text{minimize} & 1 - \tilde{P}(E_{0+}) \\ \text{over} & \tilde{P} \in \Pi(P, P'). \end{cases} \tag{3.1}$$

Stated another way, can two different càdlàg stochastic processes started from the same point be coupled to travel alongside each other for some initial segment of time?

This question has been recently studied in [65, 204], with motivations in sequential testing, Markov chain Monte Carlo (MCMC), and more. Their results include some bounds on solutions to (3.1) and some explicit calculations for the special case of Brownian motions with drift, but their methods can only go so far. The goal of this chapter is to show that the duality theory of Chapter 2 provides a robust set of tools for studying (3.1) and related problems.

An important precedent for this analysis is the following closely related problem: Writing $E_{+\infty} := \bigcup_{t>0}\{(x,x') \in D \times D : x_s = x'_s \text{ for all } s \geq t\}$ for the equivalence relation of eventual agreement, one may be interested in the $E_{+\infty}$-coupling problem:

$$\begin{cases} \text{minimize} & 1 - \tilde{P}(E_{+\infty}) \\ \text{over} & \tilde{P} \in \Pi(P, P'). \end{cases} \tag{3.2}$$

That is, can two different càdlàg stochastic processes be coupled to eventually agree?

The problem (3.2) has received a great deal of attention over many decades. Following the development of a sufficiently general theory [85, 163, 82], many authors have developed precise results about (3.2) for certain classes of processes of interest: diffusions [16, 25], integral functionals of Brownian motions and diffusions [115, 118], Brownian motion on Riemannian manifolds [116, 125], Brownian motion in Banach spaces [46], Lévy processes

[41, 178, 117], Lévy diffusions [134], and more. There are also subtle measurability issues in (3.2), and these have become an interesting question in their own right [92, 42].

We refer to the problems (3.1) and (3.2) as the *germ coupling problem* and the *tail coupling problem*, respectively. Since it is known that (3.2) is closely related to the tail σ-algebra on D, one expects that (3.1) should be closely related to the germ σ-algebra on D. This is indeed the case, and it can be made precise via the duality theory of Chapter 2. In particular, we can hope to develop precise results about (3.1) for certain classes of processes of interest, in the same way that has already been done for (3.2).

This chapter is based on the works [99, 97, 56] in which we study several aspects of the germ coupling problem (3.1). The first part characterizes when (3.1) has its optimal value equal to zero, for classes of processes including diffusions (Theorem 3) and Lévy processes (Theorem 5). The second part aims to explicitly construct solutions to (3.1) for Brownian motions with drift (Theorem 6) and pure-jump Lévy processes (Theorem 7). These results rely, in some way or another, on the sufficiently powerful duality theory for the equivalence coupling problem developed in Chapter 2.

3.1 Germ Couplings

We begin by establishing some notation. We write $D := D_0([0,\infty);\mathbb{R})$ for the space of càdlàg ("right-continuous, left-limits") paths from $[0,\infty)$ to $\mathbb{R}$, vanishing at 0, endowed with Skorokhod's J1 topology which makes this into a Polish space. For any pair of paths $x, x' \in D$, we define their *fragmentation time* as the first time that they disagree, or

$$\tau_{\text{frag}}(x, x') := \inf\{t \geq 0 : x_t \neq x'_t\}.$$

Notice that $\tau_{\text{frag}} : D \times D \to \mathbb{R}$ is a Borel measurable map. This leads us to the following:

Definition 2. *A pair $P, P' \in \mathcal{P}(D)$ is said to have the* germ coupling property (GCP) *if there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ supporting càdlàg stochastic processes X, X' with laws P, P', respectively, such that we have $\mathbb{P}(\tau_{\text{frag}}(X, X') > 0) = 1$. In this case, we say that $(\Omega, \mathcal{F}, \mathbb{P}, X, X')$* witnesses *germ coupling of P, P' or that X, X' are* germ coupled under $\mathbb{P}$.

Of course, for any $(\Omega, \mathcal{F}, \mathbb{P}, X, X')$ witnessing germ coupling of P, P', it is possible to take $\Omega = D \times D$ and to let X, X' denote the canonical coordinate processes in Ω. This is useful to keep in mind, but we will usually think of $(\Omega, \mathcal{F}, \mathbb{P})$ as an abstract probability space throughout this chapter. Also, by a slight abuse of terminology, we will say that a pairs of stochastic processes has the GCP when the pair of their laws has the GCP.

In words, a pair of stochastic processes has the GCP if they can be coupled to almost surely travel together for some positive amount of time. This notion of small-time similarity appears rather strong, and, apart from trivialities, it is not at all obvious how to construct pairs of stochastic processes with the GCP. Thus, we were greatly intrigued by the results of [65], which show that Brownian motions with any two drifts always the Brownian GCP.

One application of the duality results of the previous chapter is that we can develop a robust method of proof for this result, which has applications to many other classes of Markov processes.

To state this, we need some further notation. That is, for $t > 0$, we write D_t for the space of càdlàg functions $[0, t) \to \mathbb{R}$ and we write $\mathcal{D}_t := \sigma(x_s : 0 \le s < t)$ for the σ-algebra of all information up to time t. (Note that we do not take the completion of these σ-algebras, so $\{\mathcal{D}_t\}_{t\ge 0}$ is not right-continuous.) Then we define the *germ σ-algebra* via

$$\mathcal{D}_{0+} := \bigcap_{t>0} \mathcal{D}_t.$$

This leads us to the following fundamental result, which also explains our choice of terminology:

Proposition 1. *A pair of càdlàg stochsatic processes $P, P' \in \mathcal{P}(D)$ has the GCP if and only if we have $P(A) = P'(A)$ for all $A \in \mathcal{D}_{0+}$.*

Proof. For $t > 0$ we define the equivalence relation

$$E_t := \{(x, x') : x_s = x'_s \text{ for all } 0 \le s < t\},$$

which is smooth since $E_t = \phi_t^{-1}(\Delta)$, where $\phi_t : D \to D_t$ is the natural truncation map. Also we define $E_{0+} = \bigcup_{n\in\mathbb{N}} E_{2^{-n}}$, and we note that $\{E_{2^{-n}}\}_{n\in\mathbb{N}}$ is non-decreasing. Therefore, by Theorem 1 and Theorem 2, we see that E_{0+} is strongly dualizable. We can also check

$$E_{0+}^* = \left(\bigcup_{n\in\mathbb{N}} E_{2^{-n}}\right)^* = \bigcap_{n\in\mathbb{N}} E_{2^{-n}}^* = \bigcap_{n\in\mathbb{N}} \mathcal{D}_t = \mathcal{D}_{0+}.$$

So, by the definition of strong duality, it only remains to show that $P, P' \in \mathcal{P}(D)$ has the GCP if and only if there exists $\tilde{P} \in \Pi(P, P')$ with $\tilde{P}(E_{0+}) = 1$. For one direction, observe that, if $(\Omega, \mathcal{F}, \mathbb{P}, X, X')$ witnesses P, P' satisfying the GCP, then the joint law $\tilde{P} := \mathbb{P} \circ (X, X')^{-1}$ is exactly a coupling $\tilde{P} \in \Pi(P, P')$ with $\tilde{P}(E_{0+}) = 1$. For the other direction, suppose that $\tilde{P} \in \Pi(P, P')$ is some coupling with $\tilde{P}(E_{0+}) = 1$. Then the canonical coordinate processes X, X' on $(D \times D, \mathcal{B}(D) \otimes \mathcal{B}(D), \mathbb{P})$ are germ coupled under $\mathbb{P}$. Thus, the result holds by the definition of strong duality. □

On the one hand, this provides an authoritative answer to the question of which pairs of stochastic processes have the GCP. On the other hand, it remains to show that this equivalent condition is easy to verify in some concrete cases. In fact, the value of Proposition 1 is that the classical tools of stochastic calculus have very much to say about the germ σ-algebra $\mathcal{D}_{0+}$, and thus many soft arguments suddently become available to us. The next goal of this section is to give two examples of this.

The first example concerns time-homogeneous one-dimensional diffusions within realm of classical conditions ensuring the existence of strong solutions to a given SDE [172, Chapter IX, Theorem 2.1]. Since the GCP is spatially translation-invariant and temporally local, we lose no generality in restricting our attention to diffusions started at the origin and viewed on a finite time interval. That is, for Lipschitz functions $\mu : \mathbb{R} \to \mathbb{R}$ and $\sigma : \mathbb{R} \to [0,\infty)$, we write $P^{\mu,\sigma} \in \mathcal{P}(D_1)$ for the law of the strong solution $X = \{X_t\}_{0\le t<1}$ of the SDE

$$\begin{cases} \mathrm{d}X_t = \mu(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}B_t \text{ for } 0 \le t < 1 \\ X_0 = 0. \end{cases} \tag{3.3}$$

We also write $W \in \mathcal{P}(D)$ for the usual Wiener measure, that is, the law of a standard Brownian motion. Then we have the following:

Theorem 3. *If $\mu : \mathbb{R} \to \mathbb{R}$ and $\sigma : \mathbb{R} \to [0,\infty)$ are Lipschitz continuous, then $P^{\mu,\sigma}, W$ has the GCP if and only if $\sigma \equiv 1$ on some neighborhood of 0.*

Proof. For one direction, let $(\Omega, \mathcal{F}, \mathbb{P}, X, X')$ witness the GCP for $P^{\mu,\sigma}, W$. Then consider the event $A \in \mathcal{B}(D)$ defined via

$$A := \bigcup_{n\in\mathbb{N}} \{x \in D : \langle x, x\rangle_t = t \text{ for all } 0 \le t \le 2^{-n}\}$$

where $\langle x, x\rangle$ denotes the Itô quadratic variation of x. Since $A \in \mathcal{D}_{0+}$ and $\mathbb{P}(X' \in A) = W(A) = 1$, Proposition 1 implies that we must have $\mathbb{P}(X \in A) = 1$. Now note that the quadratic variation of X is given by $\langle X, X\rangle_t = \int_0^t (\sigma(X_s))^2\,\mathrm{d}s$, so we conclude that $\mathbb{P}$-almost surely there exists some $N \in \mathbb{N}$ with $\sigma(X_s) = 1$ for all $0 \le s \le 2^{-N}$. In particular, we conclude $\sigma(0) = 1$.

Next we define the random times $\tau^- := \inf\{t > 0 : X_t < 0\}$ and $\tau^+ := \inf\{t > 0 : X_t > 0\}$ on $(\Omega, \mathcal{F}, \mathbb{P})$, which are stopping times with respect to the natural filtration of X. We recall that the standard small-time approximation for diffusions with Lipschitz coefficients guarantees that $X_t/\sqrt{t}$ converges in distribution as $t \to 0$ to a Gaussian random variable with mean 0 and variance $\sigma(0) = 1$, hence

$$\begin{aligned} \mathbb{P}(\tau^- = 0) &= \lim_{t\to 0} \mathbb{P}(X_s < 0 \text{ for some } 0 \le s \le t) \\ &\ge \liminf_{t\to\infty} \mathbb{P}(X_t < 0) \\ &= \liminf_{t\to\infty} \mathbb{P}\left(\frac{X_t}{\sqrt{t}} < 0\right) = \frac{1}{2} > 0. \end{aligned}$$

By Blumenthal's zero-one law applied to the strong Markov process X, we see that $\mathbb{P}(\tau^- = 0) > 0$ implies $\mathbb{P}(\tau^- = 0) = 1$. The same argument applies to show $\mathbb{P}(\tau^+ = 0) = 1$.

Finally, we put all the pieces together. By the almost sure continuity of X and the fact that $\mathbb{P}(\tau^- = 0) = \mathbb{P}(\tau^+ = 0) = 1$, we conclude that $\{X_s : 0 \le s \le T\}$ contains an open neighborhood of 0 for any random variable T which is almost surely strictly positive. But we already know that there exists an $\mathbb{N}$-valued random variable N with $\sigma(X_s) = 1$ for all $0 \le s \le 2^{-N}$, so we conclude that we have $\sigma \equiv 1$ on some neighborhood of 0.

For the other direction, suppose that $\sigma \equiv 1$ on some neighborhood U of 0. Then get reals $a, b \in \mathbb{R}$ with $a < b$ and $0 \in (a,b) \subseteq [a,b] \subseteq U$. Since (3.3) admits strong solutions, we can construct $X = \{X_t\}_{t\geq 0}$ on the probability space $(D, \mathcal{B}(D), W)$ with its natural filtration $\{\mathcal{D}_t\}_{t\geq 0}$. We write $\mathbb{E}$ for the expectation on this space.

To begin, we claim that the exit time $\tau_{a,b}^X := \inf\{t > 0 : X_t \notin [a,b]\}$ has finite exponential moments of all orders. To see this, define $m := \max_{-a\leq x\leq b} \mu(x)$, which is finite by the continuity of μ on $[a,b]$. Since we have $\mu(X_t) \leq m$ for all $0 \leq t \leq \tau_{a,b}^X$ almost surely, it follows that one can construct $B^m = \{B_t^m\}_{t\geq 0}$ a Brownian motion with drift m on the same probability space in such a way that we have $X_t \leq B_t^m$ for all $0 \leq t \leq \tau_a^X$ almost surely. Then define the stopping time $\tau_a^{B^m} := \inf\{t > 0 : B_t^m \leq a\}$, and note that we have $\tau_{a,b}^X \leq \tau_a^{B^m}$ almost surely. It is known that $\tau_a^{B^m}$ has finite exponential moments of all orders, so the same must be true of $\tau_{a,b}^X$.

In particular, we have shown

$$\mathbb{E}\left[\exp\left(\frac{1}{2}\int_0^{\tau_{a,b}^X} (\mu(X_s))^2 \, \mathrm{d}s\right)\right] \leq \mathbb{E}\left[\exp\left(\frac{m^2}{2}\tau_{a,b}^X\right)\right] < \infty.$$

This stopping-time version of Novikov's condition implies a stopping-time version of Girsanov's theorem which states that $P^{\mu,\sigma}$ is mutually absolutely continuous with respect to $P_{0,\sigma}$ when restricted to the stopped σ-algebra $\mathcal{D}_{\tau_{a,b}} \subseteq \mathcal{B}(D)$, for $\tau_{a,b} := \inf\{t > 0 : x_t \notin [a,b]\}$. Importantly, note that we have $\mathcal{D}_{0+} \subseteq \mathcal{D}_{\tau_{a,b}}$ holding both $P^{\mu,\sigma}$- and $P^{0,\sigma}$-almost surely by the continuity of sample paths. Since Blumenthal's zero-one law implies that all events in $\mathcal{D}_{0+}$ have probability in $\{0,1\}$ under both $P^{\mu,\sigma}$ and $P^{0,\sigma}$ we conclude $P^{\mu,\sigma}(A) = P^{0,\sigma}(A)$ for all $A \in \mathcal{D}_{0+}$.

Finally observe that $\sigma \equiv 1$ on U implies that $P^{0,\sigma}(A) = W(A)$ for all $A \in \mathcal{D}_{\tau_U}$, where $\tau_U := \inf\{t > 0 : x_t \notin U\}$. Again applying the almost surely continuity of sample paths under both $P^{0,\sigma}$ and W, we have $\mathcal{D}_{0+} \subseteq \mathcal{D}_{\tau_U}$ almost surely, hence $P^{0,\sigma}(A) = W(A)$ for all $A \in \mathcal{D}_{0+}$. Putting this all together yields $P^{\mu,\sigma}(A) = W(A)$ for all $A \in \mathcal{D}_{0+}$, so Proposition 1 implies that the pair $P^{\mu,\sigma}, W$ has the GCP. □

As a consequence, we get the following result for Brownian motions with possibly different drift and diffusion parameters. To state it, let us write $W^{a,\sigma}$ for the law of a Brownian motion with drift a and with diffusion σ.

Corollary 1. *Brownian motions $W^{a,\sigma}, W^{a',\sigma'}$ have the GCP if and only if*

(i) $\sigma = \sigma'$, and

(ii) if $\sigma = \sigma' = 0$, then $a = a'$.

Proof. Suppose both properties (i) and (ii) are satisfied. If $\sigma = \sigma' = 0$, then $W^{0,a}$ and $W^{0,a}$ are both equal to the point mass on the path $\{at\}_{t\geq 0}$, so they can be germ coupled deterministically. If $\sigma = \sigma' > 0$ then by changing time and applying Brownian scaling, we can assume $\sigma' = 1$; in this case, the result follows from Proposition 1. Conversely, suppose that $W^{a,\sigma}, W^{a',\sigma'}$ have the GCP. If $\sigma = \sigma' = 0$, then $W^{0,a}$ and $W^{0,a}$ are respectively equal to the point masses on the paths $\{at\}_{t\geq 0}$ and $\{a't\}_{t\geq 0}$; thus, the existence of germ coupling implies $a = a'$, hence (ii). Otherwise, at least one of σ, σ' is nonzero; assume without loss of generality that $\sigma' = 1$. By changing time and applying Brownian scaling, we can again assume $\sigma' = 1$. Then Proposition 1 implies $\sigma = 1$, which is (i). □

The second example is a simple characterization of the GCP for pairs of real-valued Lévy processes in terms of their characteristic triplets. However, we begin with the case of pure-jump Lévy processes. For some notation, let us write $L_\nu \in \mathcal{P}(D)$ for the law of a Lévy processes with Lévy measure ν. We also introduce the *Hellinger distance* between Lévy measures ν, ν' as

$$H^2(\nu, \nu') := \frac{1}{2}\int_{\mathbb{R}} \left(\sqrt{\frac{\mathrm{d}\nu}{\mathrm{d}\lambda}} - \sqrt{\frac{\mathrm{d}\nu'}{\mathrm{d}\lambda}}\right)^2 \mathrm{d}\lambda$$

where λ is any σ-finite measure dominating both ν and ν', and where the value of $H^2(\nu, \nu')$ does not depend on the choice of λ. (Thus, one can take, for example, $\lambda = \nu + \nu'$.) Then we get:

Theorem 4. *A pair of pure-jump Lévy processes $L_\nu, L_{\nu'}$ has the GCP if and only if we have $H^2(\nu, \nu') < \infty$.*

Proof. For $t > 0$ we write

$$H^2_{\mathcal{D}_t}(L_\nu, L_{\nu'}) := H^2(L_\nu|_{\mathcal{D}_t}, L_{\nu'}|_{\mathcal{D}_t})$$

for the Hellinger distance between their natural restrictions of $L_\nu, L_{\nu'}$ to the space $\mathcal{P}(D_t)$. We claim that the following properties are equivalent:

(a) $L_\nu(A) = L_{\nu'}(A)$ for all $A \in \mathcal{D}_{0+}$.

(b) $\sup_{A\in\mathcal{D}_t} |L_\nu(A) - L_{\nu'}(A)| \to 0$ as $t \to 0$.

(c) $H^2_{\mathcal{D}_t}(L_\nu, L_{\nu'}) \to 0$ as $t \to 0$.

(d) $H^2(\nu, \nu') < \infty$.

Indeed, (a) is equivalent to (b) because of the relationship

$$\lim_{t\to 0} \sup_{A\in\mathcal{D}_t} |L_\nu(A) - L_{\nu'}(A)| = \sup_{A\in\mathcal{D}_{0+}} |L_\nu(A) - L_{\nu'}(A)|$$

which can be proven in the same way as our proof of (2.14). Then, (b) is equivalent to (c) because of the usual estimate relating total variation distance and Hellinger distance,

$$H^2_{\mathcal{D}_t}(L_\nu, L_{\nu'}) \leq \sup_{A\in\mathcal{D}_t} |L_\nu(A) - L_{\nu'}(A)| \leq \sqrt{2\, H^2_{\mathcal{D}_t}(L_\nu, L_{\nu'})},$$

for any $t > 0$. And, (c) is equivalent to (d) because of the classical result of Newman [152] which states that we have

$$H^2_{\mathcal{D}_t}(L_\nu, L_{\nu'}) = 1 - \exp\left(-tH^2(\nu,\nu')\right)$$

for all $t > 0$. Since (a) is equivalent to the GCP by Proposition 1, the result is proved. □

Now we get our result for general Lévy processes. To state it, we write $L_{a,\sigma,\nu}, L_{a',\sigma',\nu'} \in \mathcal{P}(D)$ for the laws of Lévy processes with characteristic triplets (a,σ,ν) and (a',σ',ν'), respectively.

Theorem 5. *A pair of Lévy processes $L_{a,\sigma,\nu}, L_{a',\sigma',\nu'}$ has the GCP if and only if*

(i) $\sigma = \sigma'$,

(ii) if $\sigma = \sigma' = 0$, *then* $a = a'$, *and*

(iii) $H^2(\nu,\nu') < \infty$.

Proof. For one direction, suppose that $L_{a,\sigma,\nu}, L_{a',\sigma',\nu'}$ has the GCP, and let $(\Omega,\mathcal{F},\mathbb{P},X,X')$ witness the germ coupling of $L_{a,\sigma,\nu}, L_{a',\sigma',\nu'}$. By the Lévy-Itô decomposition, we can write $X = X_c + X_j$ where both X_c and X_j are adapted to the natural filtration of X, and similar for X'. Then X, X' being germ coupled under $\mathbb{P}$ implies X_c, X'_c are germ coupled under $\mathbb{P}$ and X_j, X'_j are germ coupled under $\mathbb{P}$. Since X_c, X'_c have laws $W^{a,\sigma}, W^{a',\sigma'}$, we conclude (i) and (ii) from Corollary 1. Since X_j, X'_j have laws $L_\nu, L_{\nu'}$, we conclude (iii) from Theorem 4.

For the other direction, suppose that (i), (ii), and (iii) hold. By Corollary 1 and (i) and (ii), we can construct a probability space $(\Omega_c,\mathcal{F}_c,\mathbb{P}_c,X_c,X'_c)$ witnessing the germ coupling of $W_{a,\sigma}, W_{a',\sigma'}$, and, by Theorem 4 and (iii), we can construct a probability space $(\Omega_j,\mathcal{F}_j,\mathbb{P}_j,X_j,X'_j)$ witnessing the germ coupling of L_ν, L_ν. Now let $\Omega := \Omega_c \times \Omega_j, \mathcal{F} := \mathcal{F}_c \otimes \mathcal{F}_j$, and $\mathbb{P} := \mathbb{P}_c \otimes \mathbb{P}_j$, and let $X(\omega_c,\omega_j) := X_c(\omega_c) + X_j(\omega_j)$ and $X'(\omega_c,\omega_j) := X'_c(\omega_c) + X'_j(\omega_j)$. It follows that $(\Omega,\mathcal{F},\mathbb{P},X,X')$ witnesses the germ coupling of $L_{a,\sigma,\nu}, L_{a',\sigma',\nu'}$, as needed. □

In this last part of the section, we give an application where germ coupling is used in the proof of a statement that is of independent interest. To state it, recall that a càdlàg process $Y = \{Y_t\}_{t\geq 0}$ is called *self-similar with index* $\alpha > 0$ if $\{\varepsilon^{-\alpha} Y_{\varepsilon t}\}_{t\geq 0}$ and $\{Y_t\}_{t\geq 0}$ have the same distribution, for all $\varepsilon > 0$. We say that a stochastic process Y *satisfies Blumenthal's zero-one law* if its law P satisfies $P(A) \in \{0,1\}$ for all $A \in \mathcal{D}_{0+}$. Then we get the following result which generalizes the statement of [59, Lemma 4.3] where Y is a Brownian motion:

Corollary 2. *Let Y be a càdlàg process which is self-similar with index $\alpha > 0$ and which satisfies Blumenthal's zero-one law. If Y' is any càdlàg process whose law is absolutely continuous with respect to the law of Y, then*

$$\left\{\varepsilon^{-\alpha} Y'_{\varepsilon t}\right\}_{t\geq 0} \to Y$$

in distribution in D as $\varepsilon \to 0$.

Proof. Write P, P' for the distributions of Y, Y', respectively, so that $P' \ll P$. Now use absolutely continuity to see that $P(A) \in \{0,1\}$ for all $A \in \mathcal{D}_{0+}$ implies $P(A) = P'(A)$ for all $A \in \mathcal{D}_{0+}$. Thus, Proposition 1 implies that P, P' have the GCP, and we can let $(\Omega, \mathcal{F}, \mathbb{P}, X, X')$ witness the germ coupling of P, P'.

Now fix $t > 0$. Notice that the processes $X^\varepsilon := \{\varepsilon^{-\alpha} X_{\varepsilon s}\}_{s \geq 0}$ and $(X')^\varepsilon := \{\varepsilon^{-\alpha} X'_{\varepsilon s}\}_{s \geq 0}$ satisfy

$$\{X^\varepsilon_s\}_{0 \leq s \leq t} = \{(X')^\varepsilon_s\}_{0 \leq s \leq t}.$$

whenever $\varepsilon t < \tau_{\text{frag}}(X, X')$, so we have

$$\begin{aligned} &\mathbb{P}(\{\{X^\varepsilon_s\}_{0 \leq s \leq t} \in A\} \cap \{\varepsilon t < \tau_{\text{frag}}(X, X')\}) \\ &= \mathbb{P}(\{\{(X')^\varepsilon_s\}_{0 \leq s \leq t} \in A\} \cap \{\varepsilon t < \tau_{\text{frag}}(X, X')\}) \end{aligned} \tag{3.4}$$

for all $A \in \mathcal{D}_t$. Now note that $\mathbb{P}(\varepsilon t < \tau_{\text{frag}}(X, X')) \to 1$ as $\varepsilon \to 0$, and also $\mathbb{P}((X')^\varepsilon \in A) = P'(A)$ for all $\varepsilon > 0$ by self-similarity. Thus, we have

$$\mathbb{P}(\{(X')^\varepsilon \in A\} \cap \{\varepsilon t < \tau_{\text{frag}}(X, X')\}) \to P'(A)$$

as $\varepsilon \to 0$. Now it follows from (3.4) we have

$$\lim_{\varepsilon \to 0} P(\{Y^\varepsilon_s\}_{0 \leq s \leq t} \in A) = \lim_{\varepsilon \to 0} \mathbb{P}(\{X^\varepsilon_s\}_{0 \leq s \leq t} \in A) = P'(\{Y'_s\}_{0 \leq s \leq t} \in A).$$

for all $A \in D_t$.

Now we put the pieces together. Note that if $U \subseteq D$ is open, then it can be written as $U = \bigcup_{t > 0} U_t$ where $U_t := \{x \in D : \{x_s\}_{0 \leq s \leq t} \in V_t\}$ for some $\{V_t\}_{t > 0}$ such that $V_t \subseteq D_t$ is open for each $t > 0$. Also, we can assume $\{U_t\}_{t \geq 0}$ is non-decreasing. Then for $t > 0$ we have

$$\begin{aligned} \liminf_{\varepsilon \to 0} P(Y^\varepsilon \in U) &\geq \liminf_{\varepsilon \to 0} P(Y^\varepsilon \in U_t) \\ &= \liminf_{\varepsilon \to 0} P(\{Y^\varepsilon_s\}_{0 \leq s \leq t} \in V_t) \\ &= P'(\{Y'_s\}_{0 \leq s \leq t} \in V_t) \\ &= P'(Y'_s \in U_t). \end{aligned}$$

As $t > 0$ was arbitrary, we have shown

$$\liminf_{\varepsilon \to 0} P(Y^\varepsilon \in U) \geq P'(Y' \in U),$$

and this finishes the proof. □

3.2 Maximal Germ Couplings

If $P, P' \in \mathcal{P}(D)$ are laws of two stochastic processes and $(\Omega, \mathcal{F}, \mathbb{P}, X, X')$ witnesses their germ coupling, then by weak duality (Lemma 2) applied to the equivalence relation $E_t := \{(x, x') \in D \times D : x_s = x'_s \text{ for all } 0 \leq s < t\}$ for each $t > 0$, we have

$$1 - \mathbb{P}(\tau_{\text{frag}}(X, X') \geq t) \geq \max_{A \in \mathcal{D}_t} |P(A) - P'(A)|.$$

Notice that this looks a bit like a form of stochastic domination, except that neither side needs to be a distribution function. (The right side need not be right-continuous, since $\{\mathcal{D}_t\}_{t\geq 0}$ is not right-continuous.) However, by taking $s \downarrow t$ for each $t \geq 0$, we conclude

$$\mathbb{P}(\tau_{\text{frag}}(X, X') \leq t) \geq F_{P,P'}(t)$$

for all $t \geq 0$, where

$$F_{P,P'}(t) := \lim_{s\downarrow t} \max_{A\in\mathcal{D}_s} |P(A) - P'(A)| = \max_{A\in\mathcal{D}_{t+}} |P(A) - P'(A)|.$$

In other words, every probability measure in the set

$$\{\mathbb{P}(\tau_{\text{frag}}(X, X') \in \cdot\,) : (\Omega, \mathcal{F}, \mathbb{P}, X, X') \text{ witnesses germ coupling of } P, P'\}$$

is stochastically dominated by $F_{P,P'} : [0,\infty] \to [0,1]$. (Note that $\{\tau_{\text{frag}}(X, X') = \infty\} = \{X = X'\}$ can occur with positive probability.) This leads us to the main notion of this section.

Definition 3. *We say that* $(\Omega, \mathcal{F}, \mathbb{P}, X, X')$ witness maximal germ coupling *of* P, P' *if* $(\Omega, \mathcal{F}, \mathbb{P}, X, X')$ *witnesses germ coupling of* P, P' *and* $\mathbb{P}(\tau_{\text{frag}}(X, X') \leq t) = F_{P,P'}(t)$ *for all* $t \geq 0$*. Equivalently, we say that* X, X' *are* maximally germ coupled under $\mathbb{P}$.

A remarkable result of [65] is that, if P, P' have the GCP, then there exists $(\Omega, \mathcal{F}, \mathbb{P}, X, X')$ witnessing their maximal germ coupling. However, the proof in [65] is essentially non-constructive. Thus, an interesting question is to understand for which classes of process P, P' can one construct maximal germ couplings in a probabilistically meaningful way. The goal of this section is to show that this is indeed possible.

First we consider the case of Brownian motions with drift, for which is useful to restrict our path space D to the path space $C := C([0,\infty);\mathbb{R})$ In this case we can provide a completely elementary construction of the desired coupling. The construction is based on the following reflection operation:

Definition 4. *For* $\theta \in \mathbb{R}$*, let* $H^\theta : C \to C$ *denote the map defined via*

$$(H^\theta(x))_t := \begin{cases} x_t, & \text{if } t \leq \sup\{s \geq 0 : x_s = \frac{1}{2}\theta s\}, \\ \theta t - x_t, & \text{else.} \end{cases}$$

Theorem 6. *For any* $\theta \in \mathbb{R}$ *and* $t \geq 0$*, we have*

$$F_{W,W^\theta}(t) = 2\Phi\left(\frac{\sqrt{t}}{2\theta}\right) - 1,$$

where Φ *is the Gaussian distribution function. Moreover, if* $(\Omega, \mathcal{F}, \mathbb{P})$ *is a probability space on which is defined a process* B *with distribution* W*, then,* $(\Omega, \mathcal{F}, \mathbb{P}, B, H^\theta(B))$ *witnesses maximal germ coupling of* W, W^θ*.*

Proof. For some notation, let us write W_θ for the distribution of a Brownian motion started at $\theta \in \mathbb{R}$ and with drift 0 and let us write W^θ for the distribution of a Brownian motion started at 0 and with drift $\theta \in \mathbb{R}$.

For the proof, we first fix $t \geq 0$, and we lower bound $F_{W,W^\theta}(t)$. To do this, recall that we have $\mathcal{D}_t = \mathcal{D}_{t+}$ holding both W- and W^θ-almost surely by Blumenthal's zero-one law. Then we get

$$\begin{aligned} F_{W,W^\theta}(t) &= \sup_{A \in \mathcal{D}_t} |W(A) - W^\theta(A)| \\ &\geq W\left(x_t \leq \frac{\theta}{2}\sqrt{t}\right) - W^\theta\left(x_t \leq \frac{\theta}{2}\sqrt{t}\right) \\ &= \Phi\left(\frac{\theta}{2}\sqrt{t}\right) - \Phi\left(-\frac{\theta}{2}\sqrt{t}\right) \\ &= 2\Phi\left(\frac{\theta}{2}\sqrt{t}\right) - 1 \end{aligned}$$

simply by inspecting the path at its endpoint.

Now we set up the coupling. Let us write $I : C \to C$ for the time-inversion map defined via $(I(x))_t := tx_{1/t}$ for $x \in C$, and note that we have

$$\inf\left\{s \geq 0 : (I(B))_s = \frac{1}{2}\theta\right\} = \frac{1}{\sup\left\{s \geq 0 : B_s = \frac{1}{2}\theta s\right\}}$$

almost surely. Using these two representations, we see that

$$\begin{aligned} (I(H^\theta(B)))_t &= t(H^\theta(B))_{1/t} \\ &= \begin{cases} tB_{1/t}, & \text{if } \frac{1}{t} \leq \sup\{s \geq 0 : B_s = \frac{1}{2}\theta s\}, \\ t(\theta \cdot \frac{1}{t} - B_{1/t}), & \text{else.} \end{cases} \\ &= \begin{cases} (I(B))_t, & \text{if } t \geq \inf\left\{s \geq 0 : (I(B))_s = \frac{1}{2}\theta\right\}, \\ \theta - (I(B))_t, & \text{else.} \end{cases} \end{aligned}$$

By the usual time-inversion symmetry, the process $I(B)$ also has law W. Moreover, by translation symmetry, reflection symmetry, and the strong Markov property, it follows that the process $B_\theta := I(H^\theta(B))$ has the law W_θ. Therefore, since time-inversion interchanges drift and starting position, we see that that $B^\theta := H^\theta(B)$ has law W^θ. This proves the first part of the theorem.

Next, we claim that $\tau_{\text{frag}}(B, B^\theta) = \sup\{s \geq 0 : B_s = \frac{1}{2}\theta s\}$. To do this we define the *coalescence time* of two paths $x, x' \in D$ to be

$$\tau_{\text{coal}}(x, x') := \inf\{t \geq 0 : x_s = x'_s \text{ for all } s \geq t\},$$

which is of course a Borel measurable map. Then observe that we have

$$\begin{aligned}\tau_{\mathrm{coal}}(I(B^0), I(H^\theta(B^0))) &= \inf\left\{s \geq 0 : (I(B^0))_s = \frac{1}{2}\theta\right\} \\ &= \frac{1}{\sup\{s \geq 0 : B^0_s = \frac{1}{2}\theta s\}}\end{aligned}$$

hence

$$\tau_{\mathrm{frag}}(B^0, B^\theta) = \sup\left\{s \geq 0 : B^0_s = \frac{1}{2}\theta s\right\}$$

as claimed.

Finally, we show that this coupling is a maximal germ coupling. To do this, we combine the previous conclusions with the reflection principle to get

$$\begin{aligned}\mathbb{P}(\tau_{\mathrm{frag}}(B, B^\theta) > t) &= \mathbb{P}\left(\sup\left\{s \geq 0 : B_s = \frac{1}{2}\theta s\right\} > t\right) \\ &= \mathbb{P}\left(\inf\left\{s \geq 0 : B_s = \frac{1}{2}\theta\right\} < \frac{1}{t}\right) \\ &= \mathbb{P}\left(\sup_{0 \leq s \leq \frac{1}{t}} B_s > \frac{1}{2}\theta\right) \\ &= \mathbb{P}\left(\left|B_{1/t}\right| > \frac{1}{2}\theta\right) \\ &= \mathbb{P}\left(|B_0| > \frac{1}{2}\theta\sqrt{t}\right) \\ &= 2\Phi\left(\frac{\theta}{2}\sqrt{t}\right) - 1,\end{aligned}$$

for all $t \geq 0$. This shows that B, B^θ are maximally germ coupled under $\mathbb{P}$, as desired. □

Next we consider maximal germ couplings of pure-jump Lévy processes. It turns out that this problem is harder than the study of germ couplings of Lévy processes that we undertook in Theorem 4, so we need a slightly stronger condition: We require that the Lévy measures ν, ν' satisfy $\|\nu - \nu'\|_{\mathrm{TV}} < \infty$ instead of merely $H^2(\nu, \nu') < \infty$. We also have only a partial result which allows us to construct "nearly-maximal" germ couplings, in the sense that our results are only maximal in the small-time limit as $t \to 0$.

To describe our construction, suppose that Lévy measures ν, ν' satisfy $\|\nu - \nu'\|_{\mathrm{TV}} < \infty$. Then $\nu - (\nu \wedge \nu'), \nu' - (\nu \wedge \nu')$ are themselves Lévy measures, and they both have finite total activity, since

$$|\|\nu - (\nu \wedge \nu')\|_{\mathrm{TV}} - \|\nu' - (\nu \wedge \nu')\|_{\mathrm{TV}}| \leq \|\nu - \nu'\|_{\mathrm{TV}} < \infty.$$

Now we construct a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ supporting the following independent processes: First, Y is a compound Poisson process with Lévy measure $\nu - (\nu \wedge \nu')$. Second, Y'

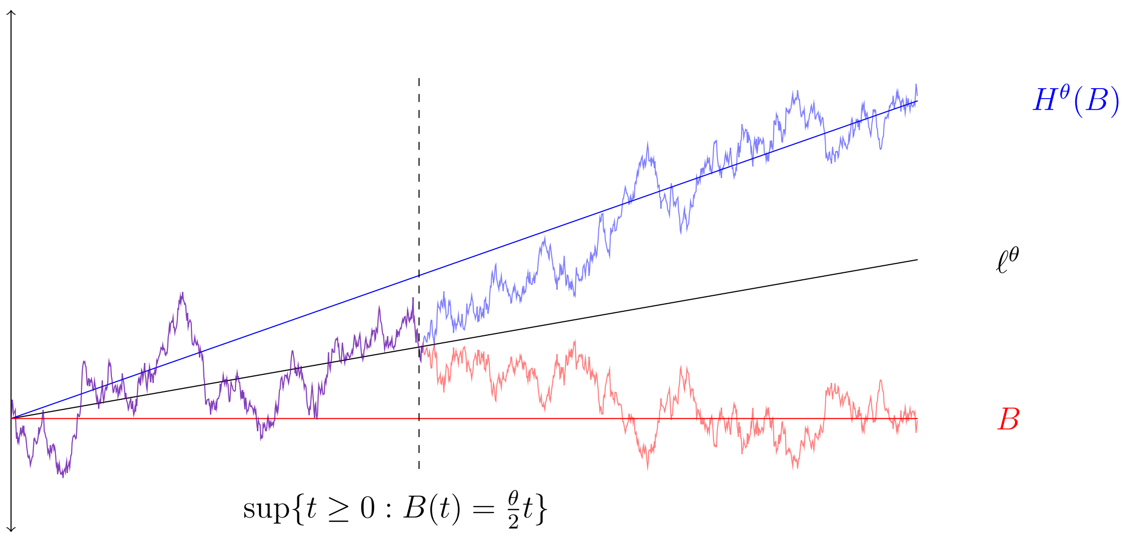

Figure 3.1: A maximal germ coupling of a Brownian motion with drift 0 and a Brownian motion with drift $\theta > 0$. First, we let B denote a Brownian motion with drift 0. Then, we let $H^\theta(B)$ be equal to the reflection of B across the line $\ell^\theta := \{(t, \frac{1}{2}\theta t) : t \geq 0\}$ after the time of its last intersection with ℓ^θ, and equal to B before this time (Definition 4).

is a compound Poisson process with Lévy measure $\nu' - (\nu \wedge \nu')$. Third, Z is a Lévy process with Lévy measure $\nu \wedge \nu'$. Finally, set $X = Y + Z$ and $X' = Y' + Z$.

Theorem 7. *If ν, ν' are Lévy measures with $\|\nu - \nu'\|_{\mathrm{TV}} < \infty$, then*

$$F_{L_\nu, L_{\nu'}}(t) \sim t\|\nu - \nu'\|_{\mathrm{TV}}$$

as $t \to 0$. Moreover, $(\Omega, \mathcal{F}, \mathbb{P}, X, X')$ described above witnesses nearly-maximal germ coupling of $L_\nu, L_{\nu'}$ in that

$$\mathbb{P}(\tau_{\mathrm{frag}}(X, X') \leq t) \sim t\|\nu - \nu'\|_{\mathrm{TV}}$$

as $t \to 0$.

Proof. First we provide a lower bound on $F_{L_\nu, L_{\nu'}}(t)$ for fixed $t \geq 0$. As above, we have $\mathcal{D}_t = \mathcal{D}_{t+}$ holding both L_ν- and $L_{\nu'}$-almost surely, by Blumenthal's zero-one law. Now by the definition of total variation, we can get a set $B \in \mathcal{B}(\mathbb{R})$ with $\nu(B) - \nu'(B) = \|\nu - \nu'\|_{\mathrm{TV}}$.

Then we bound:

$$\begin{aligned}
F_{L_\nu,L_{\nu'}}(t) &= \sup_{A\in\mathcal{D}_t} |L_\nu(A) - L_{\nu'}(A)| \\
&\geq |L_\nu(\text{a jump of size } B \text{ occurs in } [0,t)) \\
&\qquad\qquad - L_{\nu'}(\text{a jump of size } B \text{ occurs in } [0,t))| \\
&= |1 - e^{-t\nu(B)} - (1 - e^{-t\nu'(B)})| \\
&= e^{-t\nu(B)} - e^{-t\nu'(B)} \\
&= e^{-t\nu(B)}(1 - e^{-t\|\nu-\nu'\|_{\mathrm{TV}}}).
\end{aligned}$$

Dividing by $t\|\nu - \nu'\|_{\mathrm{TV}}$ and taking $t \to 0$, we have

$$\liminf_{t\to 0} \frac{F_{L_\nu,L_{\nu'}}(t)}{t\|\nu - \nu'\|_{\mathrm{TV}}} \geq \liminf_{t\to 0} e^{-t\nu(B)} \frac{1 - e^{-t\|\nu-\nu'\|_{\mathrm{TV}}}}{t\|\nu - \nu'\|_{\mathrm{TV}}} = 1$$

as claimed.

Now we provide the matching upper bound. Importantly, we have $\tau_{\text{frag}}(X, X') = \tau_{\text{frag}}(Y, Y')$ almost surely by construction. Also, the jump distributions of Y and Y' are mutually singular, so $\tau_{\text{frag}}(Y, Y')$ is exactly equal to the first time of a jump of either of Y or Y'. By construction, the first time either of Y or Y' jumps is an exponential random variable with rate $\|\nu - (\nu \wedge \nu')\|_{\mathrm{TV}} + \|\nu' - (\nu \wedge \nu')\|_{\mathrm{TV}} = \|\nu - \nu'\|_{\mathrm{TV}}$. Therefore, we have

$$\limsup_{t\to 0} \frac{\mathbb{P}(\tau_{\text{frag}}(X, X') \leq t)}{t\|\nu - \nu'\|_{\mathrm{TV}}} \leq \limsup_{t\to 0} \frac{1 - e^{-t\|\nu-\nu'\|_{\mathrm{TV}}}}{t\|\nu - \nu'\|_{\mathrm{TV}}} = 1.$$

Combining this with the first inequality finishes the proof. □

An interesting question in light of the results of this chapter is whether two processes $P, P' \in \mathcal{P}(D)$ admit a *Markovian* germ coupling, that is, a germ coupling which is a Markov process with respect to its natural joint filtration. In the case of Brownian motions with drift, our construction of a maximal germ coupling is certainly not Markovian (due to time-inversion), and it is in fact known [65] that two diffusions can never admit any Markovian germ coupling. In the case of Lévy processes, we saw that there is actually Markovian nearly-maximal germ coupling, but it is not clear whether this will be possible in the case of $H^2(\nu, \nu') < \infty$ and $\|\nu - \nu'\|_{\mathrm{TV}} = \infty$. Because of these observations, we conjecture that the condition

$$\limsup_{t\to 0} \frac{F_{P,P'}(t)}{t} < \infty.$$

may be equivalent to existence of Markovian germ coupling of P, P'.

Part II

Centering

Chapter 4

Fréchet Means in Infinite Dimensions

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space supporting an IID sequence $Y_1, Y_2, \ldots$ of random variables taking values in a metric space $(\mathcal{X}, d)$. Recall that by their empirical Fréchet mean we mean any solution to the optimization problem

$$\begin{cases} \text{minimize} & \frac{1}{n}\sum_{i=1}^{n} d^2(x, Y_i) \\ \text{over} & x \in \mathcal{X}. \end{cases} \tag{4.1}$$

As Fréchet means purport to be a sort of average, a natural question is whether they satisfy generalizations of the limit theorems that hold for for averages of real-valued random variables. Results of this form are indeed known in many cases [213, 186, 180, 94, 31, 32], and they form the basis for nearly all statistical procedures used in non-Euclidean statistics.

The most well-developed setting for probabilistic results about Fréchet means is the case where $\mathcal{X}$ is a Riemannian manifold and d is the metric induced by its Riemannian metric tensor [31, 32]; indeed, in this case, one has a law of large numbers, a central limit theorem, and more. The differentiable structure on $(\mathcal{X}, d)$ is extremely valuable here, since one can easily develop first-order conditions on (4.1) which give a priori information about Fréchet means. However, many authors have moved away from the Riemannian manifold setting, since several important application areas necessitate the analysis of Fréchet means in metric spaces which have singularities.

Yet, almost all results still rely on a "finite-dimensionality" assumption that simplifies the analysis greatly. Here, the "finite-dimensionality" of $(\mathcal{X}, d)$ usually means that $(\mathcal{X}, d)$ is a Heine-Borel space in the sense that the closed balls $\bar{B}_r(x) := \{y \in \mathcal{X} : d(x, y) \leq r\}$ are compact for all $x \in \mathcal{X}$ and $r \geq 0$. This compactness is useful for showing that Fréchet means exist and for proving, for example, the law of large numbers.

Interestingly, it is implicit in some examples that such finite-dimensionality conditions are not at all necessary. For one example, it is straightforward to show that empirical averages of IID random variables in an infinite-dimensional Hilbert space converge strongly to their Bochner expectation. For another example, it has been shown [128, 130, 48] that the empirical Fréchet means of IID random variables in the Wasserstein space $\mathcal{P}_2(\mathbb{R}^m)$ converge in the Wasserstein metric W_2 to their population Fréchet mean.

The primary goal of this chapter is to understand what geometric condition on an abstract metric space $(\mathcal{X}, d)$ ensures that the strong law of large numbers (SLLN) holds for Fréchet means. We know that Heine-Borel spaces, Hilbert spaces, and the Wasserstein space must all be included.

A secondary goal of this chapter is to understand what moment conditions are required for the SLLN to hold. Most results for Fréchet means require finite 2nd moment, but again this cannot be necessary: For $\mathcal{X} = \mathbb{R}$ with its usual metric, we know from the Khinchine SLLN that finite 1st moment is sufficient.

This chapter is based on the work [98] which is an extension of earlier work [66]. We answer the motivating questions by introducing the notion of *weak-like topology* (Definition 8) and proving a fundamental continuity result (Theorem 8). This yields the SLLN under the minimal moment condition (Theorem 9) and has other probabilistic consequences of interest. The results herein will also be used in Chapters 5, 6, and 7 where we answer some finer questions about Fréchet means.

4.1 Preliminaries

To begin, let $(\mathcal{X}, d)$ denote a metric space. Unless otherwise stated, all metric and topological notions refer to the topology generated by the metric d. We write $B_r^\circ(x) := \{y \in \mathcal{X} : d(x, y) < r\}$ and $\bar{B}_r(x) := \{y \in \mathcal{X} : d(x, y) < r\}$ for the open and closed balls of radius $r > 0$ around $x \in \mathcal{X}$. (Note, however that $\bar{B}_r(x)$ need not be the closure of $B_r^\circ(x)$ and that $B_r^\circ(x)$ need not be the interior of $\bar{B}_r(x)$.)

We write $\mathcal{P}(\mathcal{X})$ for the set of Borel probability measures on $\mathcal{X}$, where the Borel σ-algebra is generated by the topology generated by d. We write τ_{w} for the topology on $\mathcal{P}(\mathcal{X})$ such that $\{P_n\}_{n\in\mathbb{N}}$ and P in $\mathcal{P}(\mathcal{X})$ have $P_n \to P$ in τ_{w} if and only $\int_{\mathcal{X}} f \, \mathrm{d}P_n \to \int_{\mathcal{X}} f \, \mathrm{d}P$ for all bounded, continuous $f : (\mathcal{X}, d) \to \mathbb{R}$, called the *weak topology* on $\mathcal{P}(\mathcal{X})$.

Now we introduce two fundamental inequalities. The first is that for all $r \geq 0$ the constant $c_r := \max\{1, 2^{r-1}\}$ satisfies

$$d^r(x, x'') \leq c_r(d^r(x, x') + d^r(x', x'')) \tag{4.2}$$

for all $x, x', x'' \in \mathcal{X}$. Indeed, this can be easily proven by combining the triangle inequality for d with either the subadditivity or the convexity of the map $t \mapsto t^r$, corresponding to $0 \leq r \leq 1$ and $r \geq 1$, respectively. The second is that for all $p \geq 1$ we have

$$|d^p(x, x'') - d^p(x', x'')| \leq p d(x, x')(d^{p-1}(x, x'') + d^{p-1}(x', x'')) \tag{4.3}$$

all $x, x', x'' \in \mathcal{X}$. Indeed, this can be easily proven by combining the triangle inequality for d with the elementary inequality $|a^p - b^p| \leq p|a - b|(a^{p-1} + b^{p-1})$ for all $a, b \geq 0$.

For any $r > 0$, we let $\mathcal{P}_r(\mathcal{X})$ denote the set of all $P \in \mathcal{P}(\mathcal{X})$ satisfying $\int_{\mathcal{X}} d^r(x, y) \, \mathrm{d}P(y) < \infty$ for some $x \in \mathcal{X}$; by (4.2), this is equivalent to $P \in \mathcal{P}(\mathcal{X})$ satisfying $\int_{\mathcal{X}} d^r(x, y) \, \mathrm{d}P(y) < \infty$ for all $x \in \mathcal{X}$. In particular, we regard $\mathcal{P}_r(\mathcal{X})$ as the space of all distributions of an $\mathcal{X}$-valued random variable which has finite rth moment. We adopt the convention $0^0 = 1$ in this chapter, so that in particular we can consistently define $\mathcal{P}_0(\mathcal{X}) = \mathcal{P}(\mathcal{X})$.

Now fix $p \geq 1$. We define the function $W_p : \mathcal{P}_{p-1}(\mathcal{X}) \times \mathcal{X}^2 \to \mathbb{R}$ via

$$W_p(P, x, x') := \int_{\mathcal{X}} (d^p(x, y) - d^p(x', y))\, \mathrm{d}P(y),$$

which makes sense because of (4.3). By a slight abuse of notation, we use the same symbol W_p to represent the function $W_p : \mathcal{P}_p(\mathcal{X}) \times \mathcal{X} \to \mathbb{R}$ defined via

$$W_p(P, x) := \int_{\mathcal{X}} d^p(x, y)\, \mathrm{d}P(y).$$

Note that these functions satisfy the some simple identities. For $P \in \mathcal{P}_{p-1}(\mathcal{X})$ and $x, x', x'' \in \mathcal{X}$, we have $W_p(P, x, x') = -W_p(P, x', x)$ and $W_p(P, x, x'') = W_p(P, x, x') + W_p(P, x', x'')$. For $P \in \mathcal{P}_p(\mathcal{X})$ and $x, x' \in \mathcal{X}$, we additionally have $W_p(P, x, x') = W_p(P, x) - W_p(P, x')$, but note that this does not make sense when $P \in \mathcal{P}_{p-1}(\mathcal{X}) \setminus \mathcal{P}_p(\mathcal{X})$. We refer to W_p (in both of its forms above) as the *Fréchet functional*.

From these considerations we can define the main object of interest.

Definition 5. *For any metric space $(\mathcal{X}, d)$, any $p \geq 1$, and any $P \in \mathcal{P}_{p-1}(\mathcal{X})$, we let*

$$M_p(P) := \{x \in \mathcal{X} : W_p(P, x, x') \leq 0 \text{ for all } x' \in \mathcal{X}\},$$

called the Fréchet p-mean set of P.

This definition may seem a bit odd at first, since it looks quite different than the definition of Fréchet means given in the introduction where we focused on $p = 2$. The connection is made clear by the following alternative characterizations:

Lemma 5. *For any metric space $(\mathcal{X}, d)$, any $p \geq 1$, and any $P \in \mathcal{P}_{p-1}(\mathcal{X})$, the Fréchet p-mean set is the solution to the optimization problem*

$$\begin{cases} \text{minimize} & W_p(P, x, o) \\ \text{over} & x \in \mathcal{X}, \end{cases} \tag{4.4}$$

where $o \in \mathcal{X}$ is arbitrary. For any metric space $(\mathcal{X}, d)$, any $p \geq 1$, and any $P \in \mathcal{P}_p(\mathcal{X})$, the Fréchet p-mean set is the solution to the optimization problem

$$\begin{cases} \text{minimize} & W_p(P, x) \\ \text{over} & x \in \mathcal{X}. \end{cases} \tag{4.5}$$

Proof. First let us show that $M_p(P)$ is equal to the solution set of (4.4) for arbitrary $o \in \mathcal{X}$. For one direction, take $x \in M_p(P)$, and note that for any $x' \in \mathcal{X}$ we have $W_p(P, x, o) = W_p(P, x, x') + W_p(P, x', o) \leq W_p(P, x', o)$. For the other direction, suppose that $x \in \mathcal{X}$ is a solution to (4.4), and note that for any $x' \in \mathcal{X}$ we have $W_p(P, x, x') = W_p(P, x, o) - W_p(P, x', o) \leq 0$. To see that the solution set of (4.5) is equal to the solution set of (4.4) when $P \in \mathcal{P}_p(\mathcal{X})$, we simply note $W_p(P, x, o) = W_p(P, x) - W_p(P, o)$ and that the second term does not depend on $x \in \mathcal{X}$. □

It is also useful to give a name to the following quantities:

Definition 6. *For any metric space $(\mathcal{X}, d)$, any $p \geq 1$, any $P \in \mathcal{P}_p(\mathcal{X})$, we set*

$$V_p(P) := \inf_{x \in \mathcal{X}} W_p(P, x),$$

called the Fréchet p-variance of P*, and for any $P \in \mathcal{P}_p(\mathcal{X})$ and $o \in \mathcal{X}$ we set*

$$V_p(P, o) := \inf_{x \in \mathcal{X}} W_p(P, x, o),$$

called the surrogate Fréchet p-variance of P*.*

The connection between these notions is of course that the surrogate Fréchet p-variance $V_p(P, o)$ is the minimum value of the Fréchet functional $W_p(P, \cdot, o)$ and that the Fréchet p-mean $M_p(P)$ is its level set. We emphasize that $V_p(P, o)$ depends on $o \in \mathcal{X}$, while $M_p(P)$ does not.

Next we give some continuity results. To do this, we define some finer topologies on $\mathcal{P}(\mathcal{X})$. For $r > 0$, we write τ_{w}^r for the topology on $\mathcal{P}_r(\mathcal{X})$ such that $\{P_n\}_{n \in \mathbb{N}}$ and P in $\mathcal{P}(\mathcal{X})$ have $P_n \to P$ in τ_{w}^r if and only $P_n \to P$ in τ_{w} and $W_r(P_n, x) \to W_r(P, x)$ for all $x \in \mathcal{X}$. Following the usual convention we also set $\tau_{\mathrm{w}}^0 = \tau_{\mathrm{w}}$.

Our continuity results will use a common set of tricks. First, we use Skorokhod's representation theorem [104, Theorem 4.30] to represent convergence in τ_{w} as almost sure convergence on a suitable probability space. Second, we use the fact [104, Lemma 5.11] that random variables converging almost surely have converging expectations if and only if they are uniformly integrable. Lastly, we use the inequalities (4.2) and (4.3) to provide some dominations which allow us to transfer uniform integrability across different points of $(\mathcal{X}, d)$.

Lemma 6. *Suppose $(\mathcal{X}, d)$ is a separable metric space and $p \geq 1$. Then, the function $W_p : \mathcal{P}_{p-1}(\mathcal{X}) \times \mathcal{X}^2 \to \mathbb{R}$ is continuous.*

Proof. Suppose that $\{(P_n, x_n, x_n')\}_{n \in \mathbb{N}}$ and (P, x, x') in $\mathcal{P}_{p-1}(\mathcal{X}) \times \mathcal{X}^2$ satisfy $(P_n, x_n, x_n') \to (P, x, x')$ in $\tau_{\mathrm{w}}^{p-1} \times d^2$. Then, use Skorokhod's representation theorem to construct a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with expectation $\mathbb{E}$, supporting random variables $\{Y_n\}_{n \in \mathbb{N}}$ and Y with distributions $\{P_n\}_{n \in \mathbb{N}}$ and P, respectively, such that we have $Y_n \to Y$ almost surely. Note that we have

$$W_p(P_n, x_n, x_n') = \mathbb{E}[d^p(x_n, Y_n) - d^p(x_n', Y_n)]$$

and

$$W_p(P, x, x') = \mathbb{E}[d^p(x, Y) - d^p(x', Y)]$$

and also that

$$d^p(x_n, Y_n) - d^p(x'_n, Y_n) \to d^p(x, Y) - d^p(x', Y)$$

almost surely. Thus, the result is proved if we show that $\{d^p(x_n, Y_n) - d^p(x'_n, Y_n)\}_{n\in\mathbb{N}}$ is uniformly integrable. To do this, use (4.3) then (4.2) to bound

$$\begin{aligned}|d^p(x_n, Y_n) - d^p(x'_n, Y_n)| &\le pd(x_n, x'_n)(d_{p-1}(x_n, Y_n) + d^{p-1}(x'_n, Y_n)) \\ &\le pc_{p-1}d(x_n, x'_n)(d^{p-1}(x_n, x) + d^{p-1}(x'_n, x') \\ &\qquad + d^{p-1}(x, Y_n) + d^{p-1}(x', Y_n)).\end{aligned}$$

Observe that $P_n \to P$ in τ_{w}^{p-1} implies that $\{d^{p-1}(x, Y_n)\}_{n\in\mathbb{N}}$ and $\{d^{p-1}(x', Y_n)\}_{n\in\mathbb{N}}$ are uniformly integrable, so the result is proved. □

Lemma 7. *Suppose $(\mathcal{X}, d)$ is a separable metric space, that $r \ge 0$, and that $\{P_n\}_{n\in\mathbb{N}}$ and P in $\mathcal{P}_r(\mathcal{X})$ satisfy $P_n \to P$ in τ_{w}. Then, the following are equivalent:*

(i) $W_r(P_n, x) \to W_r(P, x)$ for some $x \in \mathcal{X}$.

(ii) $W_r(P_n, x) \to W_r(P, x)$ for all $x \in \mathcal{X}$.

Proof. It suffices to show that (i) implies (ii), so suppose that $W_r(P_n, x) \to W_r(P, x)$ holds for $x \in \mathcal{X}$, and let $x' \in \mathcal{X}$ be arbitrary. Use Skorokhod's representation theorem to construct a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with expectation $\mathbb{E}$, supporting random variables $\{Y_n\}_{n\in\mathbb{N}}$ and Y with distributions $\{P_n\}_{n\in\mathbb{N}}$ and P, respectively, such that we have $Y_n \to Y$ almost surely. This implies $d^r(Y_n, x') \to d^r(Y, x')$ almost surely, so it suffices to show that $\{d^r(Y_n, x')\}_{n\in\mathbb{N}}$ is uniformly integrable. To do this, simply note

$$d^r(Y_n, x') \le c_r(d^r(Y_n, x) + d^r(x, x'))$$

almost surely, by (4.2). Since $\mathbb{E}[d^r(Y_n, x)] = W_r(P_n, x) \to W_r(P, x) = \mathbb{E}[d^r(Y, x)]$ implies that $\{d^r(Y_n, x)\}_{n\in\mathbb{N}}$ is uniformly integrable, the proof is complete. □

Lemma 8. *If $(\mathcal{X}, d)$ is a metric space and $\{P_n\}_{n\in\mathbb{N}}$ and P in $\mathcal{P}(\mathcal{X})$ have $P_n \to P$ in τ_{w}, then we have*

$$\lim_{M\to\infty} \limsup_{n\to\infty} P_n(\mathcal{X} \setminus B_M(o)) \to 0$$

for all $o \in \mathcal{X}$.

Proof. Fix $o \in \mathcal{X}$, and, for $M > 0$, define

$$f_M(x) := \begin{cases} 0, & \text{if } d(x, o) \le M - 1, \\ 1 - (M - d(x, o)), & \text{if } M - 1 \le d(x, o) \le M, \\ 1, & \text{if } d(x, o) \ge M. \end{cases}$$

Observe that each f_M is a continuous function satisfying $0 \leq f_M \leq 1$ everywhere, as well as $f_M \equiv 0$ on $\bar{B}_{M-1}(o)$ and $f_M \equiv 1$ on $\mathcal{X} \setminus B_M^\circ(o)$. (Alternatively, one can use Urysohn's lemma to abtractly guarantee the existence of such a function.) Now the definition of τ_{w} gives:

$$\begin{aligned}
\limsup_{n\to\infty} P_n(\mathcal{X} \setminus B_M^\circ(o)) &= \limsup_{n\to\infty} \int_{\mathcal{X}\setminus B_M^\circ(o)} 1 \,\mathrm{d}P_n \\
&\leq \limsup_{n\to\infty} \int_{\mathcal{X}} f_M \,\mathrm{d}P_n \\
&= \lim_{n\to\infty} \int_{\mathcal{X}} f_M \,\mathrm{d}P_n \\
&= \int_{\mathcal{X}} f_M \,\mathrm{d}P \\
&\leq \int_{\mathcal{X}\setminus B_{M-1}^\circ(o)} 1 \,\mathrm{d}P \\
&= P(\mathcal{X} \setminus B_{M-1}^\circ(o)).
\end{aligned}$$

Lastly, we use $B_{M-1}^\circ(o) \uparrow \mathcal{X}$ as $M \to \infty$ and the downward continuity of P to conclude the desired result. □

Lemma 9. *If $(\mathcal{X}, d)$ is a separable metric space and $r \geq 0$, and if $\{P_n\}_{n\in\mathbb{N}}$ and P in $\mathcal{P}_r(\mathcal{X})$ have $P_n \to P$ in τ_{w}^r, then we have*

$$\lim_{M\to\infty} \limsup_{n\to\infty} \int_{\mathcal{X}\setminus B_M^\circ(o)} d^r(o,y) \,\mathrm{d}P_n(y) = 0$$

for all $o \in \mathcal{X}$.

Proof. As in the preceding proof, fix $o \in \mathcal{X}$, and define the function $f_M : \mathcal{X} \to \mathbb{R}$ for each $M > 0$. Now use Skohorkhod's representation theorem to construct a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which are defined $\mathcal{X}$-valued random variables $\{Y_n\}_{n\in\mathbb{N}}$ and Y with distributions $\{P_n\}_{n\in\mathbb{N}}$ and P, respectively, such that we have $Y_n \to Y$ holding almost surely; we let $\mathbb{E}$ denote the expectation operator on $(\Omega, \mathcal{F}, \mathbb{P})$. By construction and the assumption that $P_n \to P$ in τ_{w}^r we have that $\{d^r(o, Y_n)\}_{n\in\mathbb{N}}$ is uniformly integrable. Now observe that $d^r(o, Y_n) f_M(Y_n) \leq d^r(o, Y_n)$ almost surely for all $n \in \mathbb{N}$, so $\{d^r(o, Y_n) f_M(Y_n)\}_{n\in\mathbb{N}}$ is uniformly integrable. Since $d^r(o, Y_n) f_M(Y_n) \to d^r(o, Y) f_M(Y)$ holds almost surely as $n \to \infty$, we conclude $\mathbb{E}[d^r(o, Y_n) f_M(Y_n)] \to \mathbb{E}[d^r(o, Y) f_M(Y)]$. Putting all the pieces together, we have:

$$\begin{aligned}\limsup_{n\to\infty}\int_{\mathcal{X}\setminus B_M^\circ(o)} d^r(o,y)\,\mathrm{d}P_n(y) &\le \limsup_{n\to\infty}\int_{\mathcal{X}} d^r(o,y)f_M(y)\,\mathrm{d}P_n \\ &= \lim_{n\to\infty}\mathbb{E}[d^r(o,Y_n)f_M(Y_n)] \\ &= \mathbb{E}[d^r(o,Y)f_M(Y)] \\ &= \int_{\mathcal{X}} d^r(o,y)f_M(y)\,\mathrm{d}P(y) \\ &\le \int_{\mathcal{X}\setminus B_{M-1}^\circ(o)} d^r(o,y)\,\mathrm{d}P(y).\end{aligned}$$

Lastly, we use $P\in\mathcal{P}_r(\mathcal{X})$ and the dominated convergence theorem to get

$$\lim_{n\to\infty}\int_{X\setminus B_{M-1}^\circ(o)} d^r(o,y)\,\mathrm{d}P(y)=0,$$

and this proves the result. □

Combining these preliminaries, we prove the following which shows that the Fréchet mean sets are uniformly bounded.

Proposition 2. *Consider any separable metric space $(\mathcal{X},d)$ and any $p\ge 1$. If $\{P_n\}_{n\in\mathbb{N}}$ and P in $\mathcal{P}_{p-1}(\mathcal{X})$ have $P_n\to P$ in τ_{w}^{p-1}, then there exists a d-bounded set $B\subseteq\mathcal{X}$ satisfying $M_p(P_n)\subseteq B$ for all $n\in\mathbb{N}$.*

Proof. Fix an arbitrary $o\in\mathcal{X}$, and let us derive some initial estimates. First note by Lemma 6 that we have

$$\begin{aligned}\limsup_{n\to\infty} V_p(P_n,o) &\le \inf_{x\in\mathcal{X}}\limsup_{n\to\infty} W_p(P_n,x,o) \\ &= \inf_{x\in\mathcal{X}} W_p(P,x,o) = V_p(P,o) < \infty.\end{aligned}$$

Second, combine this with Lemma 8 and Lemma 9 to choose M sufficiently large so that we have

$$M > \max\left\{\left(\sup_{n\in\mathbb{N}} V_p(P_n,o)\right)^{1/p}, \frac{1}{16}\right\} \tag{4.6}$$

as well as

$$\int_{\mathcal{X}\setminus B_M^\circ(o)} d^{p-1}(o,y)dP_n(y) \le \frac{1}{p2^{2p+2}} \tag{4.7}$$

and

$$P_n(\mathcal{X}\setminus B_M^\circ(o)) \le \frac{1}{p2^{2p}} \le \frac{1}{2} \tag{4.8}$$

for all $n \in \mathbb{N}$.

Now we derive some lower bounds on the integrand in the Fréchet functional. First, use the elementary inequality (4.2) to get

$$\begin{aligned} d^p(x,y) - d^p(o,y) &\geq c_p^{-1} d^p(x,o) - 2d^p(o,y) \\ &\geq \frac{d^p(x,o)}{2^{p-1}} - 2d^p(o,y) \end{aligned} \tag{4.9}$$

for all $x, y \in \mathcal{X}$ and $n \in \mathbb{N}$. Second, use (4.3) then (4.2) to get

$$\begin{aligned} &|d^p(x,y) - d^p(o,y)| \\ &\leq pd(x,o)(d^{p-1}(x,y) + d^{p-1}(o,y)) \\ &\leq pd(x,o)(c_{p-1}(d^{p-1}(x,o) + d^{p-1}(o,y)) + d^{p-1}(o,y)) \\ &= pc_{p-1}d^p(x,o) + p(1 + c_{p-1})d(x,o)d^{p-1}(o,y) \\ &\leq p2^{p-1}d^p(x,o) + p2^p d(x,o)d^{p-1}(o,y) \end{aligned}$$

for all $x, y \in \mathcal{X}$. In particular, this implies

$$d^p(x,y) - d^p(o,y) \geq -p2^{p-1}d^p(x,o) - p2^p d(x,o)d^{p-1}(o,y) \tag{4.10}$$

for all $x, y, \in \mathcal{X}$.

Now we put these pieces together and claim that, for any $\{x_n\}_{n\in\mathbb{N}}$ with $x_n \in M_p(P_n)$ for all $n \in \mathbb{N}$, we have $\sup_{n\in\mathbb{N}} d(x_n, o) < 17M$. Indeed, assume towards a contradiction that there exists some $N \in \mathbb{N}$ with $d(x_N, o) > 17M$. On the one hand, this means we can use (4.9) and (4.8) to get

$$\begin{aligned} &\int_{B_M^\circ(o)} (d^p(x_N, y) - d^p(o,y))\,\mathrm{d}P_N(y) \\ &\qquad \geq \left(\frac{d^p(x_N,o)}{2^{p-1}} - 2M^p\right) P_N(B_M(o)) \\ &\qquad \geq \left(\frac{d^p(x_N,o)}{2^{p-1}} - 2M^p\right) \frac{1}{2} \\ &\qquad = \frac{d^p(x_N,o)}{2^p} - M^p, \end{aligned}$$

where in the first inequality we used $d(x_N, o) \geq 16M \geq 2M$ to see that the integrand is non-negative. On the other hand, we can use (4.10) and (4.7) to get

$$\begin{aligned}\int_{\mathcal{X}\setminus B_M^\circ(o)} &(d^p(x_N,y)-d^p(o,y))\,\mathrm{d}P_N(y)\\ &\geq -p2^{p-1}d^p(x_N,o)P_N(\mathcal{X}\setminus B_M(o))\\ &\qquad - p2^p d(x_N,o)\int_{\mathcal{X}\setminus B_M(o)} d^{p-1}(o,y)\,\mathrm{d}P_N(y)\\ &\geq -\frac{d^p(x_N,o)}{2^{p+1}}-\frac{d(x_N,o)}{2^{p+2}}\\ &\geq -\frac{d^p(x_N,o)}{2^{p+1}}-\frac{d^p(x_N,o)}{2^{p+2}},\end{aligned}$$

where in the final inequality we used $M \geq 1/16$ to see that $d(x_N,o) \geq 16M$ implies $d(x_N,o) \leq d^p(x_N,o)$. Combining the previous two displays, we have shown

$$\begin{aligned}W_p(P_N,x_N,o) &= \int_{B_M^\circ(o)} (d^p(x_N,y)-d^p(o,y))\,dP_N(y)\\ &\qquad + \int_{\mathcal{X}\setminus B_M^\circ(o)} (d^p(x_N,y)-d^p(o,y))\,dP_N(y)\\ &\geq \frac{d^p(x_N,o)}{2^p} - M^p - \frac{d^p(x_N,o)}{2^{p+1}} - \frac{d^p(x_N,o)}{2^{p+2}}\\ &= \frac{d^p(x_N,o)}{2^{p+2}} - M^p\\ &= \left(\frac{16^p}{2^{p+2}}-1\right)M^p\\ &= (2^{3p-2}-1)M^p \geq M^p > \sup_{n\in\mathbb{N}} V_p(P,o),\end{aligned}$$

which contradicts the assumption that $x_N \in M_p(P_N)$. Therefore, we must have $\sup_{n\in\mathbb{N}} d(x_n,o) < 17M$, so the result follows by taking $B := \bar{B}_{18M}(o)$. □

Corollary 3. *For any separable metric space $(\mathcal{X},d)$, any $p \geq 1$, and any $P \in \mathcal{P}_{p-1}(X)$, the set $M_p(P) \subseteq \mathcal{X}$ is closed and bounded.*

Proof. By definition we have

$$M_p(P) = \bigcap_{x'\in\mathcal{X}} \{x \in \mathcal{X} : W_p(P,x,x') \leq 0\}.$$

From Lemma 6 we see that $\{x \in X : W_p(P,x,x') \leq 0\}$ is closed for each $x' \in \mathcal{X}$, so it follows that $M_p(P)$, being an intersection of closed sets, is closed. Also, from Proposition 2 we see that $M_p(P)$ is bounded. □

In many existing works [66, 180, 94, 31, 32], the next step is to assume that $(\mathcal{X}, d)$ has the Heine-Borel property so that one can deduce that the Fréchet mean sets are compact. As we will see in the next section, this is not necessary. Instead, we provide a much weaker condition under which the Fréchet mean sets are compact, and under which we can provide our limit theorems of interest.

4.2 Weak-Like Topologies

In the previous section, the metric space $(\mathcal{X}, d)$ was always endowed with the topology generated by the metric d. In this section we will also consider $\mathcal{X}$ to be endowed with a topology weaker than the one generated by d; this serves as an analog of weak topologies that are commonly encountered in functional analysis. More precisely, we introduce the following notion:

Definition 7. *For a metric space $(\mathcal{X}, d)$, a Hausdorff topology τ on $\mathcal{X}$ is called a* weak-like topology for $(\mathcal{X}, d)$ *if it satisfies the following conditions:*

(W1) If $\{x_n\}_{n\in\mathbb{N}}$ and y in $\mathcal{X}$ satisfy $\sup_{n\in\mathbb{N}} d(x_n, y) < \infty$, then there exists a subsequence $\{n_k\}_{k\in\mathbb{N}}$ and a point $x \in \mathcal{X}$ satisfying $x_{n_k} \to x$ in τ.

(W2) If $\{x_n\}_{n\in\mathbb{N}}$ and x in $\mathcal{X}$ satisfy $x_n \to x$ in τ, then for all $y \in \mathcal{X}$ we have $d(x, y) \leq \liminf_{n\in\mathbb{N}} d(x_n, y)$.

(W3) If $\{x_n\}_{n\in\mathbb{N}}$ and x in X satisfy $x_n \to x$ in τ and satisfy $d(x_n, y) \to d(x, y)$ for some $y \in \mathcal{X}$, then we have $x_n \to x$ in d.

Note that τ is neither required to be metrizable nor second-countable. We say that $(\mathcal{X}, d)$ admits a weak-like topology if there exists a weak-like topology τ for $(\mathcal{X}, d)$.

First, we provide some interpretation for the three conditions set forth above. The first condition (W1) is to be read as "d-bounded sets are relatively (sequentially) τ-compact". The second condition (W2) is to be read as "d is (sequentially) τ-lower-semicontinuous". The third condition (W3) is somewhat difficult to translate into a concise statement using standard terminology of point-set topology, but an attempt is "d is (sequentially) τ-continuous only when it is d-continuous".

Second, we note that the term "weak-like" is meant to be suggestive of the fact that there exist many examples of weak-like topologies connected to the usual notions of weak topology in functional analysis and in probability. In fact, the reader should keep in mind the following examples:

- If $(\mathcal{X}, d)$ is any metric space, then the topology generated by d is weak-like if and only if $(\mathcal{X}, d)$ has the Heine-Borel property.

- If $(\mathcal{X}, d)$ is the metric space arising from a Banach space $(\mathcal{X}, \|\cdot\|)$, then (W1) is equivalent to $(\mathcal{X}, \|\cdot\|)$ being reflexive [148, Theorem 1.13.5], (W2) is always true [148, p. 2.5.21], and (W3) is a special property called the *Radon-Riesz property*, the *Kadec-Klee property*, or *property (H)*. In particular $(\mathcal{X}, d)$ admits a weak-like topology if $(\mathcal{X}, \|\cdot\|)$ is finite-dimensional or uniformly convex. (See [148, Proposition 5.2.15], [148, Theorem 5.2.15] and [148, Theorem 5.2.18].)

- As we show below, if $\mathcal{X}$ is the Wasserstein space $\mathcal{P}_2(\mathbb{R}^m)$ and d is the Wasserstein metric W_2, then the topology of weak convergence τ_{w} is a weak-like topology for $(\mathcal{P}_2(\mathbb{R}^m), W_2)$.

For the remainder of this chapter, we will treat $(\mathcal{X}, d)$ as an abstract metric space, imposing assumptions along the way as needed.

Let us also clarify a point of possible confusion in the notation. Since in this section there will often be two topologies at play, we will always adopt the convention that if no topology on $\mathcal{X}$ is explicitly mentioned then it will be understood to be endowed with the topology generated by d. For example, $\mathcal{P}(\mathcal{X})$ represents the space of probability measures on $(\mathcal{X}, \mathcal{B}(d))$, not on $(\mathcal{X}, \mathcal{B}(\tau))$.

Now we can prove the main result of this chapter. Note that the conclusion of this theorem does not reference the weak-like topology τ. In this sense, the existence of a weak-like topology is a geometric property of the metric space $(\mathcal{X}, d)$ which guarantees that Fréchet means therein are well-behaved.

Theorem 8. *Consider any separable metric space $(\mathcal{X}, d)$ admitting a weak-like topology and any $p \geq 1$, and suppose that $\{P_n\}_{n\in\mathbb{N}}$ and P in $\mathcal{P}_{p-1}(\mathcal{X})$ have $P_n \to P$ in τ_{w}^{p-1}. Then, for any $\{x_n\}_{n\in\mathbb{N}}$ in $\mathcal{X}$ with $x_n \in M_p(P_n)$ for all $n \in \mathbb{N}$, there exists some $\{n_k\}_{k\in\mathbb{N}}$ and $x \in M_p(P)$ with $x_{n_k} \to x$ in d.*

Proof. Let $\{x_n\}_{n\in\mathbb{N}}$ be any sequence in $\mathcal{X}$ with $x_n \in M_p(P_n)$ for all $n \in \mathbb{N}$. Then, use Proposition 2 to get that $\{x_n\}_{n\in\mathbb{N}}$ is bounded. Now let τ denote a weak-like topology on $(\mathcal{X}, d)$, and use (W1) to get that there exists a subsequence $\{n_k\}_{k\in\mathbb{N}}$ and a point $x \in \mathcal{X}$ with $x_{n_k} \to x$ in τ. It only remains to show that $x \in M_p(P)$ and that we have $x_{n_k} \to x$ in d.

First let us show $x \in M_p(P)$, and we begin by fixing an arbitrary $o \in \mathcal{X}$. Observe that by (4.3) and (4.2), we can bound, for all $y \in \mathcal{X}$ and $k \in \mathbb{N}$:

$$\begin{aligned}
&d^p(x_{n_k}, y) - d^p(o, y) \\
&\geq -pd(x_{n_k}, o)(d^{p-1}(x_{n_k}, y) + d^{p-1}(o, y)) \\
&\geq -pd(x_{n_k}, o)(c_{p-1}(d^{p-1}(x_{n_k}, o) + d^{p-1}(o, y)) + d^{p-1}(o, y)) \\
&\geq -pc_{p-1}d^p(x_{n_k}, o) - p(c_{p-1} + 1)d^{p-1}(o, y).
\end{aligned}$$

Since $\{x_n\}_{n\in\mathbb{N}}$ is bounded and $P \in \mathcal{P}_{p-1}(\mathcal{X})$, this implies that the functions $y \mapsto d^p(x_{n_k}, y) - d^p(o, y)$ posess a P-integrable lower bound, uniformly in $k \in \mathbb{N}$. This means we can apply

(W2), Fatou's lemma, and Lemma 6, to get, for arbitrary $x' \in \mathcal{X}$:

$$\begin{aligned} W_p(P,x,o) &= \int_{\mathcal{X}} (d^p(x,y) - d^p(o,y))\,\mathrm{d}P(y) \\ &\le \int_{\mathcal{X}} \liminf_{k\to\infty} (d^p(x_{n_k},y) - d^p(o,y))\,\mathrm{d}P(y) \\ &\le \liminf_{k\to\infty} \int_{\mathcal{X}} (d^p(x_{n_k},y) - d^p(o,y))\,\mathrm{d}P(y) \\ &= \liminf_{k\to\infty} W_p(P,x_{n_k},o) \\ &= \liminf_{k\to\infty} W_p(P_{n_k},x_{n_k},o) \\ &\le \liminf_{k\to\infty} W_p(P_{n_k},x',o) \\ &= W_p(P,x',o). \end{aligned} \tag{4.11}$$

In particular, by taking the infimum over all $x' \in \mathcal{X}$, we find $x \in M_p(P)$.

Towards showing $x_{n_k} \to x$ in d, we now make a short digression. For each subsequence $K = \{k_j\}_{j\in\mathbb{N}}$, we consider the set

$$A_K := \left\{ y \in \mathcal{X} : \liminf_{j\to\infty} d(x_{n_{k_j}}, y) = d(x,y) \right\}.$$

It is straightforward to see that each A_K is a d-closed subset of $\mathcal{X}$: If $\{y_\ell\}_{\ell\in\mathbb{N}}$ in A_K have $y_\ell \to y$ in d for some $y \in \mathcal{X}$, then by the triangle inequality, we have

$$\begin{aligned} &\left| \liminf_{j\to\infty} d(x_{n_{k_j}}, y) - d(x,y) \right| \\ &= \left| \liminf_{j\to\infty} d(x_{n_{k_j}}, y) - \liminf_{j\to\infty} d(x_{n_{k_j}}, y_\ell) + d(x,y_\ell) - d(x,y) \right| \\ &\le \limsup_{j\to\infty} \left| d(x_{n_{k_j}}, y) - d(x_{n_{k_j}}, y_\ell) \right| + |d(x,y_\ell) - d(x,y)| \\ &\le 2d(y_\ell, y). \end{aligned}$$

Taking $\ell \to \infty$, we get $y \in A_K$ as needed. Moreover, we claim that each A_K satisfies $P(A_K) = 1$. To show this, we use an argument identical to (4.11) above to get

$$\begin{aligned} W_p(P,x,o) &= \int_{\mathcal{X}} (d^p(x,y) - d^p(o,y))\,\mathrm{d}P(y) \\ &\le \int_{\mathcal{X}} \liminf_{j\to\infty} (d^p(x_{n_{k_j}},y) - d^p(o,y))\,\mathrm{d}P(y) \le V_p(P,o). \end{aligned}$$

Since $V_p(P,o)$ is defined as a minimum, this implies that the inequalities in the preceding display are actually equalities. In particular, we have

$$\int_{\mathcal{X}} (d^p(x,y) - d^p(o,y))\,\mathrm{d}P(y) = \int_{\mathcal{X}} \liminf_{j\to\infty} (d^p(x_{k_j},y) - d^p(o,y))\,\mathrm{d}P(y).$$

Combining this with (W2) shows

$$d^p(x,y) - d^p(o,y) = \liminf_{j\to\infty}(d^p(x_{k_j},y) - d^p(o,y))$$

for P-almost all $y \in X$, so rearranging gives $P(A_K) = 1$.

Returning to the main proof, we observe that P is a Borel measure on a second-countable topological space and that $\{A_K\}_K$ is an (arbitrary) intersection of closed sets of full P-measure. Consequently, the set $A := \bigcap_K A_K$ satisfies $P(A) = 1$. Since $P(A) = 1$ implies that A is non-empty, we can select an arbitrary $y \in A$ and we can then select a subsequence $\{k_j\}_{j\in\mathbb{N}}$ such that

$$\limsup_{k\to\infty} d(x_{n_k}, y) = \lim_{j\to\infty} d(x_{n_{k_j}}, y).$$

Finally, use $y \in A$ and (W2) to get:

$$\begin{aligned}\limsup_{k\to\infty} d(x_{n_k}, y) &= \lim_{j\to\infty} d(x_{n_{k_j}}, y)\\ &= \liminf_{j\to\infty} d(x_{n_{k_j}}, y)\\ &= d(x,y)\\ &\le \liminf_{k\to\infty} d(x_{n_k}, y).\end{aligned}$$

In other words, we have shown $d(x_{n_k}, y) \to d(x,y)$ as $k \to \infty$. Therefore, (W3) implies $x_{n_k} \to x$ in d. This completes the proof. □

Observe that (W3) was only used in the very last step of the proof of Theorem 8. Consequently the following conclusion holds if $(\mathcal{X}, d)$ is a separable metric space admitting a topology τ satisfying (W1) and (W2): If $P_n \to P$ in τ_{w}^{p-1} and $x_n \in M_p(P_n)$ for all $n \in \mathbb{N}$, then there exists some $\{n_k\}_{k\in\mathbb{N}}$ and $x \in M_p(P)$ with $x_{n_k} \to x$ in τ. We believe that weaker result may still be useful in some applications.

The form of the statement of Theorem 8 is sometimes referred to as Γ-*convergence*, and it appears to be the preferred form of continity-type results proved by several previous authors [128, 194]. However, some readers may be more familiar with other forms of consistency, stability, or approximation results. We thus observe that the main result has the following immediate consequences:

Corollary 4. *Consider any separable metric space $(\mathcal{X}, d)$ admitting a weak-like topology, and any $p \geq 1$. For any $P \in \mathcal{P}_{p-1}(\mathcal{X})$, the set $M_p(P)$ is non-empty and compact.*

Proof. Let $\{x_n\}_{n\in\mathbb{N}}$ be an arbitrary sequence in $M_p(P)$. By taking $P_n := P$, we can trivially apply Theorem 8 to get some $\{n_k\}_{k\in\mathbb{N}}$ and $x \in M_p(P)$ with $x_{n_k} \to x$ in d. This proves that $M_p(P)$ is sequentially compact, whence compact as a subset of a metric space. □

Corollary 5. *Consider any separable metric space $(\mathcal{X}, d)$ admitting a weak-like topology, and any $p \geq 1$. If $\{P_n\}_{n\in\mathbb{N}}$ and P in $\mathcal{P}_{p-1}(\mathcal{X})$ have $P_n \to P$ in τ_{w}^{p-1}, then*

$$\max_{x_n\in M_p(P_n)} \min_{x\in M_p(P)} d(x_n, x) \to 0.$$

Proof. In order to show that

$$\max_{x_n \in M_p(P_n)} \min_{x \in M_p(P)} d(x_n, x)$$

converges to 0, it suffices to show that every subsequence admits a further subsequence converging to 0. So, we let $\{n_k\}_{k\in\mathbb{N}}$ be arbitrary, and we use the compactness of Corollary 4 and the continuity of d to get, for each $k \in \mathbb{N}$, a point $x_k \in M_p(P_{n_k})$ with

$$\min_{x \in M_p(P)} d(x_k, x) = \max_{x_{n_k} \in M_p(P_{n_k})} \min_{x \in M_p(P)} d(x_{n_k}, x).$$

Now we use Theorem 8 to get a subsequence $\{k_j\}_{j\in\mathbb{N}}$ and a point $x' \in M_p(P)$ with $x_{k_j} \to x'$ in d. By construction we have

$$\max_{x_{n_k} \in M_p(P_{n_k})} \min_{x \in M_p(P)} d(x_{n_k}, x) = \min_{x \in M_p(P)} d(x_k, x) \leq d(x_{k_j}, x') \to 0,$$

which proves the claim. □

To close this section, let us prove that the Wasserstein space admits a weak-like topology. To set this up, let us fix $m \in \mathbb{N}$ and define, for $\mu, \mu' \in \mathcal{P}_2(\mathbb{R}^m)$

$$W_2(\mu, \mu') := \inf_{\tilde{\mu} \in \Pi(\mu,\mu')} \left(\int_{\mathbb{R}^m \times \mathbb{R}^m} \|x - x'\|_2^2 \, \mathrm{d}\tilde{\mu}(x, x') \right)^{1/2}.$$

Our interest is in the metric space $(\mathcal{X}, d) := (\mathcal{P}_2(\mathbb{R}^m), W_2)$ about which have the following result:

Lemma 10. *The topology τ_{w} is a weak-like topology for $(\mathcal{P}_2(\mathbb{R}^m), W_2)$.*

Proof. To prove (W1), suppose that $\{\mu_n\}_{n\in\mathbb{N}}$ form a bounded sequence in $(\mathcal{P}_2(\mathbb{R}^m), W_2)$. Then get a sequence $K_1 \subseteq K_2 \subseteq \cdots$ of compact subsets of $\mathbb{R}^m$ satisfying $\bigcup_{\ell\in\mathbb{N}} K_\ell = \mathbb{R}^m$, and assume $0 \in K_1$. Now use the triangle inequality to get $\sup_{n\in\mathbb{N}} W_2(\mu_n, \delta_0) < \infty$. Because $\Pi(\mu, \delta_0) = \{\mu \otimes \delta_0\}$, we have

$$C := \sup_{n\in\mathbb{N}} \int_{\mathbb{R}^m} \|x\|_2^2 \, \mathrm{d}\mu_n(x) < \infty.$$

At the same time, note that for each $\ell \in \mathbb{N}$ we have

$$\mathbb{1}_{\mathbb{R}^m \setminus K_\ell}(x) \leq \frac{\|x\|_2}{\inf_{x \notin K_\ell} \|x\|_2},$$

so squaring and integrating with respect to μ_n gives

$$\mu_n(\mathbb{R}^m \setminus K_\ell) \leq \frac{C}{(\inf_{x \notin K_\ell} \|x\|_2^2}.$$

Clearly, we have $\inf_{x\notin K_\ell}\|x\|_2^2 \to \infty$ as $\ell \to \infty$ and the right side above does not depend on $n \in \mathbb{N}$, so it follows that the family $\{\mu_n\}_{n\in\mathbb{N}}$ is tight. By Prokhorov's theorem [104, Theorem 16.3], this implies that there exists a subsequence $\{n_k\}_{k\in\mathbb{N}}$ and some $\mu \in \mathcal{P}(\mathbb{R}^m)$ with $\mu_{n_k} \to \mu$ in τ_{w}. Moreover, we have

$$\int_{\mathbb{R}^m} \|x\|_2^2 \,\mathrm{d}\mu(x) \le \liminf_{k\to\infty} \int_{\mathbb{R}^m} \|x\|_2^2 \,\mathrm{d}\mu_{n_k}(x) \le C$$

so it follows that $\mu \in \mathcal{P}_2(\mathbb{R}^m)$. This proves (W1).

To prove (W2), suppose that $\{\mu_n\}_{n\in\mathbb{N}}$ and μ in $\mathcal{P}_2(\mathbb{R}^m)$ have $\mu_n \to \mu$ in τ_{w}, and let $\mu' \in \mathcal{P}_2(\mathbb{R}^m)$ be arbitrary. By [166, Theorem 1] (also, see [166, Corollary 2] and the discussion thereafter), it is possible to construct a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ supporting $\mathbb{R}^m$-valued random variables $\{X_n\}_{n\in\mathbb{N}}$, X, and X', such that (X, X') is an optimal coupling of (μ, μ'), such that (X_n, X') is an optimal coupling of (μ_n, μ') for all $n \in \mathbb{N}$, and such that $X_n \to X$ holds μ-almost surely. Writing $\mathbb{E}$ for the expectation on $(\Omega, \mathcal{F}, \mathbb{P})$, we have

$$W_2(\mu, \mu') = \mathbb{E}\left[|X - X'|^2\right] \le \liminf_{n\to\infty} \mathbb{E}\left[|X_n - X'|^2\right] = \liminf_{n\to\infty} W_2(\mu_n, \mu')$$

by Fatou's lemma, which gives (W2).

Finally, let us prove (W3). To do this, suppose that $\{\mu_n\}_{n\in\mathbb{N}}$ and μ in $\mathcal{P}_2(\mathbb{R}^m)$ have $\mu_n \to \mu$ in τ_{w}, and suppose that $\mu' \in \mathcal{P}_2(\mathbb{R}^m)$ satisfies $W_2(\mu_n, \mu') \to W_2(\mu, \mu')$. Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote a probability space supporting $\mathbb{R}^m$-valued random variables $\{X_n\}_{n\in\mathbb{N}}$, X, and X' as above. We know that $\|X_n\|_2^2 \to \|X\|_2^2$ holds $\mathbb{P}$-almost surely. We also have $\|X_n\|_2^2 \le 2(\|X_n - X'\|_2^2 + \|X'\|_2^2)$ for all $n \in \mathbb{N}$; since $\mathbb{E}[\|X_n - X'\|_2^2] = W_2(\mu_n, \mu') \to W_2(\mu, \mu') < \infty$ and $\mathbb{E}[\|X'\|_2^2] < \infty$, this shows that the family $\{\|X_n\|_2^2\}_{n\in\mathbb{N}}$ is uniformly integrable. In particular, we have

$$\int_{\mathbb{R}^m} \|x\|_2^2 \,\mathrm{d}\mu_n(x) = \mathbb{E}[\|X_n\|_2^2] \to \mathbb{E}[\|X\|_2^2] = \int_{\mathbb{R}^m} \|x\|_2^2 \,\mathrm{d}\mu(x).$$

This shows $\mu_n \to \mu$ in W_2, which proves (W3). □

Note that Theorem 8 and Lemma 10 together recover the main result of [128].

4.3 Probabilistic Consequences

In this section we apply the "continuity" results of the previous section to deduce some probabilistic consequences. That is, we prove a general form of some limit theorems of interest. Throughout this section, we assume that $(\mathcal{X}, d)$ is a separable metric space admitting a weak-like topology.

First we consider the stong law of large numbers (SLLN) which concerns independent, identically-distributed (IID) sequences of $\mathcal{X}$-valued random variables. The following results connects our measure-valued theory to some probabilistic results.

Theorem 9 (Strong Law of Large Numbers)**.** *If $p \geq 1$ and $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space supporting an IID sequence $Y_1, Y_2, \ldots$ of $\mathcal{X}$-valued random variables with common distribution $P \in \mathcal{P}_{p-1}(\mathcal{X})$, then*

$$\max_{\bar{x}_n \in M_p(\bar{P}_n)} \min_{x \in M_p(P)} d(\bar{x}_n, x) \to 0$$

holds $\mathbb{P}$-almost surely.

Proof. By Theorem 8, it suffices to show $\mathbb{P}(\bar{P}_n \to P \text{ in } \tau_{\mathrm{w}}^{p-1}) = 1$. To do this, recall that there exists a sequence $\{\phi_\ell\}_{\ell \in \mathbb{N}}$ of bounded continuous functions from $(\mathcal{X}, d)$ to $\mathbb{R}$ such that $\{P_n\}_{n \in \mathbb{N}}$ and P in $\mathcal{P}(\mathcal{X})$ satisfy $P_n \to P$ in τ_{w} if and only if $\int_{\mathcal{X}} \phi_\ell \, \mathrm{d}P_n \to \int_{\mathcal{X}} \phi_\ell \, \mathrm{d}P$ for all $\ell \in \mathbb{N}$. (See [201, Theorem 3.1] and its proof.) Consequently, the event $A = \{\bar{P}_n \to P \text{ in } \tau_{\mathrm{w}}\}$ can be equivalently written as $A := \bigcap_{\ell \in \mathbb{N}} A_\ell$, where we have defined

$$A_\ell := \left\{ \int_{\mathcal{X}} \phi_\ell \, \mathrm{d}\bar{P}_n \to \int_{\mathcal{X}} \phi_\ell \, \mathrm{d}P \right\}.$$

for all $\ell \in \mathbb{N}$. Now fix $o \in \mathcal{X}$, and write $B_x := \{W_r(\bar{P}_n, x) \to W_r(P, x)\}$ for $x \in \mathcal{X}$. Using Lemma 7, we have

$$\{\bar{P}_n \to P \text{ in } \tau_{\mathrm{w}}^p\} = A \cap \bigcap_{x \in \mathcal{X}} B_x = A \cap B_o = \left(\bigcap_{\ell \in \mathbb{N}} A_\ell \right) \cap B_o.$$

By the usual SLLN, we have $\mathbb{P}(A_\ell) = 1$ for all $\ell \in \mathbb{N}$ and $\mathbb{P}(B_o) = 1$, hence the result is proved. □

This result strengthens the existing SLLNs given in [31, 32, 66, 180, 213, 186], but its true utility comes from the simplicity of the proof: We simply descended the convergence of empirical measures to convergence of the Fréchet means by virtue of the "continuity" provided by Theorem 8. This approach has two significant advantages. The first is that the structure of IID samples is unimportant; we can actually descend any almost sure convergence of empirical measures to an almost sure convergence of the Fréchet means. The second is that the structure of almost sure convergence is unimportant; we can actually descend any limit theorem for empirical measures to an analogous limit theorem for the Fréchet means. To close this chapter, we discuss some illustrations of this.

Regarding the structure of the samples, we note that we can easily prove a version of a convergence theorem for Fréchet means of suitable Markov chains:

Theorem 10 (Markov Chain Convergence Theorem)**.** *If $p \geq 1$ and $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space supporting a Harris-recurrent Markov chain $Y_1, Y_2, \ldots$ of $\mathcal{X}$-valued random variables with stationary distribution $P \in \mathcal{P}_{p-1}(\mathcal{X})$, then*

$$\max_{\bar{x}_n \in M_p(\bar{P}_n)} \min_{x \in M_p(P)} d(\bar{x}_n, x) \to 0$$

holds $\mathbb{P}$-almost surely.

Proof. The proof is identical to the proof of Theorem 9, upon replacing all appeals to the SLLN with appeals to the classical Markov chain convergence theorem. □

Regarding other types of limit theorems, we note that one can easily prove the following large deviations upper bound, which we investigate in great detail in the next chapter:

Theorem 11 (Large Deviations Upper Bound)**.** *If $(\mathcal{X}, d)$ is Polish, $p \geq 1$, and $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space supporting an IID sequence $Y_1, Y_2, \ldots$ of $\mathcal{X}$-valued random variables with common distribution $P \in \mathcal{P}(\mathcal{X})$ satisfying $\int_{\mathcal{X}} \exp(\lambda d^{p-1}(x,y))\,\mathrm{d}P(y) < \infty$ for all $\lambda \geq 0$ and $x \in \mathcal{X}$, then for all $\varepsilon > 0$ we have*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\max_{\bar{x}_n \in M_p(\bar{P}_n)} \min_{x\in M_p(P)} d(\bar{x}_n, x) \geq \varepsilon \right) \leq -c_{P,p}(\varepsilon)$$

for some $c_{P,p}(\varepsilon) > 0$.

Proof. Consider the event

$$A := \left\{ P' \in \mathcal{P}_{p-1}(X) : \max_{x'\in M_p(P')} \min_{x\in M_p(P)} d(x', x) \geq \varepsilon \right\}$$

and let us show that A is τ_{w}^{p-1}-closed. Indeed, suppose $\{P'_n\}_{n\in\mathbb{N}}$ in A and $P' \in \mathcal{P}_{p-1}(X)$ have $P'_n \to P'$ in τ_{w}^{p-1}. Then:

$$\max_{x'_n\in M_p(P'_n)} \min_{x\in M_p(P')} d(x'_n, x) + \max_{x'\in M_p(P')} \min_{x\in M_p(P)} d(x', x) \geq \max_{x'_n\in M_p(P'_n)} \min_{x\in M_p(P)} d(x'_n, x) \geq \varepsilon.$$

Note that the first term vanishes as $n \to \infty$ by Theorem 8, so $P' \in A$, as claimed.

Next, note that P is a Borel probability measure on a Polish metric space with all exponential moments finite, so we conclude via [205, Theorem 1.1] that $\{\bar{P}_n\}_{n\in\mathbb{N}}$ satisfy a large deviations principle in $(\mathcal{P}_{p-1}(X), \tau_{\mathrm{w}}^{p-1})$ with good rate function given by the relative entropy from P, denoted $H(\cdot\,|P) : \mathcal{P}_{p-1}(X) \to [0,\infty]$. In particular, the large deviations upper bound implies

$$\begin{aligned}\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\max_{\bar{x}_n\in M_p(P_n)} \min_{x\in M_p(P)} d(\bar{x}_b, x) \geq \varepsilon \right) &= \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\bar{P}_n \in A) \\ &\leq -\inf\{H(P'|P) : P' \in A\} := c_{P,p}(\varepsilon).\end{aligned}$$

Finally, assume towards a contradiction that $c_{P,p}(\varepsilon) = 0$, so that there exist $\{Q_n\}_{n\in\mathbb{N}}$ in A with $H(Q_n|P) \to 0$. As a consequence of the Donsker-Varadhan variational principle one can show that $H(Q_n|P) \to 0$ and $\int_{\mathcal{X}} \exp(\lambda d^{p-1}(x,y))\,\mathrm{d}P(y) < \infty$ for all $\lambda \geq 0$ and $x \in \mathcal{X}$ together imply $Q_n \to P$ in τ_{w}^{p-1}. Then Theorem 8 implies

$$\max_{z_n\in M_p(Q_n)} \min_{x\in M_p(P)} d(z_n, x) \to 0$$

which contradicts $Q_n \in A$ for all $n \in \mathbb{N}$. Therefore, we must have $c_{P,p}(\varepsilon) > 0$. □

Chapter 5

Applications in the Bures-Wasserstein Space

Let $m \in \mathbb{N}$ denote a fixed dimension, and write $\mathbb{K} := \mathbb{K}(m)$ for the space of all $m \times m$ real, symmetric, positive semi-definite matrices. For $\Sigma, \Sigma' \in \mathbb{K}$, we write

$$\Pi(\Sigma, \Sigma') := \sqrt{\mathrm{tr}(\Sigma) + \mathrm{tr}(\Sigma') - 2\mathrm{tr}\left((\Sigma^{1/2}\Sigma'\Sigma^{1/2})^{1/2}\right)}, \tag{5.1}$$

which makes $(\mathbb{K}, \Pi)$ into a metric space called the *Bures-Wasserstein space*. The Bures-Wasserstein space has a rich geometric structure: It is a "stratified space" meaning it is a collection of Riemannian manifolds glued together along Riemannian submanifolds of smaller dimension, it has "positive curvature" in the sense of metric geometry, and it a "uniquely geodesic space" meaning all points are connected by a unique continuous path whose arc length equals the distance between them.

The Bures-Wasserstein space and its geometry arise in several different disciplines. In quantum information theory, one restricts attention to the subset of $\Sigma \in \mathbb{K}$ with $\mathrm{tr}(\Sigma) = 1$, in which case Π determines a natural distance on pure states of a quantum system. In optimal transport, one can identify each $\Sigma \in \mathbb{K}$ with the centered Gaussian distribution $N(0, \Sigma)$, and it turns out that Π is the distance induced by the Wasserstein metric W_2. Because of these applications and more, much recent literature has studied the Bures-Wasserstein space and probabilistic and statistical aspects therein [5, 208, 128, 130, 211, 157].

In this chapter we are interested in Fréchet means in $(\mathbb{K}, \Pi)$, which are usually called *barycenters*. That is, suppose that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space on which we have an IID sequence $\Sigma_1, \Sigma_2, \ldots$ of $\mathbb{K}$-valued random variables with common distribution $P \in \mathcal{P}_2(\mathbb{K})$, and write $\{\bar{P}_n\}_{n\in\mathbb{N}}$ for their empirical measures, as usual. Under some mild conditions guaranteeing uniqueness (to be discussed below), we write $M^* := M_2(P)$ and $M_n^* := M_2(\bar{P}_n)$ for the population and empirical Bures-Wasserstein barycenters, respectively. We know from the general theory of Theorem 9 that $M_n^* \to M^*$ holds almost surely, but for several applications we require finer understanding of the speed of this convergence.

Specifically, we are interested in results of the form

$$\mathbb{P}(\Pi(M_n^*, M^*) \geq t) \approx e^{-nc_P(t)} \tag{5.2}$$

for all $t \geq 0$, where $c_P : [0, \infty) \to [0, \infty)$ is some function depending only on the distribution P. We know from the general theory of Theorem 11 that such a function c_P does exist, although, in order for (5.2) to be useful, we need to be able to evaluate c_P. At the same time, there are known finite-sample bounds for the concentration properties of Bures-Wasserstein barycenters [123], but it is not known whether these exponents are sharp. The goal of this chapter is to prove a precise large deviations principle for Bures-Wasserstein barycenters, thereby establishing (5.2) in a sharp (albeit asymptotic) manner.

This chapter is based on [101] which develops these ideas in detail. Our main results include the large deviation principle (Theorem 12), a careful study of the properties of the rate function (Proposition 3), and various geometric interpretations of these objects. We conclude the chapter by giving several examples where we can use our theory to analyze (both analytically and numerically) the exponential rate of decay of some rare of events of interest. Our method of proof for most of these results is identifying a novel notion of exponential tilting in the tangent bundle of the Bures-Wasserstein space.

5.1 Preliminaries

Let $m \in \mathbb{N}$ denote a fixed dimension. We write $\mathbb{K}$ for the space of covariances on $\mathbb{R}^m$ (equivalently, the set of real, symmetric, positive semi-definite $m \times m$ matrices), and we write $\mathbb{K}_+ \subseteq \mathbb{K}$ for the subspace of strictly positive covariances on $\mathbb{R}^m$ (equivalently, the set of real, symmetric, strictly positive definite $m \times m$ matrices). For $A, A' \in \mathbb{R}^{m\times m}$, we write $A \preceq A'$ for $A' - A \in \mathbb{K}$ and we write $A \prec A'$ for $A' - A \in \mathbb{K}_+$. Lastly, $\mathbb{S}$ denotes the space of self-adjoint operators on $\mathbb{R}^m$ (equivalently, the set of real symmetric $m \times m$ matrices). The Bures-Wasserstein metric is denoted $\Pi : \mathbb{K} \times \mathbb{K} \to [0, \infty)$; we note that, in addition to (5.1) given above, there are many equivalent formulations of the Bures-Wasserstein distance: as the Procrustes metric on the space of covariance operators [147], as the metric induced by the Wasserstein metric on the space of centered Gaussian distributions [155, 58], and more.

For $P \in \mathcal{P}_2(\mathbb{K}, \Pi)$, we say that $M \in \mathbb{K}$ *is a Bures-Wasserstein barycenter of* P if $M \in M_2(P)$. In the statistical setting we assume that $P \in \mathcal{P}(\mathbb{K}, \Pi)$ is fixed and that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space on which is defined an IID sequence $\Sigma_1, \Sigma_2, \ldots$ of $\mathbb{K}$-valued random variables with common distribution P. In this case we say that M^* *is a population Bures-Wasserstein barycenter* if $M^* \in M_2(P)$, and we say that M_n^* *is an empirical Bures-Wasserstein barycenter* if it $M_n^* \in M_2(\bar{P}_n)$. (In particular, Bures-Wasserstein barycenters are just Fréchet means in the space $(\mathbb{K}, \Pi)$.)

Throughout the remainder of the paper, we assume that $P \in \mathcal{P}(\mathbb{K}, \Pi)$ is a fixed population distribution, and that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space supporting an independent, identically-distributed (IID) sequence $\Sigma_1, \Sigma_2, \ldots$ of random variables with common distribution P. All "almost surely" statements are understood to refer to the probability measure $\mathbb{P}$. We also assume the following conditions throughout:

(i) $P\{\Sigma \in \mathbb{K} : \Sigma \succ 0\} = 1$.

(ii) For all $\lambda > 0$ and $M \in \mathbb{K}$, we have $\int_{\mathbb{K}} \exp(\lambda \Pi(\Sigma, M))\, \mathrm{d}P(\Sigma) < \infty$.

Let us briefly comment on these conditions and how they are related to analogous conditions in related literature.

The positivity condition (i) simply states that we can restrict our attention from $\mathbb{K}$ to $\mathbb{K}_+$. This assumption is weaker than the assumptions appearing in [55, Theorem 1] and [177, Assumption B(2)], but stronger than those appearing in [177, Assumption A(2)] and [123, Assumption 1] where it is only assumed that $p := P\{\Sigma \in \mathbb{K} : \Sigma \succ 0\}$ satisfies $p > 0$. As we will discuss further below, we also note that condition (i) implies that the barycenter of P exists and is unique, that is is characterized by a certain fixed-point equation, and that the same results are true for $\bar{P}_n$ almost surely.

The integrability condition (ii) simply states that P has finite exponential moments of all orders. This is easily seen to be stronger than first-moment [177, Assumption A(1)] and second-moment conditions, but weaker than sub-Gaussianity [123, Assumption 2] and boundedness [55, Theorem 1] conditions. In fact, we will later see that sub-Gaussianity assumptions lead to quantitative bounds on the large deviations behavior.

Next we recall the important Riemannian structure of $(\mathbb{K}_+, \Pi)$. For each $M \in \mathbb{K}_+$, we define the following: $\mathrm{Tan}(M)$ is the Hilbert space whose underlying set is $\mathbb{S}$ and whose inner product is given $\langle A, B\rangle_{\mathrm{Tan}(M)} := \mathrm{tr}(AMB)$ for $A, B \in \mathrm{Tan}(M)$; the *exponential map* $\exp_M : \mathrm{Tan}(M) \to \mathbb{K}$ is given by

$$\exp_M(A) := (A + I)M(A + I)$$

for $A \in \mathrm{Tan}(M)$; the *logarithm map* $\log_M : \mathbb{K} \to \mathrm{Tan}(M)$ is the inverse of the exponential map, and can be written exactly in either of two equivalent forms:

$$\begin{aligned} \log_M(\Sigma) &:= \Sigma^{1/2}(\Sigma^{1/2}M\Sigma^{1/2})^{-1/2}\Sigma^{1/2} - I \\ &:= M^{-1/2}(M^{1/2}\Sigma M^{1/2})^{1/2}M^{-1/2} - I \end{aligned} \tag{5.3}$$

for $\Sigma \in \mathbb{K}_+$. (All of these notions can still be defined when $M \in \mathbb{K} \setminus \mathbb{K}_+$, although some partial degeneracy requires $\mathrm{Tan}(M)$ to be defined as a suitable quotient of $\mathbb{S}$.)

The Riemannian logarithm map has an important interpretation through the lens of optimal transport. That is, for all $M \in \mathbb{K}_+$ and $\Sigma \in \mathbb{K}$, the matrix $t_M^\Sigma := \log_M(\Sigma) + I$ represents the optimal transport map from the centered Gaussian with covariance matrix M to the centered Gaussian with covariance Σ; more precisely, one can easily see from (5.3) that $t_M^\Sigma M t_M^\Sigma = \Sigma$. Thus, the logarithm map can be written (and should be regarded) as $\log_M(\Sigma) = t_M^\Sigma - I$, whenever M, Σ are sufficiently regular.

Next we introduce the fundamental fixed-point equation. It is known [123, Theorem 2.1] that $M \in \mathbb{K}_+$ is a Bures-Wasserstein barycenter of P if and only if

$$\int_{\mathbb{K}} (M^{1/2}\Sigma M^{1/2})^{1/2}\,\mathrm{d}P(\Sigma) = M. \tag{5.4}$$

Note that we can equivalently write (5.4) as

$$\int_{\mathbb{K}} \log_M(\Sigma)\,\mathrm{d}P(\Sigma) = 0, \tag{5.5}$$

which has a clear interpretation from a Riemannian point of view: The Bures-Wasserstein barycenter is the unique point M^* in $\mathbb{K}$ such that, if we push forward the distribution P from $(\mathbb{K}_+, \Pi)$ to $(\mathrm{Tan}(M^*), \langle\,\cdot\,,\,\cdot\,\rangle_{\mathrm{Tan}(M^*)})$, the resulting lifted distribution is centered. It is known that this is a necessary condition for Fréchet means in general Riemannian manifolds, but it is remarkable that in the Bures-Wasserstein space it is also sufficient. See Figure 5.1 for an illustration.

Next we provide some intermediate results that will be used in our main theorems. For example, it is useful to note that for any $M \in \mathbb{K}$ we have $\Pi(M, 0) = \mathrm{tr}(M^{1/2}) = \|M^{1/2}\|_2$ and that for any $M, \Sigma \in \mathbb{K}_+$ we have $\Pi(M,\Sigma) = \|(t_M^\Sigma - I)M^{1/2}\|_2$. We also need the following estimates:

Lemma 11. *For any $M, \Sigma \in \mathbb{K}$, we have*

$$\mathrm{tr}\left((M^{1/2}\Sigma M^{1/2})^{1/2}\right) \le \Pi(M,0)\Pi(\Sigma,0).$$

Proof. Using the definition of matrix square root, we have

$$\mathrm{tr}\left((M^{1/2}\Sigma M^{1/2})^{1/2}\right) = \sum_{i=1}^m \lambda_i\left((M^{1/2}\Sigma M^{1/2})^{1/2}\right) = \sum_{i=1}^m \sqrt{\lambda_i\left(M^{1/2}\Sigma M^{1/2}\right)}.$$

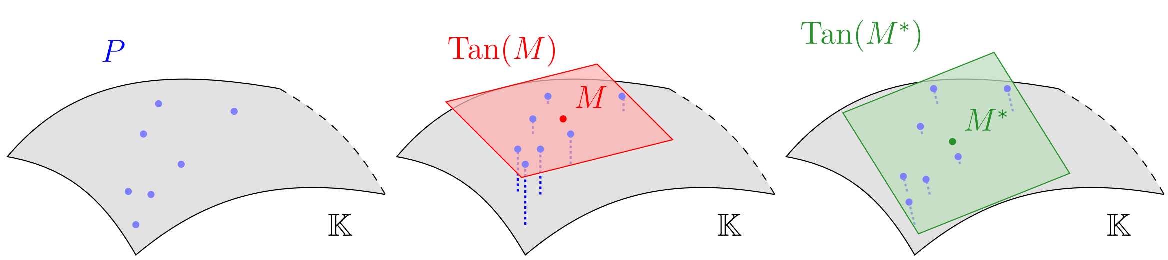

Figure 5.1: The fixed point equation for the Bures-Wasserstein barycenter of a probability measure P on $(\mathbb{K}, \Pi)$ (left). For each covariance $M \in \mathbb{K}_+$, we can push forward P by $\log_M(\,\cdot\,)$ to a probability measure on $\mathrm{Tan}(M)$. This pushforward is centered in $\mathrm{Tan}(M)$ if and only if M is a Bures-Wasserstein barycenter of P (middle, right).

Now note that, for all $i \in \{1, \ldots, m\}$, we have

$$\lambda_i\left(M^{1/2}\Sigma M^{1/2}\right) \leq \lambda_i\left(M^{1/2}\Sigma\right)\lambda_1\left(M^{1/2}\right) \leq \lambda_i\left(\Sigma\right)\left(\lambda_1\left(M^{1/2}\right)\right)^2$$

hence

$$\lambda_i\left(M^{1/2}\Sigma M^{1/2}\right) \leq \lambda_i\left(\Sigma\right)\left(\operatorname{tr}\left(M^{1/2}\right)\right)^2.$$

Combining these displays and using again the definition of matrix square root, we get

$$\begin{aligned}
\operatorname{tr}\left((M^{1/2}\Sigma M^{1/2})^{1/2}\right) &= \sum_{i=1}^{m}\sqrt{\lambda_i\left(M^{1/2}\Sigma M^{1/2}\right)} \\
&\leq \sum_{i=1}^{m}\sqrt{\lambda_i\left(\Sigma\right)}\operatorname{tr}\left(M^{1/2}\right) \\
&\leq \sum_{i=1}^{m}\lambda_i\left(\Sigma^{1/2}\right)\operatorname{tr}\left(M^{1/2}\right) \\
&= \operatorname{tr}\left(M^{1/2}\right)\operatorname{tr}\left(\Sigma^{1/2}\right) = \Pi(M,0)\Pi(\Sigma,0)
\end{aligned}$$

as claimed. □

Lemma 12. *For any $M, \Sigma \in \mathbb{K}_+$ and $A \in \mathbb{S}$, we have*

$$|\operatorname{tr}(AM(t_M^{\Sigma} - I))| \leq \|A\|_2\Pi(M,0)\Pi(M,\Sigma).$$

Proof. Use Cauchy-Schwarz and the submultiplicativity of the matrix norm $\|\cdot\|_2$ to get

$$\begin{aligned}
\operatorname{tr}(AM(t_M^{\Sigma} - I)) &\leq \|AM^{1/2}\|_2\|(t_M^{\Sigma} - I)M^{1/2}\|_2 \\
&= \|A\|_2\|M^{1/2}\|_2\|(t_M^{\Sigma} - I)M^{1/2}\|_2 \\
&= \|A\|_2\Pi(M,0)\Pi(M,\Sigma),
\end{aligned}$$

as claimed. □

Lastly, we give the following result, which we expect to be well-known to those familiar with large deviations theory but for which we could not find an appropriate reference. Here, $H(\cdot\,|\,\cdot) : (\mathcal{P}(\mathcal{X}))^2 \to [0, \infty]$ denotes the relative entropy.

Lemma 13. *Let $\mathcal{H}$ be a Hilbert space with norm $\|\cdot\|$ and inner product $\langle\cdot,\cdot\rangle$, and let p be a probability measure on $\mathcal{H}$ satisfying $\int_{\mathcal{H}} e^{\lambda\|h\|}\,\mathrm{d}p(h) < \infty$ for all $\lambda > 0$ and $h \in \mathcal{H}$. Then, for all $h \in \mathcal{H}$, the value $\Lambda_p(h) := \log(\int_{\mathcal{H}}\exp(\langle h, h'\rangle)\,\mathrm{d}p(h'))$ is well-defined, the optimization problems*

$$\begin{cases}
\text{minimize} & H(q\,|\,p) \\
\text{over} & q \in \mathcal{P}_1(\mathcal{H}) \\
\text{subject to} & \int_{\mathcal{H}} h'\,\mathrm{d}q(h') = h
\end{cases}$$

and

$$\begin{cases} \text{maximize} & \langle h'', h\rangle - \Lambda_p(h'') \\ \text{over} & h'' \in \mathcal{H}, \end{cases}$$

have the same value, and both problems are achieved.

Proof. To see that Λ_p is well-defined, simply use Cauchy-Schwarz to get

$$\int_{\mathcal{H}} e^{\langle h,h'\rangle}\,\mathrm{d}p(h') \le \int_{\mathcal{H}} e^{\|h\|\,\|h'\|}\,\mathrm{d}p(h') < \infty$$

for all $h \in \mathcal{H}$. Next, let us show that the value of the first optimization problem is bounded below by the value of the second. For arbitrary $h'' \in \mathcal{H}$ and $r > 0$, define the bounded, continuous function $\phi_{h'',r} : \mathcal{H} \to \mathbb{R}$ via

$$\phi_{h'',r}(h') = \begin{cases} r & \text{if } \langle h'', h'\rangle > r \\ \langle h'', h'\rangle & \text{if } -r \le \langle h'', h'\rangle \le r \\ -r & \text{if } \langle h'', h'\rangle < -r. \end{cases}$$

Note that $\phi_{h'',r}(h') \to \langle h'', h'\rangle$ as $r \to \infty$ for all $h' \in \mathcal{H}$. Moreover, we have $|\phi_{h'',r}(h')| \le \|h''\|\,\|h'\|$ by Cauchy-Schwarz, so we can apply dominated to convergence for any $q \in \mathcal{P}_1(\mathcal{H})$ to get

$$\int_{\mathcal{H}} \phi_{h'',r}(h')\,\mathrm{d}q(h') \to \int_{\mathcal{H}} \langle h'', h'\rangle\,\mathrm{d}q(h')$$

as $r \to \infty$. Similarly, we have $e^{\phi_{h'',r}(h')} \le e^{\|h''\|\,\|h'\|}$ by Cauchy-Schwarz, so we can use the integrability condition on p to apply dominated convergence and get

$$\int_{\mathcal{H}} e^{\phi_{h'',r}(h')}\,\mathrm{d}p(h') \to \int_{\mathcal{H}} e^{\langle h'',h'\rangle}\,\mathrm{d}p(h')$$

as $r \to \infty$. Therefore, the Donsker-Varadhan variational formula (see Lemma 6.2.13 of [60]) yields

$$\begin{aligned} H(q\,|\,p) &\ge \int_{\mathcal{H}} \phi_{h'',r}(h')\,\mathrm{d}q(h') - \log\left(\int_{\mathcal{H}} e^{\phi_{h'',r}(h')}, \mathrm{d}p(h')\right) \\ &\to \int_{\mathcal{H}} \langle h'', h'\rangle\,\mathrm{d}q(h') - \log\left(\int_{\mathcal{H}} e^{\langle h'',h'\rangle}\,\mathrm{d}p(h')\right) = \langle h'', h\rangle - \Lambda_p(h). \end{aligned}$$

As this holds for all $h'' \in \mathcal{H}$, we can take the maximum and the first direction is proved. Lastly, let us show that the value of the first optimization problem is bounded above by the value of the second. To do this, consider the probability measure $\tilde{p}$ on $\mathcal{H}$ defined via its Radon-Nikodym derivative

$$\frac{d\tilde{p}}{dp}(h') := e^{\langle h',h\rangle - \Lambda_p(h)}.$$

We can directly compute

$$
\begin{aligned}
H(\tilde{p}\,|\,p) &= \int_{\mathcal{H}} \log\left(\frac{d\tilde{p}}{dp}(h')\right) \mathrm{d}\tilde{p}(h') \\
&= \int_{\mathcal{H}} (\langle h', h\rangle - \Lambda_p(h))\,\mathrm{d}\tilde{p}(h') \\
&= \left\langle \int_{\mathcal{H}} h'\,\mathrm{d}\tilde{p}(h'), h \right\rangle - \Lambda_p(h)
\end{aligned}
$$

and this proves the claim. □

5.2 Large Deviations Theory

In this section we prove the large deviations principle for Bures-Wasserstein barycenters. We begin, in the standard fashion of large deviations theory, by defining the rate function and proving some useful properties.

Proposition 3. *The function $I_P : \mathbb{K}_+ \to [0, \infty]$ defined via*

$$
I_P(M) = \sup_{A\in\mathbb{S}} \left(\operatorname{tr}(AM) - \log \int_{\mathbb{K}} \exp\operatorname{tr}\left(AM\Sigma^{1/2}(\Sigma^{1/2}M\Sigma^{1/2})^{-1/2}\Sigma^{1/2}\right) \mathrm{d}P(\Sigma)\right) \tag{5.6}
$$

for $M \in \mathbb{K}_+$ enjoys the following properties:

(a) I_P is lower semi-continuous.

(b) I_P is coercive and satisfies the lower bound

$$
I_P(M) \geq \Pi(M, 0) - \log \int \exp\Pi(\Sigma, 0)\,\mathrm{d}P(\Sigma) \tag{5.7}
$$

for all $M \in \mathbb{K}_+$.

(c) $I_P(M^*) = 0$.

(d) For each $M \in \mathbb{K}_+$, a matrix $A \in \mathbb{S}$ achieves the supremum in the definition of $I_P(M)$ if and only if the exponentially tilted measure $P^{M\to A} \in \mathcal{P}_2(\mathbb{K}, \Pi)$ defined via

$$
\frac{\mathrm{d}P^{M\to A}}{\mathrm{d}P}(\Sigma) \propto \exp\operatorname{tr}(AM\Sigma^{1/2}(\Sigma^{1/2}M\Sigma^{1/2})^{-1/2}\Sigma^{1/2}). \tag{5.8}
$$

has Bures-Wasserstein barycenter M.

Proof. To prove (a), it suffices to show that the function

$$M \mapsto \int_{\mathbb{K}} \exp \operatorname{tr}\left(AM(t_M^\Sigma - I)\right) \mathrm{d}P(\Sigma)$$

is continuous for each fixed $A \in \mathbb{S}$. It is clear that the function

$$f(A, \Sigma, M) := \exp \operatorname{tr}\left(AM(t_M^\Sigma - I)\right)$$

is continuous as a function of $M \in \mathbb{K}_+$, for fixed $A \in \mathbb{S}$ and $\Sigma \in \mathbb{K}_+$, so it suffices to show that we can apply dominated convergence. To do this, suppose that $\{M_n\}_{n\in\mathbb{N}}$ and M in $\mathbb{K}_+$ have $\Pi(M_n, M) \to 0$ as $n \to \infty$. Then use Lemma 12 and the triangle inequality to get

$$\begin{aligned} \operatorname{tr}(AM_n(t_{M_n}^\Sigma - I)) &\leq \|A\|_2 \Pi(M_n, 0)\Pi(M_n, \Sigma) \\ &\leq \|A\|_2 \Pi(M_n, 0)(\Pi(M_n, M) + \Pi(M, \Sigma)) \end{aligned}$$

for all $\Sigma \in \mathbb{K}_+$. Now $\{M_n\}_{n\in\mathbb{N}}$ converging implies that it is bounded, hence

$$\gamma := \max\left\{\sup_{n\in\mathbb{N}} \Pi(M_n, 0), \sup_{n\in\mathbb{N}} \Pi(M_n, M)\right\} < \infty.$$

Therefore, we have the uniform upper bound:

$$\sup_{n\in\mathbb{N}} \exp \operatorname{tr}(AM_n(t_{M_n}^\Sigma - I)) \leq \exp\left(\gamma^2 \|A\|_2\right) \exp\left(\gamma \|A\|_2 \Pi(M, \Sigma)\right).$$

Condition (ii) guarantees that the right side is P-integrable over $\Sigma \in \mathbb{K}_+$, so dominated convergence applies, and this proves (a).

Next we prove the coercivity statement (b). To do this, note that for arbitrary $M \in \mathbb{K}_+$ we can take $A := I/\Pi(M, 0)$ and compute:

$$\begin{aligned} I_P(M) &\geq -\log \int \exp\left(\frac{\operatorname{tr}(M(t_M^\Sigma - I))}{\Pi(M, 0)}\right) \mathrm{d}P(\Sigma) \\ &= \Pi(M, 0) - \log \int \exp\left(\frac{\operatorname{tr}\left((M^{1/2}\Sigma M^{1/2})^{1/2}\right)}{\Pi(M, 0)}\right) \mathrm{d}P(\Sigma), \end{aligned}$$

where we used $\operatorname{tr}(M)/\Pi(M, 0) = \Pi(M, 0)$ and $\operatorname{tr}(Mt_M^\Sigma) = \operatorname{tr}((M^{1/2}\Sigma M^{1/2})^{1/2})$. Now Lemma 11 yields

$$\int \exp\left(\frac{\operatorname{tr}\left((M^{1/2}\Sigma M^{1/2})^{1/2}\right)}{\Pi(M, 0)}\right) \mathrm{d}P(\Sigma) \leq \int \exp \Pi(\Sigma, 0)\, \mathrm{d}P(\Sigma),$$

and the right side is finite by assumption (ii). We therefore have

$$I_P(M) \geq \Pi(M, 0) - \log \int \exp \Pi(\Sigma, 0)\, \mathrm{d}P(\Sigma),$$

and this shows that $I_P(M) \to \infty$ as $M \to \infty$.

For (c), we simply use Jensen's inequality and the fixed point equation for M^* to get:

$$\begin{aligned} I_P(M^*) &= \sup_{A \in \mathbb{S}} -\log \int_{\mathbb{K}} \exp \operatorname{tr}\left(AM(t_M^\Sigma - I)\right) \mathrm{d}P(\Sigma) \\ &\le \sup_{A \in \mathbb{S}} -\int_{\mathbb{K}} \operatorname{tr}\left(AM(t_M^\Sigma - I)\right) \mathrm{d}P(\Sigma) \\ &= \sup_{A \in \mathbb{S}} -\operatorname{tr}\left(AM \int_{\mathbb{K}} (t_M^\Sigma - I)\, \mathrm{d}P(\Sigma)\right) = 0, \end{aligned}$$

as claimed.

Lastly, we prove (d). To do this, notice that the optimization problem

$$\begin{cases} \text{maximize} & -\log \int_{\mathbb{K}} \exp \operatorname{tr}\left(AM(t_M^\Sigma - I)\right) \mathrm{d}P(\Sigma) \\ \text{over} & A \in \mathbb{S} \end{cases}$$

is strictly concave and hence has at most one solution. To characterize this solution through its first-order conditions, we will simply take the gradient of the objective with respect to A. Of course, we already have

$$\nabla_A f(A, \Sigma, M)(H) = \operatorname{tr}(HM(t_M^\Sigma - I)) f(A, \Sigma, M),$$

but we now want to differentiate under the integral. To do this, note that we have

$$\begin{aligned} &\left| \frac{f(A + \delta H, \Sigma, M) - f(A, \Sigma, M)}{\delta} - \operatorname{tr}(HM(t_M^\Sigma - I)) f(A, \Sigma, M) \right| \\ &= \exp \operatorname{tr}\left(AM(t_M^\Sigma - I)\right) \left| \frac{\exp(\delta \operatorname{tr}(HM(t_M^\Sigma - I))) - 1 - \delta \operatorname{tr}(HM(t_M^\Sigma - I))}{\delta} \right| \\ &\le \exp \operatorname{tr}\left(AM(t_M^\Sigma - I)\right) \frac{\delta}{2} |\operatorname{tr}(HM(t_M^\Sigma - I))|^2 \exp(\delta |\operatorname{tr}(HM(t_M^\Sigma - I))|), \end{aligned}$$

using the Taylor series remainder bound $|e^{t\delta} - 1 - t\delta| \le \frac{1}{2} t^2 \delta^2 \exp(|t|\delta)$ for all $t, \delta \in \mathbb{R}$. Now for $0 < \delta < 1$ we use the generous bound $t^2 \le e^{2|t|}$ along with Lemma 12 to further this as

$$\begin{aligned} &\sup_{0<\delta<1} \left| \frac{f(A + \delta H, \Sigma, M) - f(A, \Sigma, M)}{\delta} - \operatorname{tr}(HM(t_M^\Sigma - I)) f(A, \Sigma, M) \right| \\ &\le \frac{1}{2} \exp\left(|\operatorname{tr}\left(AM(t_M^\Sigma - I)\right)| + 3|\operatorname{tr}(HM(t_M^\Sigma - I))| \right) \\ &\le \frac{1}{2} \exp\left((\|A\|_2 + 3\|H\|_2) \Pi(M, 0) \Pi(M, \Sigma) \right). \end{aligned}$$

By the integrability condition (ii), the right side is integrable with respect to P when Σ ranges over $\mathbb{K}_+$. Therefore, we can apply dominated convergence to differentiate under the integral, yielding:

$$\nabla_A \log \int_{\mathbb{K}} \exp \operatorname{tr}\left(AM(t_M^\Sigma - I)\right) \mathrm{d}P(\Sigma) = \frac{\int_{\mathbb{K}} M(t_M^\Sigma - I) \exp \operatorname{tr}\left(AM(t_M^\Sigma - I)\right) \mathrm{d}P(\Sigma)}{\int_{\mathbb{K}} \exp \operatorname{tr}\left(AM(t_M^\Sigma - I)\right) \mathrm{d}P(\Sigma)}$$
$$= M \int_{\mathbb{K}} (t_M^\Sigma - I) \, \mathrm{d}P^{M\to A}(\Sigma).$$

Since $M \in \mathbb{K}_+$ is invertible, this shows that A is a stationary point only if

$$\int_{\mathbb{K}} (t_M^\Sigma - I) \, \mathrm{d}P^{M\to A}(\Sigma) = 0. \tag{5.9}$$

But this is exactly the fixed-point equation which states that M is the the Bures-Wasserstein barycenter of $P^{M\to A}$. □

From this result we can determine a geometric interpretation of the rate function: Fix a point $M \in \mathbb{K}_+$. Then notice that I_P can equivalently be written as

$$I_P(M) = \sup_{A\in \operatorname{Tan}(M)} -\log \int_{\mathbb{K}} \exp\langle A, \log_M(\Sigma)\rangle_{\operatorname{Tan}(M)} \, \mathrm{d}P(\Sigma).$$

That is, $I_P(M)$ is equal to the Fenchel-Legendre transform $(\Lambda_P^M)^*(0)$, where $\Lambda_P^M : \operatorname{Tan}(M) \to \mathbb{R}$ is

$$\Lambda_P^M(A) := \log \int_{\mathbb{K}} \exp\langle A, \log_M(\Sigma)\rangle_{\operatorname{Tan}(M)} \, \mathrm{d}P(\Sigma)$$

for $A \in \operatorname{Tan}(M)$. In fact, Λ_P^M is nothing more than the cumulant generating function for the pushforward measure $(\log_M)_* P \in \mathcal{P}(\operatorname{Tan}(M), \langle\,\cdot\,,\,\cdot\,\rangle_{\operatorname{Tan}(M)})$. Thus, the large deviations rate function $I_P(M)$ is very similar to the usual Euclidean large deviations rate function, after lifting P to the tangent space at M To illustrate this construction, we refer the reader to Figure 5.2.

Our main result can now be stated as follows.

Theorem 12 (LDP for BW barycenters)**.** *For all Borel measurable $S \subseteq \mathbb{K}_+$, we have*

$$-\inf\{I_P(M) : M \in S^\circ\} \le \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(M_n^* \in S)$$
$$\le \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(M_n^* \in S) \le -\inf\left\{I_P(M) : M \in \bar{S}\right\},$$

where S° and $\bar{S}$ denote the interior and closure of S with respect to Π.

Proof. By [206, Theorem 1], the moment condition on P implies that the empirical measures $\{\bar{P}_n\}_{n\in\mathbb{N}}$ satisfies a LDP in $\mathcal{P}_1(\mathbb{K})$ with good rate function $H(\,\cdot \mid P) : \mathbb{K} \to [0,\infty]$, when $\mathcal{P}_1(\mathbb{K})$ is endowed with the topology of the 1-Wasserstein metric. Now use [123, Theorem 2.1] to see that the map $M_2 : \mathcal{P}_1(\mathbb{K}_+) \to \mathbb{K}_+$ is well-defined and Theorem 8 to see that it is

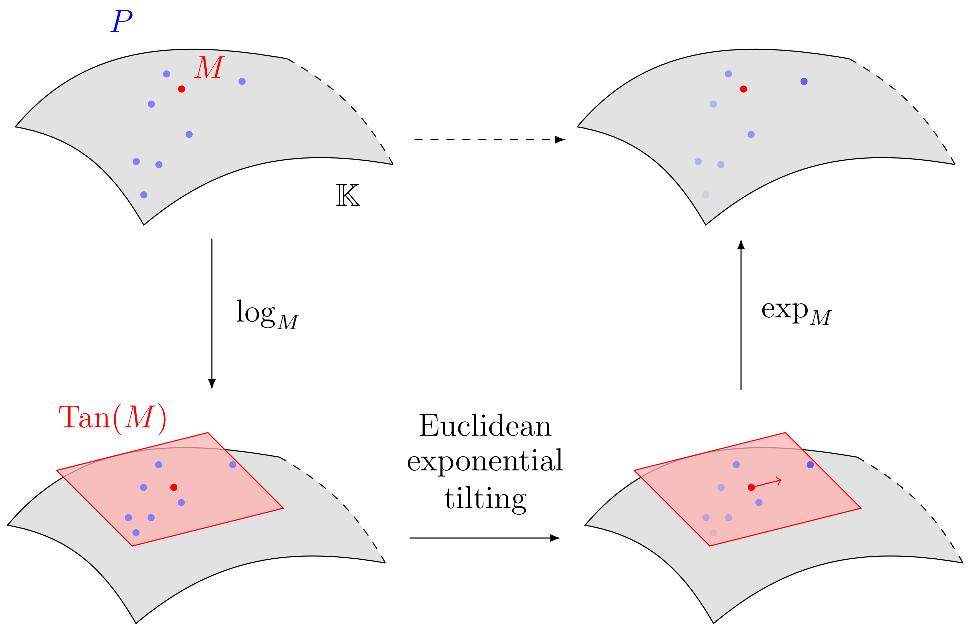

Figure 5.2: The rate function of the large deviations principle arises from exponential tilting in the tangent bundle of the Bures-Wasserstein space. For a probability measure P on $(\mathbb{K}, \Pi)$ and a point $M \in \mathbb{K}$ (top left), the value $I_P(M)$ is equal to the minimal cost of changing measure from P to a distribution whose Bures-Wasserstein barycenter is M (top right). To do this, we lift P from $(\mathbb{K}, \Pi)$ to $\mathrm{Tan}(M)$ (bottom left) and we change measure via exponential tilting in the direction of A so that the tilted lifted measure is centered (bottom right).

continuous. Since $M^* = M_2(P)$ and $M_n^* = M_2(\bar{P}_n)$ for all $n \in \mathbb{N}$, the contraction principle [60, Theorem 4.2.1] implies that $\{M_n^*\}_{n\in\mathbb{N}}$ satisfies a LDP in $(\mathbb{K}_+, \Pi)$ with good rate function J_P given by the relative entropy projection over the space of probability measures which have a fixed barycenter, that is

$$J_P(M) := \inf\{H(Q \,|\, P) : Q \in \mathcal{P}_1(\mathbb{K}_+), M_2(Q) = M\}.$$

for all $M \in \mathbb{K}_+$. In other words, $J_P(M)$ is exactly the value of the optimization problem

$$\begin{cases} \text{minimize} & H(Q \,|\, P) \\ \text{over} & Q \in \mathcal{P}_1(\mathbb{K}_+) \\ \text{subject to} & M_2(Q) = M. \end{cases} \tag{5.10}$$

It only remains to show that $J_P = I_P$, so let us fix $M \in \mathbb{K}_+$. Since $\log_M : \mathbb{K}_+ \to \mathrm{Tan}(M)$ is a well-defined bijection, we can reparameterize (5.10) in the following ways: first, we use the fixed point equation (5.5) to see that for every $Q \in \mathcal{P}_1(\mathbb{K}_+)$, the constraint $M_2(Q) = M$ is equivalent to $\int_{\mathrm{Tan}(M)} A \, \mathrm{d}((\log_M)_\# Q)(A) = 0$; second we use Lemma 12 to see that the integrability constraint $(\log_M)_\# Q \in \mathcal{P}_1(\mathrm{Tan}(M))$ is equivalent to $Q \in \mathcal{P}_1(\mathbb{K}_+)$; third, we have $H(\log_M)_\# Q \,|\, (\log_M)_\# P) = H(Q \,|\, P)$ for all $P, Q \in \mathcal{P}(\mathbb{K}_+)$. Thus, if we change variables to $\tilde{Q} := (\log_M)_\# Q \in \mathcal{P}_1(\mathrm{Tan}(M))$, then we have shown that the value of (5.10) is equal to the value of

$$\begin{cases} \text{minimize} & H(\tilde{Q} \,|\, (\log_M)_\# P) \\ \text{over} & \tilde{Q} \in \mathcal{P}_1(\mathrm{Tan}(M)) \\ \text{subject to} & \int_{\mathrm{Tan}(M)} A \, \mathrm{d}\tilde{Q}(A) = 0. \end{cases} \tag{5.11}$$

Now we apply Lemma 13 in the Hilbert space $(\mathrm{Tan}(M), \langle \cdot\,, \cdot \rangle_{\mathrm{Tan}(M)})$ to conclude. □

As an immediate application of Theorem 12, we get the following: If $K \subseteq \mathbb{K}_+$ is a compact set equal to the closure of its interior and satisfying $M^* \notin K$, then

$$\frac{1}{n} \log \mathbb{P}(M_n^* \in K) \to -\min \{I_P(M) : M \in K\} =: -I_P(K),$$

and $I_P(K) > 0$. Roughly speaking, this means we have

$$\mathbb{P}(M_n^* \in K) \approx \exp(-n I_P(K))$$

for large n, ignoring sub-exponential factors. In other words, the minimization of I_P over K exactly characterizes the exponential rate of decay of the probability of the rare event $\{M_n^* \in K\}$. In the next section we give some applications of Theorem 12 using this form.

5.3 Two Examples

To demonstrate the utility of Theorem 12, we give two extended applications of the result: one analytical for which we can derive simple upper bounds, and one numerical for which we can implement an algorithm to approximately compute the large deviations behavior.

Approximations of the Fixed-Point Equation

As always, let us assume that P satisfies conditions (i) and (ii), and let us define the map $L_P : \mathbb{K}_+ \to [0, \infty)$ via

$$L_P(M) = \left\| \int_{\mathbb{K}} (t_M^\Sigma - I) \, \mathrm{d}P(\Sigma) \right\|_{\mathrm{Tan}(M)}.$$

Recall that condition (i) implies that $L_P(M) = 0$ if and only if $M = M^*$. Thus, $L_P(M)$ is a particular way to quantify the degree to which M fails to satisfy the fixed-point equation for the barycenter of P. By straightforward continuity arguments, one can show $L_P(M_n^*) \to 0$ almost surely as $n \to \infty$, but it is also important to develop large deviations estimates for this convergence. For example, in a testing problem in which the null hypothesis is that an unknown population distribution P has some fixed Bures-Wasserstein barycenter $M_0 \in \mathbb{K}_+$, one can use the large deviations principle to analyze the asymptotic relative efficiency of a test based on the empirical Bures-Wasserstein barycenter. In this setting we can give the following result:

Theorem 13. *If $r > 0$ satisfies $P\{\Sigma \in \mathbb{K} : \Pi(M^*, \Sigma) \leq r\} = 1$, then for all $t \geq 0$ we have*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(L_P(M_n^*) \geq t) \leq -\frac{t^2}{8r^2}.$$

.

Proof. First, we claim that the effective domain of I_P (that is, the set of all points at which its value is finite) satisfies $\mathrm{dom}(I_P) \subseteq B_{2r}(M^*)$. To see this, suppose $M \in \mathbb{K}_+$ satisfies $I_P(M) < \infty$ and assume for the sake of contradiction that $\Pi(M, M^*) > 2r$. By Proposition 3(d), there exists $A \in \mathbb{S}$ such that M is the Bures-Wasserstein barycenter of $P^{M\to A}$, and we of course have $\mathrm{supp}(P) = \mathrm{supp}(P^{M\to A})$. Then we have

$$\begin{aligned} \int_{\mathbb{K}} \Pi^2(M,\Sigma)\,\mathrm{d}P^{M\to A}(\Sigma) &\geq \int_{\mathbb{K}} (\Pi(M,M^*) - \Pi(M^*,\Sigma))^2\,\mathrm{d}P^{M\to A}(\Sigma) \\ &> \int_{\mathbb{K}} (2r - r)^2\,\mathrm{d}P^{M\to A}(\Sigma) \\ &= r^2 \end{aligned}$$

and

$$\int_{\mathbb{K}} \Pi^2(M^*,\Sigma)\,\mathrm{d}P^{M\to A}(\Sigma) \leq r^2,$$

which implies M is not the Bures-Wasserstein barycenter of $P^{M\to A}$. Since this is a contradiction, we conclude $\Pi(M, M^*) \leq 2r$.

Second, we fix $t > 0$ and $M \in \mathrm{dom}(I_P)$, and we consider arbitrary $U \in \mathrm{Tan}(M)$ with $\|U\|_{\mathrm{Tan}(M)} = 1$ and $\lambda \geq 0$. Then for any $\Sigma \in \mathbb{K}$ we use Cauchy-Schwarz and $\mathrm{dom}(I_P) \subseteq B_{2r}(M^*)$ to get:

$$|\mathrm{tr}(UM(t_M^\Sigma - I)| \leq \|(t_M^\Sigma - I)M^{1/2}\|_2 = \Pi(M,\Sigma) \leq 2r.$$

In particular, the real-valued random variable $\mathrm{tr}(UM(t_M^\Sigma - I)$ lies in $[-2r, 2r]$ almost surely. Therefore, we can apply Hoeffding's lemma to get:

$$-\log \int_{\mathbb{K}} \exp(\lambda \mathrm{tr}(UM(t_M^\Sigma - I)))\,\mathrm{d}P(\Sigma) \geq -\left(\lambda \mathrm{tr}\left(UM \int_{\mathbb{K}} (t_M^\Sigma - I)\,\mathrm{d}P(\Sigma)\right) + 2\lambda^2 r^2\right)$$

for all $\lambda \in \mathbb{R}$. In particular, by taking the supremum over $\lambda \in \mathbb{R}$ and doing some calculus, we get

$$\sup_{\lambda\in\mathbb{R}} -\log \int_{\mathbb{K}} \exp(\lambda \mathrm{tr}(UM(t_M^\Sigma - I)))\,\mathrm{d}P(\Sigma) \geq \frac{1}{8r^2}\left(\mathrm{tr}\left(UM \int_{\mathbb{K}} (t_M^\Sigma - I)\,\mathrm{d}P(\Sigma)\right)\right)^2.$$

Now we consider taking the supremum over all $U \in \mathbb{S}$ with $\|UM^{1/2}\|_2 = 1$. By Cauchy-Schwarz we have exactly:

$$\sup_{\substack{U \in \mathrm{Tan}(M) \\ \|U\|_{\mathrm{Tan}(M)}=1}} \mathrm{tr}\left(UM \int_{\mathbb{K}} (t_M^\Sigma - I)\,\mathrm{d}P(\Sigma)\right) = \left\| \int_{\mathbb{K}} (t_M^\Sigma - I)\,\mathrm{d}P(\Sigma) \right\|_{\mathrm{Tan}(M)} = L_P(M)$$

hence

$$\begin{aligned} I_P(M) &= \sup_{A \in \mathrm{Tan}(M)} -\log \int_{\mathbb{K}} \exp(\lambda \mathrm{tr}(UM(t_M^\Sigma - I)))\,\mathrm{d}P(\Sigma) \\ &= \sup_{\substack{U \in \mathrm{Tan}(M) \\ \|U\|_{\mathrm{Tan}(M)}=1}} \sup_{\lambda \in \mathbb{R}} -\log \int_{\mathbb{K}} \exp(\lambda \mathrm{tr}(UM(t_M^\Sigma - I)))\,\mathrm{d}P(\Sigma) \\ &\geq \sup_{\substack{U \in \mathrm{Tan}(M) \\ \|U\|_{\mathrm{Tan}(M)}=1}} \frac{1}{8r^2} \left(\mathrm{tr}\left(UM \int_{\mathbb{K}} (t_M^\Sigma - I)\,\mathrm{d}P(\Sigma)\right)\right)^2 \\ &= \frac{(L_P(M))^2}{8r^2}, \end{aligned}$$

by monotonicity.

Finally, for any $t \geq 0$, we can bound

$$\inf\{I_P(M) : L_P(M) \geq t\} = \inf\{I_P(M) : M \in \mathrm{dom}(I_P), L_P(M) \geq t\} \geq \frac{t^2}{8r^2}$$

so Theorem 12 gives the result. □

We regard this as a Hoeffding-type concentration inequality, since it guarantees sub-Gaussian like concentration of a statistic of interest for bounded distributions P. However, we emphasize that $L_P(M_n^*)$ is a very complicated function of $\Sigma_1, \ldots, \Sigma_n$, so we are not aware of any way to prove it via elementary considerations.

Excess Trace in a Generative Model

Fix $M^* \in \mathbb{K}_+$ and suppose that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space supporting an IID sequence $T_1, T_2, \ldots$ of random matrices coming from a Wishart distribution with scale matrix I and shape k, normalized by k. (So, we have $\mathbb{E}[T_i] = (kI)/k = I$ for all $i \in \mathbb{N}$.) Then define $\Sigma_i := TM^*T$ for all $i \in \mathbb{N}$, and write P for their common distribution and $\bar{P}_n$ for their empirical measure. It follows from the fixed-point equation (5.4) that M^* is the Bures-Wasserstein barycenter of the population distribution P, and we refer to this as a *generative model* around M^*. We also note that the parameter $k \geq 0$ controls the concentration of $\bar{M}_n^*$ around M^*, since larger values of k lead to more concentration of T around I.

Now fix $t > 0$ and let us focus on the trace upper tail event given by

$$F_t^+ := \{\mathrm{tr}(M_n^*) \geq \mathrm{tr}(M^*)(1+t)\},$$

Observe that $F_t^+ = \{M_n^* \in S_t^+\}$, where $S_t^+ = \{M \in \mathbb{K}_+ : \mathrm{tr}(M) \geq \mathrm{tr}(M^*)(1+t)\}$, and that S_t^+ is equal to the closure of its interior since it is convex. Therefore, by Theorem 12 we have

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\mathrm{tr}(M_n^*) \geq \mathrm{tr}(M^*)(1+t)) = -\inf\left\{I_P(M) : \mathrm{tr}(M) \geq \mathrm{tr}(M^*)(1+t)\right\}.$$

In fact, we can numerically approximate the right side above via a form of projected Riemannian gradient descent. That is, we initialize $M_0 \in S_t^+$ arbitrarily, and, for a given stepsize $\eta > 0$, we iterate

$$M_{i+1} := \mathrm{proj}_\Pi(\exp_{M_i}(M_i - \eta \nabla_{M_i} I_P); S_t^+).$$

for $i \in \mathbb{N}$, where the projection operation is given by

$$\mathrm{proj}_\Pi(Q; S_t^+) = Q \max\left\{\frac{\mathrm{tr}(M^*)}{\mathrm{tr}(Q)}(1+t), 1\right\}.$$

Thus, it only remains to determine the gradient of the rate function.

We give a heuristic argument for how to evaluate this gradient. Since P is fully-supported, Proposition 3(d) shows that there exists a well-defined function $A_P : \mathbb{K}_+ \to \mathbb{S}$ such that M is the Bures-Wasserstein barycenter of $P^{M\to A_P(M)}$ for all $M \in \mathbb{K}_+$; in fact, one can use the implicit function theorem to show that A_P is continuously differentiable. In particular, for all $M \in \mathbb{K}_+$, we can write

$$I_P(M) = -\log \int_{\mathbb{K}_+} \exp \mathrm{tr}(A_P(M) M (t_M^\Sigma - I))\, \mathrm{d}P(\Sigma),$$

and it follows that the gradient of I_P at M is the linear operator which, when evaluated at H, yields

$$\begin{aligned}(\nabla_M I_P)(H) = &\int_{\mathbb{K}_+} \mathrm{tr}((\nabla_M A_P)(H) M (t_M^\Sigma - I))\, \mathrm{d}P^{M\to A_P(M)}(\Sigma)\\ &+ \int_{\mathbb{K}_+} \mathrm{tr}(A_P(M) H (t_M^\Sigma - I))\, \mathrm{d}P^{M\to A_P(M)}(\Sigma)\\ &+ \int_{\mathbb{K}_+} \mathrm{tr}(A_P(M) M (\nabla_M T^\Sigma)(H))\, \mathrm{d}P^{M\to A_P(M)}(\Sigma).\end{aligned}$$

Since $\int_{\mathbb{K}_+}(t_M^\Sigma - I)\, \mathrm{d}P^{M\to A_P(M)}(\Sigma) = 0$, the first two terms above vanish, and it follows that

$$(\nabla_M I_P)(H) = \mathrm{tr}\left(A_P(M) M \int_{\mathbb{K}_+} \nabla_M T^\Sigma(H))\, \mathrm{d}P^{M\to A_P(M)}(\Sigma)\right).$$

Algorithm 1 Some procedures for numerically evaluating the rate function I_P and its gradient. Here, we use capital letters to denote matrices in $\mathbb{R}^{m\times m}$ and calligraphic letters (that is, `mathcal` letters) to denote linear maps $\mathbb{R}^{m\times m} \to \mathbb{R}^{m\times m}$. We write $\{E^{ij}\}_{i,j=1,\dots,m}$ for the standard basis of $\mathbb{R}^{m\times m}$, so that E^{ij} equals 1 in the ij entry and equals 0 in all other entries.

procedure TRANSPORT
  input: covariances $M \in \mathbb{K}_+$ and $\Sigma \in \mathbb{K}$
  output: optimal transport map $t_M^\Sigma \in \mathbb{R}^{m\times m}$
  return $\Sigma^{1/2}(\Sigma^{-1/2}M\Sigma^{1/2})^{-1/2}\Sigma^{1/2}$

procedure TRANSPORTJAC
  input: covariances $M \in \mathbb{K}_+$ and $\Sigma \in \mathbb{K}$
  output: linear operator $\nabla T_M^\Sigma : \mathbb{R}^{m\times m} \to \mathbb{R}^{m\times m}$
  $D, U \leftarrow$ SPECTRALDECOMPOSITON$(\Sigma^{\frac{1}{2}} M \Sigma^{\frac{1}{2}})$ ▷ So, $U^T D U = \Sigma^{\frac{1}{2}} M \Sigma^{\frac{1}{2}}$
  for $i, j = 1, \dots, m$ **do**
    $\Delta \leftarrow U\Sigma^{\frac{1}{2}} E^{ij} \Sigma^{\frac{1}{2}} U^T$
    for $k, \ell = 1, \dots, m$ **do**
      $\bar{\Delta}_{k\ell} \leftarrow \Delta_{k\ell}/(\sqrt{D_{kk}} + \sqrt{D_{\ell\ell}})$
    $\mathcal{G}_{\cdot,ij} \leftarrow -\Sigma^{\frac{1}{2}} U^T D^{-\frac{1}{2}} \bar{\Delta} D^{-\frac{1}{2}} U \Sigma^{\frac{1}{2}}$
  return $\mathcal{G}$

procedure RATEFUNCTIONGRADIENT$_P$
  input: covariance $M \in \mathbb{K}_+$
  output: linear functional $\nabla I_P(M) : \mathbb{R}^{m\times m} \to \mathbb{R}$
  $(A, \Lambda) \leftarrow$ **minimize** $\log \int_{\mathbb{K}} \exp \operatorname{tr}(AM(\text{TRANSPORT}(M, \Sigma) - I))\, \mathrm{d}P(\Sigma)$
      over $A \in \mathbb{S}$
  $\mathcal{L} \leftarrow \int_{\mathbb{K}} \text{TRANSPORTJAC}(M, \Sigma) \exp(\operatorname{tr}(AM(\text{TRANSPORT}(M, \Sigma) - I)) - \Lambda)\, \mathrm{d}P(\Sigma)$
  for $i, j = 1, \dots, m$ **do**
    $G_{ij} \leftarrow \operatorname{tr}(AM\mathcal{L}(E^{ij}))$
  return G

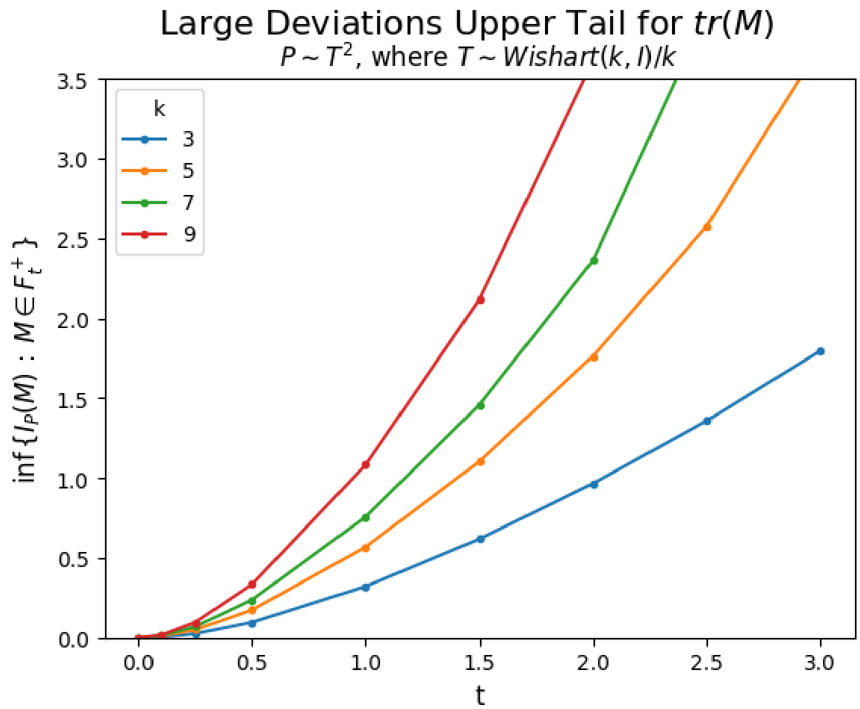

Figure 5.3: The largest deviations behavior for the excess trace in the generative model, across various values of the excess t and the scale k. We observe that the dependence on t is roughly quadratic, and that the quadratic coefficient depends on k.

In fact, the Jacobian $\nabla_M T^\Sigma$ has been computed in closed-form in [123, Lemma A.2]. Thus, we can write down Algorithm 1, which computes $\nabla_M I_P$.

Putting this all together, we can compute the large deviations exponent for a range of k and t values. We direct the reader to Figure 5.3 for a plot of the results, which show approximately quadratic dependence on the excess t, with a coefficient depending on the scale k.

Chapter 6

Fréchet Mean Set Estimation

In Theorem 9 in Chapter 4 we showed that, if $(\mathcal{X}, d)$ is a nice enough metric space and $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space supporting an IID sequence $Y_1, Y_2, \ldots$ of $\mathcal{X}$-valued random variables with common distribition $P \in \mathcal{P}_{p-1}(\mathcal{X})$ for $p \geq 1$, then we have

$$\max_{\bar{x}_n \in M_p(\bar{P}_n)} \min_{x \in M_p(P)} d(\bar{x}_n, x) \to 0 \tag{6.1}$$

holding $\mathbb{P}$-almost surely, where $\bar{P}_n$ is (as always) the empirical distribution of the first $n \in \mathbb{N}$ samples, $\bar{P}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{Y_i}$. From a statistical point of view, we regard (6.1) as an asymptotic guarantee of "no false positives" since it shows that each $\bar{x}_n \in M_p(\bar{P}_n)$ is close to some $x \in M_p(P)$. It is thus natural to wonder whether we additionally have

$$\max_{x \in M_p(P)} \min_{\bar{x}_n \in M_p(\bar{P}_n)} d(\bar{x}_n, x) \to 0 \tag{6.2}$$

holding $\mathbb{P}$-almost surely; we regard (6.2) as an asymptotic guarantee of "no false negatives" since it shows that each $x \in M_p(P)$ is close to some $\bar{x}_n \in M_p(\bar{P}_n)$.

It turns out that (6.1) and (6.2) together are equivalent to convergence in the so-called *Hausdorff metric*. To define this, let us write $\mathrm{K}(\mathcal{X})$ for the collection of all non-empty compact subsets of $\mathcal{X}$, and for $K, K' \in \mathrm{K}(\mathcal{X})$, we write

$$d_{\mathrm{H}}(K, K') := \max\left\{\max_{x \in K} \min_{x' \in K'} d(x, x'), \max_{x' \in K'} \min_{x \in K} d(x, x')\right\}.$$

Equivalently, $d_{\mathrm{H}}(K, K')$ is the smallest $r \geq 0$ such that the r-thickening of K contains K' and such that the r-thickening of K' contains K. Thus, we are interested in consistency results of the form

$$d_{\mathrm{H}}(M_p(P), M_p(\bar{P}_n)) \to 0 \tag{6.3}$$

holding $\mathbb{P}$-almost surely, which we refer to as d_{H}-*consistency*.

Several previous authors ([94, p. 1118], [180, Remark 2.5], and [198, p. 60]) have sought d_{H}-consistency results for Fréchet means, but this turns out to be a rather delicate property.

For example, it was previously unknown whether empirical Fréchet means are d_{H}-consistent or whether it was possible to construct any estimator satisfying d_{H}-consistency.

This chapter is based on the works [66, 37] and provides an authoritative answer to these questions. We begin with a negative result which shows that the empirical Fréchet mean is not a d_{H}-consistent estimator, and hence is asymptotically inadmissible for any loss function based on d_{H} (Theorem 14). We then construct the class of *relaxed empirical Fréchet mean set estimators* and we give results controlling their weak (Theorem 15) and strong (Theorem 16) error probabilities; these results are based on some functional limit theorems, including a functional central limit theorem and a functional law of the iterated logarithm. Moreover, in Chapter 7 we give a concrete application of these results to a problem in computational phylogenetics.

6.1 Preliminaries

Throughout this section (in fact, throughout the entire chapter), we assume that $p \geq 1$ and that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space supporting an IID sequence $Y_1, Y_2, \ldots$ of $\mathcal{X}$-valued random variables with common distribution $P \in \mathcal{P}(\mathcal{X})$. We write $\mathbb{E}, \mathrm{Var}$, and Cov for the expectation, variance, and covariance on this space; we add subscripts of P when necessary in order to emphasize that the distribution of Y_i is P for all $i \in \mathbb{N}$. Moreover, we will impose further geometric conditions on $(\mathcal{X}, d)$ and integrability conditions on P when we need them.

First we discuss geometric conditions on $(\mathcal{X}, d)$. As in all theory for Fréchet means, our results require certain properties of this underlying metric space $(\mathcal{X}, d)$. As we discussed in Chapter 4, a common assumption is for $(\mathcal{X}, d)$ to be a Heine-Borel space, meaning its closed balls are all compact. While we saw that this condition is not necessary for the no false positives property (6.1), it turns out that in order to achieve d_{H}-consistency (6.3) we need to assume this and slightly more. Thus, we introduce the following notion:

Definition 8. *For a metric space $(\mathcal{X}, d)$ and $\varepsilon \geq 0$, we write $\mathcal{N}_{(\mathcal{X},d)}(\varepsilon)$ or simply $\mathcal{N}_{\mathcal{X}}(\varepsilon)$ for the smallest number of d-balls of radius $\varepsilon \geq 0$ needed to cover $\mathcal{X}$. A metric space $(\mathcal{X}, d)$ is called a* Dudley space *if it is compact and*

$$\int_0^\infty \sqrt{\log \mathcal{N}_{(\mathcal{X},d)}(\varepsilon)}\, d\varepsilon < \infty.$$

A metric space is called a Heine-Borel-Dudley space *if all of its closed balls are Dudley with respect to the inherited metric.*

Evidently, every Heine-Borel-Dudley space is a Heine-Borel space so it admits a weak-like topology. In particular, the asymptotic results of Chapter 4 all hold when $(\mathcal{X}, d)$ is a Heine-Borel-Dudley space. As we will soon see, the sort of finite-dimensionality afforded by the Heine-Borel-Dudley property is primarily used to ensure that Gaussian distributions on $C(\mathcal{X})$, the space of continuous functions from $(\mathcal{X}, d)$ to $\mathbb{R}$, are sufficiently well-behaved.

Second we discuss integrability conditions on P. More precisely, we define, for each $i \in \mathbb{N}$, the random function $Z_i : \mathcal{X}^2 \to \mathbb{R}$ via

$$Z_i(x, x') := d^p(x, Y_i) - d^p(x', Y_i) - W_p(P, x, x')$$

for $x, x' \in \mathcal{X}$; of course, we need $P \in \mathcal{P}_{p-1}(\mathcal{X})$ in order for the third term to be well-defined, in which case each Z_i is a random centered function. It turns out, as the following technical result shows, that this integrability also implies several useful continuity estimates for $\{Z_i\}_{i\in\mathbb{N}}$. Here, we write $\operatorname{diam}(S) := \sup_{x,x' \in S} d(x, x')$ for the *diameter* of an arbitrary subset $S \subseteq \mathcal{X}$.

Lemma 14. *If $P \in \mathcal{P}_{p-1}(\mathcal{X})$, then for all $i \in \mathbb{N}$ and all compacts $K \subseteq \mathcal{X}$, the function $Z_i : K^2 \to \mathbb{R}$ is $\mathbb{P}$-almost surely locally Lipschitz with respect to the metric $D((x_1, x_1'), (x_2, x_2')) = d(x_1, x_1') + d(x_2, x_2')$ and its local Lipschitz constant $M_{i,K}$ satisfies*

$$M_{i,K} \leq 2pc_{p-1} \min_{o \in K} \left(d^{p-1}(o, Y_i) + W_{p-1}(P, o) + 2(\operatorname{diam}(K))^{p-1} \right)$$

$\mathbb{P}$-almost surely, where c_{p-1} is the constant appearing in (4.2).

Proof. We simply use (4.3) twice to get

$$\begin{aligned} |Z_i(x_1, x_1') - Z_i(x_2, x_2')| \leq\ & pd(x_1, x_2) \left(d^{p-1}(x_1, Y_i) + d^{p-1}(x_2, Y_i) \right) \\ & + pd(x_1', x_2') \left(d^{p-1}(x_1', Y_i) + d^{p-1}(x_2', Y_i) \right) \\ & + pd(x_1, x_2) \left(W_{p-1}(P, x_1) + W_{p-1}(P, x_2) \right) \\ & + pd(x_1', x_2') \left(W_{p-1}(P, x_1') + W_{p-1}(P, x_2') \right). \end{aligned}$$

Now for any $o \in K$ we use (4.2) to further this bound as

$$\begin{aligned} & |Z_i(x_1, x_1') - Z_i(x_2, x_2')| \\ & \leq 2pc_{p-1} D((x_1, x_2), (x_1', x_2')) \left(d^{p-1}(o, Y_i) + W_{p-1}(P, o) + 2(\operatorname{diam}(K))^{p-1} \right). \end{aligned}$$

This finishes the proof. □

Because we will also need to study the interactions between the functions $\{Z_i\}_{i\in\mathbb{N}}$, we require further integrability of P so that certain covariances are well-defined. More precisely, our next object of study is the function $R_P : \mathcal{X}^4 \to \mathbb{R}$ defined via

$$R_P(x, x', x'', x''') := \operatorname{Cov}(Z_i(x, x'), Z_i(x'', x'''))$$

for $x, x', x'', x''' \in \mathcal{X}$ and any $i \in \mathbb{N}$. For simplicity, and by a slight abuse of notation, we also write

$$R_P(x, x') := R_P(x, x', x, x') = \operatorname{Var}(Z_i(x, x'))$$

for $x, x' \in \mathcal{X}$ and any $i \in \mathbb{N}$. As before, we need $P \in \mathcal{P}_{2p-2}(\mathcal{X})$ in order for R_P to make sense, but it turns out that this condition also yields the following continuity estimates:

Lemma 15. *If $P \in \mathcal{P}_{2p-2}(\mathcal{X})$, then for all compacts $K \subseteq \mathcal{X}$, the function $R_P : K^4 \to \mathbb{R}$ is locally Lipschitz with respect to the metric $D((x_1, x_1', x_1'', x_1'''), (x_2, x_2', x_2'', x_2''')) = d(x_1, x_2) + d(x_1', x_2') + d(x_1'', x_2'') + d(x_1''', x_2''')$, and its local Lipschitz constant $M_{P,K}$ satisfies*

$$M_{P,K} \leq 48p^2 c_{p-1}^2 \mathrm{diam}(K) \left(\min_{o \in K} W_{2p-2}(P, o) + (\mathrm{diam}(K))^{2p-2} \right)$$

where c_{p-1} is the constant appearing in (4.2).

Proof. For any $x_1, x_1', x_1'', x_1''', x_2, x_2', x_2'', x_2''' \in K$ and any $i \in \mathbb{N}$, use Lemma 14 twice to get:

$$\begin{aligned}
&|Z_i(x_1, x_1')Z_1(x_1'', x_1''') - Z_i(x_2, x_2')Z_1(x_2'', x_2''')| \\
&\leq |Z_i(x_1, x_1')| \cdot |Z_i(x_1'', x_1''') - Z_i(x_2'', x_2''')| + |Z_i(x_2'', x_2''')| \cdot |Z_i(x_1, x_1') - Z_i(x_2, x_2')| \\
&\leq M_{i,K}^2 \Bigg(d(x_1, x_1')D((x_1'', x_1'''), (x_2'', x_2''')) + d(x_2'', x_2''')D((x_1, x_1'), (x_2, x_2')) \Bigg) \\
&\leq D((x_1, x_1', x_1'', x_1'''), (x_2, x_2', x_2'', x_2''')) \mathrm{diam}(K) M_{i,K}^2.
\end{aligned}$$

Since Z_i is centered, we can take the expectation above and deduce that $R_P : K^4 \to \mathbb{R}$ is locally Lipschitz with Lipschitz constant $M_{P,K}$ satisfying $M_{P,K} \leq \mathrm{diam}(K)\mathbb{E}_P[M_{i,K}^2]$. Then we can bound:

$$\begin{aligned}
\mathbb{E}_P[M_{i,K}^2] &\leq 12p^2 c_{p-1}^2 \min_{o \in K} \left(W_{2p-2}(P, o) + (W_{p-1}(P, o))^2 + 4(\mathrm{diam}(K))^{2p-2} \right) \\
&\leq 24p^2 c_{p-1}^2 \min_{o \in K} \left(\mathbb{E}_P[d^{2p-2}(x_0, Y)] + 2(\mathrm{diam}(K))^{2p-2} \right) \\
&\leq 48p^2 c_{p-1}^2 \min_{o \in K} \left(\mathbb{E}_P[d^{2p-2}(x_0, Y)] + (\mathrm{diam}(K))^{2p-2} \right),
\end{aligned}$$

where we used Cauchy-Schwarz to get $(W_{p-1}(P, o))^2 \leq W_{2p-2}(P, o)$ for all $o \in K$. □

An important consequence of these estimates is that R_P depends continuously on P:

Lemma 16. *If $\{P_n\}_{n \in \mathbb{N}}$ and P in $\mathcal{P}_{2p-2}(\mathcal{X})$ have $P_n \to P$ in τ_{w}^{2p-2}, then $R_{P_n} \to R_P$ uniformly on compact sets.*

Proof. Fix $K \subseteq \mathcal{X}^4$ compact. By Lemma 6, we have

$$\limsup_{n \in \mathbb{N}} \min_{o \in K} W_{2p-2}(P_n, o) \leq \min_{o \in K} \limsup_{n \in \mathbb{N}} W_{2p-2}(P_n, o) = \min_{o \in K} W_{2p-2}(P, o) < \infty.$$

In particular, Lemma 15 shows that $\{R_{P_n}\}_{n \in \mathbb{N}}$ are uniformly Lipschitz on K, so the Arzela-Ascoli theorem guarantees that they have a subsequential limit, in the topology of uniform convergence on K. But we already have $R_{P_n} \to R_P$ pointwise by Lemma 6, so we in fact have $R_{P_n} \to R_P$ uniformly on K. □

Now we combine these geometric considerations and integrability considerations. Our ultimate goal will be to establish limit theorems for the random continuous functions $G_n : \mathcal{X}^2 \to \mathbb{R}$ defined as

$$G_n = \sum_{i=1}^{n} Z_i,$$

for $n \in \mathbb{N}$. Of course, classical limit theorems allow us to analyze the convergence of the sequence $\{G_n(x,x')\}_{n\in\mathbb{N}}$ for any fixed $x, x' \in X$. However, for our later purposes, this will not be enough; we need sufficiently powerful functional limit theorems which allow us to analyze the convergence of the sequence $\{G_n\}_{n\in\mathbb{N}}$ across the whole domain X^2 (or compact subests thereof) simultaneously.

We begin with the functional central limit theorem, which takes place in the space $C(\mathcal{X}^2)$ of continuous functions $\mathcal{X}^2$ to $\mathbb{R}$. Because of the integrability assumption $P \in \mathcal{P}_{2p-2}(\mathcal{X})$, we see that the function R_P is a positive semi-definite kernel on the space $\mathcal{X}^2$. Thus, there exists a unique centered Gaussian measure on $C(\mathcal{X}^2)$ whose covariance structure is given by R_P; we denote this Gaussian measure by $\mathcal{G}_P$. We get the following:

Proposition 4. *If $(\mathcal{X}, d)$ is a Heine-Borel-Dudley space, $p \geq 1$, and $P \in \mathcal{P}_{2p-2}(\mathcal{X})$, then the random functions $\{n^{-1/2}G_n\}_{n\in\mathbb{N}}$ converge in distribution to $\mathcal{G}_P$ with respect to the topology of uniform convergence on compact sets.*

Proof. Fix a compact $K \subseteq \mathcal{X}$ and note that $(\mathcal{X}, d)$ being a Dudley space implies that (K^2, D) is a Dudley space, where D is the metric on $\mathcal{X}^2$ defined via $D((x_1, x_1'), (x_2, x_2')) = d(x_1, x_1') + d(x_2, x_2')$. Then use Lemma 14 to see that each random function $Z_i : K^2 \to \mathbb{R}$ for $i \in \mathbb{N}$ is $M_{i,K}$-Lipschitz, where $\mathbb{E}[M_{i,K}^2] < \infty$. In particular, this implies $\sup_{x,x'\in\mathcal{X}} \mathrm{Var}(Z_i(x, x')) < \infty$. Therefore, the desired convergence follows from [103, Theorem 1]. □

Next we study the functional law of the iterated logarithm, which requires some notation. For each compact subset $K \subseteq \mathcal{X}$ and $P \in \mathcal{P}_{2p-2}(\mathcal{X})$, we write $R_{K,P}$ and $G_{K,n}$ for the restrictions $R_{K,P} := R_P|_{K\times K}$ and $G_{K,n} := G_n|_{K\times K}$, respectively, and we let $\mathcal{H}_{K,P} \subseteq C(K \times K)$ denote the reproducing kernel Hilbert space (RKHS) with kernel $R_{K,P}$; we write $\|\cdot\|_{K,P}$ and $\langle\cdot,\cdot\rangle_{K,P}$ for the norm and inner product of $\mathcal{H}_{K,P}$, and write $\mathcal{B}_{K,P} := \{f \in \mathcal{H}_{K,P} : \|f\|_{K,P} \leq 1\} \subseteq C(K \times K)$ for the unit ball of $\mathcal{H}_{K,P}$. Then we have the following:

Proposition 5. *If $(\mathcal{X}, d)$ is a Heine-Borel-Dudley space, $p \geq 1$, and $P \in \mathcal{P}_{2p-2}(\mathcal{X})$, then, for each compact set $K \subseteq \mathcal{X}$, the random functions $\{(2n \log\log n)^{-1/2}G_{K,n}\}_{n\in\mathbb{N}}$ form a relatively compact set with closure $\mathcal{B}_{K,P}$, with respect to the topology of uniform convergence on K.*

Proof. The result follows from [131, Corollary 1.3] and the calculations appearing in the proof of Proposition 4. □

Lastly, we introduce a fundamental quantity that will appear in both of our main results.

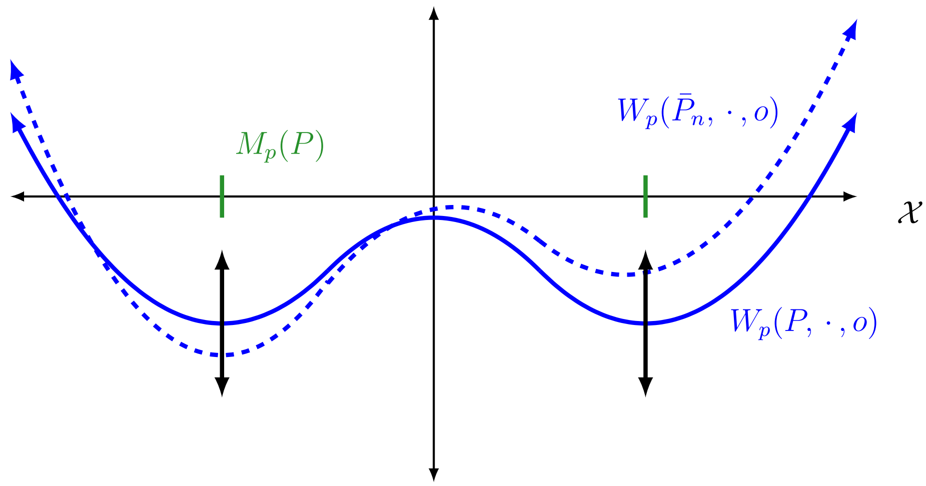

Figure 6.1: The fluctuation variation measures, up to constants, the maximum possible fluctuation of the difference between the values of the Fréchet functional on distinct points in the Fréchet mean set.

Definition 9. *For $P \in \mathcal{P}_{2p-2}(\mathcal{X})$, we write*

$$\sigma_p(P) := \sqrt{2} \sup_{x,x' \in M_p(P)} \sqrt{\mathrm{Var}_P(d^p(x, Y_1) - d^p(x', Y_1))}, \tag{6.4}$$

called the fluctuation variation of P on $M_p(P)$.

Some basic remarks about the fluctuation variation are in order. First of all, using the notation of this section, we can equivalently write $\sigma_p^2(P) := \sup_{x,x' \in M_p(P)} R_P(x, x')$. Now suppose that $(\mathcal{X}, d)$ is Heine-Borel-Dudley (or even just Heine-Borel). In this case, $M_p(P)$ is compact by Lemma 4 and R_P is continuous by Lemma 15, so the supremum is achieved and can be replaced with a maximum; in particular, $\sigma_p(P) < \infty$. Moreover, observe that if $M_p(P)$ is a singleton then we have $\sigma_p(P) = 0$, but that the converse fails; rather, we have $\sigma_p(P) = 0$ if and only if $d^p(x, \cdot) = d^p(x', \cdot)$ holds P-almost everywhere for all $x, x' \in M_p(P)$. Lastly, let us mention that our results will be sharp for the case of $0 < \sigma_p(P) < \infty$, and the analysis seems to be much more complicated when $\sigma_p(P) = 0$.

To close this section, we detail a further connection between the fluctuation variation and the RKHS of interest. It states that the fluctuation variation $\sigma_p(P)$ is equal to the the largest possible (bivariate) point evaluation of functions in the unit ball of R_P on any compact thickening of the set $M_p(P)$

Lemma 17. *If $(\mathcal{X}, d)$ is a Heine-Borel space, $p \geq 1$, and $P \in \mathcal{P}_{2p-2}(\mathcal{X})$, then for any compact $K \subseteq \mathcal{X}$ with $M_p(P) \subseteq K$, we have*

$$\sigma_p(P) = \sqrt{2} \cdot \max_{f \in \mathcal{B}_{K,P}} \max_{x,x' \in M_p(P)} f(x, x').$$

Proof. For one inequality, $f \in \mathcal{B}_{K,P}$ and $x, x' \in M_p(P)$. By the reproducing property and Cauchy-Schwarz, we have

$$f(x,x') = \langle f, R_{K,P}(x,x',\cdot,\cdot)\rangle_{K,P} \leq \|f\|_{K,P}\|R_{K,P}(x,x',\cdot,\cdot)\|_{K,P} \leq \sqrt{R_{K,P}(x,x',x,x')} \leq \frac{\sigma_p(P)}{\sqrt{2}}.$$

Thus,

$$\sqrt{2}\cdot \sup_{f\in\mathcal{B}_{K,P}} \sup_{x,x'\in M_p(P)} f(x,x') \leq \sigma_p(P).$$

For the other inequality, take arbitrary $x, x' \in M_p(P)$, and set

$$f_{x,x'} = \frac{R_{K,P}(x,x',\cdot,\cdot)}{\|R_{K,P}(x,x',\cdot,\cdot)\|_{K,P}}.$$

By construction, we have $f_{x,x'} \in \mathcal{B}_{K,P}$ and

$$f_{x,x'}(x,x') = \langle f_{x,x'}, R_{K,P}(x,x',\cdot,\cdot)\rangle_{K,P} = \sqrt{R_{K,P}(x,x',x,x')},$$

hence

$$\sqrt{2}\cdot \sup_{f\in\mathcal{B}_{K,\mu}} \sup_{x,x'\in M_p(P)} f(x,x') \geq \sigma_p(P).$$

So, the result is proved. □

We also have the following asymptotic version of this statement:

Lemma 18. *Suppose $(\mathcal{X}, d)$ is a Heine-Borel space, $p \geq 1$, and $P \in \mathcal{P}_{2p-2}(\mathcal{X})$, and that $\{K_\delta\}_{0<\delta\leq 1}$ is a collection of compact subsets of $\mathcal{X}$ satisfying $\bigcap_{0<\delta\leq 1} K_\delta = M_p(P)$. Then*

$$\sqrt{2}\cdot \sup_{f\in\mathcal{B}_{K_1,P}} \max_{x\in M_p(P)} \max_{x'\in K_\delta} f(x,x') \to \sigma_p(P)$$

as $\delta \to 0$.

Proof. It is well-known that the unit ball of a RKHS on a compact metric space is compact in the topology of uniform convergence. In particular, there exists a sequence $\{\delta_n\}_{n\in\mathbb{N}}$ with $\delta_n \to 0$, some $(f, x, x') \in \mathcal{B}_{K_1,P} \times M_p(P) \times K_1$, and some $(f_n, x_n, x'_n) \in \mathcal{B}_{K_1,P} \times M_p(P) \times K_{\delta_n}$ for each $n \in \mathbb{N}$ such that we have $f_n \to f$ uniformly, $x_n \to x$ and $x'_n \to x'$ in d, and

$$\limsup_{\delta\to 0} \sup_{f\in\mathcal{B}_{K_1,P}} \max_{x\in M_p(P)} \max_{x'\in K_\delta} f(x,x') = \lim_{n\to\infty} f_n(x_n, x'_n).$$

Of course, we have $f_n(x_n, x'_n) \to f(x,x')$ as $n \to \infty$, and the assumption of $\bigcap_{0<\delta\leq 1} K_\delta = M_p(P)$ implies $x, x' \in M_p(P)$. Therefore, using Lemma 17 we have shown

$$\sqrt{2}\cdot \limsup_{\delta\to 0} \sup_{f\in\mathcal{B}_{K_1,P}} \max_{x\in M_p(P)} \max_{x'\in K_\delta} f(x,x') \leq \sqrt{2}\cdot \sup_{f\in\mathcal{B}_{K_1,P}} \max_{x,x'\in M_p(P)} f(x,x') = \sigma_p(P).$$

Conversely, for any $0 < \delta \leq 1$ we use Lemma 17 and $K_\delta \supseteq M_p(P)$ to immediately get

$$\sqrt{2} \cdot \sup_{f \in \mathcal{B}_{K_1,P}} \max_{x \in M_p(P)} \max_{x' \in K_\delta} f(x,x') \geq \sqrt{2} \cdot \sup_{f \in \mathcal{B}_{K_1,P}} \max_{x,x' \in M_p(P)} f(x,x') = \sigma_p(P).$$

Therefore, taking the lim inf as $\delta \to 0$ completes the proof. □

6.2 Consistency Results

In this section we will prove the main results on d_{H}-consistency, so let us assume the following: $(\mathcal{X}, d)$ is a Heine-Borel-Dudley space, $p \geq 1$, and $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space supporting an IID sequence $Y_1, Y_2, \ldots$ of $\mathcal{X}$-valued random variables with common distribution $P \in \mathcal{P}_{2p-2}(\mathcal{X})$; all "in probability" and "almost surely" statements refer to $\mathbb{P}$.

Our goal is to find an estimator $\hat{M}_p : \mathcal{X}^n \to \mathrm{K}(\mathcal{X})$ which satisfies the desired consistency properties. For convenience, and in the usual fashion, let us write $\hat{M}_p^n := \hat{M}_p(Y_1, \ldots, Y_n)$. We say an estimator $\hat{M}_p^n$ is *weakly* d_{H}*-consistent* if

$$d_{\mathrm{H}}(M_p(P), \hat{M}_p) \to 0 \tag{6.5}$$

holds in probability, and that it is *strongly* d_{H}*-consistent* if (6.5) holds almost surely.

We begin with a negative result for the natural estimator, the empirical Fréchet mean $M_p(\bar{P}_n)$. While it is easy to come up with examples which show that $M_p(\bar{P}_n)$ can fail to be even weakly d_{H}-consistent, the following result shows that inconsistency is more general:

Theorem 14. *If $(\mathcal{X}, d)$ is a finite metric space, then $M_p(\bar{P}_n)$ is a strongly d_{H}-consistent estimator of $M_p(P)$ if and only if $\sigma_p(P) = 0$.*

Proof. By Theorem 9 we have

$$\max_{\bar{x}_n \in M_p(\bar{P}_n)} \min_{x \in M_p(P)} d(\bar{x}_n, x) \to 0$$

almost surely, so it suffices to show that $\sigma_p(P) = 0$ if and only if

$$\max_{x \in M_p(P)} \min_{\bar{x}_n \in M_p(\bar{P}_n)} d(x, \bar{x}_n) \to 0 \tag{6.6}$$

almost surely.

For the first direction, suppose that $\sigma_p(P) = 0$ and let us show that (6.6) holds almost surely. To do this, we first observe that $\sigma_p(P) = 0$ implies that we have $W_p(\bar{P}_n, x, o) = W_p(\bar{P}_n, x', o)$ almost surely for all $x, x' \in M_p(P)$ and $o \in \mathcal{X}$, and hence that $M_p(P) \cap M_p(\bar{P}_n) \neq \varnothing$ implies $M_p(P) \subseteq M_p(\bar{P}_n)$ almost surely. Next, we observe that since $(\mathcal{X}, d)$ is a finite metric space, (6.6) holding almost surely is equivalent to $M_p(P) \subseteq M_p(\bar{P}_n)$ for sufficiently large $n \in \mathbb{N}$ almost surely. Thus, it suffices to show that we have $M_p(P) \cap$

$M_p(\bar{P}_n) \neq \varnothing$ for sufficiently large $n \in \mathbb{N}$ almost surely. To do this, we will show that the event

$$E := \{M_p(P) \cap M_p(\bar{P}_n) = \varnothing \text{ for infinitely many } n \in \mathbb{N}\}$$

has $\mathbb{P}(E) = 0$. Indeed, observe that the following holds on $E \cap \{\bar{P}_n \to P \text{ in } \tau_{\mathrm{w}}^{p-1}\}$: There exists a subsequence $\{n_k\}_{k\in\mathbb{N}}$ such that $M_p(P) \cap M_p(\bar{P}_{n_k}) = \varnothing$ for all $k \in \mathbb{N}$, so choosing $\bar{x}_k \in M_p(\bar{P}_{n_k})$ arbitrarily for each $k \in \mathbb{N}$, we use Theorem 8 to get some $x \in M_p(P)$ and a further subsequence $\{k_j\}_{j\in\mathbb{N}}$ such that $\bar{x}_{k_j} \to x$ as $j \to \infty$. But $(\mathcal{X}, d)$ being finite implies $\bar{x}_{k_j} = x$ for sufficiently large $j \in \mathbb{N}$, so $x \in M_p(P) \cap M_p(\bar{P}_{n_{k_j}})$ is a contradiction. We therefore have

$$E \cap \{\bar{P}_n \to P \text{ in } \tau_{\mathrm{w}}^{p-1}\} = \varnothing$$

hence $\mathbb{P}(E) \leq \mathbb{P}(\Omega \setminus \{\bar{P}_n \to P \text{ in } \tau_{\mathrm{w}}^{p-1}\}) = 0$.

For the second direction, suppose that (6.6) holds almost surely, and let us show that $\sigma_p(P) = 0$. To do this, take arbitrary $x, x' \in M_p(P)$ and note that we have $W_p(P, x, x') = 0$. Now we see that the IID random variables $\{Z_i\}_{i\in\mathbb{N}}$ defined via $Z_i := d^p(x, Y_i) - d^p(x', Y_i)$ are integrable and centered, so we can define the random walk $\{S_n\}_{n\in\mathbb{N}}$ via $S_n := \sum_{i=1}^n Z_i$. Importantly, observe that $x, x' \in M_p(\bar{P}_n)$ if and only if $S_n = 0$. Thus (6.6) holding almost surely implies that $\{S_n\}_{n\in\mathbb{N}}$ is eventually equal to zero, and this can only happen if $\mathbb{P}(Z_i = 0) = 1$ for all $i \in \mathbb{N}$. This is equivalent to $\sigma_p(P) = 0$ and concludes the proof. □

This result allows us to easily generate negative examples, that is examples where the empirical Fréchet mean is not a strongly consistent estimator of the population Fréchet mean. The simplest is $\mathcal{X} = \{1, 2, \ldots, m\}$ for $m \geq 2$ where d is the discrete metric and P is the uniform measure; since every point is indistinguishable under both d and P, we certainly have $\sigma_p(P) = 0$.

In order to cook up some other estimator which is d_{H}-consistent, we introduce the notion of relaxed Fréchet means.

Definition 10. *For any metric space $(\mathcal{X}, d)$, any $p \geq 1$, and any $P \in \mathcal{P}_{p-1}(\mathcal{X})$ and $\varepsilon \geq 0$, we let*

$$M_p(P, \varepsilon) := \{x \in \mathcal{X} : W_p(P, x, x') \leq \varepsilon \text{ for all } x' \in \mathcal{X}\},$$

called the ε-relaxed Fréchet p-mean set of P.

Observe that $M_p(P, 0) = M_p(P)$ for $P \in \mathcal{P}_{p-1}(\mathcal{X})$, so the relaxation parameter $\varepsilon \geq 0$ controls which level sets of the Fréchet functional are included in this approximate Fréchet mean set. We note that we will only use the definition given in this form, but that one can also give equivalent definitions of relaxed Fréchet means, similar to what we saw for unrelaxed Fréchet means in Lemma 5.

Our proposed estimator will be of the form $M_p(\bar{P}_n, \varepsilon_n)$ for some carefully-chosen relaxation scale ε_n. A sequence of relaxation scales $\{\varepsilon_n\}_{n\in\mathbb{N}}$ applied to form the empirical relaxed Fréchet mean sets $\{M_p(\bar{P}_n, \varepsilon_n)\}_{n\in\mathbb{N}}$ will be referred to as a *relaxation rate*, and we often simply write ε_n in place of $\{\varepsilon_n\}_{n\in\mathbb{N}}$. By a *random relaxation rate* we mean a relaxation rate

ε_n in which ε_n is a random variable for each $n \in \mathbb{N}$; these can of course be deterministic or independent of $Y_1, Y_2, \ldots$, but the most interesting and important examples are *adaptive*, in the sense that ε_n is $\sigma(Y_1, \ldots, Y_n)$-measurable for all $n \in \mathbb{N}$.

This interest in relaxed Fréchet mean sets is not new. Indeed, in [180, Corollary 5.1 and Corollary 5.2] it is shown that if a relaxation rate satisfies $\varepsilon_n \to 0$, then $M_p(\bar{P}_n, \varepsilon_n)$ satisfies the no false positives property (6.1) almost surely. Interestingly, it was observed [180, Appendix A.3] in the simple example of $\mathcal{X} = \{0, 1\}, p = 2$, and $P = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$, that choosing the relaxation rate as $\varepsilon_n := n^{-1/4}$ implies that $M_2(\bar{P}_n, \varepsilon_n)$ also satisfies the no false negatives property (6.2) almost surely. In words, applying a vanishing relaxation rate does not disturb (6.1), and applying a *sufficiently slowly* vanishing relaxation rate allows obtaining (6.2). Of course, in order to be as efficient as possible, one wants to choose the relaxation rate which is the fastest possible among all sufficiently slow relaxation rates. Thus our goal is to understand the *fastest possibly sufficiently slow* relaxation rate for d_{H}-consistent estimation.

In light of the central limit theorem, a natural guess for a useful relaxation rate is something proportional to $n^{-1/2}$. As the next result shows, this is not slow enough to get even weak d_{H}-consistency, but we can obtain a Gaussian-like bound on its error probability:

Theorem 15. *Let $(\mathcal{X}, d)$ be a Heine-Borel-Dudley space, $p \geq 1$, and $P \in \mathcal{P}_{2p-2}(\mathcal{X})$. Then, there exists a constant $\mu_p(P) < \infty$ such that for any $c \geq \mu_p(P)$, the relaxation rate $\varepsilon_n = cn^{-1/2}$ satisfies*

$$\sup_{\delta>0} \limsup_{n\to\infty} \mathbb{P}(d_{\mathrm{H}}(M_p(\bar{P}_n, \varepsilon_n), M_p(P)) \geq \delta) \leq \exp\left(-\frac{(c - \mu_p(P))^2}{\sigma_p^2(P)}\right),$$

with the convention the right side is 0 if $\sigma_p^2(P) = 0$.

Proof. Fix $\delta > 0$ and denote $K^\delta := (M_p(P))^\delta$ the δ-thickening of $M_p(P)$. Since $(\mathcal{X}, d)$ is a Heine-Borel space and $M_p(P)$ is compact by Corollary 4, it follows that K^δ is compact. By Proposition 4, the sequence $\{n^{-1/2}G_{K^\delta,n}\}_{n\in\mathbb{N}}$ converges in distribution to the Gaussian process $G_{K^\delta} := \{G_{K^\delta}(x, x')\}_{x,x'\in K^\delta}$ with distribution $\mathcal{G}_P$. Because G_{K^δ} takes values in $C(K^\delta \times K^\delta)$, it almost surely bounded. Now let us define $\sigma_p^2(P; \delta) := 2 \cdot \sup_{x,x'\in K^\delta} \mathbb{E}[(G_{K^\delta}(x, x'))^2] = 2 \cdot \sup_{x,x'\in K^\delta} R_P(x, x')$ which is finite since Lemma 15 shows that R is continuous. By the Borell-TIS inequality, the quantity $\mu_p(P; \delta) := \mathbb{E}[\|G_{K^\delta}\|_\infty]$ is finite, and for all $c \geq \mu_p(P; \delta)$, we have

$$\mathbb{P}(\|G_{K^\delta}\|_\infty \geq c) \leq \exp\left(-\frac{(c - \mu_p(P; \delta))^2}{\sigma_p^2(P; \delta)}\right),$$

and also, by convergence in distribution, we have

$$\limsup_{n\to\infty} \mathbb{P}(\|n^{-1/2}G_{K^\delta,n}\|_\infty \geq c) \leq \exp\left(-\frac{(c - \mu_p(P; \delta))^2}{\sigma_p^2(P; \delta)}\right).$$

Now observe that

$$\begin{aligned}
&\{M_p(\bar{P}_n,\varepsilon_n)\subseteq K^\delta\}\cap\{\|n^{-1/2}G_{K^\delta,n}\|_\infty<c\}\\
&\subseteq\{\forall x\in M_p(P),\forall x'\in M_p(\bar{P}_n,\varepsilon_n),G_{K^\delta,n}(x,x')\le c\sqrt{n}\}\\
&\subseteq\{\forall x\in M_p(P),\forall x'\in M_p(\bar{P}_n,\varepsilon_n),W_p(\bar{P}_n,x,x')\le cn^{-1/2}\}\\
&\subseteq\{M_p(P)\subseteq M_p(\bar{P}_n,\varepsilon_n)\}.
\end{aligned}$$

Therefore, these bounds combined with Theorem 9 yield:

$$\begin{aligned}
\limsup_{n\to\infty}&\mathbb{P}(M_p(P)\not\subseteq M_p(\bar{P}_n,\varepsilon_n))\\
&\le\limsup_{n\to\infty}\left(\mathbb{P}(M_p(\bar{P}_n,\varepsilon_n)\not\subseteq K^\delta)+\mathbb{P}(\|n^{-1/2}G_{K^\delta,n}\|_\infty\ge c)\right)\\
&\le\exp\left(-\frac{(c-\mu_p(P;\delta))^2}{\sigma_p^2(P;\delta)}\right).
\end{aligned}$$

Since we of course have

$$\mathbb{P}(d_{\mathrm{H}}(M_p(\bar{P}_n,\varepsilon_n),M_p(P))\ge\delta)\le\mathbb{P}(M_p(P)\not\subseteq M_p(\bar{P}_n,\varepsilon_n)),$$

it suffices to simply take the limit as $\delta\to 0$. We already have $\sigma_p(P;\delta)\to\sigma_p(P)$ as $\delta\to 0$ by the continuity of R_P. So, we just need to show that $\mu_p(P;\delta)\to\mu_p(P):=\mathbb{E}[\|G_{M_p(P)}\|_\infty]<\infty$. Indeed, this follows easily from the sample path continuity of $\mathcal{G}_P$ and dominated convergence. □

Because of the previous result, strong d_{H}-consistency requires a relaxation rate which is slightly slower than $n^{-1/2}$ or any constant multiple thereof. As we see in the following, a logarithmic correction is what is needed:

Theorem 16. *Let $(\mathcal{X},d)$ be a Heine-Borel-Dudley space, $p\ge 1$, and $P\in\mathcal{P}_{2p-2}(\mathcal{X})$. Then, the relaxation rate $\varepsilon_n:=cn^{-1/2}(\log\log n)^{1/2}$ satisfies the following:*

(i) If $c>\sigma_p(P)$, then $M_p(\bar{P}_n,\varepsilon_n)$ is a strongly d_{H}-consistent estimator of $M_p(P)$, and $\mathbb{P}(M_p(P)\subseteq M_p(\bar{P}_n,\varepsilon_n)$ for sufficiently large $n\in\mathbb{N})=1$.

(ii) If $c<\sigma_p(P)$, then $M_p(\bar{P}_n,\varepsilon_n)$ is not a strongly d_{H}-consistent estimator of $M_p(P)$. In fact, it is strongly d_{H}-inconsistent, in that $\mathbb{P}(d_{\mathrm{H}}(M_p(P),M_p(\bar{P}_n,\varepsilon_n))\to 0)=0$.

Proof. The key observation in both cases is that our choice of relaxation parameter, along with some simple arithmetic, shows that we have

$$\{W_p(\bar{P}_n,x,x')\le\varepsilon_n\}=\left\{\frac{G_n(x,x')}{\sqrt{n\log\log n}}\le\sqrt{\frac{n}{\log\log n}}(\varepsilon_n-W_p(P,x,x'))\right\}$$

for all $x, x' \in \mathcal{X}$ and $n \in \mathbb{N}$. In particular, if $x' \in M_p(P)$, then

$$\left\{(n \log\log n)^{-1/2} G_n(x,x') > c\right\} \subseteq \{W_p(\bar{P}_n, x, x') > \varepsilon_n\} \tag{6.7}$$

for all $x \in \mathcal{X}$. Conversely, if $x \in M_p(P)$, then

$$\left\{(n \log\log n)^{-1/2} G_n(x,x') \leq c\right\} \subseteq \{W_p(\bar{P}_n, x, x') \leq \varepsilon_n\} \tag{6.8}$$

for all $x' \in \mathcal{X}$. Now we proceed to the main proof, writing $K^\delta := (M_p(P))^\delta$ for the δ-thickening of $M_p(P)$ for all $\delta > 0$; these sets are compact by the Heine-Borel property of $\mathcal{X}$ and because $M_p(P)$ is compact.

For (i), we use Theorem 9 to get

$$\max_{\bar{x}_n \in M_p(\bar{P}_n)} \min_{x \in M_p(P)} d(\bar{x}_n, x) \to 0 \tag{6.9}$$

almost surely, so it suffices to show

$$\max_{x \in M_p(P)} \min_{\bar{x}_n \in M_p(\bar{P}_n)} d(x, \bar{x}_n) \to 0 \tag{6.10}$$

almost surely; let us write E for the event that (6.9) occurs. Towards showing that (6.10) occurs almost surely, use Lemma 18 to get

$$\sqrt{2} \cdot \max_{f \in \mathcal{B}_{K^1,P}} \max_{x \in M_p(P)} \max_{x' \in K^s} f(x,x') \to \sigma_p(P)$$

as $s \to 0$, hence there exists $0 < s < 1$ such that

$$\sqrt{2} \cdot \max_{f \in \mathcal{B}_{K^1,P}} \max_{x \in M_p(P)} \max_{x' \in K^s} f(x,x') < c.$$

By Proposition 5, there is an event F with $\mathbb{P}(F) = 1$ on which $\{(2n \log\log n)^{-1/2} G_{K^1,n}\}$ is relatively compact and its set of limits is exactly $\mathcal{B}_{K^1,P}$. Since $M_p(P)$ and K^s are compact, there exists $f \in \mathcal{B}_{K^1,P}$ such that

$$\begin{aligned}\limsup_{n\to\infty} \max_{x \in M_p(P)} \max_{x' \in K^s} \frac{G_n(x,x')}{\sqrt{n \log\log n}} &= \sqrt{2} \cdot \max_{x \in M_p(P)} \max_{x' \in K^s} f(x,x') \\ &\leq \sqrt{2} \cdot \max_{f \in \mathcal{B}_{K^1,P}} \max_{x \in M_p(P)} \max_{x' \in K^s} f(x,x') < c.\end{aligned}$$

Hence, on $E \cap F$, for sufficiently large $n \in \mathbb{N}$ we have

$$\max_{x \in M_p(P)} \max_{x' \in K^s} \frac{G_n(x,x')}{\sqrt{n \log\log n}} < c.$$

Lastly observe that on E we have $M_p(\bar{P}_n) \subseteq K^s$ for sufficiently large $n \in \mathbb{N}$, so combining the previous display with (6.8) yields

$$E \cap F \subseteq E \cap \{M_p(P) \subseteq M_p(\bar{P}_n, \varepsilon_n) \text{ for sufficiently large } n \in \mathbb{N}\} \tag{6.11}$$

as claimed.

For (ii), we use Lemma 17 to get $f \in \mathcal{B}_{K^1,P}$, and $x, x' \in M_p(P)$ such that $\sqrt{2} \cdot f(x,x') = \sigma_p(P)$. Now let η satisfy $0 < \eta < \sigma_p(P) - c$. Because f is continuous, there exists $0 < r < 1$ such that we have $\sqrt{2} \cdot f(z,x') > \sigma_p(P) - \eta$ for all $z \in \bar{B}_r(x)$. Note that we selected $r < 1$ so that $\bar{B}_r(x) \subseteq K^1$. Now by Proposition 5, there exists an event E with $\mathbb{P}(E) = 1$, on which the following is true: There is a subsequence $\{n_k\}_{k\in\mathbb{N}}$ such that we have $(2n_k \log\log n_k)^{-1/2} G_{K^1,n_k} \to f$ in the topology of uniform convergence on K^1. Also recall that $G_{K^1,n}$ is just the restriction of G_n to $K^1 \times K^1$. Therefore,

$$\begin{aligned}
\limsup_{n\to\infty} \min_{z\in\bar{B}_r(x)} (n\log\log n)^{-1/2} G_n(z,x') &\geq \liminf_{k\to\infty} \min_{z\in\bar{B}_r(x)} (n_k \log\log n_k)^{-1/2} G_{n_k}(z,x') \\
&\geq \min_{z\in\bar{B}_r(x)} \sqrt{2} \cdot f(z,x') \\
&\geq \sigma_p(P) - \eta \\
&> c.
\end{aligned}$$

We combine this with (6.7) to see that on E and for fixed $x' \in M_p(P)$, there are infinitely many $n \in \mathbb{N}$ satisfying $\min_{z\in\bar{B}_r(x)} W_p(\bar{P}_n, z, x') > \varepsilon_n$. Also, observe

$$\begin{aligned}
\left\{ \min_{z\in\bar{B}_r(x)} W_p(\bar{P}_n, z, x') > \varepsilon_n \right\} &\subseteq \{\bar{B}_r(x) \cap M_p(\bar{P}_n, \varepsilon_n) = \varnothing\} \\
&\subseteq \{d_{\mathrm{H}}(M_p(\bar{P}_n, \varepsilon_n), M_p(P)) \geq r\}.
\end{aligned}$$

Therefore, on E we have $\limsup_{n\to\infty} d_{\mathrm{H}}(M_p(\bar{P}_n, \varepsilon_n), M_p(P)) \geq r$. In particular, we have shown $\mathbb{P}(d_{\mathrm{H}}(M_p(P), M_p(\bar{P}_n, \varepsilon_n)) \to 0) = 0$ as claimed. □

Let us now summarize and compare the results thus far. In light of Theorem 15, it is natural to consider the *weak asymptotic error probability* for each population distribution $P \in \mathcal{P}_{2p-2}(\mathcal{X})$ and each relaxation rate ε_n defined via

$$\mathrm{WE}_P(\varepsilon_n) = \sup_{\delta>0} \limsup_{n\to\infty} \mathbb{P}(d_{\mathrm{H}}(M_p(\bar{P}_n, \varepsilon_n), M_p(P)) \geq \delta).$$

Similarly, in light of Theorem 16, it is natural to consider the *strong asymptotic error probability* for each $P \in \mathcal{P}_{2p-2}(\mathcal{X})$ and each ε_n defined via

$$\mathrm{SE}_P(\varepsilon_n) = 1 - \mathbb{P}\left(d_{\mathrm{H}}(M_p(\bar{P}_n, \varepsilon_n), M_p(P)) \to 0\right).$$

(Note that $\mathrm{WE}_P(\varepsilon_n)$ and $\mathrm{SE}_P(\varepsilon_n)$ are shorthand for $\mathrm{WE}_P(\{\varepsilon_n\}_{n\in\mathbb{N}})$ and $\mathrm{SE}_P(\{\varepsilon_n\}_{n\in\mathbb{N}})$, respectively.) Figure 6.2 summarizes the results of Theorem 15 and Theorem 16 in terms of these weak and strong error quantities.

In words, we have shown the strong d_{H}-consistency for relaxed empirical Fréchet mean set estimators experiences a sharp transition at the relaxation rate $\varepsilon_n = \sigma_p(P) n^{-1/2} (\log\log n)^{1/2}$. The "critical" relaxation rate of $\varepsilon_n = \sigma_p(P) n^{-1/2} (\log\log n)^{1/2}$ appears to be more complicated, and we do not know what behavior to expect in this case.

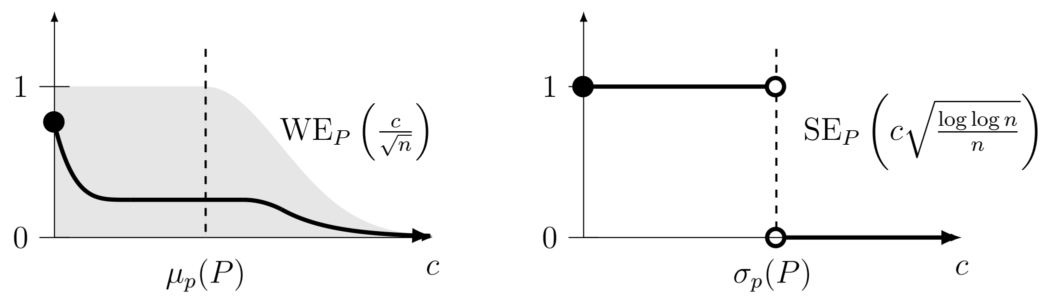

Figure 6.2: The d_{H}-consistency properties of relaxed empirical Frećhet mean set estimators, for a population distribution $P \in \mathcal{P}_{2p-2}(\mathcal{X})$. As a function of the pre-factor $c \geq 0$, the weak error of the $cn^{-1/2}$-relaxed empirical Fréchet mean set estimators decays in a Gaussian-like way (left), and the the strong error of the $cn^{-1/2}(\log\log)^{1/2}$-relaxed empirical Fréchet mean set estimators experiences a phase transition (right).

Chapter 7

Applications in Phylogenetics

For $N \in \mathbb{N}$, a labeled tree on N leaves can be identified with its distance matrix; since metric trees are known to satisfy an ultrametricity property, we define an *equidistant N-leaf tree* to be a distance matrix w on the set $[N] := \{1, 2, \ldots, N\}$ satisfying $w(i,k) \leq \max(w(i,j), w(j,k))$ for all $i,j,k \in [N]$. In particular, the space $\mathcal{U}_N$ of all equidistant N-leaf trees can be naturally embedded into the Euclidean space

$$\mathbb{R}^{\binom{N}{2}}.$$

Geometrically speaking, it is known [34, Section 2] that $\mathcal{U}_N$ is a union of $(2N-3)!! = (2N-3)(2N-5)\cdots 3\cdot 1$ orthants each of dimension $N-2$, and that each orthant corresponds to a unique binary tree topology.

The space of $\mathcal{U}_N$ inherits a natural geometry from the ambient Euclidean space by defining the distance between two points to be the smallest possible Euclidean arc length of a path in $\mathcal{U}_N$ connecting them. This idea was developed in detail in one of the seminar works of non-Euclidean statistics [34], so this metric is often called the *Billera-Holmes-Vogtmann (BHV) metric* and is denoted d_{BHV}. The work [34] pioneered the use of geometric methods in computational phylogenetics, as they showed that various statistical problems of interest could be formulated and studied using this geometric perspective, and they proved that the resulting BHV treespace $(\mathcal{U}_N, d_{\mathrm{BHV}})$ has many nice properties (Fréchet means are unique, geodesics have an explicit form, geometric quantities of interest are interpretable, etc.). However, later developments emphasized that it is notoriously difficult to do even the most basic computations in the BHV treespace.

Because of this difficulty, other authors have proposed alternative geometries on the space of phylogenetic trees which overcome these computational difficulties; primary among them is the so-called *tropical* (that is, *max-plus*) *projective treespace* [17, 209, 106, 136], which we now introduce. For $k \in \mathbb{N}$, we write $\mathbb{R}^k/\mathbb{R}\mathbf{1}$ for the Euclidean space $\mathbb{R}^k$ quotiented by the action of $\mathbb{R}\mathbf{1}$, where we define $c\mathbf{1} \cdot (x_1, \ldots, x_k) := (x_1 + c, \ldots, x_k + c)$ for $c \in \mathbb{R}$ and $(x_1, \ldots, x_k) \in \mathbb{R}^k$; in other words, $\mathbb{R}^k/\mathbb{R}\mathbf{1}$ represents Euclidean space, modulo diagonal translations. For $x = (x_1, \ldots, x_k)$ and $x' = (x'_1, \ldots, x'_k)$ in $\mathbb{R}^k$ we define

$$d_{\mathrm{tr}}(x, x') := \max\{x_i - x'_i : 1 \leq i \leq k\} - \min\{x_i - x'_i : 1 \leq i \leq k\}.$$

It is easy to see that d_{tr} is invariant under the action of $\mathbb{R}\mathbf{1}$ on $\mathbb{R}^k$; in fact, d_{tr} descends to a metric on $\mathbb{R}^k/\mathbb{R}\mathbf{1}$. By a slight abuse of notation, we also write d_{tr} for the induced metric on $\mathbb{R}^k/\mathbb{R}\mathbf{1}$, called the *tropical projective metric*. We refer to $(\mathbb{R}^k/\mathbb{R}\mathbf{1}, d_{\mathrm{tr}})$ as the *tropical projective space*. Now, for $N \in \mathbb{N}$, the space of equidistant trees $\mathcal{U}_N$ naturally embeds into the tropical projective space

$$\left(\mathbb{R}^{\binom{N}{2}}/\mathbb{R}\mathbf{1}, d_{\mathrm{tr}}\right),$$

so $\mathcal{U}_N$ can inherit the induced metric. We refer to $(\mathcal{U}_N, d_{\mathrm{tr}})$ as the *tropical projective treespace*.

For $N \in \mathbb{N}$, the authors of [135, 106] argue that $(\mathcal{U}_N, d_{\mathrm{tr}})$ is a natural metric space for phylogenetic inference. Contrary to the BHV treespace, they show that most quantities of interest in $(\mathcal{U}_N, d_{\mathrm{tr}})$ can be computed directly with the help of standard methods from convex optimization and polyhedral geometry. However, they also point out that the natural notion of central tendency—the Fréchet 1-mean M_1, often called the set of *Fermat-Weber points*—is typically non-unique. Thus, even the basic problem of mean estimation becomes very complicated.

In this chapter we assume that $N \in \mathbb{N}$ is fixed and that $(\Omega, \mathcal{F}, \mathbb{P})$ is a proability space supporting an IID sequence $Y_1, Y_2, \ldots$ of $\mathcal{U}_N$-valued random variables with common distribution $P \in \mathcal{P}(\mathcal{U}_N)$. (Note that we do not make any integrability assumption.) Our goal is to use the results of Chapter 6 to show that one can estimate the set $M_1(P)$ from the data, in the Hausdorff metric.

This work is based on [37] and contains a few results. We construct a procedure (Algorithm 2) which implements an adaptively-relaxed Frećhet mean set estimator, where, provably, we adaptively find the fastest possible relaxation rate for strongly d_{H}-consistent estimation (Theorem 18). Along the way, we also develop some novel results about tropical projective treespace which may be of independent interest. Lastly, we give several numerical experiments to demonstrate the convergence of our procedure in practice; these involve simulated data as well as real data from influenza genome sequences.

7.1 Preliminaries

In this section we show that one can specialize the general results of Chapter 6 to the case of tropical projective space and tropical projective treespace, and we also prove some novel results on tropical projective space that may be of independent interest. To begin, we show that the tropical projective space satisfies the desired regularity:

Lemma 19. *The space $(\mathbb{R}^k/\mathbb{R}\mathbf{1}, d_{\mathrm{tr}})$ is a Heine-Borel-Dudley space.*

Proof. For simplicity, we identify $\mathbb{R}^k/\mathbb{R}\mathbf{1}$ to $\mathbb{R}^{k-1}$ by setting the first coordinate to 0. For any $x, y \in \mathbb{R}^{k-1}$, we then have

$$\|x-y\|_\infty \le d_{\mathrm{tr}}(x,y) \le 2\|x-y\|_\infty.$$

Since $(\mathbb{R}^{k-1}, \|\cdot\|_\infty)$ is Heine-Borel-Dudley, the space $(\mathbb{R}^{k-1}, d_{\mathrm{tr}})$ is Heine-Borel-Dudley. □

Next we turn to the consistency results. The primary difficulty in applying them to this setting is the computation of the fluctuation variation $\sigma_1(P)$, and how to estimate it from samples. These concerns will be resolved based on the following insight:

Lemma 20. *For $P \in \mathcal{P}(\mathbb{R}^k/\mathbb{R}\mathbf{1})$ and $z \in \operatorname{supp}(P)$, the map $d_{\mathrm{tr}}(\,\cdot\,, z) : M_1(P) \to \mathbb{R}$ is affine.*

Proof. It suffices to show that $d_{\mathrm{tr}}(\cdot\,, z) : [x, x'] \to \mathbb{R}$ is affine for all $x, x' \in M_1(P)$, where $[x, x'] := \{(1-t)x + tx' : 0 \le t \le 1\}$ denotes the line segment connecting x and x'. To show this, take an arbitrary $x, x' \in M_1(P)$ and $0 \le t \le 1$, and set $x^* = (1-t)x + tx'$. Then take arbitrary $r > 0$, and use $z \in \operatorname{supp}(P)$ to get $P(B_r^\circ(z)) > 0$. Observe by construction that $d_{\mathrm{tr}}(\cdot\,, u) : M_1(P) \to \mathbb{R}$ is convex for any $u \in \mathbb{R}^k/\mathbb{R}\mathbf{1}$, hence that we have

$$0 \le (1-t)(d_{\mathrm{tr}}(x, u) - d_{\mathrm{tr}}(x^*, u)) + t(d_{\mathrm{tr}}(x', u) - d_{\mathrm{tr}}(x^*, u)). \tag{7.1}$$

Now integrate (7.1) over $(\mathbb{R}^k/\mathbb{R}\mathbf{1}) \setminus B_r^\circ(z)$ with respect to P to get

$$\begin{aligned} 0 \le (1-t) &\int_{(\mathbb{R}^k/\mathbb{R}\mathbf{1})\setminus B_r^\circ(z)} (d_{\mathrm{tr}}(x, u) - d_{\mathrm{tr}}(x^*, u))\, \mathrm{d}P(u) \\ &+ t \int_{(\mathbb{R}^k/\mathbb{R}\mathbf{1})\setminus B_r^\circ(z)} (d_{\mathrm{tr}}(x', u) - d_{\mathrm{tr}}(x^*, u))\, \mathrm{d}P(u). \end{aligned} \tag{7.2}$$

Next add

$$P(B_r^\circ(z))\Bigg((1-t)(d_{\mathrm{tr}}(x, z) - d_{\mathrm{tr}}(x^*, z)) + t(d_{\mathrm{tr}}(x', z) - d_{\mathrm{tr}}(x^*, z))\Bigg) \tag{7.3}$$

to both sides of (7.2), use $x, x' \in M_1(P)$, rearrange, and use the triangle inequality to get

$$\begin{aligned} P(B_r^\circ(z))&\Bigg((1-t)(d_{\mathrm{tr}}(x, z) - d_{\mathrm{tr}}(x^*, z)) + t(d_{\mathrm{tr}}(x', z) - d_{\mathrm{tr}}(x^*, z))\Bigg) \\ &\le (1-t)W_1(P, x, x^*) + tW_1(P, x', x^*) \\ &+ (1-t) \int_{B_r^\circ(z)} ((d_{\mathrm{tr}}(x, z) - d_{\mathrm{tr}}(x^*, z)) - (d_{\mathrm{tr}}(x, u) - d_{\mathrm{tr}}(x^*, u)))\, \mathrm{d}P(u) \\ &+ t \int_{B_r^\circ(z)} ((d_{\mathrm{tr}}(x', z) - d_{\mathrm{tr}}(x^*, z)) - (d_{\mathrm{tr}}(x', u) - d_{\mathrm{tr}}(x^*, u)))\, \mathrm{d}P(u) \\ &\le (1-t) \int_{B_r^\circ(z)} (r - (-r))\, \mathrm{d}P(u) \\ &+ t \int_{B_r^\circ(z)} (r - (-r))\, \mathrm{d}P(u) \\ &\le 2P(B_r^\circ(z)) \cdot r. \end{aligned}$$

Dividing the above by $P(B_r^\circ(z)) > 0$ implies

$$(1-t)(d_{\mathrm{tr}}(x,z) - d_{\mathrm{tr}}(x^*,z)) + t(d_{\mathrm{tr}}(x',z) - d_{\mathrm{tr}}(x^*,z) \leq 2r. \tag{7.4}$$

Finally, by taking $r \to 0$ in (7.4) and combining with (7.1) for $u = z$, we get

$$d_{\mathrm{tr}}(x^*,z) = (1-t)d_{\mathrm{tr}}(x,z) + td_{\mathrm{tr}}(x',z). \tag{7.5}$$

This proves the result. □

As a consequence, it leads to the following:

Lemma 21. *If $P \in \mathcal{P}(\mathbb{R}^k/\mathbb{R}\mathbf{1})$, then the function $\mathrm{Var}_P(d_{\mathrm{tr}}(\cdot\,,Y_1) - d_{\mathrm{tr}}(\cdot\,,Y_1)) : M_1(P) \times M_1(P) \to [0,\infty)$ is convex, hence $\sigma_1(P)$ is equal to the value of the optimization problem*

$$\begin{cases} \text{maximize} & \sqrt{2}\cdot\sqrt{\mathrm{Var}_P(d_{\mathrm{tr}}(x,Y_1) - d_{\mathrm{tr}}(x',Y_1))} \\ \text{over} & x,x' \in \mathrm{ex}(M_1(P)). \end{cases} \tag{7.6}$$

Proof. Since $\mathbb{R}^k/\mathbb{R}\mathbf{1}$ is a separable metric space, one has $P(\mathrm{supp}(P)) = 1$. Also, for $x,x' \in M_1(P)$, we have $W_1(P,x,x') = W_1(P,x',x) = 0$. Thus:

$$\begin{aligned} &\mathrm{Var}_P(d_{\mathrm{tr}}(x,Y_1) - d_{\mathrm{tr}}(x',Y_1)) \\ &= \int_{\mathbb{R}^k/\mathbb{R}\mathbf{1}} (d_{\mathrm{tr}}(x,u) - d_{\mathrm{tr}}(x',u))^2\,\mathrm{d}P(u) - \left(\int_{\mathbb{R}^k/\mathbb{R}\mathbf{1}} (d_{\mathrm{tr}}(x,u) - d_{\mathrm{tr}}(x',u))\,\mathrm{d}P(u)\right)^2 \\ &= \int_{\mathrm{supp}(P)} (d_{\mathrm{tr}}(x,u) - d_{\mathrm{tr}}(x',u))^2\,\mathrm{d}P(u). \end{aligned}$$

By Lemma 3.2, for any $u \in \mathrm{supp}(P)$, the function $(x,x') \mapsto d_{\mathrm{tr}}(x,u) - d_{\mathrm{tr}}(x',u)$ is affine on $M_1(P)\times M_1(P)$, hence its square is convex. This shows that the map $(x,x') \mapsto \mathrm{Var}_P(d_{\mathrm{tr}}(x,Y_1) - d_{\mathrm{tr}}(x',Y_1))$ is a convex combination of convex functions, hence convex. □

In statistical applications, the fluctuation variation $\sigma_1(P)$ is typically not known, so one needs to estimate it in order to determine the optimal pre-factor for a relaxed Fermat-Weber set estimator. One idea is simply to consider the plug-in estimator $\sigma_1(\bar{P}_n)$, but it can be shown in simple examples that the convergence $\sigma_1(\bar{P}_n) \to \sigma_1(P)$ can plainly fail. Instead, we require the following notion:

Definition 11. *For $P \in \mathcal{P}(\mathcal{U}_N)$ and a compact convex set $A \subseteq \mathbb{R}^k/\mathbb{R}\mathbf{1}$, we write*

$$\hat{\sigma}_1(P,A) := \sqrt{2}\sup_{x,x'\in A}\sqrt{\mathbb{E}_P\left[(d_{\mathrm{tr}}(x,Y_1) - d_{\mathrm{tr}}(x',Y_1))^2\right]}, \tag{7.7}$$

called the approximate fluctuation variation of P on A.

The key observation will be that for any consistent but sub-optimal relaxation rate ε_n, we can consistently estimate the fluctuation variation of P on $M_1(P)$ through the approximate fluctuation variation of $\bar{P}_n$ on $M_1(\bar{P}_n, \varepsilon_n)$. In particular, it suffices to take $\varepsilon_n := n^{-1/2} \log\log n$ for the desired conclusion.

Unfortunately, the convexity of the variance functional that we used in Lemma 21 does not extend to the relaxed Fermat-Weber set, so computing the approximate fluctuation variation may be difficult. To get around this, we need to introduce some notation: For each $z \in \mathbb{R}^k/\mathbb{R}\mathbf{1}$, there exists a covering of $\mathbb{R}^k/\mathbb{R}\mathbf{1}$ by finitely many polyhedra $\mathscr{P}_z = \{A_{z,\ell} : \ell \in L_z\}$ such that $d_{\mathrm{tr}}(\cdot\,, z) : \mathbb{R}^k/\mathbb{R}\mathbf{1} \to \mathbb{R}$ is linear on each $A_{z,\ell}$; for a set $S \subseteq \mathbb{R}^k/\mathbb{R}\mathbf{1}$, we write $\mathscr{P}_S := \{A_{S,\ell} : \ell \in L_S\}$ for the coarsest common refinement of $\{\mathscr{P}_z\}_{z\in S}$.

Proposition 6. *Suppose that $A \subseteq \mathbb{R}^k/\mathbb{R}\mathbf{1}$ is a compact convex set with extreme points $\{v_i : i \in I\} := \mathrm{ex}(A)$ and that $P \in \mathcal{P}(\mathbb{R}^k/\mathbb{R}\mathbf{1})$, and write $\{v_{A,j} : j \in I_A\} := A \cap \bigcup_{\ell\in L_{\mathrm{supp}(P)}} \mathrm{ex}(A_{\mathrm{supp}(P),\ell})$ for the set of extreme points of the polyhedra in $\mathscr{P}_{\mathrm{supp}(P)}$ which fall in A. Then, $\hat{\sigma}_1(P, A)$ is equal to the value of the optimization problem*

$$\begin{cases} \text{maximize} & \sqrt{2}\cdot\sqrt{\mathbb{E}_P\left[|d_{\mathrm{tr}}(x, Y_1) - d_{\mathrm{tr}}(x', Y_1)|^2\right]} \\ \text{over} & x, x' \in \{v_i : i \in I\} \cup \{v_{A,j} : j \in J\}. \end{cases} \tag{7.8}$$

Proof. By the same argument as in the proof of Lemma 21, the function $(x, x') \mapsto \mathbb{E}[(d_{\mathrm{tr}}(x, z) - d_{\mathrm{tr}}(x', z))^2]$ is convex on $A_{\mathrm{supp}(P),\ell} \times A_{\mathrm{supp}(P),\ell'}$ for all $\ell, \ell' \in L_{\mathrm{supp}(P)}$, so its maximum must be achieved on $\mathrm{ex}(A_{\mathrm{supp}(P)\ell} \times A_{\mathrm{supp}(P),\ell'}) = \mathrm{ex}(A_{\mathrm{supp}(P),\ell}) \times \mathrm{ex}(A_{\mathrm{supp}(P),\ell'})$. Because of the intersection with A, the result follows. □

Next, we recall [136] which shows that $M_1(\bar{P}_n)$ is a compact convex set, and the extreme points of the polyhedra in $\mathscr{P}_{\{Y_1,\ldots,Y_n\}}$ can be found as the projection onto the first k coordinates of the extreme points of the following polyhedron

$$\left\{(v, c) \in \mathbb{R}^k \times \mathbb{R}^n : c_l \geq v_i - v_j - (Y_l)_i + (Y_l)_j,\ l \in [n], i, j \in [k]\right\}.$$

As a result, for any $\varepsilon \geq 0$, we can find the extreme points of $M_1(P, \varepsilon)$ and the extreme points of the partition $\mathscr{P}_{\{Y_1,\ldots,Y_n\}}$ falling in $M_1(P, \varepsilon)$ by taking the projection onto the first k coordinates of the extreme points of the polyhedron

$$\left\{(v, c) \in \mathbb{R}^k \times \mathbb{R}^n : \frac{1}{n}\sum_{l\in[n]} c_l \leq V_1(\bar{P}_n) + \varepsilon, c_l \geq v_i - v_j - (Y_l)_i + (Y_l)_j, \forall l \in [n], i, j \in [k]\right\}.$$

Thus Proposition 6 and some standard subroutines from polyhedral geometry allow us to exactly compute $\hat{\sigma}_1(P, M_1(P, \varepsilon))$.

Lastly, let us mention that, for a probability measure $P \in \mathcal{P}(\mathcal{U}_N)$, it is possible to think of the Fermat-Weber set M_1 as being computed in either of

$$(\mathbb{R}^{\binom{N}{2}}/\mathbb{R}\mathbf{1}, d_{\mathrm{tr}}) \text{ or } (\mathcal{U}_N, d_{\mathrm{tr}}).$$

In general, Fréchet means do not behave well with respect to subspace restriction, but in this case it turns out that the two problems are closely related. To state a precise result, let us write $M_1^S(P,\varepsilon)$ for the ε-relaxed Fermat-Weber set, computed in the space (S, d_{tr}), where S is an arbitrary closed subset of the tropical projective space. Then we have the following:

Theorem 17. *For $N \in \mathbb{N}$ and $P \in \mathcal{P}(\mathbb{R}^{\binom{N}{2}}/\mathbb{R}\mathbf{1})$ with $P(\mathcal{U}_N) = 1$, and for any $\varepsilon \geq 0$, we have*

$$M_1^{\mathcal{U}_N}(P,\varepsilon) = M_1^{\mathbb{R}^{\binom{N}{2}}/\mathbb{R}\mathbf{1}}(P,\varepsilon) \cap \mathcal{U}_N.$$

Proof. It suffices to show $M_1(P) \cap \mathcal{U}_N \neq \varnothing$, since then the Fréchet functional attains the same minimum on both the space of equidistant phylogenetic trees $\mathcal{U}_N$ and the ambient tropical projective space. To do this, we write $E = \{\{i,j\} : 1 \leq i < j \leq N\}$ for the set of all pairs of leaves. Then we consider the optimization problem

$$\begin{cases} \text{minimize} & \sum_{e\in E}(u_e - u_{\{1,2\}}) \\ \text{over} & u \in M_1(P). \end{cases} \tag{7.9}$$

(Note that the objective is invariant under the action of $\mathbb{R}\mathbf{1}$, as it must be.) Recall that $M_1(P)$ is compact (Corollary 4) and the objective is continuous, so there exists a minimizer u^* of (7.9).

We claim $u^* \in \mathcal{U}_N$. Assume for the sake of contradiction that this is not the case, so that there exist distinct $i,j,k \in [N]$ with $u^*_{\{i,k\}} > \max(u^*_{\{i,j\}}, u^*_{\{j,k\}})$. Now define $u^{**} \in \mathbb{R}^E$ via

$$u_e^{**} := \begin{cases} u_e & \text{if } e \neq \{i,k\}, \\ \max(u^*_{\{i,j\}}, u^*_{\{j,k\}}) & \text{if } e = \{i,k\}. \end{cases}$$

Observe that, for any $w \in \mathcal{U}_N$, we have

$$\begin{aligned} u^*_{\{i,k\}} - w_{\{i,k\}} &> \max(u^*_{\{i,j\}}, u^*_{\{j,k\}}) - \max(w_{\{i,j\}}, w_{\{j,k\}}) \\ &\geq \min(u^*_{\{i,j\}} - w_{\{i,j\}}, u^*_{\{j,k\}} - w_{\{j,k\}}). \end{aligned}$$

This means that, for both $u \in \{u^*, u^{**}\}$, the map $e \mapsto u_e - w_e$ is minimized at a coordinate different than $e = \{i,k\}$. Since $u^{**}_{\{i,k\}} < u^*_{\{i,k\}}$ and since u^* and u^{**} coincide on all coordinates other than $e = \{i,k\}$, we conclude

$$\begin{aligned} d_{\mathrm{tr}}(u^*, w) &= \max_{e\in E}(u_e^* - w_e) - \min_{e\in E}(u_e^* - w_e) \\ &\geq \max_{e\in E}(u_e^{**} - w_e) - \min_{e\in E}(u_e^{**} - w_e) = d_{\mathrm{tr}}(u^{**}, w) \end{aligned}$$

for all $w \in \mathcal{U}_N$. Now note that, for all $u \in \mathcal{U}_N$:

$$\begin{aligned} W_1(P, u^{**}, u) &= \int_{\mathcal{U}_N} (d_{\mathrm{tr}}(u^{**}, w) - d_{\mathrm{tr}}(u,w))\, \mathrm{d}P(w) \\ &= \int_{\mathcal{U}_N} (d_{\mathrm{tr}}(u^*, w) - d_{\mathrm{tr}}(u,w))\, \mathrm{d}P(w) = W_1(P, u^*, u) \leq 0. \end{aligned}$$

This implies $u^{**} \in M_1(P)$. However, we have

$$\sum_{e\in E}(u_e^* - u_{\{1,2\}}^*) > \sum_{e\in E} u_e^{**} - u_{\{1,2\}}^{**},$$

which contradicts the optimality of u^* for the optimization problem (7.9). This shows $u^* \in M_1(P) \cap \mathcal{U}_N$, which ends the proof of the result. □

7.2 The Estimation Procedure

In this section we construct a procedure which estimates the Fermat-Weber set of an unknown population distribution on the basis of IID samples.

To motivate this, let us recall that in Theorem 16 we saw that $cn^{-1/2}(\log\log n)^{1/2}$-relaxed Fermat-Weber set estimators are d_{H}-consistent estimators of $M_1(P)$, provided that $c > \sigma_1(P)$. Also, we saw that we can consistently estimate $\sigma_1(P)$ via $\hat{\sigma}_1(P, M_1(P, \varepsilon_n))$ where ε_n is a consistent but sub-optimal relaxation rate. This motivates a certain sort of "two-step estimator": first, compute the relaxed Fermat-Weber set for the relaxation rate $\varepsilon_{1,n} := n^{-1/2}\log\log n$, and use this to compute the approximate fluctuation variation $c_n := \hat{\sigma}_1(P, M_1(P, \varepsilon_{1,n}))$; second, use the adaptively-chosen relaxation rate $c_n n^{-1/2}(\log\log n)^{1/2}$ to compute a better estimate. These considerations lead us to Algorithm 2, where the multi-step procedure AdaptRelaxFermatWeberSet can be seen as a natural extension of the two-step procedure described above.

Theorem 18. *For any $P \in \mathcal{P}(\mathbb{R}^k/\mathbb{R}\mathbf{1})$, AdaptRelaxFermatWeberSet$(Y_1, \dots, Y_n)$ is a strongly d_{H}-consistent estimator of $M_1(P)$. Moreover, the terminal relaxation rate $\varepsilon_{s^*,n,\delta}$ is asymptotically optimal in that it satisfies*

$$\lim_{\delta\to 0}\lim_{n\to\infty} \varepsilon_{s^*,n,\delta}\sqrt{\frac{n}{\log\log n}} = \sigma_1(P)$$

almost surely.

Proof. We write $s^* \in \mathbb{N}$ for the last index appearing in the **repeat** loop of line 22.

We begin by deriving some lower bounds on $\varepsilon_{s^*,n,\delta}$. For $0 < \eta < \delta$, write $\varepsilon_{n,\eta} := (1+\eta)^{1/2}\sigma_1(P)n^{-1/2}(\log\log n)^{1/2}$. Also, for $\varepsilon \geq 0$ let us define

$$\sigma_1(P, \varepsilon) := \sqrt{2} \sup_{x,x' \in M_1(P,\varepsilon)} \sqrt{\mathrm{Var}_P(d_{\mathrm{tr}}(x, Y_1) - d_{\mathrm{tr}}(x', Y_1))}. \tag{7.10}$$

Using the same argument as in the proof of Theorem 16 leading to (6.11), we see that, for $0 < \eta < \delta$, the event

$$E_\eta := \{M_1(P) \subseteq M_1(\bar{P}_n, \varepsilon_{n,\eta}) \text{ for sufficiently large } n \in \mathbb{N}\} \cap \left\{\liminf_{n\to\infty} \sigma_1(\bar{P}_n, \varepsilon_{n,\eta}) \geq \sigma_1(P)\right\}$$

Algorithm 2 In the tropical projective space $(\mathbb{R}^k/\mathbb{R}\mathbf{1}, d_{\mathrm{tr}})$, we define a procedure AdaptRelaxFermatWeberSet which (by Theorem 18) adaptively finds the optimal relaxation rate for strongly d_{H}-consistent estimation.

1: **procedure** FermatWeberValue
2: **input:** data $Y_1, \ldots, Y_n \in \mathbb{R}^k$
3: **output:** optimal objective $V_1(\bar{P}_n) \geq 0$
4: $V_n \leftarrow$ **minimize** $\frac{1}{n}\sum_{l=1}^{n} c_l$
5: **over** $(v, c) \in \mathbb{R}^{k \times n}$
6: **with** $c_l \geq v_i - v_j - (Y_l)_i + (Y_l)_j$ for all $l \in [n],\ i, j \in [k]$
7: **return** V_n
8: **procedure** FermatWeberSet
9: **input:** data $Y_1, \ldots, Y_n \in \mathbb{R}^k$ and relaxation scale $\varepsilon_n \geq 0$
10: **output:** extreme point set $\{v_j : j \in I_n\}$ of $M_1(\bar{P}_n, \varepsilon_n)$
11: $V_n \leftarrow$ FermatWeberValue$(Y_1, \ldots, Y_n)$
12: $S \leftarrow \{(v, c) \in \mathbb{R}^k \times \mathbb{R}^n : \frac{1}{n}\sum_{l=1}^{n} c_l \leq V_n + \varepsilon_n$, and
13: $c_l \geq v_i - v_j - (Y_l)_i + (Y_l)_j$ for all $l \in [n],\ i, j \in [k]\}$
14: $\{(v_j, c_j) \in \mathbb{R}^k \times \mathbb{R}^n : j \in I_n\} \leftarrow$ ExtremePoints(S)
15: **return** $\{v_j : j \in I_n\}$
16: **procedure** AdaptRelaxFermatWeberSet
17: **input:** data $Y_1, \ldots, Y_n \in \mathbb{R}^k$ and small constant $\delta > 0$
18: **output:** subset $M_n^* \subseteq \mathbb{R}^k$
19: $V_n \leftarrow$ FermatWeberValue$(Y_1, \ldots, Y_n)$
20: $\varepsilon_{1,n,\delta} \leftarrow V_n n^{-1/2} \log\log n$
21: $s \leftarrow 0$
22: **repeat**
23: $s \leftarrow s + 1$
24: $\{v_{s,j} : j \in I_{s,n}\} \leftarrow$ FermatWeberSet$(Y_1, \ldots, Y_n, \varepsilon_{s,n,\delta})$
25: $c_{s,n} \leftarrow$ **maximize** $(\frac{1}{n}\sum_{i=1}^{n} |d_{\mathrm{tr}}(v_{s,j}, Y_i) - d_{\mathrm{tr}}(v_{s,j'}, Y_i)|^2)^{1/2}$
26: **over** $j, j' \in I_{s,n}$
27: $\varepsilon_{s+1,n,\delta} \leftarrow (2 + \delta)^{1/2} c_{s,n} n^{-1/2} (\log\log n)^{1/2}$
28: **until** $\varepsilon_{s+1,n,\delta} \geq \varepsilon_{s,n,\delta}$
29: **return** ConvexHull$(\{v_{s,j} : j \in I_{s,n}\})$

has full probability. Now we claim that, on E_η, we have $\varepsilon_{s,n,\delta} \geq \varepsilon_{n,\eta}$ for sufficiently large $n \in \mathbb{N}$. We prove this by induction, for which the base case $s = 1$ follows immediately from $\varepsilon_{1,n,\delta} = \omega(n^{-1/2}(\log\log n)^{1/2})$. For the inductive step assume that the result holds for $s \in \mathbb{N}$. Then we can lower bound the constant $c_{s,n}$ appearing in the **maximize/over** statement in lines 25–26 by using Proposition 6, the inductive hypothesis, Jensen's inequality, and the two events comprising E_η:

$$\begin{aligned} c_{s,n} &= \hat{\sigma}_1(\bar{P}_n, \varepsilon_{s,n,\delta}) \\ &= \sqrt{\max_{x,x' \in M_1(\bar{P}_n, \varepsilon_{s,n,\delta})} \frac{1}{n} \sum_{i=1}^{n} (d_{\mathrm{tr}}(x, Y_i) - d_{\mathrm{tr}}(x', Y_i))^2} \\ &\geq \sqrt{\max_{x,x' \in M_1(\bar{P}_n, \varepsilon_{n,\eta})} \frac{1}{n} \sum_{i=1}^{n} (d_{\mathrm{tr}}(x, Y_i) - d_{\mathrm{tr}}(x', Y_i))^2} \\ &\geq \frac{\sigma_1(\bar{P}_n, \varepsilon_{n,\eta})}{\sqrt{2}} \\ &\geq \sqrt{\frac{1+\eta}{2+\delta}} \sigma_1(P), \end{aligned}$$

for sufficiently large $n \in \mathbb{N}$. As a result, we obtain

$$\varepsilon_{s+1,n,\delta} = c_{s,n} \sqrt{\frac{(2+\delta)\log\log n}{n}} \geq \sigma_1(P) \sqrt{\frac{(1+\eta)\log\log n}{n}} = \varepsilon_{n,\eta},$$

for sufficiently large $n \in \mathbb{N}$, and the induction is complete. In particular, the claim holds for the terminal index s^*, hence

$$E_\eta \subseteq \{\varepsilon_{s^*,n,\delta} \geq \varepsilon_{n,\eta} \text{ for sufficiently large } n \in \mathbb{N}\}.$$

Consequently, on E_η, we have $M_1(P) \subseteq M_1(\bar{P}_n, \varepsilon_{n,\eta}) \subseteq M_1(\bar{P}_n, \varepsilon_{s^*,n,\delta})$ for sufficiently large $n \in \mathbb{N}$, as well as

$$\liminf_{n\to\infty} \varepsilon_{s^*,n,\delta} \sqrt{\frac{n}{\log\log n}} \geq \liminf_{n\to\infty} \varepsilon_{n,\eta} \sqrt{\frac{n}{\log\log n}} = \sqrt{\frac{2+\eta}{2}} \sigma_1(P).$$

Now define the event $E := \bigcap_{\eta \in (0,\delta)\cap\mathbb{Q}} E_\eta$ which has full probability. For any $\delta > 0$, we have and $M_1(P) \subseteq M_1(\bar{P}_n, \varepsilon_{s^*,n,\delta})$ for sufficiently large $n \in \mathbb{N}$, as well as

$$\liminf_{n\to\infty} \varepsilon_{s^*,n,\delta} \sqrt{\frac{n}{\log\log n}} \geq \sqrt{\frac{2+\delta}{2}} \sigma_1(P), \tag{7.11}$$

on E.

For an upper bound on $\varepsilon_{s^*,n,\delta}$, we use Theorem 16 to see that the event

$$F = \left\{ \limsup_{n\to\infty} \sigma_1(\bar{P}_n, \varepsilon_{1,n,\delta}) \leq \sigma_1(P) \right\}$$

has full probability. Next, recall by construction of the relaxed Fréchet mean set that the Fréchet functional takes values in $[V_1(\bar{P}_n), V_1(\bar{P}_n) + \varepsilon_{1,n,\delta}]$ for any $x, x' \in M_1(\bar{P}_n, \varepsilon_{1,n,\delta})$ and consequently

$$c_{1,n} = \sqrt{\max_{x,x' \in M_1(\bar{P}_n, \varepsilon_{1,n,\delta})} \frac{1}{n} \sum_{i=1}^{n} (d_{\mathrm{tr}}(x, Y_i) - d_{\mathrm{tr}}(x', Y_i))^2} \leq \sqrt{\frac{\sigma_1(\bar{P}_n, \varepsilon_{1,n,\delta})^2}{2} + (\varepsilon_{1,n,\delta})^2}$$

by Proposition 6. In particular, on F, we obtain

$$\limsup_{n\to\infty} \varepsilon_{2,n,\delta} \sqrt{\frac{n}{\log\log n}} \leq \sqrt{\frac{2+\delta}{2}} \limsup_{n\to\infty} \sigma_1(\bar{P}_n, \varepsilon_{1,n,\delta}) \leq \sqrt{\frac{2+\delta}{2}} \sigma_1(P).$$

Now note that we always have $\varepsilon_{s^*,n,\delta} \leq \varepsilon_{2,n,\delta}$, so we obtain

$$\limsup_{n\to\infty} \varepsilon_{s^*,n,\delta} \sqrt{\frac{n}{\log\log n}} \leq \sqrt{\frac{2+\delta}{2}} \sigma_1(P) \tag{7.12}$$

on F.

Now we put the pieces together. By combining (7.11) and (7.12), we get

$$\lim_{n\to\infty} \varepsilon_{s^*,n,\delta} \sqrt{\frac{n}{\log\log n}} = \sqrt{\frac{2+\delta}{2}} \sigma_1(P)$$

on $E \cap F$. In particular, we have

$$\lim_{\delta\to 0} \lim_{n\to\infty} \varepsilon_{s^*,n,\delta} \sqrt{\frac{n}{\log\log n}} = \sigma_1(P)$$

almost surely, as claimed. Additionally, from (7.11) and part (i) of Theorem 16, we conclude $d_{\mathrm{H}}(M_1(P), M_1(\bar{P}_n, \varepsilon_{s^*,n,\delta})) \to 0$ almost surely, as claimed. □

It is also useful to mention some heuristic modifications which can be applied to Algorithm 2; while these modifications render our optimality and consistency proofs invalid, we find that they make no difference in practice. One modification concerns line 15: Instead of returning the set $\{v_j : j \in I_n\}$, one can instead **return** ExtremePoints(ConvexHull($\{v_j : j \in I_n\}$)). This shrinks the resulting set of vertices, which significantly improves the speed of the **maximize**/**over** optimization problem in lines 25–26. Another modification concerns line 27: We find that one can simply take $\delta = 0$.

7.3 Experiments

In this section we perform some experiments to test the d_{H}-consistency results we have just proved about AdaptRelaxFermatWeberSet. The experiments are based on simulated data and real data, contained in the next two subsections.

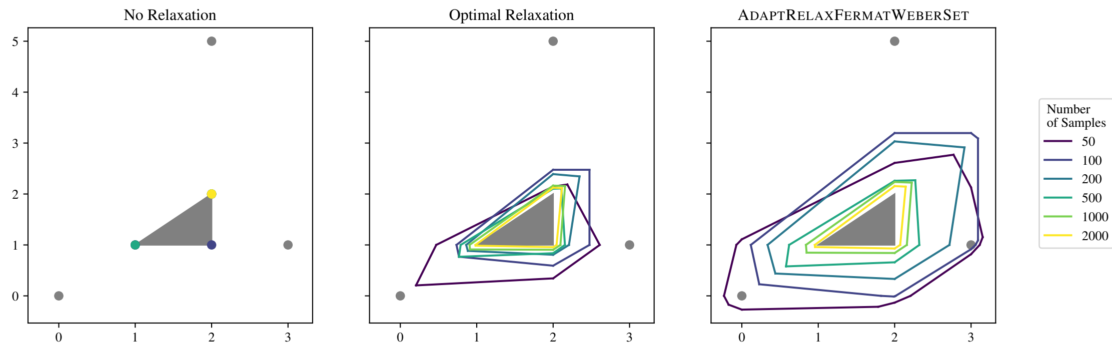

Figure 7.1: A toy model for Fermat-Weber set estimation in tropical projective space.

Simulated data

We consider the tropical projective space $\mathbb{R}^3/\mathbb{R}\mathbf{1}$. Because of the quotient, every point $\mathbb{R}^3/\mathbb{R}\mathbf{1}$ can be represented by a point in $\mathbb{R}^2$, if we simply set the first coordinate to be equal to zero. Hence, in what follows, we identify $\mathbb{R}^3/\mathbb{R}\mathbf{1}$ with $\mathbb{R}^2$.

Now we consider the example given in [136, Example 7], where the population distribution is

$$P = \frac{1}{3}\delta_{(0,0,0)} + \frac{1}{3}\delta_{(0,3,1)} + \frac{1}{3}\delta_{(0,2,5)},$$

and for which the Fermat-Weber set is the triangle with vertices $(0,1,1)$, $(0,2,1)$, and $(0,2,2)$. As before, we let $Y_1, Y_2, \ldots$ denote IID samples from P, and we consider the question of estimating $M_1(P)$ from the data $Y_1, \ldots, Y_n$ alone.

In Figure 7.1, we consider 3 possible estimators, computed on simulated data for number of samples equal to $n \in \{50, 100, 200, 500, 1000, 2000\}$. The first is the unrelaxed Fermat-Weber set. The second is the relaxed Fréchet mean set estimator with optimal relaxation rate $2\cdot 3^{-1/2}\cdot n^{-1/2}(\log\log n)^{1/2}$; note that we can exactly compute the critical pre-factor to be $\sigma_1(P) = 2\cdot 3^{-1/2}$ since in this case we know the population distribution, but that in general we do not have access to this. The third is the estimator AdaptRelaxFermatWeber-Set from Algorithm 2 which, by Theorem 18, adaptively finds the asymptotically optimal relaxation rate.

We make the following observations about this example. First, the empirical unrelaxed Fermat-Weber set is typically a single point: It jumps around the vertices of the population Fermat-Weber set, but it does not, for a fixed value of n, appear to give a reasonable estimator. Second, we note that the empirical adaptively-relaxed Fermat-Weber set converges somewhat slowly to its population counterpart, when compared to the empirical optimally-relaxed Fermat-Weber set. We found this behavior to be extremely stable across multiple trials of the same experiment.

Influenza data

In this subsection, we consider estimating the Fermat-Weber set of an unknown population distribution of evolutionary trees of hemagglutinin genome sequences. We take the data set from [106], which processes the publicly-available GI-SAID EpiFlu$^{\text{TM}}$ data from 1995 in New York state. We refer the reader to [106, Section 6.3] for details of the pre-processing used to construct the data set. For our purposes, the data consists of tens of thousands of phylogenetic trees, each of which has 4 or 5 leaves.

Estimation in tropical projective space

First we consider estimating the Fermat-Weber set of our tree data in $(\mathbb{R}^{\binom{N}{2}}/\mathbb{R}\mathbf{1}, d_{\text{tr}})$, the ambient tropical projective space. For this part, we consider the 4-leaf data set, i.e., $N = 4$, so the data naturally lies in dimension $\binom{4}{2} = 6$; by setting a specified coordinate to 0, we can identify $\mathbb{R}^6/\mathbb{R}\mathbf{1}$ with $\mathbb{R}^5$.

We compute the output of both the unrelaxed Fermat-Weber set and the procedure AdaptRelaxFermatWeberSet for a number of data points $n \in \{5, 10, 20, 30, 40, 50\}$, and the results are shown in Figure 7.3. In order to visualize these estimators, we plot the projection of the resulting 5-dimensional polytopes onto 2-dimensional subspaces chosen uniformly at random. The results for 3 random subspaces are shown in Figure 7.3. We observe that the output of AdaptRelaxFermatWeberSet seems to provide a quite conservative outer estimate of $M_1(P)$ compared to the small (but somewhat unstable) unrelaxed Fermat-Weber set.

Estimation in equidistant tree space

Next we consider estimating the Fermat-Weber set of our tree data in $(\mathcal{U}_N, d_{\text{tr}})$, the space of equidistant trees. For this part, we consider the 5-leaf data set. This means that $N = 5$, so we can think of our data as lying in a union of $(2 \cdot 5 - 3)!! = 105$ orthants each of dimension $5 - 2 = 3$.

We compute the output of both the unrelaxed Fermat-Weber set and the procedure AdaptRelaxFermatWeberSet for $n = 12$ data points. For the unrelaxed Fermat-Weber set, we find that it contains a single equidistant tree. For AdaptRelaxFermatWeberSet, we find that the output only intersects 3 orthants of $\mathcal{U}_5$, which means that only 3 different tree topologies are represented in the estimated set; in order to visualize this, we sample an extreme point uniformly at random from each polytope that results from intersecting the estimated set with each non-trivial topology. The results can be seen in Figure 7.3. While the estimator from AdaptRelaxFermatWeberSet can be more difficult to interpret, it robustly identifies a few notable qualitative features: 4 and 5 share the most recent common ancestor, 1 and 2 share the most distant common ancestor, and 3 lies somewhere in between.

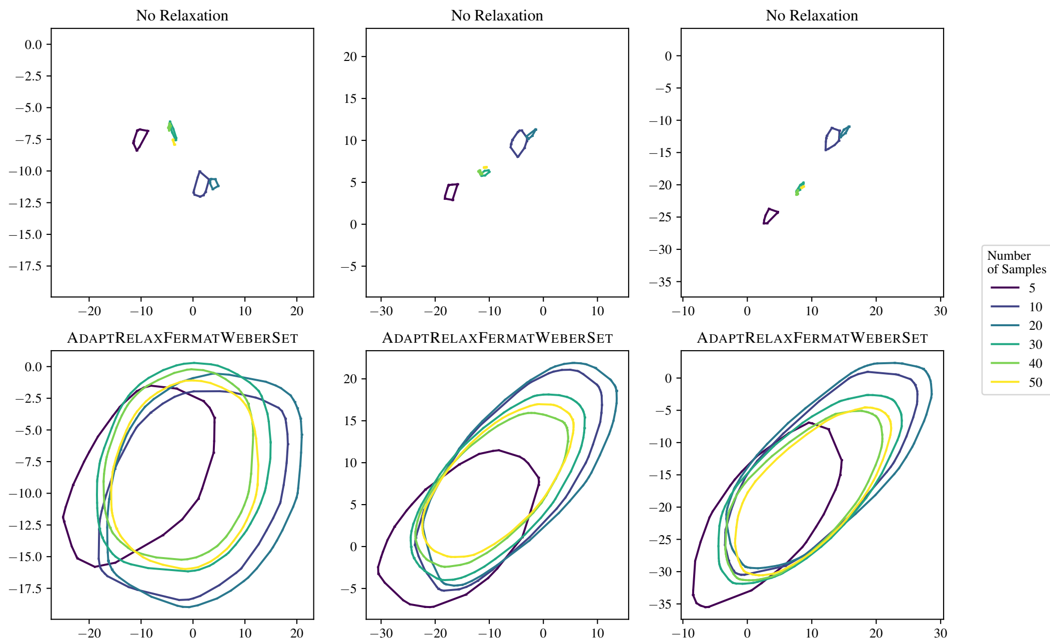

Figure 7.2: Estimating the Fermat-Weber set $M_1(P)$ of the 4-leaf inluenza data, in the ambient tropical projective space $\mathbb{R}^6/\mathbb{R}\mathbf{1}$, using the empirical Fermat-Weber mean set (top row) and the output of AdaptRelaxFermatWeberSet (bottom row). To visualize the set estimators, we plot random projections onto 2-dimensional subspaces chosen uniformly at random (one projection per column).

Discussion

In this example of tropical projective space, we have seen that the abstract optimality result of Theorem 16 can be implemented as AdaptRelaxFermatWeberSet in Algorithm 2 in order to estimate an unknown population Fermat-Weber set. On simulated and real data, we have seen that this estimator provides a different view of the estimand compared to the unrelaxed empirical Fermat-Weber set.

Let us make some basic remarks. On the positive side, we see that relaxation methods provide increased stability for the estimation problem. This is contrasted with the unrelaxed empirical Fermat-Weber set, which can be highly sensitive to even a single data point. On the negative side, it seems that even the asymptotically optimal procedure AdaptRelaxFermatWeberSet provides a very conservative outer estimate. In this way, relaxation methods can be seen as a complement to, rather than a replacement of, unrelaxed estimators.

We also make some comments about computation. For one, the main bottleneck of Algorithm 2 is the combinatorial optimization problem in line 25–26; its time complexity is quadratic in the number of extreme points of the previous Fermat-Weber set estima-

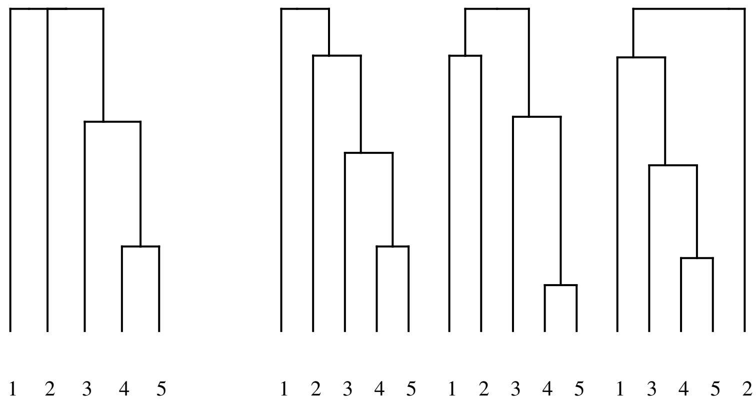

Figure 7.3: Estimating the Fermat-Weber set $M_1(P)$ of the 5-leaf inluenza data, in tropical projective treespace $\mathcal{U}_5$. We find that the empirical Fermat-Weber set is a singleton (left) while the output of AdaptRelaxFermatWeberSet intersects 3 different tree topologies (right); for the latter, we plot one extremal tree (sampled uniformly at random) from each represented topology.

tor, and this number of extreme points appears to grow exponentially with the dimension of the problem. For example, the computation of AdaptRelaxFermatWeberSet in Part 7.3 above already has roughly 3,400 extreme points. Another comment is that, in many settings one needs n to be very large in order for the asymptotically adaptively optimal rate in $\Theta(n^{-1/2}(\log\log n)^{1/2})$ to become smaller than the asymptotically sub-optimal rate of $V_1(\bar{P}_n)n^{-1/2}\log\log n$. For these reasons, we believe that a useful heuristic in practice is just to compute the relaxed empirical Fermat-Weber set for the relaxation rate $V_1(\bar{P}_n)n^{-1/2}\log\log n$.

Part III

Clustering

Chapter 8

Limit Theorems for Adaptive Clustering

The asymptotic theory of k-means clustering was initiated by the work of Pollard [165] and Abaya and Wise [1], and has seen many developments in recent years. However, existing results are lacking in a key way: They do not provide any guarantees for "adaptive" variants of clustering that are used in practice by most applied statisticians and data scientists. To explain these procedures further, let $\mathcal{H}$ denote a (possibly infinite-dimensional) Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$, and let $Y_1, Y_2, \ldots$ be a sequence of points in $\mathcal{H}$.

The first of these procedures is a slight generalization of k-means.

***R*-restricted *k*-means.** Fix $k \in \mathbb{N}$ and $R \subseteq \mathcal{H}$. Then consider

$$\begin{cases} \text{minimize} & \frac{1}{n}\sum_{i=1}^{n} \min_{x\in S} \|x - Y_i\|^2 \\ \text{over} & S \subseteq R \\ \text{with} & 1 \leq \#S \leq k. \end{cases} \tag{8.1}$$

When $R = \mathcal{H}$, this is exactly k-means clustering, but the set R provides a uniform way to restrict the feasible set of cluster centers. Such considerations might arise, for example, in scientific settings where a priori domain knowledge dictates that cluster centers must lie in a "meaningful region", or in engineering settings where physical constraints dictate that cluster centers must lie in an "attainable region".

Although the setting of R-restricted k-means is slightly different from the class of procedures considered by Pollard in [165], the methods therein can be easily modified to show strong consistency under the analogous technical hypotheses; the limit of any sequence of solutions turns out to be, as expected, a solution to the analogous population problem. In the setting above we call k the *number of clusters* and R the *domain of the cluster centers*.

By an *adaptive* clustering procedure we mean one where at least one of the number of clusters k or the domain of the cluster centers R is not fixed but rather is taken to be a measurable function of the data. The most important adaptive clustering procedures, from the point of view of applications in machine learning, are as follows:

***k*-medoids.** Fix $k \in \mathbb{N}$, and consider

$$\begin{cases} \text{minimize} & \frac{1}{n}\sum_{i=1}^{n} \min_{x\in S} \|x - Y_i\|^2 \\ \text{over} & S \subseteq \{Y_1, \dots Y_n\} \\ \text{with} & 1 \le \#S \le k. \end{cases} \tag{8.2}$$

By constraining the domain of the clusters centers in this way, we force the cluster centers to be bona fide data points rather than abstract points $\mathcal{H}$; such procedures are particularly important in interpretable machine learning where it is sometimes desirable that each cluster center be an actual datum which serves as a prototype for the whole cluster.

Elbow-method *k*-means. For each $k \in \mathbb{N}$, set

$$V_k := \inf_{\substack{S\subseteq\mathcal{H} \\ 1\le\#S\le k}} \frac{1}{n}\sum_{i=1}^{n} \min_{x\in S} \|x - Y_i\|^2,$$

which is the minimal objective achievable by any set of cluster centers. Then define the discrete second derivative for $k \ge 2$ via $\Delta^2 V_k := V_{k+1} + V_{k-1} - 2V_k$, and set

$$k^{\text{elb}} := \min\{\arg\max\{\Delta^2 V_k : k \in \mathbb{N}, k \ge 2\}\}.$$

Note that the restriction $k \ge 2$ can be understood as taking the convention that $V_0 = \infty$, which equivalently means that $k = 1$ will never be selected as the number of clusters. Now we consider

$$\begin{cases} \text{minimize} & \frac{1}{n}\sum_{i=1}^{n} \min_{x\in S} \|x - Y_i\|^2 \\ \text{over} & S \subseteq \mathcal{H} \\ \text{with} & 1 \le \#S \le k^{\text{elb}}. \end{cases} \tag{8.3}$$

This is a naive formalism of the well-known method for choosing k, where one inspects the graph of $\{V_k\}_{k\in\mathbb{N}}$ and chooses the value of k for the which the added model complexity experiences maximally diminishing returns. Such procedures are ubiquitous in exploratory data science, where the number of clusters is not known and fixed ahead of time, but rather must itself be estimated. See Figure 8.1 for an illustration.

In this chapter, we prove a general consistency result which shows that all of these clustering procedures converge to a suitable population limit (Proposition 9), and we give several probabilistic applications. Our work largely parallels the results of Chapter 4 although we have to be much more careful with the set-valued analysis considerations. The main technical innovation which allows us to deal with the adaptive case is a general-purpose "continuity" result which shows that the set of sets of k-means cluster centers depends continuously on the number of cluster centers and on the domain of the cluster centers.

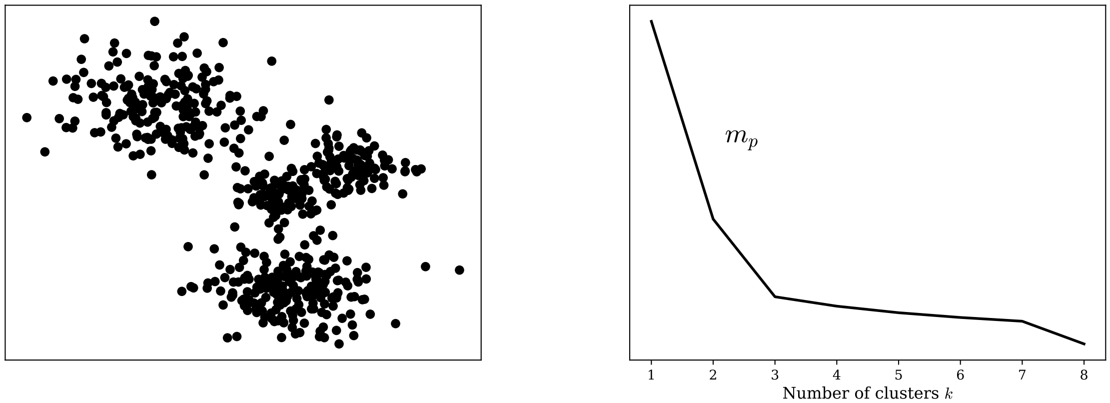

Figure 8.1: The elbow-method for choosing the number of clusters in k-means clustering. We compute the minimal objective achieveable by k-means clustering when k ranges over $\{1, 2, \ldots, 8\}$, and we plot the resulting graph. We select k to be the "elbow" of the plot, which is the point that maximizes the discrete second derivative.

8.1 Preliminaries

In this section we develop the basic results which will combine in the next sections to prove various limit theorems for a wide class of clustering procedures. Many of these results are slight generalizations of results from Chapter 4, where we now focus on the set-valued setting.

To begin, we write $\mathrm{C}(\mathcal{H})$ and $\mathrm{C}^w(\mathcal{H})$ for the collections of (norm) closed and weakly closed subsets of $\mathcal{H}$, respectively. Additionally, we write $\mathrm{K}(\mathcal{H})$ for the collection of all non-empty (norm) compact subsets of $\mathcal{H}$. In this subsection we introduce some notions of convergence for such spaces of subsets and establish some basic properties that we will later need. In order to clarify a possible confusion in the notation, let us emphasize that the symbol w always represents the topology of weak convergence of measures and that the symbol w always the topology of weak convergence in a Hilbert space.

To begin we introduce and review some basic properties of "Kuratowski convergence", sometimes called the "Kuratowski-Painlevé convergence". For $\{C_n\}_{n\in\mathbb{N}}$ arbitrary subsets of $\mathcal{H}$ we define the sets

$$\operatorname*{Ls}_{n\to\infty} C_n := \left\{ x \in \mathcal{H} : \begin{array}{c} \text{for all (norm) open neighborhoods } U \text{ of } x, \\ U \cap C_n \neq \varnothing \text{ for infinitely many } n \in \mathbb{N} \end{array} \right\}$$
$$\operatorname*{Li}_{n\to\infty} C_n := \left\{ x \in \mathcal{H} : \begin{array}{c} \text{for all (norm) open neighborhoods } U \text{ of } x, \\ U \cap C_n \neq \varnothing \text{ for large enough } n \in \mathbb{N} \end{array} \right\}.$$

called the *(norm) Kuratowski upper limit* and *(norm) Kuratowski lower limit*, respectively. More concretely, for a point $x \in \mathcal{H}$, we have $x \in \mathrm{Ls}_{n\in\mathbb{N}} C_n$ if and only if there exists some subsequence $\{n_j\}_{j\in\mathbb{N}}$ and some $x_j \in C_{n_j}$ for each $j \in \mathbb{N}$ such that $\|x_j - x\| \to 0$, and we have $x \in \mathrm{Li}_{n\in\mathbb{N}} C_n$ if and only if for any subsequence $\{n_j\}_{j\in\mathbb{N}}$ there exists a further subsequence $\{j_i\}_{i\in\mathbb{N}}$ and some $x_i \in C_{n_{j_i}}$ for each $i \in \mathbb{N}$ with $\|x_i - x\| \to 0$. We also write $\mathrm{Ls}^w_{n\in\mathbb{N}} C_n$ and $\mathrm{Li}^w_{n\in\mathbb{N}} C_n$ for the analogous notions for weak convergence, which can be defined mutatis mutandis.

Next we recall some notions of "Hausdorff convergence", which we already discussed briefly in Chapter 6. That is, for $x \in \mathcal{H}$ and $C' \in \mathrm{K}(\mathcal{H})$, write

$$d(x, C') := \min_{x'\in C'} \|x - x'\|$$

for the shortest distance from the point x to the set C'. Observe that C' being non-empty and compact implies that $d(x, C') < \infty$ and that the infimum is achieved. Now for $C, C' \in \mathrm{K}(\mathcal{H})$, write

$$\vec{d}_{\mathrm{H}}(C, C') := \max_{x\in C} d(x, C') = \max_{x\in C} \min_{x'\in C'} \|x - x'\|$$

for the largest possible shortest distance from a point in C to the set C'. Observe that C and C' being non-empty and compact imply that $\vec{d}_{\mathrm{H}}(C, C') < \infty$ and that the supremum and infimum are both achieved. Although $\vec{d}_{\mathrm{H}}$ is not a metric (it is not symmetric), it it easy to show that it satisfies the following type of triangle inequality

$$\vec{d}_{\mathrm{H}}(C, C'') \le \vec{d}_{\mathrm{H}}(C, C') + \vec{d}_{\mathrm{H}}(C', C'') \tag{8.4}$$

for $C, C', C'' \in \mathrm{K}(\mathcal{H})$. Additionally, it is easy to see that we have a version of (4.2): For any $r \ge 0$, the constant $c_r := \max\{1, 2^{r-1}\}$ satisfies

$$(\vec{d}_{\mathrm{H}}(C, C''))^r \le c_r \left((\vec{d}_{\mathrm{H}}(C, C'))^r + (\vec{d}_{\mathrm{H}}(C', C''))^r \right) \tag{8.5}$$

for $C, C', C'' \in \mathrm{K}(\mathcal{H})$.

We refer to $\vec{d}_{\mathrm{H}}$ as the *one-sided Hausdorff distance* (although it is not a metric), and its relationship to the Hausdorff metric of Chapter 6 is just

$$d_{\mathrm{H}}(C, C') := \max\{\vec{d}_{\mathrm{H}}(C, C'), \vec{d}_{\mathrm{H}}(C', C)\}$$

for $C, C' \in \mathrm{K}(\mathcal{H})$.

Note that The definition of $\vec{d}_{\mathrm{H}}$ immediately extends from to the case of $C, C' \in \mathrm{K}(\mathcal{H})$ to the case of non-empty $C, C' \in \mathrm{C}^w(\mathcal{H})$, provided that we replace the maximum and minimum with supremum and infimum and that we allow it to take values in the extended real half-line, $[0, \infty]$. So, if $\{C_n\}_{n\in\mathbb{N}}$ and C in $\mathrm{C}^w(\mathcal{H})$ are assumed only to be non-empty then the expression $\vec{d}_{\mathrm{H}}(C_n, C) \to 0$ is taken to mean that $\vec{d}_{\mathrm{H}}(C_n, C)$ is finite for sufficiently large $n \in \mathbb{N}$ and that it converges to zero as $n \to \infty$.

Now we give some important auxiliary results.

Lemma 22. *If $\{R_n\}_{n\in\mathbb{N}}$ in $\mathrm{C}(\mathcal{H})$ and C in $\mathrm{K}(\mathcal{H})$ have $C \subseteq \mathrm{Li}_{n\in\mathbb{N}} R_n$ and $\#C < \infty$, then, there exists a subsequence $\{n_j\}_{j\in\mathbb{N}}$ and $C_j \subseteq R_{n_j}$ with $\#C_j = \#C$ for all $j \in \mathbb{N}$ such that we have $d_{\mathrm{H}}(C_j, C) \to 0$.*

Proof. Write $C = \{x_1, \ldots, x_k\}$ for $k := \#C$, and set $\varepsilon := \min\{d(x_\ell, x_{\ell'}) : 1 \le \ell < \ell' \le k\} > 0$. Then apply $C \subseteq \mathrm{Li}^d_{n\in\mathbb{N}} R_n$ iteratively k times to get a subsequence $\{n_j\}_{j\in\mathbb{N}}$ and sets $C_j := \{x_{1,j}, \ldots, x_{k,j}\} \subseteq R_{n_j}$ for each $j \in \mathbb{N}$ such that we have $x_{\ell,j} \to x_\ell$ in d as $j \to \infty$ for all $1 \le \ell \le k$.

Next, observe that we have $\max_{1\le\ell\le k} d(x_{\ell,j}, x_\ell) < \varepsilon/3$ for sufficiently large $j \in \mathbb{N}$. Also by the triangle inequality, we have $\varepsilon \le d(x_\ell, x_{\ell'}) \le d(x_\ell, x_{\ell,j}) + d(x_{\ell,j}, x_{\ell',j}) + d(x_{\ell',j}, x_{\ell'})$ for all $1 \le \ell < \ell' \le k$. It follows that, for sufficiently large $j \in \mathbb{N}$ and all $1 \le \ell < \ell' \le k$, we have $0 < \frac{\varepsilon}{3} \le d(x_{\ell,j}, x_{\ell',j})$. This implies that $\#C_j = k$ for sufficiently large $j \in \mathbb{N}$.

By construction we have $\vec{d}_{\mathrm{H}}(C, C_j) \to 0$, so it only remains to show $C_j \to C$ in d_{H} as $j \to \infty$. To do this, note that for sufficiently large $j \in \mathbb{N}$ we have $\vec{d}_{\mathrm{H}}(C, C_j) < \varepsilon/2$ hence we can construct a map $\phi_j : C \to C_j$ by sending each $x \in C$ to some $\phi_j(x) \in C_j$ such that $d(x, \phi_j(x)) < \varepsilon/2$. Of course, if $\phi_j(x) = \phi_j(x')$ for $x, x' \in C$ then we have $d(x, x') \le d(x, \phi_j(x)) + d(x', \phi_j(x')) < \varepsilon$, hence $x = x'$ by the minimality of ε, hence $\phi_j : C \to C_j$ is injective. Since an injective map between finite sets of the same cardinality is automatically a bijection, there exists, for sufficiently large $j \in \mathbb{N}$, a well-defined function $\phi_j^{-1} : C_j \to C$ such that for all $x_j \in C_j$ we have $d(x_j, \phi_j^{-1}(x_j)) < \varepsilon/2$. This means we have $\vec{d}_{\mathrm{H}}(C_j, C) < \varepsilon/2$ for sufficiently large $j \ge N$, as needed. □

Lemma 23. *If $\{y_n\}_{n\in\mathbb{N}}$ and y in $\mathcal{H}$ have $y_n \to y$ and $\{S_n\}_{n\in\mathbb{N}}$ and S in $\mathrm{K}(\mathcal{H})$ have $S_n \to S$ in d_{H}, then $d(y_n, S_n) \to d(y, S)$.*

Proof. By (8.4), we have

$$d(y_n, S_n) \le d(y_n, S) + \vec{d}_{\mathrm{H}}(S, S_n)$$

and

$$d(y_n, S) \le d(y_n, S_n) + \vec{d}_{\mathrm{H}}(S_n, S).$$

Thus, the result follows from the well-known continuity of $d(\,\cdot\,, S) : \mathcal{H} \to \mathbb{R}$. □

Next we define the objective function which will be minimized in k-means clustering and its variants. That is, for $p \ge 1$, we define the function $W_p : \mathrm{K}(\mathcal{H}) \times \mathcal{P}_p(\mathcal{H}) \to [0, \infty)$ via

$$W_p(S, P) := \int_{\mathcal{H}} d^p(y, S)\, \mathrm{d}P(y) = \int_X \min_{x\in S} \|x - y\|^p\, \mathrm{d}P(y).$$

In order to study W_p and its convergence properties, we will (as in the case of Fréchet means from Chapter 4) need to employ various tricks with uniform integrability. The core to these is the following extension of Skorokhod's representation theorem:

Lemma 24. *For $p \ge 1$, suppose that $\{P_n\}_{n\in\mathbb{N}}$ and P in $\mathcal{P}_p(\mathcal{H})$ have $P_n \to P$ in τ_{w}^p. Then, there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with expectation $\mathbb{E}$, on which are defined random variables $\{Y^n\}_{n\in\mathbb{N}}$ and Y with laws $\{P_n\}_{n\in\mathbb{N}}$ and P, respectively, such that we have $Y^n \to Y$ almost surely and $\mathbb{E}[\|Y^n - Y\|^p] \to 0$ as $n \to \infty$.*

Proof. By the standard Skorokhod theorem [104, Theorem 4.30], we get a sequence of random variables with desired laws and with the desired almost sure convergence property, so it only remains to show that this coupling satisfies $\mathbb{E}[\|Y^n - Y\|^p] \to 0$ as $n \to \infty$. Indeed, recall from the definition of τ_{w}^p that there exists some $x \in \mathcal{H}$ with $\mathbb{E}[\|x - Y^n\|^p] \to \mathbb{E}[\|x - Y\|^p]$, hence $\{\|x - Y^n\|^p\}_{n\in\mathbb{N}}$ is uniformly integrable. Also, from (4.2) we have:

$$\|Y^n - Y\|^p \le c_p \left(\|x - Y^n\|^p + \|x - Y\|^p\right).$$

This implies that $\{\|Y^n - Y\|^p\}_{n\in\mathbb{N}}$ is uniformly integrable, and it also converges to zero $\mathbb{P}$-almost surely, hence we have $\mathbb{E}[\|Y^n - Y\|^p] \to 0$. □

From this, we get two important approximation results:

Lemma 25. *The function* $W_p : (\mathrm{K}(\mathcal{H}) \times \mathcal{P}_p(\mathcal{H}), d_{\mathrm{H}} \times \tau_{\mathrm{w}}^p) \to [0, \infty)$ *is continuous.*

Proof. Suppose $\{(S_n, P_n)\}_{n\in\mathbb{N}}$ and (S, P) in $\mathrm{K}(\mathcal{H}) \times \mathcal{P}_p(\mathcal{H})$ have $(S_n, P_n) \to (S, P)$ in $d_{\mathrm{H}} \times \tau_{\mathrm{w}}^p$, and let $(\Omega, \mathcal{F}, \mathbb{P})$ be as in Lemma 24. By Lemma 23, we have $d^p(Y^n, S_n) \to d^p(Y, S)$ almost surely. Now fix $x \in S$ arbitrarily and note that $\{\|x - Y^n\|^p\}_{n\in\mathbb{N}}$ is uniformly integrable. Then we use (8.5) to get

$$d^p(Y^n, S_n) \le c_p(\|x - Y^n\|^p + (\vec{d}_{\mathrm{H}}(S, S_n))^p),$$

hence $\{d^p(Y^n, S_n)\}_{n\in\mathbb{N}}$ is uniformly integrable. This implies $W_p(S_n, P_n) = \mathbb{E}[d^p(Y^n, S_n)] \to \mathbb{E}[d^p(Y, S)] = W_p(S, P)$ as $n \to \infty$, as claimed. □

Lemma 26. *If* $\{S_n\}_{n\in\mathbb{N}}$ *in* $\mathrm{K}(\mathcal{H})$ *and* $\{P_n\}_{n\in\mathbb{N}}$ *and* P *in* $\mathcal{P}_p(\mathcal{H})$ *have* $P_n \to P$ *in* τ_{w}^p*, then*

$$\liminf_{n\to\infty} W_p(S_n, P_n) = \liminf_{n\to\infty} W_p(S_n, P)$$

and

$$\limsup_{n\to\infty} W_p(S_n, P_n) = \limsup_{n\to\infty} W_p(S_n, P).$$

Proof. Get a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ as in Lemma 24. Then, for arbitrary $\varepsilon > 0$, use [202, equation (6.10)] to get a constant $C_\varepsilon > 0$ such that we have

$$\begin{aligned} W_p(S_n, P_n) &= \mathbb{E}\left[d^p(Y^n, S_n)\right] \\ &\le C_\varepsilon \mathbb{E}\left[\|Y^n - Y\|^p\right] + (1+\varepsilon)\mathbb{E}\left[d^p(Y, S_n)\right] \\ &= C_\varepsilon \mathbb{E}\left[\|Y^n - Y\|^p\right] + (1+\varepsilon) W_p(S_n, P). \end{aligned}$$

Then take $n \to \infty$ and $\varepsilon \to 0$ to get

$$\liminf_{n\to\infty} W_p(S_n, P_n) \le \liminf_{n\to\infty} W_p(S_n, P). \tag{8.6}$$

Similarly, we can bound, for any $\varepsilon > 0$:

$$\begin{aligned} W_p(S_n, P) &= \mathbb{E}\left[d^p(Y, S_n)\right] \\ &\leq C_\varepsilon \mathbb{E}\left[\|Y - Y^n\|^p\right] + (1+\varepsilon)\mathbb{E}\left[d^p(Y^n, S_n)\right] \\ &= C_\varepsilon \mathbb{E}\left[\|Y - Y^n\|^p\right] + (1+\varepsilon) W_p(S_n, P_n). \end{aligned}$$

Thus,

$$\liminf_{n\to\infty} W_p(S_n, P) \leq \liminf_{n\to\infty} W_p(S_n, P_n). \tag{8.7}$$

Combining (8.6) and (8.7) gives the desired lim inf equality, and the lim sup equality is proved in the same way. □

8.2 Analysis of the Clustering Map

In this section, we show that various natural maps used in the construction of clustering procedures are "continuous" with respect to various topologies of spaces of measures, space of sets, and on spaces of sets of sets. In particular, we show Proposition 9 and Proposition 10 which provide "continuity" of the clustering map, and we show Proposition 11 which provides sufficient conditions for the continuity of the adaptive choice of k arising in the elbow-method. Throughout this section, as before, $\mathcal{H}$ is a separable (possilbly infinite-dimensional) Hilbert space, and $p \geq 1$ an arbitrary exponent.

To begin, we give the basic notions of the clustering procedures of interest.

Definition 12. *For $P \in \mathcal{P}_p(\mathcal{H})$, $k \in \mathbb{N}$, and $R \in \mathrm{C}^w(\mathcal{H})$, set*

$$V_{k,p}(P, R) := \inf_{\substack{S' \subseteq R \\ 1 \leq \#S' \leq k}} W_p(S', P), \tag{8.8}$$

and also set $C_p(P, k, R)$ to be the set of all $S \subseteq R$ with $1 \leq \#S \leq k$ satisfying

$$W_p(S, P) \leq V_{k,p}(P, R). \tag{8.9}$$

If a set $S \subseteq X$ has $S \subseteq R$ and $1 \leq \#S \leq k$ it is called feasible *and if it achieves* (8.9) *it is called* optimal. *Note that $C_p(P, k, R)$ is empty if there are no optimal sets or if R is empty. We refer to $C_p(P, k, R)$ as the* set of sets of R-restricted (k, p)-means clustering centers. *For $\varepsilon \geq 0$, we also set $C_p(P, k, R; \varepsilon)$ to be the set of all $S \subseteq R$ with $1 \leq \#S \leq k$ satisfying*

$$W_p(S, P) \leq V_{k,p}(P, R) + \varepsilon, \tag{8.10}$$

called the set of sets of ε-relaxed R-restricted (k, p)-means clustering centers.

Proposition 7. *Suppose that $\{R_n\}_{n\in\mathbb{N}}$ and R in $\mathrm{C}^w(\mathcal{H})$ satisfy $\mathrm{Ls}^w_{n\in\mathbb{N}} R_n \subseteq R$, and that $\{S_n\}_{n\in\mathbb{N}}$ satisfy $S_n \subseteq R_n$ and $\#S_n = k$ for all $n \in \mathbb{N}$. Then, for any labeling $S_n = \{a_1^n, \ldots, a_k^n\}$ for all $n \in \mathbb{N}$, there exists a subsequence $\{n_j\}_{j\in\mathbb{N}}$ such that, for each $1 \leq \ell \leq k$, exactly one of*

- $\{a_\ell^{n_j}\}_{j\in\mathbb{N}}$ *is unbounded, or*
- $\{a_\ell^{n_j}\}_{j\in\mathbb{N}}$ *converges weakly to some* $a_\ell \in R$

holds. Consequently, the set $S := \{a_\ell : 1 \le \ell \le k, \{a_\ell^{n_j}\}_{j\in\mathbb{N}} \text{ is bounded}\}$ *satisfies* $S \subseteq R$ *and* $\#S \le k$*, and we have*

$$d^p(y, S) \le \liminf_{j\to\infty} d^p(y, S_{n_j})$$

for all $y \in \mathcal{H}$*, hence*

$$W_p(S, P) \le \liminf_{j\to\infty} W_p(S_{n_j}, P)$$

for all $P \in \mathcal{P}_p(\mathcal{H})$*.*

Proof. We construct $\{n_j\}_{j\in\mathbb{N}}$ by iteratively applying the Banach-Alaoglu theorem: First, if $A_1 := \{a_1^n\}_{n\in\mathbb{N}}$ is bounded, we use Banach-Alaoglu to get $\{n_{1,j}\}_{j\in\mathbb{N}}$ and $a_1 \in \mathcal{H}$ with $a_1^{n_{1,j}} \to a_1$ weakly. Then recursively for $1 < \ell \le k$, if $A_\ell := \{a_\ell^{n_{\ell-1,j}}\}_{n\in\mathbb{N}}$ is bounded, we use Banach-Alaoglu to get $\{n_{\ell,j}\}_{j\in\mathbb{N}}$ and $a_\ell \in \mathcal{H}$ with $a_\ell^{n_{\ell,j}} \to a_\ell$ weakly. Now we take $\{n_j\}_{j\in\mathbb{N}} := \{n_{k,j}\}_{j\in\mathbb{N}}$, and we set

$$S'_j := \{a_\ell^{n_j} : 1 \le \ell \le k, A_\ell \text{ is bounded}\}$$

for $j \in \mathbb{N}$. Observe that we immediately have $S \subseteq R$ and $\#S \le k$. It follows that we have $S'_j \subseteq S_{n_j}$ for all $j \in \mathbb{N}$, as well as

$$\liminf_{j\to\infty} d^p(y, S'_j) = \liminf_{j\to\infty} d^p(y, S_{n_j})$$

for all $y \in \mathcal{H}$, since all elements of $S_{n_j} \setminus S'_j$ come from unbounded sequences as $j \to \infty$.

To show the inequalities, we take arbitrary $y \in \mathcal{H}$. Let $\{j_i\}_{i\in\mathbb{N}}$ be a subsequence satisfying

$$\liminf_{j\to\infty} d^p(y, S'_j) = \lim_{i\to\infty} d^p(y, S'_{j_i}).$$

By the pigeonhole principle, there exists some $1 \le \ell \le k$ and a further subsequence $\{i_u\}_{u\in\mathbb{N}}$ such that

$$d^p(y, S'_{j_{i_u}}) = \left\| a_\ell^{n_{j_{i_u}}} - y \right\|^p$$

for all $u \in \mathbb{N}$. Therefore, by (W2) and the construction, we get:

$$\begin{aligned} d^p(y, S) \le \|a_\ell - y\|^p \le \liminf_{u\to\infty} \left\| a_\ell^{n_{j_{i_u}}} - y \right\|^p &= \lim_{u\to\infty} d^p(y, S'_{j_{i_u}}) \\ &= \liminf_{j\to\infty} d^p(y, S'_j) \\ &= \liminf_{j\to\infty} d^p(y, S_{n_j}). \end{aligned}$$

Finally, we apply Fatou to get

$$\begin{aligned} W_p(S,P) &\leq \int_{\mathcal{H}} \liminf_{j\to\infty} d^p(y,S_{n_j})\,\mathrm{d}P(y) \\ &\leq \liminf_{j\to\infty} \int_{\mathcal{H}} d^p(y,S_{n_j})\,\mathrm{d}P(y) \leq \liminf_{j\to\infty} W_p(S_{n_j},P), \end{aligned}$$

as claimed. ☐

Lemma 27. *For any $(P,k,R) \in \mathcal{P}_p(\mathcal{H}) \times \mathbb{N} \times \mathrm{C}^w(\mathcal{H})$, the set $C_p(P,k,R)$ is non-empty.*

Proof. For each $n \in \mathbb{N}$, let S_n be a set of 2^{-n}-relaxed R-restricted (k,p)-means cluster centers. That is, we have $S_n \subseteq R$ and $1 \leq \#S_n \leq k$, as well as $W_p(S_n,P) \leq V_{k,p}(P,R) + 2^{-n}$ for all $n \in \mathbb{N}$. We can also assume that $\#S_n = k$ by adding more points if necessary, since this cannot increase the objective. Now get $\{n_j\}_{j\in\mathbb{N}}$ and $S \subseteq R$ as in Proposition 7, and note that this implies

$$W_p(S,P) \leq \liminf_{j\to\infty} W_p(S_{n_j},P) \leq \liminf_{j\to\infty} (V_{k,p}(P,R) + 2^{-n_j}) = V_{k,p}(P,R).$$

This shows $S \in C_p(P,k,R)$, so $C_p(P,k,R)$ is non-empty. ☐

The remainder of our results require an important notion of non-singularity, which has been introduced in [133, 161, 160]. To understand it, consider any $(P,k,R) \in \mathcal{P}_p(\mathcal{H}) \times \mathbb{N} \times \mathrm{C}^w(\mathcal{H})$. Notice that the infimum in (8.8) can equivalently be taken over all $S' \subseteq R$ with $\#S' = k$, since adding points to a set of cluster centers can never increase its objective. However, it is possible that a set of cluster centers $S' \subseteq R$ with $\#S' < k$ is already optimal. The following notion excludes this possibility:

Definition 13. *We say that $(P,k,R) \in \mathcal{P}_p(\mathcal{H}) \times \mathbb{N} \times \mathrm{C}^w(\mathcal{H})$ is* non-singular *if $V_{1,p}(P,R) > V_{2,p}(P,R) > \cdots > V_{k,p}(P,R)$ and* singular *otherwise.*

Observe for $R = \varnothing$ that we have $m_{k,p}(P,\varnothing) = \infty$ for all $k \in \mathbb{N}$, so $(P,k,\varnothing)$ can never be non-singular. In other words, (P,k,R) being non-singular implies that R is non-empty. We also give the following simple sufficient condition for non-singularity:

Lemma 28. *If $(P,k,R) \in \mathcal{P}_p(\mathcal{H}) \times \mathbb{N} \times \mathrm{C}^w(\mathcal{H})$ has $k \leq \#(R \cap \mathrm{supp}(P))$, then (P,k,R) is non-singular.*

Proof. Fix $1 < \ell \leq k$, and use Lemma 27 to get $S \in C_p(P,\ell-1,R)$. Then $\#S \leq \ell - 1 < k$ and $\#(R \cap \mathrm{supp}(P)) \geq k$ together imply that there is some $z \in (R \cap \mathrm{supp}(P)) \setminus S$. In particular, the set $S \cup \{z\}$ satisfies $S \cup \{z\} \subseteq R$ and $1 \leq \#(S \cup \{z\}) \leq \ell$. Thus, the proof is complete if we can show that we have $W_p(S \cup \{z\},P) < W_p(S,P) = V_{\ell-1,p}(P,R)$. To do this, assume for the sake of contradiction that we have $W_p(S \cup \{z\},P) = W_p(S,P)$. Since we of course have $d^p(y,S \cup \{z\}) \leq d^p(y,S)$ for all $y \in \mathcal{H}$, we conclude that we must have $d^p(y,S \cup \{z\}) = d^p(y,S)$ for all P-almost all $y \in \mathcal{H}$. Thus, choosing $r > 0$ small enough so that we have $d^p(y,S \cup \{z\}) = d^p(z,y)$ for all $y \in B_r^\circ(z)$, we find $P(B_r^\circ(z)) = 0$, and this contradicts $z \in \mathrm{supp}(P)$. ☐

One important consequence of non-singularity is that it guarantees the following uniform boundedness property; this will help us verify the hypotheses of the last part of Proposition 7 in our later work. Its proof is just a slight strengthening of [165, 133], so we defer the details to the supplementary material.

Proposition 8. *Suppose that* $\{(P_n, k_n, R_n)\}_{n\in\mathbb{N}}$ *and* (P, k, R) *in* $\mathcal{P}_p(\mathcal{H}) \times \mathbb{N} \times \mathrm{C}^w(\mathcal{H})$ *are such that* (P, k, R) *is non-singular, and*

- $P_n \to P$ *in* τ_{w}^p,
- $k_n \neq k$ *for finitely many* $n \in \mathbb{N}$, *and*
- $R \subseteq \mathrm{Li}_{n\in\mathbb{N}} R_n$.

Also suppose that non-negative constants $\{\varepsilon_n\}_{n\in\mathbb{N}}$ *have* $\varepsilon_n \to 0$. *Then, there exists some* $z \in \mathcal{H}$ *and* $r > 0$ *such that for sufficiently large* $n \in \mathbb{N}$ *and all* $S_n \in C_p(P_n, k_n, R_n; \varepsilon_n)$, *we have* $\#S_n = k$ *and* $S_n \subseteq \bar{B}_r(z)$.

Proof. Let L denote the set of all $\ell \in \{0, 1, \ldots, k\}$ for which there exists $z_\ell \in X$ and $r_\ell > 0$ such that for all sufficiently large $n \in \mathbb{N}$ and all $S_n \in C_p(P_n, k_n, R_n; \varepsilon_n)$, we have $\#(S_n \cap \bar{B}_{r_\ell}(x_\ell)) \geq \ell$. We claim that $L = \{0, 1, \ldots, k\}$; if this holds, then the result follows by taking $z := z_k$ and $r := r_k$. Since clearly $0 \in L$, we see that L is non-empty. Now, we assume for the sake of contradiction that $L \neq \{0, 1, \ldots, k\}$.

Under this assumption, we can let ℓ denote the largest element of L. Thus, $\ell \in L$ and $\ell + 1 \neq L$. Now use $\ell \in L$ to get some $z_\ell \in X$ and $r_\ell > 0$ such that for all sufficiently large $n \in \mathbb{N}$ and all $S_n \in C_p(P_n, k_n, R_n; \varepsilon_n)$, we have,

$$\#(S_n \cap \bar{B}_{r_\ell}(z_\ell)) \geq \ell.$$

Next, use $\ell + 1 \notin L$ inductively to get a sequence $\{n_j\}_{j\in\mathbb{N}}$ with the following property: For each $j \in \mathbb{N}$ there exists $S_j \in C_p(P_{n_j}, k_{n_j}, R_{n_j}; \varepsilon_{n_j})$ with

$$\#(S_j \cap \bar{B}_{r_\ell + j}(z_\ell)) < \ell + 1.$$

On the other hand, we also have

$$\#(S_j \cap \bar{B}_{r_\ell + j}(z_\ell)) \geq \#(S_n \cap \bar{B}_{r_\ell}(z_\ell)) \geq \ell.$$

Since these finite cardinalities are all integers, this means we have

$$\#(S_j \cap \bar{B}_{r_\ell + j}(z_\ell)) = \ell$$

for all $j \in \mathbb{N}$.

Next, for $j \in \mathbb{N}$ we define $T_j := S_j \cap \bar{B}_{r_\ell}(z_\ell)$, and we investigate the asymptotics of the excess loss $W_p(T_j, P_{n_j}) - W_p(S_j, P_{n_j})$ as $j \to \infty$. To do this, construct a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ as in Lemma 24. By construction we have $S_j \setminus T_j \subseteq \mathcal{H} \setminus \bar{B}_{r_\ell + j}(z_\ell)$, hence

$$|d^p(Y, T_j) - d^p(Y, S_j)| \to 0$$

almost surely. Also, we have

$$|d^p(Y^{n_j}, S_j) - d^p(Y, S_j)| \leq \|Y^{n_j} - Y\|^p \quad \text{and} \quad |d^p(Y^{n_j}, T_j) - d^p(Y, T_j)| \leq \|Y^{n_j} - Y\|^p,$$

so this implies

$$|d^p(Y^{n_j}, T_j) - d^p(Y^{n_j}, S_j)| \to 0$$

almost surely. Since (4.2) shows

$$|d^p(Y^{n_j}, T_j) - d^p(Y^{n_j}, S_j)| \leq c_p r_\ell^p + c_p\|z_\ell - Y^{n_j}\|^p,$$

we get that $\{d^p(Y^{n_j}, T_j) - d^p(Y^{n_j}, S_j)\}_{n\in\mathbb{N}}$ is uniformly integrable. Thus,

$$\lim_{n\to\infty}(W_p(T_j, P_{n_j}) - W_p(S_j, P_{n_j})) = 0$$

or

$$\limsup_{n\to\infty} W_p(T_j, P_{n_j}) \leq \liminf_{n\to\infty} W_p(S_j, P_{n_j})$$

upon rearranging.

Finally, let us show that this analysis gives a contradiction. To do this, take an arbitrary $S' \subseteq R$ with $1 \leq \#S' \leq k$. By Lemma 22 we can get a subsequence $\{j_i\}_{i\in\mathbb{N}}$ and a set $S_i' \subseteq R_{n_{j_i}}$ with $\#S_i' = \#S_{j_i}$ for each $i \in \mathbb{N}$, such that we have $S_i' \to S'$ in d_{H}. Then Lemma 25 implies

$$W_p(S_{j_i}, P_{n_{j_i}}) \leq W_p(S_i', P_{n_{j_i}}) + \varepsilon_{n_{j_i}} \to W_p(S', P)$$

as $i \to \infty$. Putting it all together, we get

$$\begin{aligned}\limsup_{j\to\infty} W_p(T_j, P) &= \limsup_{j\to\infty} W_p(T_j, P_{n_j}) \\ &\leq \liminf_{j\to\infty} W_p(S_j, P_{n_j}) \\ &\leq \liminf_{i\to\infty} W_p(S_{j_i}, P_{n_{j_i}}) \\ &\leq W_p(S', P).\end{aligned}$$

Taking the infimum over all feasible S' and using the non-singularity of (P, k, R), we have

$$\limsup_{i\to\infty} W_p(T_{j_i}, P_{n_{j_i}}) \leq V_{k,p}(P, R) < V_{\ell,p}(P, R).$$

But this is a contradiction since we have $\#T_{j_i} = \ell$ for all $i \in \mathbb{N}$ by construction. Hence, we must have $L = \{0, 1, \ldots, k\}$, and the result is proved. □

Now we can prove the main result of this section.

Proposition 9. *Let* $\{(P_n, k_n, R_n)\}_{n\in\mathbb{N}}$ *and* (P, k, R) *in* $\mathcal{P}_p(\mathcal{H}) \times \mathbb{N} \times \mathrm{C}^w(\mathcal{H})$ *all be non-singular, and suppose*

- $P_n \to P$ *in* τ_{w}^p,
- $k_n \neq k$ *for finitely many* $n \in \mathbb{N}$, *and*
- $\mathrm{Ls}^w_{n\in\mathbb{N}} R_n \subseteq R \subseteq \mathrm{Li}_{n\in\mathbb{N}} R_n$.

Then, for any $S_n \in C_p(P_n, k_n, R_n)$ *for all* $n \in \mathbb{N}$, *there exists* $\{n_j\}_{j\in\mathbb{N}}$ *and* $S \in C_p(P, k, R)$ *such that* $d_{\mathrm{H}}(S_n, S) \to 0$ *as* $j \to \infty$.

Proof. The non-singularity of (P_n, k_n, R_n) for each $n \in \mathbb{N}$ implies that $\#S_n = k$. Also, the non-singularity of (P, k, R) and Proposition 8 imply that $\{S_n\}_{n\in\mathbb{N}}$ are uniformly bounded. Thus, we can get $\{n_j\}_{j\in\mathbb{N}}$ and $S \subseteq R$ with $\#S \leq k$ as in Proposition 7. By construction, S is feasible. To show that S is also optimal, we let $J := \{j_i\}_{i\in\mathbb{N}}$ be an arbitrary subsequence, and we use the inequality of Proposition 7 along with Lemma 26 to get

$$\begin{aligned} W_p(S, P) &\leq \int_{\mathcal{H}} \liminf_{i\to\infty} d^p(y, S_{n_{j_i}})\, \mathrm{d}P(y) \\ &\leq \liminf_{i\to\infty} \int_{\mathcal{H}} d^p(y, S_{n_{j_i}})\, \mathrm{d}P(y) \\ &= \liminf_{i\to\infty} W_p(S_{n_{j_i}}, P) = \liminf_{i\to\infty} W_p(S_{n_{j_i}}, P_{n_{j_i}}). \end{aligned} \tag{8.11}$$

Now take arbitrary $S' \subseteq R$ with $\#S' = k$. By Lemma 22 we can get a subsequence $\{i_u\}_{u\in\mathbb{N}}$ and a set $S'_u \subseteq R_{n_{j_{i_u}}}$ with $\#S'_u = \#S' = k$ for all $u \in \mathbb{N}$ with $d_{\mathrm{H}}(S'_u, S') \to 0$. Therefore, Lemma 25 implies

$$\begin{aligned} \liminf_{i\to\infty} W_p(S_{n_{j_i}}, P_{n_{j_i}}) &\leq \liminf_{u\to\infty} W_p(S_{n_{j_{i_u}}}, P_{n_{j_{i_u}}}) \\ &\leq \liminf_{u\to\infty} W_p(S'_u, P_{n_{j_{i_u}}}) = W_p(S', P). \end{aligned}$$

Taking the infimum over all such S', we have proven

$$W_p(S, P) \leq V_{k,p}(P, R).$$

Therefore, S is optimal, hence $S \in C_p(P, k, R)$. By the non-singularity of (P, k, R), this also implies $\#S = k$. In other words, we can write $S_{n_j} = \{a_1^j, \ldots, a_k^j\}$ for all $j \in \mathbb{N}$ and $S = \{a_1, \ldots, a_k\}$ so that we have $a_\ell^j \to a_\ell$ weakly for all $1 \leq \ell \leq k$.

Before we can prove the second claim, we must make some preparations. Let us define, for each subsequence $J := \{j_i\}_{i\in\mathbb{N}}$, the set $B_J := \{y \in \mathcal{H} : d(y, S) = \liminf_{i\to\infty} d(y, S_{n_{j_i}})\}$. Crucially, observe that the optimality of S implies that the inequalities of (8.11) are in fact equalities, and hence that $P(B_J) = 1$ for each J. While of course the uncountable intersection of full-measure sets need not have full measure, it suffices to further establish that each set is closed (in norm). To see that B_J is indeed closed, suppose that $\{y_m\}_{m\in\mathbb{N}}$ in

B_J have $y_m \to y \in \mathcal{H}$. Then:

$$\begin{aligned}
&\left|d(y,S) - \liminf_{i\to\infty} d(y,S_{n_{j_i}})\right| \\
&= |d(y,S) - d(y_m,S)| + \left|\liminf_{i\to\infty} d(y_m,S_{n_{j_i}}) - \liminf_{i\to\infty} d(y,S_{n_{j_i}})\right| \\
&= |d(y,S) - d(y_m,S)| + \limsup_{i\to\infty} |d(y_m,S_{n_{j_i}}) - d(y,S_{n_{j_i}})| \le 2\|y_m - y\| \to 0.
\end{aligned}$$

Consequently, the set $B := \bigcap_J B_J$ satisfies $P(B) = 1$.

Next, we define the sets $V_\ell := \{y \in \mathcal{H} : \|y - a_\ell\| = d(y,S)\}$ for $1 \le \ell \le k$. We claim that for all $1 \le \ell \le k$ there exists $y \in V_\ell \cap B$ satisfying $d(y,S_{n_j}) = \|y - a_\ell^j\|$ for sufficiently large $j \in \mathbb{N}$. If not, then there exists $1 \le \ell \le k$ such that for all $y \in V_\ell \cap B$ there exists $\ell(y) \in \{1,\ldots,k\} \setminus \{\ell\}$ such that we have $d(y,S_{n_j}) = \|y - a_{\ell(y)}^j\|$ for infinitely many $j \in \mathbb{N}$. Denoting by $\{j_i\}_{i\in\mathbb{N}}$ such a subsequence, we can use the lower semi-continuity of the norm and $y \in V_\ell \cap B$ to bound:

$$\begin{aligned}
d\left(y, S \setminus \{a_\ell\}\right) \le \left\|y - a_{\ell(y)}\right\| &\le \liminf_{j\to\infty} \left\|y - a_{\ell(y)}^j\right\| \\
&\le \liminf_{i\to\infty} \left\|y - a_{\ell(y)}^{j_i}\right\| \\
&= \liminf_{i\to\infty} \left\|y - S_{n_{j_i}}\right\| = d(y,S) = \|y - a_\ell\|.
\end{aligned}$$

In other words, $V_\ell \cap B \subseteq \bigcup_{\ell' \ne \ell} V_{\ell'}$. However, this implies that $S \setminus \{a_\ell\}$ is optimal, which contradicts the non-singularity of (P,k,R).

Now we put all the pieces together. For each $1 \le \ell \le k$, we use the above to get some $y \in V_\ell \cap B$ satisfying $d(y,S_{n_j}) \in \|y - a_\ell^j\|$ for sufficiently large $j \in \mathbb{N}$. Then let $\{j_i\}_{i\in\mathbb{N}}$ be any subsequence satisfying

$$\limsup_{j\to\infty} d\left(y, S_{n_j}\right) = \lim_{i\to\infty} d\left(y, S_{n_{j_i}}\right),$$

and use the lower semi-continuity of the norm and $y \in V_\ell \cap B$ to get:

$$\begin{aligned}
\limsup_{j\to\infty} \left\|y - a_\ell^j\right\| = \limsup_{j\to\infty} d\left(y, S_{n_j}\right) &= \lim_{i\to\infty} d\left(y, S_{n_{j_i}}\right) \\
&= \liminf_{i\to\infty} d\left(y, S_{n_{j_i}}\right) \\
&= d(y,S) = \|y - a_\ell\| \le \liminf_{j\to\infty} \left\|y - a_\ell^j\right\|.
\end{aligned}$$

This shows $\|y - a_\ell^j\| \to \|y - a_\ell\|$ as $j \to \infty$. Combining this with $a_\ell^j \to a_\ell$ weakly as $j \to \infty$, this implies $a_\ell^j \to a_\ell$ in norm as $j \to \infty$. Thus we have shown $S_{n_j} \to S$ in d, and this completes the proof. □

The next goal is to show that Proposition 9 implies a uniform convergence result with respect to the Hausdorff metric d_{H}. In order to make this precise, we need the following form of regularity.

Lemma 29. *If $(P,k,R) \in \mathcal{P}_p(\mathcal{H})\times\mathbb{N}\times\mathrm{C}^w(\mathcal{H})$ is non-singular, then $C_p(P,k,R)$ is non-empty and d_{H}-compact.*

Proof. Lemma 27 gives non-emptiness, and Proposition 9 gives compactness. □

Lemma 30. *If $\{(P_n,R_n)\}_{n\in\mathbb{N}}$ in $\mathcal{P}_p(\mathcal{H})\times\mathrm{C}^w(\mathcal{H})$ and (P,k,R) in $\mathcal{P}_p(\mathcal{H})\times\mathbb{N}\times\mathrm{C}^w(\mathcal{H})$ are all non-singular and*

- *$P_n \to P$ in τ_{w}^p, and*
- *$\mathrm{Ls}^w_{n\in\mathbb{N}} R_n \subseteq R \subseteq \mathrm{Li}_{n\in\mathbb{N}} R_n$,*

then $V_{\ell,p}(P_n,R_n)\to V_{\ell,p}(P,R)$ for all $1\le \ell\le k$.

Proof. Let us show that any subsequence of $\{V_{\ell,p}(P_n,R_n)\}_{n\in\mathbb{N}}$ has a further subsequence converging to $V_{\ell,p}(P,R)$. Indeed, take arbitrary $\{n_j\}_{j\in\mathbb{N}}$, and, for each $j\in\mathbb{N}$ use Lemma 27 to get some $S_j\in C_p(P_{n_j},\ell,R_{n_j})$. Since (P,k,R) being non-singular certainly implies that (P,ℓ,R) is non-singular, we can apply Proposition 9 to get a further subsequence $\{j_i\}_{i\in\mathbb{N}}$ and some $S\in C_p(P,\ell,R)$ such that $d_{\mathrm{H}}(S_{j_i},S)\to 0$ as $i\to\infty$. Finally, note that Lemma 25 gives

$$V_{\ell,p}(P_{n_{j_i}},R_{n_{j_i}}) = W_p(S_{j_i},P_{n_{j_i}}) \to W_p(S,P) = V_{\ell,p}(P,R)$$

whence the result. □

Now our uniform convergence result follows:

Proposition 10. *If $\{(P_n,k_n,R_n)\}_{n\in\mathbb{N}}$ and (P,k,R) in $\mathcal{P}_p(\mathcal{H})\times\mathbb{N}\times\mathrm{C}^w(\mathcal{H})$ are all non-singular, and*

- *$P_n\to P$ in τ_{w}^p,*
- *$k_n\ne k$ for finitely many $n\in\mathbb{N}$, and*
- *$\mathrm{Ls}^w_{n\in\mathbb{N}} R_n \subseteq R \subseteq \mathrm{Li}_{n\in\mathbb{N}} R_n$,*

then we have

$$\max_{S_n\in C_p(P_n,k_n,R_n)}\ \min_{S\in C_p(P,k,R)} d_{\mathrm{H}}(S_n,S)\to 0$$

as $n\to\infty$.

Proof. Lemma 29 implies that $C_p(P_n,k_n,R_n)$ are non-empty and d_{H}-compact, so the supremum and infimum can, in fact, be replaced with a maximum and minimum. Now to prove the claim it suffice to show that for each subsequence $\{n_j\}_{j\in\mathbb{N}}$ there exists a further subsequence $\{j_i\}_{i\in\mathbb{N}}$ satisfying

$$\max_{S_i\in C_p(P_{n_{j_i}},k_{n_{j_i}},R_{n_{j_i}})}\ \min_{S\in C_p(P,k,R)} d_{\mathrm{H}}(S_i,S)\to 0$$

as $i \to \infty$. To do this, take arbitrary $\{n_j\}_{j\in\mathbb{N}}$, use Lemma 29 to get $S'_j \in C_p(P_{n_j}, k_{n_j}, R_{n_j})$ satisfying

$$\min_{S\in C_p(P,k,R)} d_{\mathrm{H}}(S'_j, S) = \max_{S_j\in C_p(P_{n_j},k_{n_j},R_{n_j})} \min_{S\in C_p(P,k,R)} d_{\mathrm{H}}(S_j, S).$$

Finally, use Proposition 9 to get $S' \in C_p(P, k, R)$ and $\{j_i\}_{i\in\mathbb{N}}$ such that we have $d_{\mathrm{H}}(S'_{j_i}, S') \to 0$ as $i \to \infty$, and note that this implies

$$\begin{aligned}\max_{S_i\in C_p(P_{n_{j_i}},k_{n_j},R_{n_j})} \min_{S\in C_p(P,k,R)} d_{\mathrm{H}}(S'_{j_i}, S) &= \min_{S\in C_p(P,k,R)} d_{\mathrm{H}}(S'_{j_i}, S)\\ &\le d_{\mathrm{H}}(S'_{j_i}, S') \to 0,\end{aligned}$$

as needed. □

In addition to the preceding continuity result, we describe the continuity of the adaptive choice of k arising in the elbow method. (Note that in this case, our "continuity" statement is a bona fide continuity statement.) To set this up, we let $P \in \mathcal{P}_p(\mathcal{H})$ be arbitrary, and we define

$$\Delta^2 V_{k,p}(P) := V_{k+1,p}(P) + V_{k-1,p}(P) - 2V_{k,p}(P)$$

for $k \ge 2$; the restriction to $k \ge 2$ is equivalent to adopting the convention $V_{0,p}(P) = \infty$. Then we define the function $k_p^{\mathrm{elb}}(P) : \mathcal{P}_p(\mathcal{H}) \to \mathbb{N} \cup \{\infty\}$ via

$$k_p^{\mathrm{elb}}(P) := \min\{\arg\max\{\Delta^2 V_{k,p}(P) : k \in \mathbb{N}\}\}$$

for $P \in \mathcal{P}_p(\mathcal{H})$. This leads us to the following result.

Proposition 11. *Suppose that $P \in \mathcal{P}_p(\mathcal{H})$ satisfies $\#\mathrm{supp}(P) = \infty$ and $\#\arg\max\{\Delta^2 V_{k,p}(P) : k \in \mathbb{N}, k \ge 2\} = 1$. Then, the function $k_p^{\mathrm{elb}} : \mathcal{P}_p(\mathcal{H}) \to \mathbb{N}$ is continuous at P.*

Proof. First, we claim that we have $V_{k,p}(P) \to 0$ as $k \to \infty$. To do this, use Prokhorov's theorem to get, for each $\eta > 0$, a compact $L_\eta \subseteq \mathcal{H}$ such that $P(\mathcal{H} \setminus L_\eta) \le \eta$. In particular, we have $\mathbb{1}\{y \notin L_\eta\} \to 0$ as $\eta \to 0$ holding P-almost surely for all $y \in \mathcal{H}$. Now let N_η be the union of $\{0\}$ with an $\eta^{1/p}$-net of L_η. It follows that we have

$$\begin{aligned}\int_{\mathcal{H}} d^p(y, N_\eta)\,\mathrm{d}P(y) &= \int_{L_\eta} d^p(y, N_\eta)\,\mathrm{d}P(y) + \int_{\mathcal{H}\setminus L_\eta} d^p(y, N_\eta)\,\mathrm{d}P(y)\\ &\le \eta \cdot P(L_\eta) + \int_{\mathcal{H}\setminus L_\eta} \|y\|^p\,\mathrm{d}P(y) \to 0\end{aligned}$$

as $\eta \to 0$, where we applied dominated convergence to the last term. Writing $\ell_\eta := \#N_\eta$ for $\eta > 0$, this implies $V_{\ell_\eta,p}(P) \to 0$ as $\eta \to 0$. Since $\{\ell_\eta\}_{\eta>0}$ is non-decreasing as $\eta \to 0$, this further implies $V_{k,p}(P) \to 0$ as $k \to \infty$.

Now we consider the value $M := \max\{\Delta^2 V_{k,p}(P) : k \in \mathbb{N}, k \ge 2\}$. It is clear that $V_{k,p}(P) \to 0$ implies $\Delta^2 V_{k,p}(P) \to 0$ hence $M < \infty$. Next let us also show $M > 0$. To do this, assume for the sake of contradiction that $M \leq 0$, so $\Delta^2 V_{k,p}(P) \leq 0$ for all $k \geq 2$. Then for all $2 \leq j \leq j'$, we have

$$V_{j,p}(P) - V_{j+1,p}(P) - (V_{j',p}(P) - V_{j'+1,p}(P)) = \sum_{k=j}^{j'} \Delta^2 V_{k,p}(P) \leq 0.$$

Sending $j' \to \infty$, we conclude $V_{j,p}(P) \leq V_{j+1,p}(P)$ for all $j \geq 2$. In other words $j \mapsto V_{j,p}(P)$ is non-decreasing for $j \geq 2$. But it is obviously non-increasing by definition, so it must in fact be constant. Now note that the non-singularity of $(P, 3, \mathcal{H})$ implies $V_{2,p}(P) > 0$, hence $\liminf_{k\to\infty} V_{k,p}(P) > 0$. This contradicts the conclusion of the preceding paragraph, so we must have $0 < M < \infty$. Furthermore, by assumption there is a unique $k_* := k_p^{\mathrm{elb}}(P) \in \mathbb{N}$ such that $M = \Delta^2 V_{k_*,p}(P)$.

Now suppose that $\{P_n\}_{n\in\mathbb{N}}$ in $\mathcal{P}_p(\mathcal{H})$ have $P_n \to P$ in τ_{w}^p. The result is proved if we can show that for each subsequence $\{n_j\}_{j\in\mathbb{N}}$ there exists a further subsequence $\{j_i\}_{i\in\mathbb{N}}$ satisfying $k_p^{\mathrm{elb}}(P_{n_{j_i}}) \to k_*$ as $i \to \infty$. To do this, we use the construction from above to choose $\eta > 0$ sufficiently small so that

$$\int_{\mathcal{H}\setminus L_\eta} \|y\|^p \,\mathrm{d}P(y) \leq \frac{M}{16}.$$

Then, we set $K := L_\eta$, we let N denote the union of $\{0\}$ with a $\frac{1}{2}(\frac{M}{8})^{1/p}$-net of K, and we define $\ell := \#N$. Now construct $(\Omega, \mathcal{F}, \mathbb{P})$ as in Lemma 24. Observe that for any subsequence $\{n_j\}_{j\in\mathbb{N}}$ we can choose a further subsequence $\{j_i\}_{i\in\mathbb{N}}$ satisfying both

$$\limsup_{j\to\infty} \sup_{k\geq\ell} V_{k,p}(P_{n_j}) = \lim_{i\to\infty} \sup_{k\geq\ell} V_{k,p}(P_{n_{j_i}}) \tag{8.12}$$

and

$$\mathbb{E}[\|Y^{n_{j_i}} - Y\|^p] \leq 2^{-i} \qquad \text{for all} \qquad i \in \mathbb{N}. \tag{8.13}$$

Now write $\delta_i := i^2 2^{-i}$, and observe by Markov's inequality that we have

$$\sum_{i\in\mathbb{N}} \mathbb{P}\left(\|Y^i - Y\|^p \geq \delta_i\right) \leq \sum_{i\in\mathbb{N}} \frac{\mathbb{E}\left[\|Y^i - Y\|^p\right]}{\delta_i} \leq \sum_{i\in\mathbb{N}} \frac{1}{i^2} < \infty.$$

Thus, by Borel-Cantelli, we have $\mathbb{P}(\|Y^i - Y\|^p \leq \delta_i$ for sufficiently large $i \in \mathbb{N}) = 1$.

Moreover, if we define $K_i := \{y \in \mathcal{H} : d^p(y, K) \leq \delta_i\}$ for each $i \in \mathbb{N}$, then we have

$$\limsup_{i\to\infty} \mathbb{1}\{Y^i \notin K_i\} \leq \mathbb{1}\{Y \notin K\}$$

almost surely. So, we can apply Fatou's lemma to get

$$\begin{aligned}
\limsup_{i\to\infty} \int_{\mathcal{H}\setminus K_i} \|y\|^p \,\mathrm{d}P_{n_{j_i}}(y) &= \limsup_{i\to\infty} \mathbb{E}\left[\|Y^i\|^p \mathbb{1}\{Y^i \notin K_i\}\right] \\
&\le \mathbb{E}\left[\limsup_{i\to\infty} \|Y^i\|^p \mathbb{1}\{Y^i \notin K_i\}\right] \\
&\le \mathbb{E}\left[\|Y\|^p \mathbb{1}\{Y \notin K\}\right] \\
&= \int_{\mathcal{H}\setminus K} \|y\|^p \,\mathrm{d}P(y).
\end{aligned}$$

By (4.2), we also have $d^p(y, N) \le \frac{M}{16} + 2^{p-1}\delta_i$ for all $y \in K_i$. Therefore,

$$\begin{aligned}
\sup_{k\ge\ell} V_{k,p}(P_{n_{j_i}}) &= \sup_{k\ge\ell} \inf_{\substack{S\subseteq\mathcal{H} \\ 1\le\#S\le k}} \int_{\mathcal{H}} d^p(y, S) \,\mathrm{d}P_{n_{j_i}}(y) \\
&\le \int_{\mathcal{H}} d^p(y, N) \,\mathrm{d}P_{n_{j_i}}(y) \\
&= \int_{K_i} d^p(y, N) \,\mathrm{d}P_{n_{j_i}}(y) + \int_{\mathcal{H}\setminus K_i} d^p(y, N) \,\mathrm{d}P_{n_{j_i}}(y) \\
&\le P_{n_{j_i}}(K_i) \left(\frac{M}{16} + 2^{p-1} a_i\right) + \int_{\mathcal{H}\setminus K_i} \|y\|^p \,\mathrm{d}P_{n_{j_i}}(y).
\end{aligned}$$

Thus, taking $i \to \infty$ yields

$$\limsup_{j\to\infty} \sup_{k\ge\ell} V_{k,p}(P_{n_j}) = \lim_{i\to\infty} \sup_{k\ge\ell} V_{k,p}(P_{n_{j_i}}) \le \frac{M}{8}.$$

Now simply apply the triangle inequality to get

$$\limsup_{j\to\infty} \sup_{k\ge\ell} \Delta^2 V_{k,p}(P_{n_j}) \le \frac{M}{2}.$$

This means that, for sufficiently large $j \in \mathbb{N}$, the maximum of $\{\Delta^2 V_{k,p}(P_{n_j}) : k \in \mathbb{N}, k \ge 2\}$ is not achieved on $\{\ell, \ell+1, \ldots\}$.

Finally, set

$$\varepsilon := \min\left\{|\Delta^2 V_{k_*,p}(P) - \Delta^2 V_{k,p}(P)| : k \le \ell, k \ne k_*\right\} > 0.$$

By Lemma 30, there is sufficiently large $j \in \mathbb{N}$ such that we have $|V_{k,p}(P_{n_j}) - V_{k,p}(P)| < \frac{\varepsilon}{8}$ for all $k \le \ell$. Combining this with the above and the triangle inequality shows that, for sufficiently large $j \in \mathbb{N}$, we have

$$\arg\max\{\Delta^2 V_{k,p}(P_{n_j}) : k \in \mathbb{N}\} = \arg\max\{\Delta^2 V_{k,p}(P) : k \in \mathbb{N}\} = \{k_*\},$$

hence $k_p^{\mathrm{elb}}(P_{n_j}) = k_*$. ∎

The condition $\#\arg\max\{\Delta^2 V_{k,p}(P) : k \in \mathbb{N}, k \geq 2\} = 1$ should be interpreted as saying that the distribution P has a uniquely-defined number of clusters, in the sense of the elbow method. Thus, the preceding result says that, if $P_n \to P$ in τ_{w}^p and P has a uniquely-defined number of clusters, then the same is true for P_n for sufficiently large $n \in \mathbb{N}$.

8.3 Probabilistic Consequences

We now show how the considerations of the previous section can be used to prove a number of limit theorems for clustering procedures of interest in statistical applications. For this section, we fix $p \geq 1$. An underlying probability space will be denoted $(\Omega, \mathcal{F}, \mathbb{P})$ and will be assumed to be complete.

Strong consistency for IID data

Suppose that $Y_1, Y_2, \ldots$ is an IID sequence of random variables with common distribution $P \in \mathcal{P}_p(\mathcal{H})$, and define the empirical measures via $\bar{P}_n := \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$ for all $n \in \mathbb{N}$. First, we have the following fundamental strong consistency result, which follows easily from our preparations.

Theorem 19. *Suppose $P \in \mathcal{P}_p(\mathcal{H})$, that $\{R_n\}_{n\in\mathbb{N}}$ and R are random weakly closed subsets of $\mathcal{H}$, and that $\{k_n\}_{n\in\mathbb{N}}$ and k are random positive integers. Then, we have*

$$\max_{S_n \in C_p(\bar{P}_n, k_n, R_n)} \min_{S \in C_p(P,k,R)} d_{\mathrm{H}}(S_n, S) \to 0$$

almost surely on

$$\{k \leq \#(R \cap \mathrm{supp}(P))\} \cap \{k_n \leq \#(R_n \cap \mathrm{supp}(\bar{P}_n)) \text{ for sufficiently large } n \in \mathbb{N}\} \\ \cap \{k_n \neq k \text{ for finitely many } n \in \mathbb{N}\} \cap \{\mathrm{Ls}^w_{n\in\mathbb{N}} R_n \subseteq R \subseteq \mathrm{Li}_{n\in\mathbb{N}} R_n\}.$$

Proof. As in the proof of Theorem 9, we have $\bar{P}_n \to P$ in τ_{w}^p almost surely. Thus, the result follows from Proposition 10. □

To see that this result has immediate consequences for strong consistency of (k,p)-means suppose that $P \in \mathcal{P}_p(\mathcal{H})$ has $\mathrm{supp}(P) = \infty$. Then fix $k \in \mathbb{N}$, and take $k_n = k$ and $R_n = R = X$ for all $n \in \mathbb{N}$. It immediately follows that we have

$$\max_{S_n \in C_{k,p}(\bar{P}_n)} \min_{S \in C_{k,p}(P)} d_{\mathrm{H}}(S_n, S) \to 0$$

almost surely. In words, this says that (k,p)-means is strongly consistent.

For (k,p)-medoids, we fix $k_n = k \in \mathbb{N}$ for $n \in \mathbb{N}$, and we take $R_n = \mathrm{supp}^w(\bar{P}_n) = \{Y_1, \ldots, Y_n\}$ for $n \in \mathbb{N}$ and $R = \mathrm{supp}^w(P)$, where $\mathrm{supp}^w(\,\cdot\,)$ denotes the support of a probability measure with respect to the topology of weak convergence. In order to apply Theorem 19,we require the following:

Lemma 31. *We have* $\mathrm{Ls}^w_{n\in\mathbb{N}}\mathrm{supp}^w(\bar{P}_n) \subseteq \mathrm{supp}^w(P) \subseteq \mathrm{Li}_{n\in\mathbb{N}}\mathrm{supp}^w(\bar{P}_n)$ *almost surely.*

Proof. Recall that $\mathrm{supp}^w(\bar{P}_n) = \{Y_1, \ldots, Y_n\}$ almost surely. To show $\mathrm{Ls}^w_{n\in\mathbb{N}}\mathrm{supp}^w(\bar{P}_n) \subseteq \mathrm{supp}^w(P)$ almost surely, consider any $x \notin \mathrm{supp}^w(P)$. This means there exists a weakly open set $U \subseteq \mathcal{H}$ with $x \in U$ with $P(U) = 0$. In particular, this implies $Y_n \notin U$ for all $n \in \mathbb{N}$ almost surely, which implies that x cannot be the limit point of any weakly convergent subsequence of elements of $\{\mathrm{supp}^w(\bar{P}_n)\}_{n\in\mathbb{N}}$. To show $\mathrm{supp}^w(P) \subseteq \mathrm{Li}_{n\in\mathbb{N}}\mathrm{supp}^w(\bar{P}_n)$ almost surely, note that we almost surely have $\bar{P}_n \to P$ in τ_{w}. Thus, the result follows. □

Consequently, we get

$$\max_{S_n \in C^{\mathrm{med}}_{k,p}(\bar{P}_n)} \min_{S \in C^{\mathrm{med}}_{k,p}(P)} d_{\mathrm{H}}(S_n, S) \to 0$$

almost surely, which states that (k,p)-medoids is strongly consistent. Note that we used $\mathrm{supp}^w(P)$ instead of $\mathrm{supp}(P)$ in this result, and that we have $\mathrm{supp}^w(P) \supseteq \mathrm{supp}(P)$ in general.

For (k,p)-means where k is chosen adaptively according to the elbow method, we take $k_n = k^{\mathrm{elb}}_p(\bar{P}_n)$ for all $n \in \mathbb{N}$ as well as $k = k^{\mathrm{elb}}_p(P)$, and $R_n = R = X$ for all $n \in \mathbb{N}$. If we assume additionally that $\#\arg\max\{\Delta^2 V_{k,p}(\mu) : k \in \mathbb{N}, k \geq 2\} = 1$, then Proposition 11 implies

$$\max_{S_n \in C^{\mathrm{elb}}_{p}(\bar{P}_n)} \min_{S \in C^{\mathrm{elb}}_{p}(P)} d_{\mathrm{H}}(S_n, S) \to 0$$

almost surely. In words, (k,p)-means, where k is chosen adaptively according to the elbow method, is strongly consistent provided that P has a uniquely-defined number of clusters in the sense of the elbow method.

Strong consistency for MC data

Suppose now that $Y_1, Y_2, \ldots$ is an aperiodic Harris-recurrent Markov chain (MC) on $\mathcal{H}$. Let K denote its transition kernel and P its unique stationary distribution. As always, write $\bar{P}_n := \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$ for the empirical measures of the first $n \in \mathbb{N}$ data points. Then we get the following:

Theorem 20. *If* $P \in \mathcal{P}_p(\mathcal{H})$ *and* $k \leq \#\mathrm{supp}(P)$, *then we have*

$$\max_{S_n \in C_{k,p}(\bar{P}_n)} \min_{S \in C_{k,p}(P)} d_{\mathrm{H}}(S_n, S) \to 0$$

almost surely.

Proof of Theorem 20. By [174, Theorem 4] we have $\bar{P}_n \to P$ in total variation, and by [174, Fact 5] we have $\frac{1}{n}\sum_{i=1}^n \|x - Y_i\|^p \to \int_{\mathcal{H}} \|x - y\|^p \,\mathrm{d}P(y)$, both holding almost surely; in particular, we have $\bar{P}_n \to P$ in τ^p_{w} almost surely. Thus, the result follows from Theorem 19 by taking $R_n = R = \mathcal{H}$ and $k_n = k$ for all $n \in \mathbb{N}$. □

In words, this result guarantees that (k, p)-means is strongly consistent when applied to data coming from a suitable MC. A typical application of interest in which data assumed to follow such a Markovian structure is the setting where $Y_1, Y_2, \ldots$ are samples from a posterior distribution in Bayesian statistics which are computed via Markov chain Monte Carlo (MCMC). In this setting, one typically has $\mathcal{H} = \mathbb{R}^m$ with its usual metric, which certainly satisfies our hypothesis. Moreover, if P and $K(x, \cdot)$ for all $x \in \mathbb{R}^m$ have strictly positive densities with respect to the Lebesgue measure, then the only hypothesis that one needs to check is the simple $\int_{\mathbb{R}^m} \|y\|^p \, \mathrm{d}P(y) < \infty$.

The case of (k, p)-medoids applied to data coming from a MC is more subtle. For the sake of simplicity, we focus on the case of MCs on finite state spaces. Thus, let us make the following assumption for the remainder of this subsection: $X \subseteq \mathcal{H}$ is a finite set that can be decomposed into $X = X_0 \sqcup X_1$ where X_0 are the inessential states with respect to K (that is, the states that are almost surely visited finitely often) and X_1 are the essential states with respect to P (that is, the states that are almost surely visited infinitely often), and K is aperiodic and irreducible on X_1.

It is illustrative to see what can go wrong in the simplest possible setting: Consider $X = \{-1, 0, 1\}$ with the metric inherited from the real line. Then let $Y_1, Y_2, \ldots$ be a MC with $Y_1 = 0$ and with the transition matrix

$$P = \begin{pmatrix} 1/2 & 0 & 1/2 \\ 1/3 & 1/3 & 1/3 \\ 1/2 & 0 & 1/2 \end{pmatrix}.$$

In words, this MC stays at state 0 for a geometric amount of time, then subsequently visits $\{-1, 1\}$ independently and uniformly at random. It is clear that the only stationary distribution for this MC is $P = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$, and that it is aperiodic and Harris-recurrent.

Now write $I := \max\{i \in \mathbb{N} : Y_i = 0\}$ for the number of data equal to 0 and $Z_n := \sum_{i=1}^n Y_i$ for the cumulative sum of the data. Observe in particular that $I < \infty$ almost surely and that we have $\bar{P}_n = \frac{1}{2n}(n - Z_n - I)\delta_{-1} + \frac{I}{n}\delta_0 + \frac{1}{2n}(n + Z_n - I)\delta_1$ for $n > I$. Moreover, $\{Z_n\}_{n>I}$ is a simple symmetric random walk. Next notice that on $\{n > I, Z_n = 0\}$ we have

$$\int_{\mathbb{R}} \| \pm 1 - y\|^p \, \mathrm{d}\bar{P}_n(y) = \frac{I}{n} + \left(1 - \frac{I}{n}\right) 2^{p-1} > 1 - \frac{I}{n} = \int_{\mathbb{R}} \|y\|^p \, \mathrm{d}\bar{P}_n(y),$$

hence $C^{\mathrm{med}}_{1,p}(\bar{P}_n) = \{\{0\}\}$. Also, $C^{\mathrm{med}}_{1,p}(P) = \{\{-1\}, \{1\}\}$. In particular, we have shown

$$\max_{S_n \in C^{\mathrm{med}}_{1,p}(\bar{P}_n)} \min_{S \in C^{\mathrm{med}}_{1,p}(P)} d_{\mathrm{H}}(S_n, S) = 1$$

on $\{n > I, Z_n = 0\}$. Finally, notice that we have $\mathbb{P}(n > I, Z_n = 0$ for infinitely many $n \in \mathbb{N}) = 1$ by the recurrence of the simple random walk, hence we are able to conclude that $\max_{S_n \in C^{\mathrm{med}}_{1,p}(\bar{P}_n)} \min_{S \in C^{\mathrm{med}}_{1,p}(P)} d_{\mathrm{H}}(S_n, S) \to 0$ occurs with probability zero. In words, (k, p)-medoids for data coming from this MC is strongly inconsistent.

We now introduce a method to repair this apparent deficiency. The difficulty in the preceding example is that the adaptively-chosen domain of the cluster centers includes some states not included in the support of the stationary distribution, so, to get around this, we allow our clustering procedure to "forget" some initial segment of states. This is similar to giving the MC a suitable "burn-in" period.

Indeed, let $f = \{f_n\}_{n\in\mathbb{N}}$ be any integer sequence with $0 \le f_n \le n$ for all $n \in \mathbb{N}$. Then define

$$\bar{P}_n^f := \frac{1}{n - f_n} \sum_{i=f_n+1}^{n} \delta_{Y_i}$$

for $n \in \mathbb{N}$; that is, $\{\bar{P}_n^f\}_{n\in\mathbb{N}}$ are the empirical measures of only the most recent data points, where we forget initial segments of sizes determined by f.

Lemma 32. *If $f_n/n \to 0$ and $f_n \to \infty$, then*

(i) $\bar{P}_n^f \to \mu$ in τ_{w} almost surely, and

(ii) $\mathrm{Ls}_{n\in\mathbb{N}}\mathrm{supp}(\bar{P}_n^f) = \mathrm{supp}(P) \subseteq \mathrm{Li}_{n\in\mathbb{N}}\mathrm{supp}(\bar{P}_n^f)$ almost surely.

Proof. For (i), note that X being finite means that convergence in τ_{w} is equivalent to convergence in the total variation norm, $\|\cdot\|_{\mathrm{TV}}$. To use this, write

$$\bar{P}_n^f = \frac{n}{n - f_n} \cdot \frac{1}{n} \sum_{i=1}^{n} \delta_{Y_i} - \frac{1}{n - f_n} \sum_{i=1}^{f_n} \delta_{Y_i}$$

hence

$$\begin{aligned} \|\bar{P}_n^f - \bar{P}_n\|_{\mathrm{TV}} &\le \left|\frac{n}{n - f_n} - 1\right| \frac{1}{n} \sum_{i=1}^{n} \|\delta_{Y_i}\|_{\mathrm{TV}} + \frac{1}{n - f_n} \sum_{i=1}^{f_n} \|\delta_{Y_i}\|_{\mathrm{TV}} \\ &\le \left|\frac{n}{n - f_n} - 1\right| + \frac{f_n}{n - f_n}. \end{aligned}$$

Now note that $f_n/n \to 0$ implies that the right side goes to zero, hence $\|\bar{P}_n^f - \bar{P}_n\|_{\mathrm{TV}} \to 0$. Since $\|\bar{P}_n - P\|_{\mathrm{TV}} \to 0$ almost surely by the classical ergodic theorem, we conclude $\|\bar{P}_n^f - P\|_{\mathrm{TV}} \to 0$ whence (i). For (ii), note that X being finite means that convergence in norm and weak convergence are equivalent. By the same argument in Lemma 31, we have $\mathrm{supp}(P) \subseteq \mathrm{Li}_{n\in\mathbb{N}}\mathrm{supp}(\bar{P}_n^f)$ almost surely. For the converse, suppose that $x \in \mathrm{Ls}_{n\in\mathbb{N}}\mathrm{supp}(\bar{P}_n^f)$. Since X is discrete, this means that there is a subsequence $\{n_j\}_{j\in\mathbb{N}}$ with $x \in \mathrm{supp}(\bar{P}_{n_j}^f)$ for all $j \in \mathbb{N}$. Consequently, there is some sequence $\{\ell_j\}_{j\in\mathbb{N}}$ (not necessary non-decreasing) with $f_{n_j} + 1 \le \ell_j \le n_j$ and $Y_{\ell_j} = x$ for all $j \in \mathbb{N}$. Since $f_n \to \infty$, this means $Y_1, Y_2, \ldots$ visits x infinitely often. But X_0 is inessential with respect to K, so we must have $x \notin X_0$. This implies $x \in X_1$ hence $x \in \mathrm{supp}(P)$ since X_1 is irreducible and aperiodic with respect to K. We have shown $\mathrm{Ls}_{n\in\mathbb{N}}\mathrm{supp}(\bar{P}_n^f) \subseteq \mathrm{supp}(P)$ almost surely, so combining with the first part gives (ii). ∎

This, in particular, gives the following strong consistency.

Theorem 21. *If $f_n/n \to 0$ and $f_n \to \infty$, then we have*

$$\max_{S_n \in C_{k,p}^{\mathrm{med}}(\bar{P}_n^f)} \min_{S \in C_{k,p}^{\mathrm{med}}(P)} d_{\mathrm{H}}(S_n, S) \to 0$$

almost surely.

Proof. Immediate by Theorem 19 and Lemma 32. □

To be concrete in the setting above, one can take $f_n := \lfloor \log n \rfloor$ for $n \in \mathbb{N}$.

Large deviations for IID data

Suppose again that $Y_1, Y_2, \ldots$ is an IID sequence of random variables with common distribution P. As always, let us define the empirical measures via $\bar{P}_n := \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$ for all $n \in \mathbb{N}$. We have already established the almost sure convergence

$$\max_{S_n \in C_{k,p}(\bar{P}_n)} \min_{S \in C_{k,p}(P)} d_{\mathrm{H}}(S_n, S) \to 0,$$

and this of course implies the convergence in probability

$$\mathbb{P}\left(\max_{S_n \in C_{k,p}(\bar{P}_n)} \min_{S \in C_{k,p}(P)} d_{\mathrm{H}}(S_n, S) \geq \varepsilon\right) \to 0,$$

for all $\varepsilon > 0$. Presently, we address the question of determining the rate at which these probabilities decay to zero.

Theorem 22. *Suppose that $P \in \mathcal{P}_p(\mathcal{H})$ satisfies $\int_{\mathcal{H}} \exp(\alpha\|y\|^p)\,\mathrm{d}P(y) < \infty$ for all $\alpha > 0$. Then, for all $k \leq \#\mathrm{supp}(P)$ and $\varepsilon > 0$, we have*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\sup_{S_n \in C_{k,p}(\bar{P}_n)} \inf_{S \in C_{k,p}(P)} d_{\mathrm{H}}(S_n, S) \geq \varepsilon\right) \leq -c_{k,p}(P, \varepsilon),$$

for some $c_{k,p}(P, \varepsilon) > 0$.

Proof of Theorem 12. Consider the set

$$\begin{aligned} A &:= \left\{Q \in \mathcal{P}_p(\mathcal{H}) : \sup_{T \in C_{k,p}(Q)} \inf_{S \in C_{k,p}(P)} d_{\mathrm{H}}(T, S) \geq \varepsilon\right\} \\ &= \left\{Q \in \mathcal{P}_p(\mathcal{H}) : \vec{D}_{\mathrm{H}}(C_{k,p}(Q), C_{k,p}(P)) \geq \varepsilon\right\}. \end{aligned}$$

Here, we write $\vec{D}_{\mathrm{H}}$ for the one-sided Hausdorff distance on $(\mathrm{K}(\mathcal{H}), d_{\mathrm{H}})$. Now let us show that A is τ_{w}^p-closed. Indeed, suppose $\{Q_n\}_{n\in\mathbb{N}}$ in A and $Q \in \mathcal{P}_p(\mathcal{H})$ have $Q_n \to Q$ in τ_{w}^p. Then apply (8.4) to get:

$$\vec{D}_{\mathrm{H}}(C_{k,p}(Q_n), C_{k,p}(Q)) + \vec{D}_{\mathrm{H}}(C_{k,p}(Q), C_{k,p}(P)) \geq \vec{D}_{\mathrm{H}}(C_{k,p}(Q_n), C_{k,p}(P)) \geq \varepsilon.$$

Now use Proposition 10 with $k_n = k$ and $R_n = \mathcal{H}$ for all $n \in \mathbb{N}$, which guarantees that we have $\vec{D}_{\mathrm{H}}(C_{k,p}(Q_n), C_{k,p}(Q)) \to 0$, hence $\vec{D}_{\mathrm{H}}(C_{k,p}(Q), C_{k,p}(P)) \geq \varepsilon$, by the above. In particular, $Q \in A$, thus A is τ_{w}^p-closed.

Next, note that P is a Borel probability measure on a Polish metric space with all exponential moments finite, so we conclude via [205, Theorem 1.1] that $\{\bar{P}_n\}_{n\in\mathbb{N}}$ satisfy a large deviations principle in $(\mathcal{P}_p(\mathcal{H}), \tau_{\mathrm{w}}^p)$ with good rate function $H(\,\cdot\,|P) : \mathcal{P}_p(\mathcal{H}) \to [0,\infty]$. In particular, the large deviations upper bound implies

$$\begin{aligned}\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\vec{D}_{\mathrm{H}}(C_{k,p}(\bar{P}_n), C_{k,p}(P)) \geq \varepsilon\right) &= \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\bar{P}_n \in A) \\ &\leq -\inf\{H(Q|P) : Q \in A\} := c_{k,p}(P,\varepsilon).\end{aligned}$$

Finally, assume towards a contradiction that $c_{k,p}(P,\varepsilon) = 0$, so that there exist $\{Q_n\}_{n\in\mathbb{N}}$ in A with $H(Q_n|P) \to 0$. As a consequence of the Donsker-Varadhan variational principle one can show that $H(Q_n|P) \to 0$ and $\int_{\mathcal{H}} \exp(\alpha\|y\|^p)\,\mathrm{d}P(y) < \infty$ for all $\alpha > 0$ together imply $Q_n \to P$ in τ_{w}^p, so Proposition 10 implies $\vec{D}_{\mathrm{H}}(C_{k,p}(Q_n), C_{k,p}(P)) \to 0$. This is impossible since $Q_n \in A$ for all $n \in \mathbb{N}$, hence we must have $c_{k,p}(P,\varepsilon) > 0$. □

Now, we make some remarks on possible limitations and extensions of this result. First of all, the constant $c_{k,p}(P,\varepsilon)$ appearing as the exponential rate of decay has an exact characterization as

$$c_{k,p}(P,\varepsilon) := \inf\{H(Q|P) : Q \in \mathcal{P}_p(\mathcal{H}), \vec{D}_{\mathrm{H}}(C_{k,p}(Q), C_{k,p}(P)) \geq \varepsilon\}. \tag{8.14}$$

From this form we have shown $c_{k,p}(P,\varepsilon) > 0$ for all $\varepsilon > 0$, but it appears difficult to say much else. We believe it would be interesting to try to understand for which distributions $P \in \mathcal{P}_p(\mathcal{H})$ the quantity $c_{k,p}(P,\varepsilon)$ can be exactly or approximately computed. For example, if P is compactly-supported, then do we have $c_{k,p}(P,\varepsilon) \lesssim \varepsilon^{-2}$, which can be interpreted as a sort of asymptotically sub-Gaussian concentration?

Bibliography

[1] Efren F. Abaya and Gary L. Wise. "Convergence of vector quantizers with applications to optimal quantization". In: *SIAM J. Appl. Math.* 44.1 (1984), pp. 183–189. ISSN: 00361399. URL: http://www.jstor.org/stable/2101312 (visited on 03/29/2023).

[2] Allan F. Abrahamse. "The tail field of a Markov chain". In: *Ann. Math. Statist.* 40 (1969), pp. 127–136. ISSN: 0003-4851. DOI: 10.1214/aoms/1177697809. URL: https://doi.org/10.1214/aoms/1177697809.

[3] Michał Adamaszek, Henry Adams, and Florian Frick. "Metric reconstruction via optimal transport". In: *SIAM Journal on Applied Algebra and Geometry* 2.4 (2018), pp. 597–619.

[4] Bijan Afsari. "Riemannian L^p center of mass: existence, uniqueness, and convexity". In: *Proc. Amer. Math. Soc.* 139.2 (2011), pp. 655–673. ISSN: 0002-9939. DOI: 10.1090/S0002-9939-2010-10541-5. URL: https://doi.org/10.1090/S0002-9939-2010-10541-5.

[5] Martial Agueh and Guillaume Carlier. "Barycenters in the Wasserstein space". In: *SIAM Journal on Mathematical Analysis* 43.2 (2011), pp. 904–924.

[6] Martial Agueh and Guillaume Carlier. "Vers un théorème de la limite centrale dans l'espace de Wasserstein?" In: *Comptes Rendus. Mathématique* 355.7 (2017), pp. 812–818.

[7] A. Ahidar-Coutrix, T. Le Gouic, and Q. Paris. "Convergence rates for empirical barycenters in metric spaces: curvature, convexity and extendable geodesics". In: *Probab. Theory Related Fields* 177.1-2 (2020), pp. 323–368. ISSN: 0178-8051. DOI: 10.1007/s00440-019-00950-0. URL: https://doi.org/10.1007/s00440-019-00950-0.

[8] David J. Aldous and Hermann Thorisson. "Shift-coupling". In: *Stochastic Process. Appl.* 44.1 (1993), pp. 1–14. ISSN: 0304-4149. DOI: 10.1016/0304-4149(93)90034-2. URL: https://doi.org/10.1016/0304-4149(93)90034-2.

[9] Daniel Aloise et al. "NP-hardness of Euclidean sum-of-squares clustering". In: *Machine Learning* 75.2 (2009), pp. 245–248. DOI: 10.1007/s10994-009-5103-0. URL: https://doi.org/10.1007/s10994-009-5103-0.

[10] Ethem Alpaydin. *Introduction to Machine Learning*. 2nd. The MIT Press, 2010. ISBN: 026201243X.

[11] Jason Altschuler et al. "Averaging on the Bures–Wasserstein manifold: dimension-free convergence of gradient descent". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 22132–22145.

[12] Pedro C. Álvarez-Esteban et al. "Wide consensus aggregation in the Wasserstein space. Application to location-scatter families". In: *Bernoulli* 24.4A (2018), pp. 3147 –3179. DOI: 10.3150/17-BEJ957. URL: https://doi.org/10.3150/17-BEJ957.

[13] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. "Gradient flows with metric and differentiable structures, and applications to the Wasserstein space". In: *Atti della Accademia Nazionale dei Lincei. Classe di Scienze Fisiche, Matematiche e Naturali. Rendiconti Lincei. Matematica e Applicazioni* 15.3-4 (2004), pp. 327–343.

[14] Vivek Bagaria et al. "Medoids in Almost-Linear Time via Multi-Armed Bandits". In: ed. by Amos Storkey and Fernando Perez-Cruz. Vol. 84. Proceedings of Machine Learning Research. Playa Blanca, Lanzarote, Canary Islands: PMLR, 2018, pp. 500–509. URL: http://proceedings.mlr.press/v84/bagaria18a.html.

[15] Tavor Baharav and David Tse. "Ultra Fast Medoid Identification via Correlated Sequential Halving". In: *Advances in Neural Information Processing Systems*. 2019, pp. 3655–3664.

[16] S. Banerjee and W. S. Kendall. "Rigidity for Markovian maximal couplings of elliptic diffusions". In: *Probab. Theory Related Fields* 168.1-2 (2017), pp. 55–112. ISSN: 0178-8051,1432-2064. DOI: 10.1007/s00440-016-0706-4. URL: https://doi.org/10.1007/s00440-016-0706-4.

[17] David Barnhill and Ruriko Yoshida. "Clustering methods over the tropical projective torus". In: *Mathematics* 11.15 (2023), p. 3433. ISSN: 2227-7390. DOI: 10.3390/math11153433. URL: http://dx.doi.org/10.3390/math11153433.

[18] E. del Barrio et al. "Robust clustering tools based on optimal transportation". In: *Statistics and Computing* 29.1 (2019), pp. 139–160. DOI: 10.1007/s11222-018-9800-z. URL: https://doi.org/10.1007/s11222-018-9800-z.

[19] Miroslav Bačák. "Computing medians and means in Hadamard spaces". In: *SIAM J. Optim.* 24.3 (2014), pp. 1542–1566. ISSN: 1052-6234. DOI: 10.1137/140953393. URL: https://doi.org/10.1137/140953393.

[20] Mathias Beiglböck, Pierre Henry-Labordère, and Friedrich Penkner. "Model-independent bounds for option prices—a mass transport approach". In: *Finance Stoch.* 17.3 (2013), pp. 477–501. ISSN: 0949-2984. DOI: 10.1007/s00780-013-0205-8. URL: https://doi.org/10.1007/s00780-013-0205-8.

[21] Mathias Beiglböck, Christian Léonard, and Walter Schachermayer. "A general duality theorem for the Monge-Kantorovich transport problem". In: *Studia Math.* 209.2 (2012), pp. 151–167. ISSN: 0039-3223. DOI: 10.4064/sm209-2-4. URL: https://doi.org/10.4064/sm209-2-4.

[22] Mathias Beiglböck, Christian Léonard, and Walter Schachermayer. "On the duality theory for the Monge-Kantorovich transport problem". In: *Optimal transportation.* Vol. 413. London Math. Soc. Lecture Note Ser. Cambridge Univ. Press, Cambridge, 2014, pp. 216–265.

[23] Mathias Beiglböck, Marcel Nutz, and Nizar Touzi. "Complete duality for martingale optimal transport on the line". In: *Ann. Probab.* 45.5 (2017), pp. 3038–3074. ISSN: 0091-1798. DOI: 10.1214/16-AOP1131. URL: https://doi.org/10.1214/16-AOP1131.

[24] Mathias Beiglböck et al. "Optimal and better transport plans". In: *J. Funct. Anal.* 256.6 (2009), pp. 1907–1927. ISSN: 0022-1236. DOI: 10.1016/j.jfa.2009.01.013. URL: https://doi.org/10.1016/j.jfa.2009.01.013.

[25] Gérard Ben Arous, Michael Cranston, and Wilfrid S. Kendall. "Coupling constructions for hypoelliptic diffusions: two examples". In: *Stochastic analysis (Ithaca, NY, 1993).* Vol. 57. Proc. Sympos. Pure Math. Amer. Math. Soc., Providence, RI, 1995, pp. 193–212. ISBN: 0-8218-0289-5. DOI: 10.1090/pspum/057/1335472. URL: https://doi.org/10.1090/pspum/057/1335472.

[26] Patrizia Berti and Pietro Rigo. "Sufficient conditions for the existence of disintegrations". In: *J. Theoret. Probab.* 12.1 (1999), pp. 75–86. ISSN: 0894-9840. DOI: 10.1023/A:1021792409934. URL: https://doi.org/10.1023/A:1021792409934.

[27] Jean Bertoin. *Lévy processes.* Vol. 121. Cambridge Tracts in Mathematics. Cambridge University Press, Cambridge, 1996, pp. x+265. ISBN: 0-521-56243-0.

[28] Abhishek Bhattacharya. "Statistical analysis on manifolds: a nonparametric approach for inference on shape spaces". In: *Sankhyā* 70.2, Ser. A (2008), pp. 223–266. ISSN: 0972-7671.

[29] Abhishek Bhattacharya and Rabi Bhattacharya. *Nonparametric inference on manifolds.* Vol. 2. Institute of Mathematical Statistics (IMS) Monographs. With applications to shape spaces. Cambridge University Press, Cambridge, 2012, pp. xiv+237. ISBN: 978-1-107-01958-4. DOI: 10.1017/CBO9781139094764. URL: https://doi.org/10.1017/CBO9781139094764.

[30] Rabi Bhattacharya and Lizhen Lin. "Omnibus CLTs for Fréchet means and nonparametric inference on non-Euclidean spaces". In: *Proc. Amer. Math. Soc.* 145.1 (2017), pp. 413–428. ISSN: 0002-9939. DOI: 10.1090/proc/13216. URL: https://doi.org/10.1090/proc/13216.

[31] Rabi Bhattacharya and Vic Patrangenaru. "Large sample theory of intrinsic and extrinsic sample means on manifolds. I". In: *Ann. Statist.* 31.1 (2003), pp. 1–29. ISSN: 0090-5364. DOI: 10.1214/aos/1046294456. URL: https://doi.org/10.1214/aos/1046294456.

[32] Rabi Bhattacharya and Vic Patrangenaru. "Large sample theory of intrinsic and extrinsic sample means on manifolds. II". In: *Ann. Statist.* 33.3 (2005), pp. 1225–1259. ISSN: 0090-5364. DOI: 10.1214/009053605000000093. URL: https://doi.org/10.1214/009053605000000093.

[33] Gérard Biau, Luc Devroye, and Gábor Lugosi. "On the performance of clustering in Hilbert spaces". In: *IEEE Trans. Inform. Theory* 54 (Mar. 2008), pp. 781–790. DOI: 10.1109/TIT.2007.913516.

[34] Louis J. Billera, Susan P. Holmes, and Karen Vogtmann. "Geometry of the space of phylogenetic trees". In: *Adv. in Appl. Math.* 27.4 (2001), pp. 733–767. ISSN: 0196-8858.

[35] David Blackwell and Persi Diaconis. "A non-measurable tail set". In: *Statistics, probability and game theory.* Vol. 30. IMS Lecture Notes Monogr. Ser. Inst. Math. Statist., Hayward, CA, 1996, pp. 1–5. DOI: 10.1214/lnms/1215453560. URL: https://doi.org/10.1214/lnms/1215453560.

[36] David Blackwell and Lester E. Dubins. "On existence and non-existence of proper, regular, conditional distributions". In: *Ann. Probability* 3.5 (1975), pp. 741–752. ISSN: 0091-1798. DOI: 10.1214/aop/1176996261. URL: https://doi.org/10.1214/aop/1176996261.

[37] Moïse Blanchard and Adam Quinn Jaffe. "Fréchet Mean Set Estimation in the Hausdorff Metric, via Relaxation". https://arxiv.org/abs/2212.12057.

[38] T. S. Blyth. *Lattices and ordered algebraic structures.* Universitext. Springer-Verlag London, Ltd., London, 2005, pp. x+303. ISBN: 1-85233-905-5.

[39] Hans-Hermann Bock. "Origins and extensions of the k-means algorithm in cluster analysis." eng. In: *J Électron. Hist. Probab. Stat.* 4.2 (2008), pp. 1–18. URL: http://eudml.org/doc/130880.

[40] V. I. Bogachev. *Measure theory. Vol. I, II.* Springer-Verlag, Berlin, 2007, Vol. I: xviii+500 pp., Vol. II: xiv+575. ISBN: 978-3-540-34513-8; 3-540-34513-2. DOI: 10.1007/978-3-540-34514-5. URL: https://doi.org/10.1007/978-3-540-34514-5.

[41] Björn Böttcher, René L. Schilling, and Jian Wang. "Constructions of coupling processes for Lévy processes". In: *Stochastic Process. Appl.* 121.6 (2011), pp. 1201–1216. ISSN: 0304-4149,1879-209X. DOI: 10.1016/j.spa.2011.02.007. URL: https://doi.org/10.1016/j.spa.2011.02.007.

[42] Krzysztof Burdzy and Wilfrid S. Kendall. "Efficient Markovian couplings: examples and counterexamples". In: *Ann. Appl. Probab.* 10.2 (2000), pp. 362–409. ISSN: 1050-5164,2168-8737. DOI: 10.1214/aoap/1019487348. URL: https://doi.org/10.1214/aoap/1019487348.

[43] Ofer Busani. *Diffusive Scaling Limit of the Busemann Process in Last Passage Percolation*. Preprint available at https://arxiv.org/abs/2110.03808. 2023. arXiv: 2110.03808 [math.PR].

[44] Ofer Busani, Timo Seppäläinen, and Evan Sorensen. *The Stationary Horizon and Semi-Infinite Geodesics in the Directed Landscape*. Preprint available at https://arxiv.org/abs/2203.13242. 2023. arXiv: 2203.13242 [math.PR].

[45] Brittny Calsbeek and Charles J. Goodnight. "Empirical comparison of G matrix test statistics: finding biologically relevant change". In: *Evolution: International Journal of Organic Evolution* 63.10 (2009), pp. 2627–2635.

[46] Elisabetta Candellero and Wilfrid S. Kendall. "Coupling of Brownian motions in Banach spaces". In: *Electron. Commun. Probab.* 23 (2018), Paper No. 9, 13. ISSN: 1083-589X. DOI: 10.1214/18-ECP109. URL: https://doi.org/10.1214/18-ECP109.

[47] Yueqi Cao, Prudence Leung, and Anthea Monod. *k-means clustering for persistent homology*. arxiv preprint. 2023. arXiv: 2210.10003 [stat.AP].

[48] Guillaume Carlier, Enis Chenchene, and Katharina Eichinger. "Wasserstein medians: robustness, PDE characterization and numerics". https://arxiv.org/abs/2307.01765. 2023.

[49] Guillaume Carlier, Katharina Eichinger, and Alexey Kroshnin. "Entropic-Wasserstein barycenters: PDE characterization, regularity, and CLT". In: *SIAM Journal on Mathematical Analysis* 53.5 (2021), pp. 5880–5914.

[50] René Carmona and François Delarue. *Probabilistic theory of mean field games with applications I*. Vol. 83. Probability Theory and Stochastic Modelling. Mean field FBSDEs, control, and games. Springer, Cham, 2018, pp. xxv+713. ISBN: 978-3-319-56437-1; 978-3-319-58920-6.

[51] E. Cartan. *Leçons Sur la Géométrie Des Espaces de Riemann*. Gauthier-Villars, 1963. URL: https://books.google.com/books?id=3jp_AQAACAAJ.

[52] Frédéric Cazals, Bernard Delmas, and Timothée O'Donnell. "Fréchet mean and p-mean on the unit circle: characterization, decidability, and algorithm". In: (2020).

[53] Rudrasis Chakraborty and Baba C. Vemuri. "Recursive Fréchet Mean Computation on the Grassmannian and its Applications to Computer Vision". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2015.

[54] Rudrasis Chakraborty and Baba C. Vemuri. "Statistics on the Stiefel manifold: theory and applications". In: *Ann. Statist.* 47.1 (2019), pp. 415–438. ISSN: 0090-5364. DOI: 10.1214/18-AOS1692. URL: https://doi.org/10.1214/18-AOS1692.

[55] Sinho Chewi et al. "Gradient descent algorithms for Bures-Wasserstein barycenters". In: *Proceedings of Thirty Third Conference on Learning Theory*. Ed. by Jacob Abernethy and Shivani Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, 2020, pp. 1276–1304. URL: http://proceedings.mlr.press/v125/chewi20a.html.

[56] Yang Chu and Adam Q. Jaffe. *Nearly-maximal germ couplings of Lévy processes*. In progress.

[57] J. A. Cuesta and C. Matrán. "The strong law of large numbers for k-means and best possible nets of Banach valued random variables". In: *Probab. Theory Related Fields* 78.4 (1988), pp. 523–534. ISSN: 0178-8051. DOI: 10.1007/BF00353875. URL: https://doi.org/10.1007/BF00353875.

[58] J. A. Cuesta-Albertos, C. Matrán-Bea, and A. Tuero-Diaz. "On lower bounds for the L2-Wasserstein metric in a Hilbert space". In: *Journal of Theoretical Probability* 9.2 (1996), pp. 263–283. DOI: 10.1007/BF02214649. URL: https://doi.org/10.1007/BF02214649.

[59] Duncan Dauvergne, Sourav Sarkar, and Bálint Virág. "Three-halves variation of geodesics in the directed landscape". In: *Ann. Probab.* 50.5 (2022), pp. 1947–1985. ISSN: 0091-1798. DOI: 10.1214/22-aop1574. URL: https://doi.org/10.1214/22-aop1574.

[60] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Stochastic Modelling and Applied Probability. Springer Berlin Heidelberg, 2009. ISBN: 9783642033117. URL: https://books.google.com/books?id=iT9JRlGPx5gC.

[61] Ian L. Dryden, Alexey Koloydenko, and Diwei Zhou. "Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging". In: *The Annals of Applied Statistics* 3.3 (2009), pp. 1102–1123.

[62] Gabriele Eichfelder, Thomas Hotz, and Johannes Wieditz. "An algorithm for computing Fréchet means on the sphere". In: *Optimization Letters* 13.7 (2019), pp. 1523–1533.

[63] H.-J. Engelbert, M. A. Urusov, and M. Walther. "A canonical setting and separating times for continuous local martingales". In: *Stochastic Process. Appl.* 119.4 (2009), pp. 1039–1054. ISSN: 0304-4149. DOI: 10.1016/j.spa.2008.05.003. URL: https://doi.org/10.1016/j.spa.2008.05.003.

[64] R. Engelking. "Sur l'impossibilité de définir la limite topologique inférieure à l'aide des opérations dénombrables de l'algèbre de Boole et de l'opération de fermeture". In: *Bull. Acad. Polon. Sci. Cl. III.* 4 (1956), pp. 659–662.

[65] Philip A. Ernst et al. "MEXIT: maximal un-coupling times for stochastic processes". In: *Stochastic Process. Appl.* 129.2 (2019), pp. 355–380. ISSN: 0304-4149. DOI: 10.1016/j.spa.2018.03.001. URL: https://doi.org/10.1016/j.spa.2018.03.001.

[66] Steven N. Evans and Adam Q. Jaffe. “Limit Theorems for Fréchet mean sets”. In: *Bernoulli* 30.1 (2024), pp. 419–447.

[67] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. “Testing the manifold hypothesis”. In: *J. Amer. Math. Soc.* 29 (2016), pp. 983–1049. DOI: 10.1090/jams/852.

[68] Jacob Feldman and Calvin C. Moore. “Ergodic equivalence relations, cohomology, and von Neumann algebras. I”. In: *Trans. Amer. Math. Soc.* 234.2 (1977), pp. 289–324. ISSN: 0002-9947. DOI: 10.2307/1997924. URL: https://doi.org/10.2307/1997924.

[69] Aasa Feragen et al. “Means in spaces of tree-like shapes”. In: *2011 International Conference on Computer Vision*. IEEE. 2011, pp. 736–746.

[70] Daniel Ferguson and François G. Meyer. “On the Number of Edges of the Fréchet Mean and Median Graphs”. In: *Network Science*. Ed. by Pedro Ribeiro et al. Cham: Springer International Publishing, 2022, pp. 26–40. ISBN: 978-3-030-97240-0.

[71] Daniel Ferguson and François G Meyer. “Theoretical analysis and computation of the sample Fréchet mean of sets of large graphs for various metrics”. In: *Information and Inference: A Journal of the IMA* 12.3 (Mar. 2023), pp. 1347–1404. ISSN: 2049-8772. DOI: 10.1093/imaiai/iaad002. eprint: https://academic.oup.com/imaiai/article-pdf/12/3/1347/51603021/iaad002.pdf. URL: https://doi.org/10.1093/imaiai/iaad002.

[72] Simone Fiori. “Learning the Fréchet mean over the manifold of symmetric positive-definite matrices”. In: *Cognitive computation* 1.4 (2009), p. 279.

[73] Bernard D. Flury. “Estimation of principal points”. In: *J. Roy. Statist. Soc. Ser. C* 42.1 (1993), pp. 139–151. ISSN: 0035-9254,1467-9876. DOI: 10.2307/2347416. URL: https://doi.org/10.2307/2347416.

[74] Bernhard A. Flury. “Principal points”. In: *Biometrika* 77.1 (1990), pp. 33–41. ISSN: 0006-3444,1464-3510. DOI: 10.1093/biomet/77.1.33. URL: https://doi.org/10.1093/biomet/77.1.33.

[75] H. Föllmer. “Tail structure of Markov chains on infinite product spaces”. In: *Z. Wahrsch. Verw. Gebiete* 50.3 (1979), pp. 273–285. ISSN: 0044-3719. DOI: 10.1007/BF00534151. URL: https://doi.org/10.1007/BF00534151.

[76] E. W. Forgy. “Cluster analysis of multivariate data : efficiency versus interpretability of classifications”. In: *Biometrics* 21 (1965), pp. 768–769.

[77] Joshua Frisch et al. *Realizations of countable Borel equivalence relations*. 2021. DOI: 10.48550/ARXIV.2109.12486. URL: https://arxiv.org/abs/2109.12486.

[78] Hans-Otto Georgii. “Orbit coupling”. In: *Ann. Inst. H. Poincaré Probab. Statist.* 33.2 (1997), pp. 253–268. ISSN: 0246-0203. DOI: 10.1016/S0246-0203(97)80123-9. URL: https://doi.org/10.1016/S0246-0203(97)80123-9.

[79] Aurelien Géron. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. Sebastopol, CA: O'Reilly Media, 2017. ISBN: 978-1491962299.

[80] Promit Ghosal, Marcel Nutz, and Espen Bernton. "Stability of entropic optimal transport and Schrödinger bridges". In: *J. Funct. Anal.* 283.9 (2022), Paper No. 109622, 22. ISSN: 0022-1236. DOI: 10.1016/j.jfa.2022.109622. URL: https://doi.org/10.1016/j.jfa.2022.109622.

[81] Cedric E Ginestet. "Strong Consistency of Fréchet Sample Mean Sets for Graph-Valued Random Variables". In: *arXiv preprint arXiv:1204.3183* (2012).

[82] Sheldon Goldstein. "Maximal coupling". In: *Z. Wahrsch. Verw. Gebiete* 46.2 (1978/79), pp. 193–204. ISSN: 0044-3719. DOI: 10.1007/BF00533259. URL: https://doi.org/10.1007/BF00533259.

[83] Oscar Gonzalez et al. "Absolute versus relative entropy parameter estimation in a coarse-grain model of DNA". In: *Multiscale Modeling & Simulation* 15.3 (2017), pp. 1073–1107.

[84] Charles J. Goodnight and James M. Schwartz. "A Bootstrap Comparison of Genetic Covariance Matrices". In: *Biometrics* 53.3 (1997), pp. 1026–1039. ISSN: 0006341X, 15410420. URL: http://www.jstor.org/stable/2533561 (visited on 01/20/2024).

[85] David Griffeath. "A maximal coupling for Markov chains". In: *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* 31 (1974/75), pp. 95–106. DOI: 10.1007/BF00539434. URL: https://doi.org/10.1007/BF00539434.

[86] Karsten Grove and Hermann Karcher. "How to conjugate C^1-close group actions". In: *Mathematische Zeitschrift* 132.1 (1973), pp. 11–20.

[87] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Second. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[88] Greg Hjorth. "Borel equivalence relations". In: *Handbook of set theory. Vols. 1, 2, 3*. Springer, Dordrecht, 2010, pp. 297–332. DOI: 10.1007/978-1-4020-5764-9_5. URL: https://doi.org/10.1007/978-1-4020-5764-9_5.

[89] J. Hoffmann-Jø rgensen. "Existence of conditional probabilities". In: *Math. Scand.* 28 (1971), 257–264 (1972). ISSN: 0025-5521. DOI: 10.7146/math.scand.a-11023. URL: https://doi.org/10.7146/math.scand.a-11023.

[90] Frank den Hollander. *Probability Theory: The Coupling Method*. 2012.

[91] Thomas Hotz et al. "Sticky central limit theorems on open books". In: *Ann. Appl. Probab.* 23.6 (2013), pp. 2238–2258. ISSN: 1050-5164. DOI: 10.1214/12-AAP899. URL: https://doi.org/10.1214/12-AAP899.

[92] Elton P. Hsu and Karl-Theodor Sturm. "Maximal coupling of Euclidean Brownian motions". In: *Commun. Math. Stat.* 1.1 (2013), pp. 93–104. ISSN: 2194-6701. DOI: 10.1007/s40304-013-0007-5. URL: https://doi.org/10.1007/s40304-013-0007-5.

[93] Stephan Huckemann, Thomas Hotz, and Axel Munk. "Intrinsic Shape Analysis: Geodesic PCA for Riemannian manifolds modulo isometric Lie group actions". In: *Statistica Sinica* 20.1 (2010), pp. 1–58. ISSN: 10170405, 19968507. URL: http://www.jstor.org/stable/24308976 (visited on 10/03/2023).

[94] Stephan F. Huckemann. "Intrinsic Inference on the Mean Geodesic of Planer Shapes and Tree Discrimination by Leaf Growth". In: *The Annals of Statistics* 39.2 (2011), pp. 1098–1124. ISSN: 00905364, 21688966. URL: http://www.jstor.org/stable/29783668 (visited on 09/23/2022).

[95] Stephan F. Huckemann. "(Semi-)Intrinsic Statistical Analysis on Non-Euclidean Spaces". In: *Advances in Complex Data Modeling and Computational Methods in Statistics*. Ed. by Anna Maria Paganoni and Piercesare Secchi. Cham: Springer International Publishing, 2015, pp. 103–118. ISBN: 978-3-319-11149-0. DOI: 10.1007/978-3-319-11149-0_7. URL: https://doi.org/10.1007/978-3-319-11149-0_7.

[96] Stephan F. Huckemann. "Stickiness and Smeariness". In: *Mini-Workshop: Asymptotic Statistics on Stratified Spaces*. Mathematisches Forschungsinstitut Oberwolfach. Zurich : EMS Pub. House, 2004-, 2014.

[97] Sebastian Hummel and Adam Q. Jaffe. *Constructing Maximal Germ Couplings of Brownian Motions with Drift*. Under review at *Electronic Communications in Probability*. 2023.

[98] Adam Q. Jaffe. *Fréchet means in infinite dimensions*. In preparation. 2024.

[99] Adam Quinn Jaffe. "A strong duality principle for equivalence couplings and total variation". In: *Electron. J. Probab.* 28 (2023), pp. 1–33. ISSN: 1083-6489. DOI: 10.1214/23-EJP1016.

[100] Adam Quinn Jaffe. "Strong Consistency for a Class of Adaptive Clustering Procedures". https://arxiv.org/abs/2202.13423. 2022.

[101] Adam Quinn Jaffe and Leonardo Santoro. *Large Deviations Principle for Bures Wasserstein Barycenters*. In progress.

[102] Anil K. Jain. "Data clustering: 50 years beyond K-means". In: *Pattern Recognit. Lett.* 31.8 (2010). Award winning papers from the 19th International Conference on Pattern Recognition (ICPR), pp. 651–666. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2009.09.011. URL: https://www.sciencedirect.com/science/article/pii/S0167865509002323.

[103] Naresh C Jain and Michael B Marcus. "Central limit theorems for $C(S)$-valued random variables". In: *J. Funct. Anal.* 19.3 (1975), pp. 216–231.

[104] Olav Kallenberg. *Foundations of modern probability*. Second. Probability and its Applications (New York). Springer-Verlag, New York, 2002, pp. xx+638. ISBN: 0-387-95313-2. DOI: 10.1007/978-1-4757-4015-8. URL: https://doi.org/10.1007/978-1-4757-4015-8.

[105] Olav Kallenberg. *Random measures, theory and applications*. Vol. 77. Probability Theory and Stochastic Modelling. Springer, Cham, 2017, pp. xiii+694. ISBN: 978-3-319-41596-3; 978-3-319-41598-7. DOI: 10.1007/978-3-319-41598-7. URL: https://doi.org/10.1007/978-3-319-41598-7.

[106] Qiwen Kang et al. *Tropical geometry of phylogenetic tree space: a statistical perspective*. 2018. arXiv: 1805.12400 [math.MG].

[107] Hermann Karcher. "Riemannian center of mass and mollifier smoothing". In: *Communications on pure and applied mathematics* 30.5 (1977), pp. 509–541.

[108] Leonard Kaufman and Peter J Rousseeuw. "Clustering by means of medoids: statistical Data Analysis based on the L^1 Norm". In: *Y. Dodge, Ed* (1987), pp. 405–416.

[109] Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley, 1990. ISBN: 978-0-47031680-1.

[110] Alexander S. Kechris. *Classical descriptive set theory*. Vol. 156. Graduate Texts in Mathematics. Springer-Verlag, New York, 1995, pp. xviii+402. ISBN: 0-387-94374-9. DOI: 10.1007/978-1-4612-4190-4. URL: https://doi.org/10.1007/978-1-4612-4190-4.

[111] Alexander S. Kechris. *The Theory of Countable Borel Equivalence Relations*. 2019.

[112] Alexander S. Kechris and Christian Rosendal. "Turbulence, amalgamation, and generic automorphisms of homogeneous structures". In: *Proc. Lond. Math. Soc. (3)* 94.2 (2007), pp. 302–350. ISSN: 0024-6115. DOI: 10.1112/plms/pdl007. URL: https://doi.org/10.1112/plms/pdl007.

[113] Hans G. Kellerer. "Duality theorems for marginal problems". In: *Z. Wahrsch. Verw. Gebiete* 67.4 (1984), pp. 399–432. ISSN: 0044-3719. DOI: 10.1007/BF00532047. URL: https://doi.org/10.1007/BF00532047.

[114] John L. Kelley. *General topology*. D. Van Nostrand Company, Inc., Toronto-New York-London, 1955, pp. xiv+298.

[115] Wilfrid S. Kendall. "Coupling all the Lévy stochastic areas of multidimensional Brownian motion". In: *Ann. Probab.* 35.3 (2007), pp. 935–953. ISSN: 0091-1798,2168-894X. DOI: 10.1214/009117906000001196. URL: https://doi.org/10.1214/009117906000001196.

[116] Wilfrid S. Kendall. "Nonnegative Ricci curvature and the Brownian coupling property". In: *Stochastics* 19.1-2 (1986), pp. 111–129. ISSN: 0090-9491. DOI: 10.1080/17442508608833419. URL: https://doi.org/10.1080/17442508608833419.

[117] Wilfrid S. Kendall, Mateusz B. Majka, and Aleksandr Mijatović. *Optimal Markovian coupling for finite activity Lévy processes*. Preprint available at https://arxiv.org/abs/2210.11251. 2022. arXiv: 2210.11251 [math.PR].

[118] Wilfrid S. Kendall and Catherine J. Price. "Coupling iterated Kolmogorov diffusions". In: *Electron. J. Probab.* 9 (2004), no. 13, 382–410. ISSN: 1083-6489. DOI: 10.1214/EJP.v9-201. URL: https://doi.org/10.1214/EJP.v9-201.

[119] Ali Khezeli. "Shift-coupling of random rooted graphs and networks". In: *Unimodularity in randomly generated graphs*. Vol. 719. Contemp. Math. Amer. Math. Soc., [Providence], RI, [2018] ©2018, pp. 175–211. DOI: 10.1090/conm/719/14474. URL: https://doi.org/10.1090/conm/719/14474.

[120] Yegor Klochkov, Alexey Kroshnin, and Nikita Zhivotovskiy. "Robust k-means clustering for distributions with two moments". In: *Ann. Statist.* 49.4 (2021), pp. 2206–2230. ISSN: 0090-5364. DOI: 10.1214/20-aos2033. URL: https://doi.org/10.1214/20-aos2033.

[121] Eric D. Kolaczyk et al. "Averages of unlabeled networks: geometric characterization and asymptotic behavior". In: *Ann. Statist.* 48.1 (2020), pp. 514–538. ISSN: 0090-5364. DOI: 10.1214/19-AOS1820. URL: https://doi.org/10.1214/19-AOS1820.

[122] Krzysztof Krakowski, Knut Hüper, and J Manton. "On the computation of the Karcher mean on spheres and special orthogonal groups". In: *Conference Paper, Robomat*. Citeseer. 2007.

[123] Alexey Kroshnin, Vladimir Spokoiny, and Alexandra Suvorikova. "Statistical inference for Bures–Wasserstein barycenters". In: *The Annals of Applied Probability* 31.3 (2021), pp. 1264 –1298. DOI: 10.1214/20-AAP1618. URL: https://doi.org/10.1214/20-AAP1618.

[124] K. Kuratowski. *Topology. Vol. I*. New edition, revised and augmented. Translated from the French by J. Jaworowski. Academic Press, New York-London; Państwowe Wydawnictwo Naukowe, Warsaw, 1966, pp. xx+560.

[125] Kazumasa Kuwada and Karl-Theodor Sturm. "A counterexample for the optimality of Kendall-Cranston coupling". In: *Electron. Comm. Probab.* 12 (2007), pp. 66–72. ISSN: 1083-589X. DOI: 10.1214/ECP.v12-1160. URL: https://doi.org/10.1214/ECP.v12-1160.

[126] T. Laloë. "L_1-quantization and clustering in Banach spaces". In: *Math. Methods Stat.* 19.2 (2010), pp. 136–150. ISSN: 1066-5307. DOI: 10.3103/S1066530710020031. URL: https://doi.org/10.3103/S1066530710020031.

[127] Huiling Le and Alfred Kume. "The Fréchet mean shape and the shape of the means". In: *Advances in Applied Probability* (2000), pp. 101–113.

[128] Thibaut Le Gouic and Jean-Michel Loubes. "Existence and consistency of Wasserstein barycenters". In: *Probability Theory and Related Fields* 168.3 (2017), pp. 901–917. DOI: 10.1007/s00440-016-0727-z. URL: https://doi.org/10.1007/s00440-016-0727-z.

[129] Thibaut Le Gouic and Quentin Paris. "A notion of stability for k-means clustering". In: *Electronic journal of statistics* 12.2 (2018), pp. 4239–4263.

[130] Thibaut Le Gouic et al. "Fast convergence of empirical barycenters in Alexandrov spaces and the Wasserstein space". In: *J. Eur. Math. Soc. (JEMS)* 25.6 (2023), pp. 2229–2250. ISSN: 1435-9855,1435-9863. DOI: 10.4171/jems/1234. URL: https://doi.org/10.4171/jems/1234.

[131] Michel Ledoux and M Talagrand. "Characterization of the law of the iterated logarithm in Banach spaces". In: *Ann. Probab.* (1988), pp. 1242–1264.

[132] John M. Lee. *Introduction to topological manifolds*. Second. Vol. 202. Graduate Texts in Mathematics. Springer, New York, 2011, pp. xviii+433. ISBN: 978-1-4419-7939-1. DOI: 10.1007/978-1-4419-7940-7. URL: https://doi.org/10.1007/978-1-4419-7940-7.

[133] Jüri Lember. "On minimizing sequences for k-centres". In: *J. Approx. Theory* 120.1 (2003), pp. 20–35. ISSN: 0021-9045. DOI: 10.1016/S0021-9045(02)00010-2. URL: https://doi.org/10.1016/S0021-9045(02)00010-2.

[134] Mingjie Liang, René L. Schilling, and Jian Wang. "A unified approach to coupling SDEs driven by Lévy noise and some applications". In: *Bernoulli* 26.1 (2020), pp. 664–693. ISSN: 1350-7265,1573-9759. DOI: 10.3150/19-BEJ1148. URL: https://doi.org/10.3150/19-BEJ1148.

[135] Bo Lin, Anthea Monod, and Ruriko Yoshida. "Tropical geometric variation of tree shapes". In: *Discrete Comput. Geom.* 68.3 (2022), pp. 817–849. URL: https://doi.org/10.1007/s00454-022-00410-y.

[136] Bo Lin and Ruriko Yoshida. "Tropical Fermat–Weber points". In: *SIAM J. Discrete Math.* 32.2 (2018), pp. 1229–1245. eprint: https://doi.org/10.1137/16M1071122.

[137] Zhenhua Lin and Fang Yao. "Intrinsic Riemannian functional data analysis". In: (2019).

[138] T. Linder. "On the training distortion of vector quantizers". In: *IEEE Trans. Inform. Theory* 46.4 (2000), pp. 1617–1623. DOI: 10.1109/18.850705.

[139] Torgny Lindvall. *Lectures on the coupling method*. Corrected reprint of the 1992 original. Dover Publications, Inc., Mineola, NY, 2002, pp. xiv+257. ISBN: 0-486-42145-7.

[140] Torgny Lindvall. "On Strassen's theorem on stochastic domination". In: *Electron. Comm. Probab.* 4 (1999), pp. 51–59. ISSN: 1083-589X. DOI: 10.1214/ECP.v4-1005. URL: https://doi.org/10.1214/ECP.v4-1005.

[141] Torgny Lindvall. "W. Doeblin 1915–1940". In: *Ann. Probab.* 19.3 (1991), pp. 929–934. ISSN: 0091-1798.

[142] S. Lloyd. "Least squares quantization in PCM". In: *IEEE Trans. Inform. Theory* 28.2 (1982), pp. 129–137. DOI: 10.1109/TIT.1982.1056489.

[143] Artur O. Lopes and Jairo K. Mengue. "Duality theorems in ergodic transport". In: *J. Stat. Phys.* 149.5 (2012), pp. 921–942. ISSN: 0022-4715. DOI: 10.1007/s10955-012-0626-3. URL: https://doi.org/10.1007/s10955-012-0626-3.

[144] Aaron Lou et al. "Differentiating through the Fréchet Mean". In: *arXiv preprint arXiv:2003.00335* (2020).

[145] J. MacQueen. "Some methods for classification and analysis of multivariate observations". In: *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66).* Univ. California Press, Berkeley, Calif., 1967, Vol. I: Statistics, pp. 281–297.

[146] Kanti V. Mardia et al. "Principal component analysis and clustering on manifolds". In: *J. Multivariate Anal.* 188 (2022). 50th Anniversary Jubilee Edition, p. 104862. ISSN: 0047-259X. DOI: https://doi.org/10.1016/j.jmva.2021.104862. URL: https://www.sciencedirect.com/science/article/pii/S0047259X21001408.

[147] Valentina Masarotto, Victor M. Panaretos, and Yoav Zemel. "Procrustes Metrics on Covariance Operators and Optimal Transportation of Gaussian Processes". In: *Sankhya A* 81.1 (2019), pp. 172–213. DOI: 10.1007/s13171-018-0130-1. URL: https://doi.org/10.1007/s13171-018-0130-1.

[148] Robert E. Megginson. *An Introduction to Banach Space Theory.* First. Graduate Texts in Mathematics. Springer-Verlag, New York, 1998. ISBN: 978-1-4612-6835-2. DOI: 10.1007/978-1-4612-0603-3. URL: https://doi.org/10.1007/978-1-4612-0603-3.

[149] François G. Meyer. "The Fréchet Mean of Inhomogeneous Random Graphs". In: *Complex Networks & Their Applications X.* Ed. by Rosa Maria Benito et al. Cham: Springer International Publishing, 2022, pp. 207–219. ISBN: 978-3-030-93409-5.

[150] Nina Miolane et al. "Geomstats: A Python Package for Riemannian Geometry in Machine Learning". In: *arXiv preprint arXiv:1805.08308* (2018).

[151] Christoph Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable.* 2nd ed. 2022. URL: https://christophm.github.io/interpretable-ml-book.

[152] Charles M. Newman. "The inner product of path space measures corresponding to random processes with independent increments". In: *Bull. Amer. Math. Soc.* 78 (1972), pp. 268–271. ISSN: 0002-9904. DOI: 10.1090/S0002-9904-1972-12952-5. URL: https://doi.org/10.1090/S0002-9904-1972-12952-5.

[153] Marcel Nutz and Johannes Wiesel. "Entropic optimal transport: convergence of potentials". In: *Probab. Theory Related Fields* 184.1-2 (2022), pp. 401–424. ISSN: 0178-8051. DOI: 10.1007/s00440-021-01096-8. URL: https://doi.org/10.1007/s00440-021-01096-8.

[154] Neil O'Connell and Marc Yor. "Brownian analogues of Burke's theorem". In: *Stochastic Process. Appl.* 96.2 (2001), pp. 285–304. ISSN: 0304-4149,1879-209X. DOI: 10.1016/S0304-4149(01)00119-3. URL: https://doi.org/10.1016/S0304-4149(01)00119-3.

[155] I. Olkin and F. Pukelsheim. "The distance between two random vectors with given dispersion matrices". In: *Linear Algebra and its Applications* 48 (1982), pp. 257–263. ISSN: 0024-3795. DOI: https://doi.org/10.1016/0024-3795(82)90112-4. URL: https://www.sciencedirect.com/science/article/pii/0024379582901124.

[156] Hirofumi Osada and Shota Osada. "Discrete approximations of determinantal point processes on continuous spaces: tree representations and tail triviality". In: *J. Stat. Phys.* 170.2 (2018), pp. 421–435. ISSN: 0022-4715. DOI: 10.1007/s10955-017-1928-2. URL: https://doi.org/10.1007/s10955-017-1928-2.

[157] Victor M Panaretos and Yoav Zemel. *An invitation to statistics in Wasserstein space.* Springer Nature, 2020.

[158] Hae-Sang Park and Chi-Hyuck Jun. "A simple and fast algorithm for K-medoids clustering". In: *Expert Syst. Appl.* 36.2, Part 2 (2009), pp. 3336–3341. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2008.01.039. URL: https://www.sciencedirect.com/science/article/pii/S095741740800081X.

[159] K. Pärna. "On the existence and weak convergence of k-centres in Banach spaces". In: *Tartu Riikl. Ül. Toimetised* 893 (1990), pp. 17–28.

[160] K. Pärna. "On the stability of k-means clustering in metric spaces". In: *Tartu Riikl. Ül. Toimetised* 798 (1988), pp. 19–36.

[161] K. Pärna. "Strong consistency of k-means clustering criterion in separable metric spaces". In: *Tartu Riikl. Ül. Toimetised* 733 (1986), pp. 86–96.

[162] Davide Pigoli et al. "Distances and inference for covariance operators". In: *Biometrika* 101.2 (2014), pp. 409–422. ISSN: 00063444. URL: http://www.jstor.org/stable/43305622 (visited on 01/20/2024).

[163] J. W. Pitman. "On coupling of Markov chains". In: *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* 35.4 (1976), pp. 315–322. DOI: 10.1007/BF00532957. URL: https://doi.org/10.1007/BF00532957.

[164] D. Pollard. "Quantization and the method of k-means". In: *IEEE Trans. Inform. Theory* 28.2 (1982), pp. 199–205. DOI: 10.1109/TIT.1982.1056481.

[165] David Pollard. "Strong consistency of k-means clustering". In: *Ann. Statist.* 9.1 (1981), pp. 135–140.

[166] Luca Pratelli and Pietro Rigo. "A Strong Version of the Skorohod Representation Theorem". In: *Journal of Theoretical Probability* 36.1 (2023), pp. 372–389. DOI: 10.1007/s10959-022-01161-5. URL: https://doi.org/10.1007/s10959-022-01161-5.

[167] Chris Preston. *Some Notes on Standard Borel and Related Spaces*. 2007.

[168] Daniel Raban and Adam Q. Jaffe. *Kantorovich duality for some topologically irregular zero-one costs*. In progress.

[169] Svetlozar T. Rachev and Ludger Rüschendorf. *Mass transportation problems. Vol. I*. Probability and its Applications (New York). Theory. Springer-Verlag, New York, 1998, pp. xxvi+508. ISBN: 0-387-98350-3.

[170] Alexander Rakhlin and Andrea Caponnetto. "Stability of k-means clustering". In: *Adv. in Neural Inf. Process. Syst.* 19 (2007), p. 1121.

[171] D. Ramachandran and L. Rüschendorf. "A general duality theorem for marginal problems". In: *Probab. Theory Related Fields* 101.3 (1995), pp. 311–319. ISSN: 0178-8051. DOI: 10.1007/BF01200499. URL: https://doi.org/10.1007/BF01200499.

[172] Daniel Revuz and Marc Yor. *Continuous martingales and Brownian motion*. Third. Vol. 293. Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer-Verlag, Berlin, 1999, pp. xiv+602. ISBN: 3-540-64325-7. DOI: 10.1007/978-3-662-06400-9. URL: https://doi.org/10.1007/978-3-662-06400-9.

[173] Pietro Rigo. "A note on duality theorems in mass transportation". In: *J. Theoret. Probab.* 33.4 (2020), pp. 2337–2350. ISSN: 0894-9840. DOI: 10.1007/s10959-019-00932-x. URL: https://doi.org/10.1007/s10959-019-00932-x.

[174] Gareth O. Roberts and Jeffrey S. Rosenthal. "General state space Markov chains and MCMC algorithms." eng. In: *Probab. Surv.* 1 (2004), pp. 20–71. URL: http://eudml.org/doc/224487.

[175] L. C. G. Rogers and J. W. Pitman. "Markov functions". In: *Ann. Probab.* 9.4 (1981), pp. 573–582. ISSN: 0091-1798,2168-894X.

[176] Sheldon M Ross. *Introduction to Probability Theory*. Elsevier, 2014, pp. 1–19. DOI: 10.1016/b978-0-12-407948-9.00001-3.

[177] Leonardo V. Santoro and Victor M. Panaretos. *Large sample theory for Bures-Wasserstein barycenters*. arxiv pre-print. 2023. arXiv: 2305.15592 [math.PR].

[178] René L. Schilling and Jian Wang. "On the coupling property of Lévy processes". In: *Ann. Inst. Henri Poincaré Probab. Stat.* 47.4 (2011), pp. 1147–1159. ISSN: 0246-0203,1778-7017. DOI: 10.1214/10-AIHP400. URL: https://doi.org/10.1214/10-AIHP400.

[179] Christof Schötz. "Convergence rates for the generalized Fréchet mean via the quadruple inequality". In: *Electron. J. Stat.* 13.2 (2019), pp. 4280–4345. DOI: 10.1214/19-EJS1618. URL: https://doi.org/10.1214/19-EJS1618.

[180] Christof Schötz. "Strong laws of large numbers for generalizations of Fréchet mean sets". In: *Statistics* 56.1 (2022), pp. 34–52. DOI: 10.1080/02331888.2022.2032063. eprint: https://doi.org/10.1080/02331888.2022.2032063. URL: https://doi.org/10.1080/02331888.2022.2032063.

[181] Forte Shinko. "Equidecomposition in cardinal algebras". In: *Fund. Math.* 253.2 (2021), pp. 197–204. ISSN: 0016-2736. DOI: 10.4064/fm922-6-2020. URL: https://doi.org/10.4064/fm922-6-2020.

[182] Sean Skwerer, Scott Provan, and J. Marron. "Relative Optimality Conditions and Algorithms for Treespace Fréchet Means". In: *SIAM Journal on Optimization* 28 (Jan. 2018), pp. 959–988. DOI: 10.1137/15M1050914.

[183] S. M. Srivastava. *A course on Borel sets*. Vol. 180. Graduate Texts in Mathematics. Springer-Verlag, New York, 1998, pp. xvi+261. ISBN: 0-387-98412-7. DOI: 10.1007/978-3-642-85473-6. URL: https://doi.org/10.1007/978-3-642-85473-6.

[184] V. Strassen. "The existence of probability measures with given marginals". In: *Ann. Math. Statist.* 36 (1965), pp. 423–439. ISSN: 0003-4851. DOI: 10.1214/aoms/1177700153. URL: https://doi.org/10.1214/aoms/1177700153.

[185] Karl-Theodor Sturm. "Probability measures on metric spaces of nonpositive curvature". In: *Heat kernels and analysis on manifolds, graphs, and metric spaces (Paris, 2002)*. Vol. 338. Contemp. Math. Amer. Math. Soc., Providence, RI, 2003, pp. 357–390. DOI: 10.1090/conm/338/06080. URL: https://doi.org/10.1090/conm/338/06080.

[186] Harald Sverdrup-Thygeson. "Strong law of large numbers for measures of central tendency and dispersion of random variables in compact metric spaces". In: *Ann. Statist.* 9.1 (1981), pp. 141–145. ISSN: 0090-5364.

[187] Yoichiro Takahashi. "Absolute continuity of Poisson random fields". In: *Publ. Res. Inst. Math. Sci.* 26.4 (1990), pp. 629–647. ISSN: 0034-5318. DOI: 10.2977/prims/1195170849. URL: https://doi.org/10.2977/prims/1195170849.

[188] Asuka Takatsu. "On Wasserstein geometry of Gaussian measures". In: *Probabilistic approach to geometry*. Mathematical Society of Japan, 2010, pp. 463–472.

[189] Thaddeus Tarpey and Kimberly K. J. Kinateder. "Clustering functional data". In: *J. Classif.* 20.1 (2003), pp. 093–114. DOI: 10.1007/s00357-003-0007-3. URL: https://doi.org/10.1007/s00357-003-0007-3.

[190] Hermann Thorisson. "Coupling". In: *Probability towards 2000 (New York, 1995)*. Vol. 128. Lect. Notes Stat. Springer, New York, 1998, pp. 319–339. ISBN: 0-387-98458-5. DOI: 10.1007/978-1-4612-2224-8_19. URL: https://doi.org/10.1007/978-1-4612-2224-8_19.

[191] Hermann Thorisson. *Coupling, stationarity, and regeneration*. Probability and its Applications (New York). Springer-Verlag, New York, 2000, pp. xiv+517. ISBN: 0-387-98779-7. DOI: 10.1007/978-1-4612-1236-2. URL: https://doi.org/10.1007/978-1-4612-1236-2.

[192] Hermann Thorisson. "On maximal and distributional coupling". In: *Ann. Probab.* 14.3 (1986), pp. 873–876. ISSN: 0091-1798.

[193] Hermann Thorisson. "Transforming random elements and shifting random fields". In: *Ann. Probab.* 24.4 (1996), pp. 2057–2064. ISSN: 0091-1798. DOI: 10.1214/aop/1041903217. URL: https://doi.org/10.1214/aop/1041903217.

[194] Matthew Thorpe et al. "Convergence of the k-means minimization problem using Γ-convergence". In: *SIAM J. Appl. Math.* 75.6 (2015), pp. 2444–2474. ISSN: 0036-1399. DOI: 10.1137/140974365. URL: https://doi.org/10.1137/140974365.

[195] Robert Tibshirani, Guenther Walther, and Trevor Hastie. "Estimating the number of clusters in a data set via the gap statistic". In: *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 63.2 (2001), pp. 411–423. ISSN: 13697412, 14679868. URL: http://www.jstor.org/stable/2680607.

[196] Mo Tiwari et al. "BanditPAM: almost linear time k-medoids clustering via multi-armed bandits". In: *Adv. in Neural Inf. Process. Syst.* Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 10211–10222. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/73b817090081cef1bca77232f4532c5d-Paper.pdf.

[197] A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York, 2008. ISBN: 9780387790527. URL: https://books.google.com/books?id=mwB8rUBsbqoC.

[198] Katharine Turner et al. "Fréchet means for distributions of persistence diagrams". In: *Discrete & Computational Geometry* 52.1 (2014), pp. 44–70.

[199] M. A. Urusov and A. S. Chernyĭ. "Separating times for measures on spaces with filtration". In: *Teor. Veroyatnost. i Primenen.* 48.2 (2003), pp. 416–427. ISSN: 0040-361X. DOI: 10.1137/S0040585X97980488. URL: https://doi.org/10.1137/S0040585X97980488.

[200] V. S. Varadarajan. "On the convergence of sample probability distributions". In: *Sankhyā* 19 (1958), pp. 23–26. ISSN: 0036-4452.

[201] V. S. Varadarajan. "Weak convergence of measures on separable metric spaces". In: *Sankhyā* 19 (1958), pp. 15–22. ISSN: 0036-4452.

[202] Cédric Villani. *Optimal transport*. Vol. 338. Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Old and new. Springer-Verlag, Berlin, 2009, pp. xxii+973. ISBN: 978-3-540-71049-3. DOI: 10.1007/978-3-540-71050-9. URL: https://doi.org/10.1007/978-3-540-71050-9.

[203] Cédric Villani. *Topics in optimal transportation*. Vol. 58. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2003, pp. xvi+370. ISBN: 0-8218-3312-X. DOI: 10.1090/gsm/058. URL: https://doi.org/10.1090/gsm/058.

[204] Florian Völlering. *On maximal agreement couplings*. 2016. DOI: 10.48550/ARXIV.1608.01511. URL: https://arxiv.org/abs/1608.01511.

[205] Ran Wang, Xinyu Wang, and Li-Ming Wu. "Sanov's theorem in the Wasserstein distance: A necessary and sufficient condition". In: *Statistics & Probability Letters* 80 (Mar. 2010), pp. 505–512. DOI: 10.1016/j.spl.2009.12.003.

[206] Ran Wang, Xinyu Wang, and Liming Wu. "Sanov's theorem in the Wasserstein distance: a necessary and sufficient condition". In: *Statistics & Probability Letters* 80.5-6 (2010), pp. 505–512.

[207] Dongkuan Xu and Yingjie Tian. "A comprehensive survey of clustering algorithms". In: *Ann. Data Sci.* 2.2 (2015), pp. 165–193. DOI: 10.1007/s40745-015-0040-1. URL: https://doi.org/10.1007/s40745-015-0040-1.

[208] Zhenhua Lin Yaqing Chen and Hans-Georg Müller. "Wasserstein Regression". In: *Journal of the American Statistical Association* 118.542 (2023), pp. 869–882. DOI: 10.1080/01621459.2021.1956937. eprint: https://doi.org/10.1080/01621459.2021.1956937. URL: https://doi.org/10.1080/01621459.2021.1956937.

[209] Ruriko Yoshida. "Tropical balls and its applications to K nearest neighbor over the space of phylogenetic trees". In: *Mathematics* 9.7 (2021), p. 779. ISSN: 2227-7390.

[210] D. A. Zaev. "On the Monge-Kantorovich problem with additional linear constraints". In: *Mat. Zametki* 98.5 (2015), pp. 664–683. ISSN: 0025-567X. DOI: 10.4213/mzm10896. URL: https://doi.org/10.4213/mzm10896.

[211] Yoav Zemel and Victor M. Panaretos. "Fréchet means and Procrustes analysis in Wasserstein space". In: *Bernoulli* 25.2 (2019), pp. 932 –976. DOI: 10.3150/17-BEJ1009. URL: https://doi.org/10.3150/17-BEJ1009.

[212] Hang Zhou, Zhenhua Lin, and Fang Yao. "Intrinsic Wasserstein Correlation Analysis". In: *arXiv preprint arXiv:2105.15000* (2021). arxiv pre-print.

[213] Herbert Ziezold. "On expected figures and a strong law of large numbers for random elements in quasi-metric spaces". In: *Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes and of the Eighth European Meeting of Statisticians (Tech. Univ. Prague, Prague, 1974),Vol. A*. 1977, pp. 591–602.