UNIFORM LARGE SAMPLE THEORY OF GENERALIZED FRÉCHET MEANS.

by

Andrea Aveni

Department of Statistical Science
Duke University

Defense Date: March 12, 2025

Approved:

Galen Reeves, Supervisor

Sayan Mukherjee

Jonathan Mattingly

Filippo Ascolani

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2025

ABSTRACT

UNIFORM LARGE SAMPLE THEORY OF GENERALIZED FRÉCHET MEANS.

by

Andrea Aveni

Department of Statistical Science
Duke University

Defense Date: March 12, 2025

Approved:

Galen Reeves, Supervisor

Sayan Mukherjee

Jonathan Mattingly

Filippo Ascolani

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University
2025

Abstract

Motivated by the complexity of modern data and the need to fully take advantage of their geometric structure, in this thesis, we study a flexible notion of centrality for data not necessarily lying on a linear space. In particular, we consider a broad generalization of the Fréchet mean. The classical Fréchet mean is obtained minimizing a quadratic notion of dispersion. Our generalization instead considers a more general definition of dispersion indexed by an increasing and convex function ϕ. We start by showing the main properties of these means on linear spaces, and then we study their properties when defined on metric spaces. In this setting, we find necessary and sufficient conditions for ϕ-means to be well-defined and provide a tight bound for the diameter of the ϕ-mean set. We also provide sufficient conditions for all the ϕ-means to coincide in a single point.

In one of our main contributions, we prove the consistency of the sample ϕ-mean of iid observations to its population analogue. Moreover, motivated by the need for robust estimation, we consider a setting in which multiple notions of dispersions are considered jointly. For this purpose, we find conditions under which classes of sample ϕ-means converge uniformly.

Secondly, we quantify more precisely the speed of convergence of the sample ϕ means, for this we need to restrict to Riemannian manifolds where the locally linear structure of these space allows us to prove a CLT. Our result is able to capture the slowing impact of the possible presence of probability mass on the cut locus of the population ϕ-mean. This phenomenon is known in the literature as smeariness and we effectively isolate its impact on the asymptotic variance in our CLT. As in the case of Consistency, we deal with the problem of the joint behaviors of classes of ϕ-means and provide a functional CLT in this case.

We illustrate applications of our results in specific settings that can be of use in practical applications, like when the ambient space is linear, a Sphere, a projective space, or the Lie group of 3d rotations, SO_3. We also briefly describe the algorithms that can be used when dealing with generalized Fréchet means.

Dedication

To Lipschitz and Hölder.

Contents

List of Figures

List of Abbreviations and Symbols

Symbols

$\mathbf{Z}$	Integers
$\mathbf{R}$	Real Numbers
$\mathbf{C}$	Complex Number
$\mathbf{H}$	Quaternions
SO_n	Special Orthogonal Group
id	The identity map
$\mathbb{E}$	Expected value
$\mathbb{V}$	(Co)Variance
$\mathcal{N}_m(\mu, \Sigma)$	m-dimensional Normal distribution with mean μ and covariance Σ
$\mathrm{T}_x\mathcal{X}$	The tangent space of the Riemmanian manifold $\mathcal{X}$ at point x
C_x	The cut locus of point x
H_x	The points in the cut locus C_x for which there are precisely two minimizing geodesic from x.
vol	The volume measure on a Riemannian Manifold
$\nabla f(x)$	Gradient of the function f at point x
$\mathrm{Hess} f(x)(u, v)$	Hessian of the function f at the point x evaluated at the vectors $u, v \in T_x\mathcal{X}$.
I_n	$n \times n$ Identity matrix
$\mathrm{GP}(\mu(\cdot), c(\cdot))$	Gaussian process with mean $\mu(\cdot)$ and covariance $c(\cdot)$

Abbreviations

diam	diameter
iid	independent and identically distributed
pdf	probability density function
GP	Gaussian Process
SLLN	Strong Law of Large Numbers
CLT	Central Limit Theorem

Acknowledgements

I would like to thank all the wonderful people I met during my PhD at Duke.

First of all, Sayan Mukherjee for being my advisor during the years I spent at Duke. It was certainly a challenge and an honor to work with him.

I would also like to thank the several Duke faculty members I had the pleasure of meeting, Galen Reeves, Jonathan Mattingly, Filippo Ascolani, Peter Hoff, Andrea Agazzi, Li Ma, Robert Wolpert, Rick Durrett, Paul Aspinwall, Nicholas A Cook, Ingrid Daubechies, Alexander Dunlap, Saman Habibi Esfahani, Mark A. Stern, Xiaobai Sun, Stacy Sabraw, and especially Lenhard Ng.

I would also mention some of the other people from Duke or outside Duke that have been an important part of my life in these years. Shreya Arya, Youngsoo Baek, Berfin Bastug, Benedetta Bruni, Michele Caprio, Simone Cerreia-Vioglio, Silla Collesei, Ludovico Crippa, Ursula D'Elia, Alex Dombowsky, Marzieh Eidi, Mane Estevez, Beatrice Ferraio, Dejan Govc, Marta Guidolin, Jessie Jiang, Jürgen Jost, Jennifer Kampe, Zeki Kazan, Rohit Kumar, Debottam Kundu, Antonio Lijoi, Paolo Leonetti, Fabio Angelo Maccheroni, Marta Micheli, Ben Menzner, Katerina Papagiannouli, Nicola Pavoni, Rick Presman, Giulio Principi, Eric Roldan, Erika Roldan, Emily Tallman, Ed Tam, Alex Tong, Nahid Torbati, Lucrezia Villa, Ziyi Wang, Qianhe Yin, Alessandro Zito, Giulio Zucal among many others.

A special thanks goes to Giulio Principi who has supported me more than anyone.

1. Introduction

1.1 Motivation

Motivated by practical applications, a broad objective in my research has been the need to design theoretically sound statistical procedures for complex data displaying certain geometric or algebraic regularities. In particular, I have contributed towards developing a statistics theory for the large sample behaviour of a large class of non-parametric estimators on Metric Spaces and Manifolds. In particular, along with my advisor, I considered a broad generalization of Frechét means which derives from simple and flexible hypotheses on the loss function, and I provided consistency results not just of each such estimator, but also studied the joint behaviour of classes of these estimators. In our first work Aveni and Mukherjee, 2024, we provided a uniform consistency result reminiscent of Glivenko-Cantelli class results. In a second work Aveni and Mukherjee, 2025 we restricted to Riemannian Manifolds and we considered Central Limit Theorems for our generalized Frechét means, and again we studied their joint behaviour in a way reminiscent of Donsker class results.

The concept of Generalized Féchet mean naturally finds an application even in a discrete setting when dealing with ranking aggregation. This research area, on which I worked with Ludovico Crippa and Giulio Principi, is similarly driven by the need for techniques that are at the same time theoretically motivated, flexible, and computationally feasible. In Aveni et al., 2024 we provided a unified axiomatic approach to the problem of rank aggregation with applications in several fields, in particular with Mallows Models. And, from the computational standpoint, we generalized an essential result regarding the computation of a particular case of Generalized Fréchet mean by Diaconis Diaconis and Graham, 1977 to our class of distances.

1.2 Main Results

In applied data analysis and mathematical statistics, the increasing complexity of modern data has motivated the development of statistical inference on more geometric objects such as Riemannian manifolds or metric spaces – examples include data that are shapes, positive definite matrices, or linear subspaces. These kinds of data appear routinely in Geoscience, Astrophysics, Elementary Physics, Radiology, and Data Visualization. The Fréchet mean, defined as the point in a metric space for which the sum of squared distances is minimized (see Fréchet, 1948; Grove and Karcher, 1973), naturally extends the notion of the mean and centrality from Euclidean spaces to these general settings.

The recent literature in applied probability and mathematical statistics on Fréchet means has focused on two aspects of Fréchet means: large sample asymptotics and generalization of the definition of the mean itself. The work of Bhattacharya R. Bhattacharya and Patrangenaru, 2003, R. Bhattacharya and Patrangenaru, 2005, R. Bhattacharya and Lin, 2017 originated a vast theoretical literature on the large sample asymptotics of Fréchet means on metric spaces for both intrinsic and extrinsic means recently extended to stratified spaces in Mattingly et al., 2024. The other direction has been the generalization of the Fréchet mean, an illustrative example is moving from the *squared* distance to other exponents $p \geqslant 1$ A. Bhattacharya and Bhattacharya, 2012. The argument made for the generalizations has been that using the *squared* distance is restrictive and different.

In my contributions, I considered both aspects – large sample asymptotics and generalization of the Fréchet mean. Previous generalizations of the Fréceht mean have focused on particular loss functions or a single parameter family of loss functions. We consider less restrictive generalizations of Fréchet means. In addition, we fill a gap in the literature studying the *joint* behaviour of classes of Generalized Fréchet means. Instead, I consider a broad generalization of the Fréchet means indexed not by a single real parameter but by a convex function ϕ and whose large sample behaviour is explored jointly. In particular, let $\phi : \mathbf{R}_+ \to \mathbf{R}_+$ be an increasing convex function, then we define the ϕ-mean of a Borel probability measure μ on a metric space $(\mathcal{X}, \rho)$ as the set

$$\mathfrak{m}^{\phi}(\mu) := \operatorname*{arg\,min}_{x \in \mathcal{X}} \int_{\mathcal{X}} \phi(\rho(x, y))\mu(\mathrm{d}y).$$

We show that under mild conditions, this set is well-defined, non-empty, and compact. Furthermore, we characterize its diameter. Our main focus is, however, the convergence of the sample ϕ-mean to its population counterpart. Thus, let $\mathfrak{m}_n^\phi := \mathfrak{m}^\phi(\frac{1}{n}\sum_{j=1}^n \delta_{X_j})$ be the random variable obtained from a sequence of iid observations X_j from μ. We study how well $\mathfrak{m}_n^\phi$ approximates $\mathfrak{m}^\phi(\mu)$.

In our first paper Aveni and Mukherjee, 2024 we focus on consistency, and our first result is a Strong Law that characterizes the μ and ϕ for which

$$\rho_H(\mathfrak{m}_n^\phi, \mathfrak{m}^\phi(\mu)) \underset{n\to\infty}{\overset{a.s.}{\to}} 0,$$

where ρ_H denotes the asymmetric Hausdorff distance which reduced to ρ when both arguments are singletons.

But our strongest result deals with uniform consistency of families of ϕ-means. In particular, using techniques from Empirical Processes (Vaart and Wellner, 1996), we find conditions under which μ and a family Ψ of functions ϕ are such that

$$\sup_{\phi\in\Psi} \rho_H(\mathfrak{m}_n^\phi, \mathfrak{m}^\phi(\mu)) \underset{n\to\infty}{\overset{a.s.}{\to}} 0.$$

One condition for this second result to hold is that $\mathfrak{m}^\phi(\mu)$ needs to be a singleton. For this reason, we investigate conditions that guarantee this to happen. We find that if $\mathcal{X}$ is a Riemannian symmetric space (described in Cartan, 1926) and μ is continuous with a density radially decaying around a point $x \in \mathcal{X}$, then $\mathfrak{m}^\phi(\mu) = \{x\}$ is the same singleton independently on ϕ.

In our second paper Aveni and Mukherjee, 2025, we focused on the case in which $\mathcal{X}$ is a Riemannian manifold and we always assumed $\mathfrak{m}^\phi(\mu)$ to be a singleton. In this case, we can quantify the speed of convergence of $\mathfrak{m}_n^\phi$ to $\mathfrak{m}^\phi := \mathfrak{m}^\phi(\mu)$ with a Central Limit Theorem (CLT). In order to do so, we need to linearize $\mathcal{X}$ in the vicinity of the population ϕ-mean. The Riemannian structure is precisely what enables us to perform this linearization. In particular, when $\mathfrak{m}_n^\phi$ is sufficiently close to $\mathfrak{m}^\phi$, we can project it onto the tangent plane $T_{\mathfrak{m}^\phi}\mathcal{X}$ at $\mathfrak{m}^\phi$ through the log map $\exp_{\mathfrak{m}^\phi}^{-1} : \mathcal{X} \to T_{\mathfrak{m}^\phi}\mathcal{X}$ which is a well defined radial isometry (see do Carmo, 2013). Using this technique we obtain a vector $\exp_{\mathfrak{m}^\phi}^{-1}(\mathfrak{m}_n^\phi)$ which we can now inflate with a function depending on n. The first result we obtained, is that if $\phi'(0) = 0$ and under some mild conditions

$$\sqrt{n}\exp^{-1}_{\mathfrak{m}^\phi}(\mathfrak{m}^\phi_n) \xrightarrow[n\to\infty]{w} \mathcal{N}(0, V^\phi),$$

where w denotes the convergence in distribution, $\mathcal{N}$ is the multivariate normal distribution and V^ϕ is a positive definite symmetric function depending on ϕ and μ. Our expression for V^ϕ comprises two terms; a standard term generalizing that corresponds to the Variance in the Euclidean and quadratic CLT and a non-standard term that only appears when the cut locus of the manifold $\mathcal{X}$ at $\mathfrak{m}^\phi$ is topologically significant (in particular when it has at least codimension 2). The impact of the cut locus has previously been noted in special cases and it is known in the literature as *smeariness.* (Described for the first time in Eltzner, 2020 and analyzed in the context of CLTs in Hotz et al., 2022)

Moreover, as in the case of consistency, we also study the joint behaviour of multiple ϕ-means. In particular, we conjecture conditions for families of functions Ψ under which

$$\sup_{\phi\in\Psi} W\left(\sqrt{n}\exp^{-1}_{\mathfrak{m}^\phi}(\mathfrak{m}^\phi_n), \mathcal{N}(0, V^\phi)\right) \xrightarrow[n\to\infty]{} 0,$$

where W is the Wasserstein distance between (the distribution of) the random variable $\sqrt{n}\exp^{-1}_{\mathfrak{m}^\phi}(\mathfrak{m}^\phi_n)$ and the normal distribution $\mathcal{N}(0, V^\phi)$. This allows us to also strengthen our previous CLT to the case in which $\phi'(0) > 0$ as in the case of the median.

1.3 Structure of this Thesis

In the first Chapter, we provide an overview of our results and we explain their novelty in comparison to the existing literature.

In the second Chapter, we provide the basic definitions of our object of study, the generalized Fréchet means. We explain some basic properties of these objects, and in particular we illustrate the generality of our definition with general results and with numerical examples.

In the third Chapter, we start considering the statistical problem of estimating the population ϕ-mean of a distribution with its sample counterpart. In particular, in this Chapter, we work on metric spaces and we establish our Consistency results. The first one for a fixed function ϕ and the second for classes of these functions.

In the fourth Chapter, we restrict ourselves to Riemannian spaces, and we, after introducing the relevant notions about manifolds (with particular emphasis on the cut locus) we state and prove our CLT. Finally, we expose our conjecture for a uniform CLT.

In the fifth Chapter, we illustrate our results for particular classes of manifolds of particular interest in applications. We also illustrate some algorithms that can be used to effectively compute generalized Fréchet means.

In the sixth and last Chapter, we consider an applied problem of particular interest where ϕ means arise naturally. In particular, in this Chapter we deal with ranking aggregation.

1.4 Literature Review

The Fréchet mean extended the notion of the mean and centrality from Euclidean spaces to more general settings like Riemannian manifolds or metric spaces. In applied data analysis and mathematical statistics, the increasing complexity of modern data has motivated the development of statistical inference on more geometric objects such as Riemannian manifolds or metric spaces – examples include data that are shapes, positive definite matrices, or linear subspaces. The Fréchet mean, defined as the point in a metric space for which the sum of squared distances is minimised (see Fréchet, 1948; Grove and Karcher, 1973) is a natural extension of the arithmetic mean to arbitrary metric spaces.

The recent literature in applied probability and mathematical statistics on Fréchet means has focused on two aspects of Fréchet means: large sample asymptotics and generalization of the definition of the mean itself. The work of Battacharia R. Bhattacharya and Patrangenaru, 2003 originated a vast theoretical literature on the large sample asymptotics of Fréchet means on metric spaces for both intrinsic and extrinsic means. The other direction has been the generalization of the Fréchet mean, an illustrative example is moving from the *squared* distance to other exponents $p \geqslant 1$ A. Bhattacharya and Bhattacharya, 2012. The argument made for the generalizations has been that the use of the *squared* distance is restrictive so one should consider at least a one-parameter family of Fréchet means.

This paper considers both aspects – large sample asymptotics and generalisation of the Fréchet mean. Previous generalizations of the Fréceht mean have focused on particular loss functions or a single parameter family of loss functions. We consider less restrictive definitions of Fréchet means. In addition, we fill a gap in the literature studying the *joint* behaviour of classes of Generalized Fréchet means. This paper considers a broad generalization of the Fréchet means indexed not by a single real parameter but by a convex function ϕ. In addition, we study the joint behaviour of particular classes of Generalized Fréchet means.

The standard notion of a mean in linear vector spaces is the arithmetic mean, however, other notions of centrality such as the geometric median are used for robustness with respect to outliers, see Narula and Wellington, 1982 for a discussion in the context of linear models. An illuminating discussion of the concept of mean is provided in the papers by De Finetti de Finetti, 1931 and Kolmogorov Kolmogorov, 1930. Again, Fréchet means generalize the usual notion of the arithmetic mean, typically for Riemannian manifolds or metric spaces.

The large sample properties of Fréchet means and generalisations of Fréchet means in metric spaces have been an area of active research. The two main large sample properties considered have been (strong) law of large numbers (SLLN) and central limit theorems (CLT). Note the CLT provides information regarding the speed of convergence of the empirical Fréchet mean to its population counterpart. The starting point of this literature is in the influential papers by Bhattacharya R. Bhattacharya and Patrangenaru, 2003 and R. Bhattacharya and Patrangenaru, 2005, where an extensive theory of convergence of the Fréchet mean on manifolds was laid out for both the intrinsic and extrinsic means. The Fréchet mean on the circle has been examined in detail with the goal of finding necessary and sufficient conditions for the Fréchet mean to be a singleton Charlier, 2012; Hotz and Huckemann, 2015. The consistency of Fréchet means where distances are given by the

p-norm (the 2-norm is the Fréchet mean) has been studied A. Bhattacharya and Bhattacharya, 2012; Evans and Jaffe, 2020. For all the above definitions of the Fréchet mean, the distance function is either fixed or a parametric family, typically a one-parameter family. Recently, positive results on the consistency of non-parametric generalizations of the Fréchet means Romon and Brunel, 2023; Schötz, 2018 have been provided under suitable conditions. For central limit theorems the influential paper R. Bhattacharya and Lin, 2017 proved a CLT for a large class of spaces but assumed strong conditions on the support of the probability measures. Most of the hypotheses on the support of the probability measures were removed in a recent paper Hotz et al., 2022 that proved a CLT for Fréchet means on Riemannian manifolds. A Donsker theorem for p-means on manifolds also exists Le, 2016. In a different direction, a CLT for Fréchet means was recently provided for stratified spaces Mattingly et al., 2024 – stratified spaces are unions and intersections of manifolds.

The strong law of large numbers results for Fréchet means stated above all consider the convergence of a single generalized Fréchet mean and do not consider the problem of joint consistency across a class of generalized Fréchet means. This joint analysis in addition to a necessary and sufficient condition on the class to ensure the joint consistency, is the main contribution of our paper. In a companion paper Aveni and Mukherjee, 2025, we will restrict the space to Riemannian manifolds and provide a uniform CLT for classes of ϕ-means.

2. Definitions

2.1 Definition of Generalized Fréchet Means

Given a (Borel) probability measure μ, the classical notion of centrality is the mean, which can be defined on a linear space $\mathcal{X}$ as follows

$$\mathfrak{m}_\mu := \int_{\mathcal{X}} y\mu(\mathrm{d}y)$$

This definition, however, cannot be easily extended to non-linear spaces, because in the absence of a vector structure, the operations of summing points and multiplying them with a scalar (indeed the above integral is precisely a weighted sum of vectors) are not defined. However, it is possible to characterize the mean on linear spaces in a way that does not rely on the vector structure of these spaces, but rather as the solution of a minimization problem only involving the distance ρ between points. In particular, if we define the quadratic loss function $F_\mu : \mathcal{X} \to \mathbf{R}_+$ as

$$F_\mu(x) = \int_{\mathcal{X}} (\rho(x, y))^2 \mu(\mathrm{d}y),$$

then we have the following.

Proposition 1. *Given any probability measure μ with finite second moment, we have that*

$$\mathfrak{m}_\mu = \arg\min_{x \in \mathcal{X}} F_\mu(x)$$

Since the definition of F_μ only depends on the metric structure of the space X, by taking this characterization as the definition of the mean, we can extend it to any metric space. The resulting object is the Fréchet mean defined for the first time in Fréchet, 1948. This concept adheres to the classical notion of mean on linear space and for probability measures in $\mathcal{L}^2$, but can exhibit exotic behaviours on general metric spaces. In particular, it may not be given by a unique point, but instead by a set. A typical example would be a uniform distribution on a circle; in the case the Fréchet mean set is the entire circle.

The Fréchet mean is thus defined as the point (or the collection of points) on a metric space, that minimizes the average squared distance from all the points in the space weighted according to a certain measure μ. This definition immediately leads to a notion of dispersion for a distribution. In particular, the Fréchet dispersion of a probability measure can be defined as

$$\mathfrak{s}_\mu := \min_{x \in \mathcal{X}} F_\mu(x) = F_\mu(\mathfrak{m}_\mu)$$

This notion of dispersion reduces to the definition of the Variance when $\mathcal{X} = \mathbf{R}$ and to the so called total variance (that is the trace of the covariance matrix) in the case of $\mathcal{X} = \mathbf{R}^n$.

From this review of the Fréchet mean, we can see how the notion of mean is intimately linked to the choice of a particular loss function that depends uniquely on the distance between points. In the case of the Fréchet mean, the choice is to consider the average squared distance. This is a reasonable choice since the square function is an increasing function, and thus the resulting loss will increase the further away we are from the "center" of our distribution μ. Moreover, the square is a convex function and this leads to a unique minimizer (at least in the case of linear spaces). However, this choice forces us to consider only those measures finite second moment and moreover tend to give high importance to points in the support of μ that are far away from its "center". In many contexts, a subquadratic loss function is preferable, especially in the presence of outliers.

Our generalization of the Fréchet mean is similarly defined starting from a loss function depending only on the distance between points but is much more general in that, instead of the square distance, it considers a generic increasing and convex function ϕ of the distance, which is then similarly averaged according to μ. Let Φ denote the collection of all functions $\phi : \mathbf{R} \to \mathbf{R}_+$ that are strictly increasing, convex and with $\phi(0) = 0$, then we defined the generalized Fréchet mean as

Definition 2. *Given $\mu \in \mathcal{P}(\mathcal{X})$ and $\phi \in \Phi$, we define the ϕ-mean of μ as*

$$\mathfrak{m}_\mu^\phi := \arg\min_{x \in \mathcal{X}} F_\mu^\phi(x),$$

where, $F^{\phi}_{\mu} : \mathcal{X} \to \mathbf{R}_+$ is the ϕ-loss function of μ,

$$F^{\phi}_{\mu}(x) := \int_{\mathcal{X}} \phi(\rho(x,y))\mu(\mathrm{d}y).$$

Moreover, we define the ϕ-dispersion of μ as

$$\mathfrak{s}^{\phi}_{\mu} = \min_{x \in \mathcal{X}} F^{\phi}_{\mu}(x) = F^{\phi}_{\mu}(\mathfrak{m}^{\phi}_{\mu}).$$

As usual in the literature, when $\mathfrak{m}^{\phi}_{\mu}$ is a singleton, we do not distinguish between the singleton and its sole member.

We also define a notion of ϕ-mean for a collection of observations $(x_j)_{j=1}^n$. Such a collection of observations corresponds uniquely to an empirical probability measure $\hat{\mu}((x_j)_{j=1}^n) := \frac{1}{n}\sum_{j=1}^n \delta_{x_j}$ and the ϕ-mean of the observations is defined as the ϕ-mean of the induced empirical measure. That is

$$F^{\phi}_{(x_j)_{j=1}^n}(x) := F^{\phi}_{\hat{\mu}((x_j)_{j=1}^n)}(x) = \frac{1}{n}\sum_{j=1}^n \phi(\rho(x,x_j)),$$

$$\mathfrak{m}^{\phi}((x_j)_{j=1}^n) := \mathfrak{m}^{\phi}_{\hat{\mu}((x_j)_{j=1}^n)} = \arg\min_{x \in \mathcal{X}} \frac{1}{n}\sum_{j=1}^n \phi(\rho(x,x_j)),$$

$$\mathfrak{s}^{\phi}((x_j)_{j=1}^n) := \mathfrak{s}^{\phi}_{\hat{\mu}((x_j)_{j=1}^n)} = \min_{x \in \mathcal{X}} \frac{1}{n}\sum_{j=1}^n \phi(\rho(x,x_j)).$$

Often when the observations are clear from the context, we simply write F^{ϕ}_n, $\mathfrak{m}^{\phi}_n$ and $\mathfrak{s}^{\phi}_n$ instead of $F^{\phi}_{(x_j)_{j=1}^n}(x)$, $\mathfrak{m}^{\phi}((x_j)_{j=1}^n)$ and $\mathfrak{s}^{\phi}((x_j)_{j=1}^n)$.

Our definitions of ϕ-mean and ϕ-dispersion make sense on general metric spaces, however, in order to guarantee the ϕ-mean set to be well behaved (for instance to be non-empty and compact), we will restrict ourselves to particularly nice metric spaces. In particular, we work on a complete, locally-compact, length-metric space $(\mathcal{X}, \rho)$. By length-metric spaces, we mean metric spaces where the distance between any two points is equal to the infimum of the length of all the continuous curves connecting the two points. Notice that, in particular, length-spaces are path connected. By the general version of the Hopf-Rinow Theorem proven by Cohn-Vossen in 1935 (see Gromov, 2007), such a space has the Heine-Borel property meaning that closed and bounded subsets of $\mathcal{X}$ are compact. An important class of spaces with these properties is given by connected and complete Riemannian manifolds equipped with their geodesic metric. (For further details see Chapter 6 of Lee, 2021 or Chapter 3 of do Carmo, 2013.) We endow $\mathcal{X}$ with its Borel sigma algebra Σ_ρ and all the probability measures we consider will belong to $\mathcal{P}(\mathcal{X})$, the collection of all Borel probability measures on $(\mathcal{X}, \Sigma_\rho)$.

As particular cases, by considering $\phi : t \mapsto t^2$ we recover the classical Fréchet mean. Instead, by taking $\phi : t \mapsto |t|$, we obtain a generalization of the median on general metric spaces, known as the geometric median.

As we said previously, the requirement of ϕ being increasing is essential in order to obtain a measure of centrality. Indeed, we need a loss function that increases when we get progressively further away from the "center" of our distribution. Convexity of ϕ is also required, both for mathematical traceability, but also for reasons that will be explored in section 2.2. Finally, given that ϕ is increasing and convex, it must be bounded from below and thus, the assumption of $\phi(0) = 0$ is wlog because adding any constant to it does not modify the minimization problem and thus does not affect $\mathfrak{m}^\phi_\mu$. Moreover, this arbitrary choice is essential in order to guarantee that ϕ-dispersion will be non negative and that degenerate distributions will have zero ϕ-dispersion.

2.2 Basic properties of generalized Fréchet means

In this section, we consider the main properties that each ϕ-mean for $\phi \in \Phi$ has on Euclidean spaces $\mathcal{X} = \mathbf{R}^m$. This will help elucidate the importance of the requirement of functions in Φ to be convex. We denote by Φ^+ the collection of functions in Φ that are strictly convex, meaning that for any $x < y$ we have that $\phi((x+y)/2) < (\phi(x) + \phi(y))/2$.

Theorem 3. *Let $\phi \in \Phi$, the the ϕ-mean $\mathfrak{m}^\phi$ has the following properties that holds for any collection $(x_j)_{j=1}^n$ of points.*

- ***Identity.*** *For all $x \in \mathcal{X}$,*

$$\mathfrak{m}^{\phi}((x)_{j=1}^{n}) = x.$$

- ***Symmetry.*** *For all $x \in \mathcal{X}$, $\sigma \in S_n$,*

$$\mathfrak{m}^{\phi}((x_j)_{j=1}^{n}) = \mathfrak{m}^{\phi}((x_{\sigma(j)})_{j=1}^{n}).$$

- ***Coherence.*** *For any number of repetitions m,*

$$\mathfrak{m}^{\phi}(((x_j)_{j=1}^{n})_{i=1}^{m}) = \mathfrak{m}^{\phi}((x_j)_{j=1}^{n}).$$

- ***Stability.*** *If $x \in \mathfrak{m}^{\phi}((x_j)_{j=1}^{n})$, then*

$$\mathfrak{m}^{\phi}((x_j)_{j=1}^{n}, (x)_{j=1}^{m}) = x.$$

- ***Equivariance.*** *For any isometry $f : \mathcal{X} \to \mathcal{X}$,*

$$\mathfrak{m}^{\phi}((f(x_j))_{j=1}^{n}) = f(\mathfrak{m}^{\phi}((x_j)_{j=1}^{n})).$$

Moreover, if $\mathcal{X} = \mathbf{R}^n$ is an Euclidean space and $\phi \in \Phi^{+}$, then

- ***Uniqueness.*** *The ϕ-mean is well defined it is a unique point.*
- ***Convexity.*** *Let* co *denotes the convex hull, then*

$$\mathfrak{m}^{\phi}((x_j)_{j=1}^{n}) \in \operatorname{co}((x_j)_{j=1}^{n}).$$

- ***Monotonicity.*** *For any direction $w \in \mathbf{R}^n$,*

$$\langle w, \partial_h|_{h=0}\, \mathfrak{m}^{\phi}(x_1 + hw, (x_j)_{j=2}^{n}) \rangle \geqslant 0.$$

Proof. The first three properties are immediate observations. As for stability, we have that the only difference between the corresponding losses is given by

$$(n+m)F^{\phi}_{((x_j)_{j=1}^{n},(x)_{j=1}^{m})}(y) - nF^{\phi}_{(x_j)_{j=1}^{n}}(y) = m\phi(\rho(x,y)).$$

And this quantity is non-negative and strictly positive for any $y \neq x$. This implies that the argmin of $F^{\phi}_{((x_j)_{j=1}^{n},(x)_{j=1}^{m})}$ is now reduced to x.

For equivariance, we have that

$$F^{\phi}_{(f(x_j))_{j=1}^n}(y) = \frac{1}{n}\sum_{j=1}^{n}\phi(\rho(y, f(x_j)))$$

$$= \frac{1}{n}\sum_{j=1}^{n}\phi(\rho(f^{-1}(y), x_j)) = F^{\phi}_{(x_j)_{j=1}^n}(f^{-1}(y)).$$

And thus,

$$\mathfrak{m}^{\phi}((f(x_j))_{j=1}^n) = \arg\min_{y\in\mathcal{X}} F^{\phi}_{(f(x_j))_{j=1}^n}(y)$$

$$= \arg\min_{y\in\mathcal{X}} F^{\phi}_{(x_j)_{j=1}^n}(f^{-1}(y))$$

$$= f\left(\arg\min_{z\in\mathcal{X}} F^{\phi}_{(x_j)_{j=1}^n}(z)\right).$$

In the case in which $\mathcal{X} = \mathbf{R}^n$, each loss $F^{\phi}_{(x_j)_{j=1}^n}$ is a sum of convex functions, and it is itself convex, thus the ϕ-mean will be a closed convex set. In particular, if $\phi \in \Phi^+$, then the loss function will be strictly convex and the minimum will be unique.

Since each function $x \mapsto \phi(\|x - x_j\|)$ is continuous and convex, its subgradient is defined everywhere and given by

$$\nabla_x\phi(\|x - x_j\|) = (x - x_j)\frac{\phi'(\|x - x_j\|)}{\|x - x_j\|} = (x - x_j)h(\|x - x_j\|),$$

where f' is the subdifferential of the convex function $t \mapsto \phi(|t|)$ and $h : t \mapsto f'(t)/t$ is well definite and finite everywhere, except possibly at zero, where the expression $(x - x_j)h(\|x - x_j\|)$ (interpreted in a limiting way) is nonetheless finite. Thus, the ϕ-mean is given by the unique point x such that

$$0 \in \nabla_x F^{\phi}_{(x_j)_{j=1}^n}(x) = \frac{1}{n}\sum_{j=1}^{n}\nabla_x\phi(\|x - x_j\|) = \frac{1}{n}\sum_{j=1}^{n}(x - x_j)\frac{\phi'(\|x - x_j\|)}{\|x - x_j\|}.$$

But then, by rewriting these terms, we obtain that the unique minimum can be expressed as a convex linear combination of the observations $(x_j)_{j=1}^n$.

$$x = \left(\sum_{j=1}^n x_j \frac{\phi'(\|x - x_j\|)}{\|x - x_j\|}\right) \Big/ \left(\sum_{j=1}^n \frac{\phi'(\|x - x_j\|)}{\|x - x_j\|}\right) = \frac{\sum x_j h(\|x - x_j\|)}{\sum h(\|x - x_j\|)}.$$

Where, notice that the weight of each observation only depends on its distance from the "center". As for monotonicity, we need to apply the implicit function theorem to the above expression. Let $v_j := x - x_j$ and $t_j := \|v_1\|$, then the Jacobian of the function $x_1 \mapsto x = \mathfrak{m}^\phi((x_j)_{j=1}^n)$ is given by the matrix

$$J = M_1 \left(\sum_{j=1}^n M_j\right)^{-1}, \quad M_j := \left(I - \frac{v_j v_j^T}{t_j^2}\right) \frac{\phi'(t_j)}{t_j} + \frac{v_j v_j^T}{t_j^2} \phi''(t_j).$$

We can generally assume that $t_j > 0$, then all the matrices involved are positive definite with determinant

$$\det(M_j) = \left(\frac{\phi'(t_j)}{t_j}\right)^{d-1} \phi''(t_j).$$

which is positive precisely when $\phi \in \Phi^+$. Finally, the quantity of interest is given by,

$$\langle w, \partial_h|_{h=0}\, \mathfrak{m}^\phi(x_1 + hw, (x_j)_{j=2}^n)\rangle = w^T J w \geqslant 0.$$

$\square$

The monotonicity property indicates that when one observation is moved in a certain direction w, the resulting ϕ-mean also moves approximately in the same direction at least in the sense that the inner product between w and the derivative of the ϕ-mean when an observation is moved infinitesimally in the direction of w is non-negative.

Notice that in the case of the mean, when $\phi : t \mapsto t^2$, we have that $M_j = I$ and thus $\partial_h|_{h=0}\, \mathfrak{m}^\phi(x_1 + hw, (x_j)_{j=2}^n)$ is precisely equal to w/n and the two directions are parallel. For a generic ϕ we cannot expect these two vectors to be parallel, but at least to move in similar directions. We illustrate this with a figure.

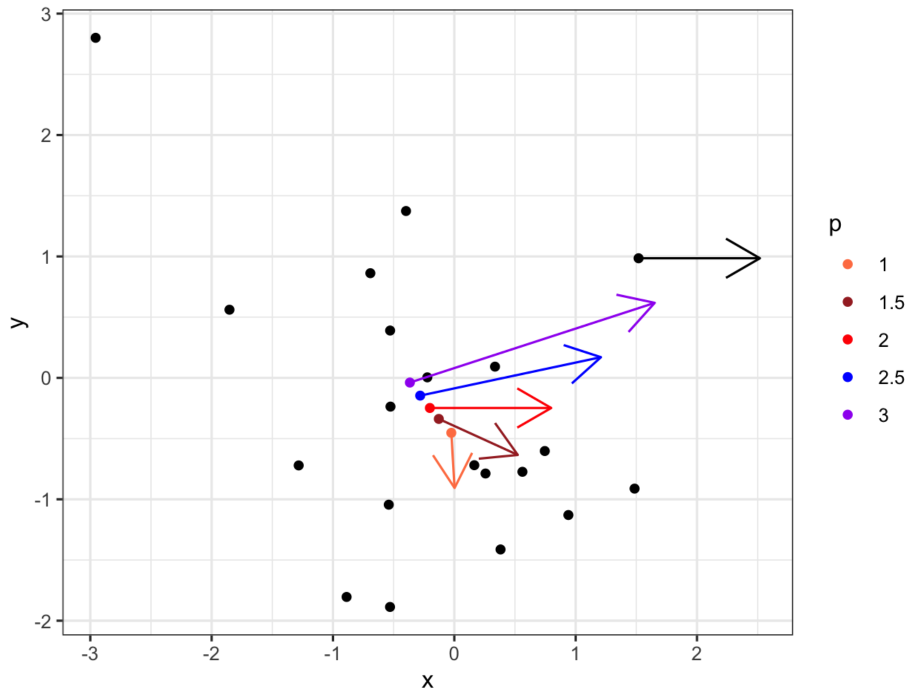

FIGURE 2.1: We drew 20 points in black, and we perturbed one observation according to the black arrow. The colored points are the ϕ-means when $\phi : t \mapsto t^p$ for different values of p. Each of these is accompanied by an arrow that shows the effect of perturbing the back observation. These effects are multiplied by 20, to ease their comparison with the black arrow. As it is possible to see, all ϕ-means move in a similar direction as the perturbed observation. However, only the classical Fréchet mean moves precisely as the black observation. We can also see how higher values of p tend to give more weight to outliers, and vice-versa, how values of p closer to one lead to more robust ϕ-means.

2.2.1 The case of two points

When we have just one point, it is immediate to see that $\mathfrak{m}^\phi(x) = x$ independently of ϕ and of the metric space $\mathcal{X}$. The case of two points is slightly more interesting, and it will provide strong reasons to always take convex functions ϕ. Given any two distinct points x, y on a complete, locally-compact, length-metric space, there is always at least one minimal geodesic connecting them. We will show that, according to our intuition, when ϕ is strictly convex, the ϕ-mean of (x, y) consists of the collection of midpoints of all these minimal geodesics. On the contrary, if ϕ is just convex, the ϕ-mean will consist of some "central" section of these geodesics; finally if ϕ is simply increasing, it may be the case that the ϕ-mean is given by the endpoints x, y, violating our intuition. More precisely, we have the following statement.

Proposition 4. *Let $\mathcal{X}$ be a complete, locally-compact, length-metric space and let $x, y \in \mathcal{X}$ with $d = \rho(x, y)$. We denote by Γ, the collection of all unit speed minimizing geodesics $\gamma : [0, d] \to \mathcal{X}$ with $\gamma(0) = x$ and $\gamma(d) = y$. Then,*

- *If $\phi \in \Phi^+$, then*
$$\mathfrak{m}^{\phi}((x, y)) = \bigcup_{\gamma \in \Gamma} \gamma(d/2).$$
- *If $\phi \in \Phi$, then there is some $a \in [0, d/2]$ depending only on $\rho(x, y)$ such that*
$$\mathfrak{m}^{\phi}((x, y)) = \bigcup_{\gamma \in \Gamma} \gamma([a, d/2 - a]).$$
- *If ϕ is a non-negative continuous increasing function, then there is a subset $A \subseteq [0, d]$ with $d/2 - A = A$ such that*
$$\mathfrak{m}^{\phi}((x, y)) = \bigcup_{\gamma \in \Gamma} \gamma(A).$$

Proof. Let $d = \rho(x, y)$. We start by proving the third point, of which the other are particular cases. Assume by contradiction, that there is some $z \in \mathcal{X} \backslash \bigcup_{\gamma \in \Gamma} \gamma([0, d])$ among the ϕ-means. Let us say that $\rho(z, x) = t > 0$. Since $w \in \bigcup_{\gamma \in \Gamma} \gamma([0, d])$ iff $\rho(x, w) + \rho(w, y) = \rho(x, y)$, we must have that $\rho(z, y) > d - t$. Now, if $w \in \bigcup_{\gamma \in \Gamma} \gamma([0, d])$ is any point with $\rho(w, x) = \min(t, d)$, we have that

$$2F^{\phi}_{(x,y)}(z) = \phi(t) + \phi(\rho(z, y)) > \phi(t) + \phi((d - t)^+) = F^{\phi}_{(x,y)}(w).$$

This shows that $\mathfrak{m}^{\phi}((x, y)) \subseteq \bigcup_{\gamma \in \Gamma} \gamma([0, d])$. The fact that it must be of the form in the proposition is clear by the symmetry of the sum and by the independence of the problem on the specific geodesic considered. In particular, we have that

$$A = \underset{t \in [0,d]}{\arg\min}\, \phi(t) + \phi(d - t).$$

Now that we have established the third point, we can see that, when ϕ is convex, the same is true for $\phi(t) + \phi(d-t)$ (which is the value of the loss function at any point of the form $\gamma(t)$). This is a symmetric function around $d/2$ and thus, in this case, $A = [a, d/2 - a]$ for some $a \in [0, d/2]$. Finally, if $\phi \in \Phi^+$, we have that A must be a singleton, implying part one. $\square$

We illustrate the situation by considering $d = 1$ and functions ϕ of the form $\phi : t \mapsto |t|^p$ for $p \in [0, \infty)$. In the following figure we plot the loss function evaluated on points of the form $\gamma(t)$ with $t \in [0,1]$. As we can see, when ϕ is strictly convex, we get $\mathfrak{m}^\phi = \gamma(1/2)$,

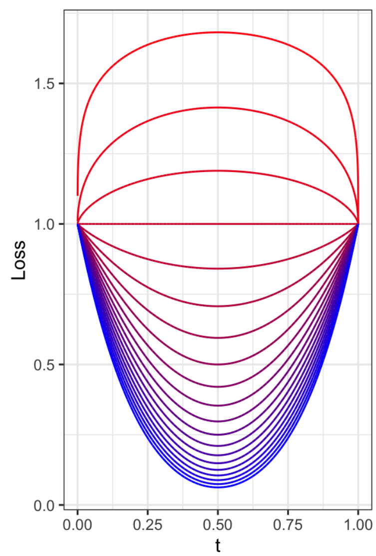

FIGURE 2.2: The loss function $F^\phi_{(x,y)}(\gamma(t))$ for $\phi : t \mapsto t^p$ and $p = (1, \ldots, 20)/4$. As can be seen, these losses are minimized at the midpoint of the interval only when $p \geqslant 1$.

when $\phi = \mathrm{id}$, we get $\mathfrak{m}^\phi = \gamma([0,1])$ and when ϕ is not convex, we get $\mathfrak{m}^\phi = \gamma(\{0,1\})$. Another pathology that we obtain when ϕ is not convex is sketched in the next subsection.

2.2.2 Distribution of maximal dispersion

Given a (sub)set $\mathcal{X}$, we may want to find the most ϕ-dispersed distribution, that is

$$\underset{\mu \in \mathcal{P}(\mathcal{X})}{\arg\max}\, \mathfrak{s}^\phi_\mu.$$

We consider the case of $\mathcal{X}$ being a compact interval $[a, b]$ (which is equivalent to considering any geodesically convex geodesic segment in complete, locally compact, length-metric space). In this case, it is natural to assume that the most dispersed distribution will be given by $(\delta_a + \delta_b)/2$, however, this is the case only when ϕ is convex.

Proposition 5. *Let $\phi \in \Phi$, then the only distribution of maximal ϕ-dispersion is $(\delta_a + \delta_b)/2$.*

Proof. Wlog, we can take $\mathcal{X} = [-1, 1]$. Recall that when ϕ is convex (and we are in a convex subset of a linear space), the loss function is convex too, thus $\mathfrak{m}_\mu^\phi$ will be a closed interval. If moreover $\phi \in \Phi^+$, this will be a point.

First of all, assume that we have some $\mu \in \arg\max_{\mu \in \mathcal{P}(\mathcal{X})} \mathfrak{s}_\mu^\phi$ which is not symmetric (meaning that $\mu(A) \neq \mu(-A)$ for some $A \in \mathcal{B}(\mathcal{X})$). We define two new probability measures, $\tilde{\mu}(\cdot) = \mu(-\cdot)$ and $\nu := (\mu + \tilde{\mu})/2$. Clearly, we have that ν is symmetric and, for any $x \in [-1, 1]$, we have that

$$
\begin{aligned}
F_\nu^\phi(x) &= \int_{-1}^{1} \phi(|y - x|) \frac{\mu + \tilde{\mu}}{2}(\rho y) \\
&= \int_{-1}^{1} \frac{\phi(|y - x|) + \phi(|y + x|)}{2} \mu(\rho y) \\
&\geqslant \int_{-1}^{1} \phi(|y|) \mu(\rho y) \\
&= F_\mu^\phi(0) \geqslant \mathfrak{s}_\mu^\phi
\end{aligned}
$$

but then,

$$
\mathfrak{s}_\nu^\phi = \underset{x \in [-1,1]}{\arg\min} F_\nu^\phi(x) \geqslant \mathfrak{s}_\mu^\phi
$$

Thus, if μ is a solution, ν is a solution too. Moreover, this also shows that if μ is a solution, it must have $0 \in \mathbb{E}_\phi[\mu]$, otherwise, ν would be strictly better than μ.

Therefore, any solution μ, must have $0 \in \mathbb{E}_\phi[\mu]$, but then

$$
\mathfrak{s}_\mu^\phi = \int_{-1}^{1} \phi(|y|) \mu(\rho y) \leqslant \phi(1).
$$

And the equality is obtained only if $\mu = p\delta_{-1} + (1-p)\delta_1$ for some $p \in [0,1]$. But of these solutions, only the one with $p = 1/2$ has $0 \in \mathfrak{m}_\mu^\phi$ and this concludes the proof. $\square$

Space of sub-exponential convex functions

For our main results we consider a particular class of functions in Φ for which the loss F_μ^ϕ satisfies continuity properties under mild hypotheses (see Lemmas 14 and 15).

Definition 6. *We define the class of subexponentially growing functions in Φ as*

$$\Phi_{ex} := \{\phi \in \Phi : \gamma_\phi < \infty\}, \text{ where } \gamma_\phi := \sup_{x \in \mathbb{R}_+} \frac{\phi(1+x)}{\phi(1)+\phi(x)}.$$

For all $\phi \in \Phi$, the quantity γ_ϕ is a measure of the growth rate of the function ϕ and Φ_{ex} is the collection of functions in Φ with finite growth. This is a large class of functions, including, for $p > 1$, $\phi : t \mapsto t^p$ and the exponential functions $\phi : t \mapsto p^t - 1$. In the appendix, we provide alternative characterizations of this class of functions and prove several properties employed in next sections.

We recall that by Φ we denote the collection of all bijections of $\mathbb{R}_+ := [0,\infty)$ that are convex. Thus, every function $\phi \in \Phi$ is strictly increasing and continuous. Moreover, we defined

$$\Phi_{ex} := \left\{\phi \in \Phi : \gamma_\phi := \sup_{x \in \mathbf{R}_+} \frac{\phi(x+1)}{\phi(x)+\phi(1)} < \infty\right\}.$$

Notice that for any $\phi \in \Phi$, the function $x \mapsto \phi(1+x)/(\phi(1)+\phi(x))$ is continuous. Thus, an equivalent characterization of this set is given by

$$\Phi_{ex} = \left\{\phi \in \Phi : \limsup_{x\to\infty} \frac{\phi(1+x)}{\phi(1)+\phi(x)} < \infty\right\} = \left\{\phi \in \Phi : \limsup_{x\to\infty} \frac{\phi(1+x)}{\phi(x)} < \infty\right\}$$

We also have that $\phi(1+x)/(\phi(1)+\phi(x))$ is equal to 1 at $x = 0$ so that $\gamma_\phi \geqslant 1$. It is easy to see that $\gamma_\phi = 1$ only when ϕ is linear, that is $\phi(x) = x\phi(1)$. Examples of functions in this class are given by $\phi : x \mapsto x^p$ for $p \geqslant 1$; in this case $\gamma_\phi = 2^{p-1}$. Another class is given by $\phi : x \mapsto \frac{p^x-1}{p-1}$ for $p \geqslant 1$ in this case $\gamma_\phi = p$.

Both Φ and Φ_{ex} are closed under multiplication, sum and positive multiples. In particular we have that, for every $c > 0$ and $\phi, \psi \in \Phi_{ex}$,

$$\gamma_{c\phi} = \gamma_{\phi}, \qquad \gamma_{\phi\psi} \leqslant 2\gamma_{\phi}\gamma_{\psi}, \qquad \gamma_{\phi+\psi} \leqslant \max(\gamma_{\phi}, \gamma_{\psi}),$$

The first equality is clear, as for the second,

$$\begin{aligned}
\gamma_{\phi\psi} &= \sup_{x\in\mathbb{R}_+} \frac{\phi(x+1)\psi(x+1)}{\phi(x)\psi(x)+\phi(1)\psi(1)} \\
&= \sup_{x\in\mathbb{R}_+} \frac{\phi(x+1)\psi(x+1)}{(\phi(x)+\phi(1))(\psi(x)+\psi(1))}\left(1+\frac{\phi(1)\psi(x)+\psi(1)\phi(x)}{\phi(x)\psi(x)+\phi(1)\psi(1)}\right) \\
&\leqslant \left(1+\sup_{x\in\mathbb{R}_+}\frac{\phi(1)\psi(x)+\psi(1)\phi(x)}{\phi(x)\psi(x)+\phi(1)\psi(1)}\right)\sup_{x\in\mathbb{R}_+}\frac{\phi(x+1)}{\phi(x)+\phi(1)}\sup_{x\in\mathbb{R}_+}\frac{\psi(x+1)}{\psi(x)+\psi(1)} \\
&\leqslant 2\gamma_{\phi}\gamma_{\psi}.
\end{aligned}$$

Using elementary calculus, it can be seen that $\frac{\phi(1)\psi(x)+\psi(1)\phi(x)}{\phi(x)\psi(x)+\phi(1)\psi(1)}$ is zero at $x = 0$, it increases up to 1 at $x = 1$ and then decreases back to 0 for $x \to \infty$. Finally, for the third inequality, we have that

$$\begin{aligned}
\gamma_{\phi+\psi} &= \sup_{x\in\mathbb{R}_+} \frac{\phi(x+1)+\psi(x+1)}{\phi(x)+\psi(x)+\phi(1)+\psi(1)} \\
&= \sup_{x\in\mathbb{R}_+} \frac{\phi(x+1)}{\phi(x)+\phi(1)}\frac{\phi(x)+\phi(1)}{\phi(x)+\phi(1)+\psi(x)+\psi(1)}+ \\
&\quad + \frac{\psi(x+1)}{\psi(x)+\psi(1)}\left(1-\frac{\phi(x)+\phi(1)}{\phi(x)+\phi(1)+\psi(x)+\psi(1)}\right) \\
&\leqslant \sup_{x\in\mathbb{R}_+} \frac{\phi(x)+\phi(1)}{\phi(x)+\phi(1)+\psi(x)+\psi(1)}\gamma_{\phi} \\
&\quad + \left(1-\frac{\phi(x)+\phi(1)}{\phi(x)+\phi(1)+\psi(x)+\psi(1)}\right)\gamma_{\psi} \\
&\leqslant \sup_{\alpha\in[0,1]} \alpha\gamma_{\phi}+(1-\alpha)\gamma_{\psi} = \max(\gamma_{\phi},\gamma_{\psi}).
\end{aligned}$$

We now want to remark that the choice of 1 in the definition of Φ_{ex} is arbitrary, in the sense that

Lemma 7. *For any $t > 0$ we can alternative characterize Φ_{ex} as*

$$\Phi_{ex} = \left\{ \phi \in \Phi : \gamma_\phi^{(t)} := \sup_{x \in \mathbb{R}_+} \frac{\phi(x+t)}{\phi(x) + \phi(t)} < \infty \right\} = \left\{ \phi \in \Phi : \limsup_{x \to \infty} \frac{\phi(x+t)}{\phi(x)} < \infty \right\}$$

Proof. We prove this by showing that if $s, t > 0$, then the conditions $\gamma_\phi^{(t)} < \infty$ and $\gamma_\phi^{(s)} < \infty$ are equivalent. Without loss of generality, we assume that $0 < s < t$. Let now $\phi \in \Phi$ with $\gamma_\phi^{(t)} < \infty$, then

$$\begin{aligned} \gamma_\phi^{(s)} &= \sup_{x \in \mathbb{R}_+} \frac{\phi(x+s)}{\phi(x) + \phi(s)} \\ &\leqslant \sup_{x \in \mathbb{R}_+} \frac{\phi(x+t)}{\phi(x) + \phi(t)} \frac{\phi(t) + \phi(x)}{\phi(s) + \phi(x)} \\ &\leqslant \frac{\phi(t)}{\phi(s)} \sup_{x \in \mathbb{R}_+} \frac{\phi(x+t)}{\phi(x) + \phi(t)} \leqslant \gamma_\phi^{(t)} \frac{\phi(t)}{\phi(s)} < \infty \end{aligned}$$

On the contrary, if $\phi \in \Phi$ has $\gamma_\phi^{(s)} < \infty$, then

$$\begin{aligned} \gamma_\phi^{(t)} &= \sup_{x \in \mathbb{R}_+} \frac{\phi(x+t)}{\phi(x) + \phi(t)} \leqslant \sup_{x \in \mathbb{R}_+} \frac{\phi(x + \lceil t/s \rceil s) + \phi(s)}{\phi(x) + \phi(s)} \\ &= \sup_{x \in \mathbb{R}_+} \prod_{n=1}^{\lceil t/s \rceil} \frac{\phi(x + ns) + \phi(s)}{\phi(x + (n-1)s) + \phi(s)} \leqslant \prod_{n=1}^{\lceil t/s \rceil} \sup_{x \in \mathbb{R}_+} \frac{\phi(x + ns) + \phi(s)}{\phi(x + (n-1)s) + \phi(s)} \\ &\leqslant \left(1 + \gamma_\phi^{(s)}\right)^{\lceil t/s \rceil} < \infty \end{aligned}$$

$\square$

The following lemma motivates the name of Φ_{ex}.

Lemma 8. *Functions in Φ_{ex} are dominated by exponentials, in particular, if $\phi \in \Phi_{ex}$ with $\gamma = \gamma_\phi > 1$, then*

$$\forall x \in \mathbb{R}_+, \qquad \phi(x) \leqslant \gamma \frac{\gamma^x - 1}{\gamma - 1} \phi(1).$$

Proof. Let $\phi \in \Phi_{ex}$ with $\gamma = \gamma_\phi > 1$, then for any $x \geqslant 0$, we prove by induction on $\lfloor x \rfloor$ that

$$\phi(x) \leqslant \phi(1) \sum_{j=1}^{\lfloor x \rfloor} \gamma^j + \gamma^{\lfloor x \rfloor} \phi(x - \lfloor x \rfloor)$$

Indeed, if $\lfloor x \rfloor = 0$, then $x \in [0, 1)$, and $\phi(x) = \phi(1)0 + 1\phi(x - \lfloor x \rfloor)$. So let now assume the formula to hold for all numbers with floor part equal to n, and that y is such that $\lfloor y \rfloor = n + 1$, then $x := y - 1$ is such that $\lfloor x \rfloor = n$ and $x - \lfloor x \rfloor = y - \lfloor y \rfloor$, thus

$$\begin{aligned}
\phi(y) &= \phi(x+1) \leqslant \gamma(\phi(1) + \phi(x)) \\
&\leqslant \gamma \left(\phi(1) + \phi(1) \sum_{j=1}^{\lfloor x \rfloor} \gamma^j + \gamma^{\lfloor x \rfloor} \phi(x - \lfloor x \rfloor) \right) \\
&= \phi(1) \sum_{j=1}^{\lfloor y \rfloor} \gamma^j + \gamma^{\lfloor y \rfloor} \phi(y - \lfloor y \rfloor)
\end{aligned}$$

By convexity, and by the fact that $\phi(0) = 0$, we have that

$$\phi(x) \leqslant \phi(1) \sum_{j=1}^{\lfloor x \rfloor} \gamma^j + \gamma^{\lfloor x \rfloor} \phi(x - \lfloor x \rfloor) \leqslant \gamma \frac{\gamma^{\lfloor x \rfloor} - 1}{\gamma - 1} \phi(1) + \gamma^{\lfloor x \rfloor}(x - \lfloor x \rfloor)\phi(1)$$

Now, for any $g > 1$ and $y \geqslant 0$, we have that $y \leqslant g \frac{g^y - 1}{g - 1}$ thus we conclude that

$$\begin{aligned}
\phi(x) - \frac{\gamma^x - 1}{\gamma - 1} \gamma \phi(1) &\leqslant \frac{\gamma \phi(1)}{\gamma - 1} (\gamma^{\lfloor x \rfloor} - \gamma^x) + \gamma^{\lfloor x \rfloor}(x - \lfloor x \rfloor)\phi(1) \\
&= \gamma^{\lfloor x \rfloor} \phi(1) \left(x - \lfloor x \rfloor - \gamma \frac{\gamma^{x - \lfloor x \rfloor} - 1}{\gamma - 1} \right) \leqslant 0
\end{aligned}$$

□

Also notice that if $\gamma_\phi = 1$, then $\phi(x) = x\phi(1) = \phi(1) \lim_{\gamma \downarrow 1} \gamma \frac{\gamma^x - 1}{\gamma - 1}$.

We prove some basic properties of functions in Φ_{ex} that we use extensively.

Lemma 9. *If $\phi \in \Phi_{ex}$ with $\gamma = \gamma_\phi$, then for any $h \in (0, 1]$,*

$$\phi(x+h) - \phi(x) \leqslant h(\gamma \phi(1) + (\gamma - 1)\phi(x))$$

Moreover, if a, b, c are positive numbers such that $a \leqslant b + c$, then

$$2\phi(a/2) \leqslant \phi(b) + \phi(c)$$

Proof. We have that $\phi(x+1) \leqslant \gamma(\phi(x) + \phi(1))$, so that $\phi(x+1) - \phi(x) \leqslant \gamma\phi(1) + (\gamma - 1)\phi(x)$. And now, because of convexity, for any $h \in (0, 1]$,

$$\phi(x+h) - \phi(x) \leqslant h(\phi(x+1) - \phi(x)) \leqslant h(\gamma\phi(1) + (\gamma - 1)\phi(x)).$$

We want to show that for any $a > 0$ and for any $\phi \in \Phi$, we have that for any $b, c \geqslant 0$ such that $a \leqslant b + c$, then $\phi(a) \leqslant \frac{\phi(a)}{2\phi(a/2)}(\phi(b) + \phi(c))$. It holds that

$$\min_{b+c \geqslant a}(\phi(b) + \phi(c)) = \min_{b+c \geqslant a, \min(b,c) \leqslant a}(\phi(b) + \phi(c)) = \min_{b \in [0,a]}(\phi(b) + \phi(a-b)) = 2\phi(a/2)$$

Then, we have that

$$\phi(a) = \frac{\phi(a)}{2\phi(a/2)} 2\phi(a/2) = \frac{\phi(a)}{2\phi(a/2)} \min_{b+c \geqslant a}(\phi(b) + \phi(c)) \leqslant \frac{\phi(a)}{2\phi(a/2)}(\phi(b) + \phi(c))$$

Notice that, because of convexity and because $\phi(0) = 0$, we must have that $\frac{\phi(a)}{2\phi(a/2)} \geqslant 1$. Now by simplifying, we get our desired result. □

Also, notice that this result implies that

$$\limsup_{x \to \infty} \frac{\phi'_+(x)}{\phi(x)} \leqslant \limsup_{x \to \infty} \frac{(\gamma_\phi - 1)\phi(x) + \gamma_\phi\phi(1)}{\phi(x)} = \gamma_\phi - 1.$$

3. Consistency

In this section we study one essential aspect of the sample ϕ-mean as an estimator for the population ϕ-mean. First we consider this problem for a given function $\phi \in \Phi_{ex}$ and then, we consider the joint behavior of ϕ-means when ϕ belongs to a subcollection $\Psi \subseteq \Phi_{ex}$. We now introduce the probabilistic framework.

3.1 Probabilistic Framework

In this section $(\mathcal{X}, \rho)$ will always be a metric space with the Heine-Borel property and it will be equipped with its Borel σ-algebra Σ_ρ. All probability measures will be defined on this σ-algebra.

We consider *iid* samples of points from a probability measure μ, and employ a notion of sample ϕ-mean. Starting from a random variable $X : (\Omega, \Sigma) \to (\mathcal{X}, \Sigma_\rho)$ with probability law μ, we consider an infinite sample of *iid* observations distributed according to μ, so that $(X_j)_{j\geqslant 1}$ is a random variable from $(\boldsymbol{\Omega} := \Omega^\infty, \boldsymbol{\Sigma} := \Sigma^{\otimes\infty})$ to $(\mathcal{X}^\infty, \Sigma_\rho^{\otimes\infty})$ with law given by $\boldsymbol{\mu} := \mu^{\otimes\infty}$. Given $\boldsymbol{\omega} \in \boldsymbol{\Omega}$ and $n \geqslant 0$, the empirical measure is denoted by

$$\hat{\mu}_n : \boldsymbol{\omega} \mapsto \frac{1}{n}\sum_{j=1}^{n} \delta_{X_j(\boldsymbol{\omega})}.$$

Now we define the empirical loss function and the corresponding empirical ϕ-mean.

Definition 10. *Let $\boldsymbol{\omega} \in \boldsymbol{\Omega}$, $n \geqslant 0$, and $(X_j)_{j\geqslant 1}$ iid according to $\mu \in \mathcal{P}(\mathcal{X})$,*

$$F_n^\phi(\boldsymbol{\omega}) = F_{\hat{\mu}_n(\boldsymbol{\omega})}^\phi : x \mapsto \frac{1}{n}\sum_{j=1}^{n} \phi(\rho(X_j(\boldsymbol{\omega}), x)), \qquad \mathfrak{m}_n^\phi(\boldsymbol{\omega}) = \mathfrak{m}_{\hat{\mu}_n(\boldsymbol{\omega})}^\phi.$$

For the sake of simplicity, we avoid any reference to $\boldsymbol{\Omega}$ and we write F_n^ϕ and $\mathfrak{m}_n^\phi$. Finally, by **almost surely**, we always mean $\boldsymbol{\mu}$-almost surely.

In this section, we state our two main results concerning the consistency of the sample ϕ-means as estimators for the population ϕ-means. Since $\mathfrak{m}^\phi$ and $\mathfrak{m}_n^\phi$ may not be singletons, we now introduce a notion of distance between sets, called the one-sided Hausdorff distance.

Definition 11. *For all compact sets $A, B \subseteq \mathcal{X}$ we define*

$$\rho_\infty(A,B) := \sup_{a\in A} \inf_{b\in B} \rho(a,b)$$

It is easy to see that $\rho_\infty(A,B) < \varepsilon$ iff A is contained in an ϵ-thickening[1] of B as $\rho_\infty(A,B) = \sup_{a\in A} \rho(a,B)$.

3.2 Main Results

We now state our first consistency result, which provides conditions under which, given a specific $\phi \in \Phi_{ex}$, the set $\mathfrak{m}_n^\phi$ converges almost surely to $\mathfrak{m}^\phi$ in the sense of one-sided Hausdorff distance.

Theorem 12. *Let $(\mathcal{X}, \rho)$ be a metric space with the Heine-Borel property, $\phi \in \Phi_{ex}$, and $\mu \in \mathcal{P}(\mathcal{X})$ be such that F_μ^ϕ is finite somewhere. Then, as $n \to \infty$,*

$$\rho_\infty\left(\mathfrak{m}_n^\phi, \mathfrak{m}^\phi\right) \overset{\mu-\text{as}}{\to} 0$$

This implies that, if ξ_n is any measurable selection from $\mathfrak{m}_n^\phi$, then under the hypotheses of the theorem, $\rho(\xi_n, \mathfrak{m}^\phi) \overset{as}{\to} 0$, highlighting the applicability of the result.

Next, we strengthen the previous result, under a few additional assumptions, to prove uniform consistency. In particular, we study the collective behaviour of classes of sample ϕ-means and their uniform (almost sure) convergence to the respective population ϕ-means. Before stating the result, we need to make some further notations.

A collection $\Psi \subseteq \Phi_{ex}$ is called

- **dominated from above**, if $\sup_{\phi\in\Psi} \gamma_\phi < \infty$ and $\sup_{\phi\in\Psi} \phi(1) < \infty$. According to Lemma 8 this implies that there is some $\psi \in \Psi_{ex}$ such that, for all $\phi \in \Psi$ and $t \geqslant 0$, $\phi(t) < \psi(t)$,
- **compact** if for any $k > 0$, the collection $\{\phi|_{[0,k]} : \phi \in \Psi\}$ is compact under the sup norm on $C([0,k])$.

[1] That is the collection of points at distance less than ε from B.

An example is given by $\{\phi : t \mapsto t^p : p \in [1, P]\}$ for any fixed $P > 1$.

Theorem 13. *Let $(\mathcal{X}, \rho)$ be a metric space with the Heine-Borel property, let $\mu \in \mathcal{P}(\mathcal{X})$ and $\Psi \subseteq \Phi_{ex}$ be compact, dominated from above with $\psi \in \Phi_{ex}$ be such that $\phi(t) < \psi(x)$ for all $\phi \in \Psi$, and such that F_μ^ϕ is finite with a unique minimizer for every $\phi \in \Psi$. Then as $n \to \infty$,*

$$\sup_{\phi \in \Psi} \rho_\infty(\mathfrak{m}_n^\phi, \mathfrak{m}^\phi) \xrightarrow{as} 0.$$

Notice that under the hypothesis that $\mathfrak{m}^\phi$ is a singleton, we have that $\rho_\infty(\mathfrak{m}_n^\phi, \mathfrak{m}^\phi) = \rho(\mathfrak{m}_n^\phi, \mathfrak{m}^\phi)$. Theorem 31 provides a large collection of examples where the hypotheses of Theorem 13 are satisfied. This theorem hints at an analogy between certain collections of sample ϕ-means and Glivenko-Cantelli classes.

3.3 Proofs of Theorems 12 and 13

In order to prove these theorems, we need results concerning finiteness, continuity, and compactness.

3.3.1 Finiteness and Continuity

We now study the properties of ϕ-means, when $\phi \in \Phi_{ex}$. First of all, it is important to determine whether and where the loss function F_μ^ϕ is finite. This depends on the space $\mathcal{X}$, the probability measure μ and the function ϕ. It is immediate to see that if $\mathcal{X}$ is compact, then F_μ^ϕ is always finite. Thus, we focus on the case in which $\mathcal{X}$ is not compact. In this case, we prove that if $\phi \in \Phi_{ex}$, then F_μ^ϕ is either infinite everywhere or finite everywhere. Notably, a converse is also provided.

Lemma 14. *Let $\mathcal{X}$ be a metric space with the Heine Borel property. If $\mu \in \mathcal{P}(\mathcal{X})$ and $\phi \in \Phi_{ex}$, then F_μ^ϕ is finite somewhere iff it is finite everywhere. Conversely, if $\mathcal{X}$ is an unbounded, complete, locally compact, length metric space, $\phi \in \Phi$ and for all $\mu \in \mathcal{P}(\mathcal{X})$, F_μ^ϕ is finite everywhere iff it is finite somewhere, then $\phi \in \Phi_{ex}$.*

Proof. If $\phi \in \Phi_{ex}$, let $p, q \in \mathcal{X}$ be such that $F_\mu^\phi(p) < \infty$ and $\rho(p,q) \leqslant 1$, then, by Lemma 9,

$$F_\mu^\phi(q) = \int \phi(\rho(x,q))\mu(\mathrm{d}x) \leqslant \int \phi(\rho(x,p) + \rho(p,q))\mu(\mathrm{d}x)$$

$$\leqslant \gamma\phi(1) + \gamma \int \phi(\rho(x,p))\mu(\mathrm{d}x) = \gamma\phi(1) + \gamma F_\mu^\phi(p) < \infty$$

By induction, we obtain finiteness on the entire space. Of course, if $\mathcal{X}$ is compact or μ has compact support, then F_μ^ϕ will always be finite everywhere for any $\phi \in \Phi$ and indeed in this case any $\phi \in \Phi$ can be replaced by an equivalent [2] $\psi \in \Phi_{ex}$ in such a way that $\phi(x) = \psi(x)$ for every $x \in [0, \mathrm{diam}(\mathcal{X})]$.

If $\mathcal{X}$ is unbounded and $\phi \in \Phi \backslash \Phi_{ex}$, then we can find a probability measure $\mu \in \mathcal{P}(\mathcal{X})$ such that F_μ^ϕ is finite only on a nontrivial subset of $\mathcal{X}$. Indeed, it is enough to prove this claim on an unbounded interval $[-7, \infty)$, since any unbounded, complete, locally compact, length metric space has the Heine-Borel property by the Hopf–Rinow theorem (see Gromov, 2007 Section 1B) and it contains a subset isometric to $[-7, \infty)$. We then consider the following probability measure

$$\mu(\mathrm{d}x) := \frac{g(x)\mathbf{1}_{[0,\infty)}(x)}{1 + \phi(x)} \frac{\mathrm{d}x}{c},$$

$$g(x) := \sum_{n=1}^{\infty} \frac{\mathbf{1}_{A_n}(x)}{n^2 \lambda(A_n)},$$

$$A_n := \left\{ x : \frac{\phi(1+x)}{\phi(1) + \phi(x)} \in [n, n+1) \right\}$$

where $c = \int_0^\infty \frac{g(x)\mathrm{d}x}{1+\phi(x)} \in (0, \infty)$ is the integration constant and λ is the Lebesgue measure. It is now possible to see that F_μ^ϕ is finite in some points but not in others, indeed

$$F_\mu^\phi(0) = \int_0^\infty \frac{\phi(x)}{\phi(1) + \phi(x)} \frac{g(x)}{c} \mathrm{d}x \leqslant \frac{1}{c} \int_0^\infty g(x)\mathrm{d}x = \frac{1}{c} \sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6c} < \infty,$$

$$F_\mu^\phi(-1) = \int_0^\infty \frac{\phi(x+1)}{\phi(1) + \phi(x)} \frac{g(x)}{c} \mathrm{d}x \geqslant \sum_{n=1}^{\infty} \int_0^\infty \frac{\mathbf{1}_{A_n}(x)}{n\lambda(A_n)} \frac{\mathrm{d}x}{c} \geqslant \frac{1}{c} \sum_{n=1}^{\infty} \frac{1}{n} = \infty.$$

[2] that is such that for any probability measure μ, $F_\mu^\phi = F_\mu^\psi$

This concludes the proof since $\mathcal{X}$ is trivially closed in itself. $\square$

Notice that for certain measures μ there may not exist any $\phi \in \Phi_{ex}$ for which F_μ^ϕ is finite. The typical example would be to consider $\mathcal{X} = \mathbf{R}$ and μ a Cauchy probability measure, that is $\mu(\mathrm{d}x) = \frac{1}{1+x^2}\frac{\mathrm{d}x}{\pi}$.

Now we study the continuity properties of the loss function.

Lemma 15. *Let $\mathcal{X}$ be a metric space with the Heine Borel property and $\mu \in \mathcal{P}(\mathcal{X})$. For all $\phi \in \Phi_{ex}$, if F_μ^ϕ is finite somewhere, then it is finite and continuous everywhere. In particular, F_μ^ϕ is locally Lipschitz continuous.*

Proof. Global finiteness of F_μ^ϕ follows from Lemma 14. Assume $p, q \in \mathcal{X}$ be such that $\rho(p,q) \leqslant 1$. By triangular inequality, for any $x \in \mathcal{X}$, $|\rho(p,x) - \rho(x,q)| \leqslant \rho(p,q) \leqslant 1$. Thus, by Lemma 9,

$$
\begin{aligned}
|F_\mu^\phi(p) - F_\mu^\phi(q)| &\leqslant \int |\phi(\rho(p,x)) - \phi(\rho(x,q))|\mu(\mathrm{d}x) \\
&\leqslant \int \min_{u \in \{p,q\}} (\phi(\rho(x,u) + \rho(p,q)) - \phi(\rho(x,u)))\mu(\mathrm{d}x) \\
&\leqslant \rho(p,q)\left(\gamma_\phi \phi(1) + (\gamma_\phi - 1)\min_{u \in p,q} \int \phi(d(x,u))\mu(\mathrm{d}x)\right) \\
&\leqslant \rho(p,q)\left(\gamma_\phi \phi(1) + (\gamma_\phi - 1)\min_{u \in \{p,q\}} F_\mu^\phi(u)\right)
\end{aligned}
$$

This shows that on any ball $B_{1/2}(p)$, the loss function F_μ^ϕ is Lipschitz continuous with constant smaller or equal than

$$
\gamma_\phi \phi(1) + (\gamma_\phi - 1) \sup_{q \in B_{1/2}(p)} F_\mu^\phi(q) < \infty.
$$

$\square$

Notice that when $\gamma_\phi = 1$ (that is when $\phi(x) = x\phi(1)$), F_μ^ϕ is globally Lipschitz continuous with constant bounded by $\phi(1)$. Another case in which F_μ^ϕ is globally Lipschitz continuous is, of course, when $\mathcal{X}$ is compact. However, in general, it is not. For instance, if $\mathcal{X} = \mathbf{R}$ and μ is compactly supported, then it is easy to verify that F_μ^ϕ is globally Lipschitz iff ϕ is globally Lipschitz.

With Lipschitz continuity, we can prove the compactness of the ϕ-mean set.

Theorem 16. *Let $\mathcal{X}$ be a metric space with the Heine Borel property, $\mu \in \mathcal{P}(\mathcal{X})$, and $\phi \in \Phi_{ex}$. If F_μ^ϕ is finite somewhere, then $\mathfrak{m}_\mu^\phi$ is a non-empty compact set with a diameter bounded by*

$$\operatorname{diam}(\mathfrak{m}_\mu^\phi) \leqslant 2\phi^{-1}\left(\inf_{x\in\mathcal{X}} F_\mu^\phi(x)\right).$$

Proof. Since $\mathcal{X}$ has the Heine-Borel property, we just need to show that $\mathfrak{m}^\phi(\mu)$ is closed and bounded. Closedness is a consequence of the continuity of F_μ^ϕ which is in turn guaranteed by Lemma 14. Boundedness is clear if $\mathcal{X}$ is compact, so we prove boundedness assuming $\mathcal{X}$ to be unbounded. Let $p \in \mathcal{X}$, we know that $F_\mu^\phi(p)$ is finite (and positive), there is an $r_p > 0$ such that

$$\int_{B_{r_p}(p)} \phi(\rho(x,p))\mu(\mathrm{d}x) \geqslant \frac{1}{2}F_\mu^\phi(p).$$

But now, if $q \in \mathcal{X}$ is such that $\rho(p,q) \geqslant kr_p$, and $k \geqslant 3$, then

$$\begin{aligned} F_\mu^\phi(q) &\geqslant \int_{B_{r_p}(p)} \phi(\rho(q,x))\mu(\mathrm{d}x) \geqslant \int_{B_{r_p}(p)} \phi((k-1)\rho(p,q))\mu(\mathrm{d}x) \\ &\overset{*}{\geqslant} (k-1)\int_{B_{r_p}(p)} \phi(\rho(p,q))\mu(\mathrm{d}x) \geqslant (k-1)\int_{B_{r_p}(p)} \phi(k\rho(p,x))\mu(\mathrm{d}x) \\ &\overset{*}{\geqslant} k(k-1)\int_{B_{r_p}(p)} \phi(\rho(x,p))\mu(\mathrm{d}x) = \frac{k(k-1)}{2}F_\mu^\phi(p) \geqslant 3F_\mu^\phi(p). \end{aligned}$$

Where (*) hold because if ϕ is convex, with $\phi(0) = 0$ and $c > 1$, then $\phi(cx) \geqslant c\phi(x)$ for every $x \geqslant 0$. This shows that (for any $p \in \mathcal{X}$) $\mathfrak{m}^\phi(\mu) \subseteq B_{3r_p}(p)$ so we obtain boundedness. If $\mathfrak{m}^\phi(\mu)$ is a singleton, then the result is obvious, so let $y, z \in \mathfrak{m}^\phi(\mu)$ be distinct. We consider the triangle inequality with a generic point $x \in \mathcal{X}$, $\rho(y,z) \leqslant \rho(y,x) + \rho(x,z)$, but then from Lemma 9 we have that

$$2\phi(\rho(y,z)/2) \leqslant \phi(\rho(y,x)) + \phi(\rho(z,x))$$

And now, integrating by $\mu(\mathrm{d}x)$, we conclude that

$$2\phi(\rho(y,z)/2) \leqslant F_\mu^\phi(y) + F_\mu^\phi(z) = 2 \inf_{x\in\mathcal{X}} F_\mu^\phi(x).$$

$\square$

This proof also shows that if $\mathcal{X}$ is unbounded, then F_μ^ϕ is unbounded too, and in particular, for any $p \in \mathcal{X}$ and $m > 0$, we can find a radius $s_{p,m}$ such that

$$\forall q \notin B_{s_{p,m}}(p), \quad F_\mu^\phi(q) \geqslant m.$$

Notice that $\inf_{x\in\mathcal{X}} F_\mu^\phi(x)$ is zero iff μ is a Dirac distribution. In this case Lemma 16 yields that $\mathfrak{m}_{\delta_{\hat{x}}}^\phi = \{\hat{x}\}$.

We now show with an example that this bound can be tight. Let $\mathcal{X} = S^1$, $\phi = \mathrm{Id}$ and μ be the normalized Haar measure. In this case we have that F_μ^ϕ is constant and equal to

$$F_\mu^\phi(p) = \int_{-\pi}^{\pi} \phi(|x|)\frac{\mathrm{d}x}{2\pi} = \frac{1}{\pi}\int_0^\pi x\mathrm{d}x = \frac{\pi}{2}$$

Thus, in this case, $\mathfrak{m}_\mu^\phi = \mathcal{X}$ and $\mathrm{diam}(\mathfrak{m}_\mu^\phi) = \mathrm{diam}(S^1) = \pi = 2\phi^{-1}(\inf_{p\in\mathcal{X}} F_\mu^\phi(p))$.

Also notice that Lemma 16 implies that, whenever μ has compact support,

$$\mathrm{diam}(\mathfrak{m}_\mu^\phi) \leqslant \mathrm{diam}(\mathrm{supp}(\mu)).$$

3.3.2 Proof of Consistency

The proof of Theorem 12 is based on the following two lemmas.

Lemma 17. *For any compact set $K \subseteq \mathcal{X}$, as $n \to \infty$,*

$$\sup_{x\in K} \left| F_n^\phi(x) - F^\phi(x) \right| \overset{\mu-\mathrm{as}}{\to} 0$$

Proof. Let $\gamma := \gamma_\phi$. Fix an arbitrary $\varepsilon > 0$ and $k := \lceil \mathrm{diam}(K) \rceil < \infty$. For any $z \in K$,

$$\sup_{x\in K} F_n^\phi(x) = \sup_{x\in K} \frac{1}{n}\sum_{j=1}^n \phi(\rho(X_j, x)) \leqslant \sup_{x\in K} \frac{1}{n}\sum_{j=1}^n \phi(\rho(X_j, z) + \rho(z, x))$$

$$\leqslant \frac{1}{n}\sum_{j=1}^n \phi(\rho(X_j, z) + k) \overset{\text{Lemma 9}}{\leqslant} \gamma^k \frac{1}{n}\sum_{j=1}^n \phi(\rho(X_j, z)) = \gamma^k F_n^\phi(z)$$

As $\mathbb{E}[F_n^\phi(z)] = F^\phi(z) < \infty$ is finite by Lemma 14, we can apply the SLLN to find an almost surely finite random integer N such that for any $n \geqslant N$, $F_n^\phi(z) \leqslant F^\phi(z) + 1$. Therefore,

$$\forall n \geqslant N, \quad \sup_{x\in K} F_n^\phi(x) \overset{\mu-\text{as}}{\leqslant} \gamma^k \left(F^\phi(z) + 1\right) =: \alpha_K$$

Moreover, for any $\delta_1 \leqslant 1$ and $n \geqslant N$,

$$\sup_{x,y\in K:\rho(x,y)\leqslant\delta_1} \left|F_n^\phi(x) - F_n^\phi(y)\right|$$

$$\leqslant \sup_{x,y\in K:\rho(x,y)\leqslant\delta_1} \frac{1}{n}\sum_{j=1}^n \left|\phi(\rho(x, X_j)) - \phi(\rho(y, X_j))\right|$$

$$\leqslant \sup_{x,y\in K:\rho(x,y)\leqslant\delta_1} \frac{1}{n}\sum_{j=1}^n |\rho(x, X_j) - \rho(y, X_j)|(\gamma\phi(1) + (\gamma-1)\phi(\rho(x, X_j)))$$

$$\leqslant \delta_1 \left(\gamma\phi(1) + (\gamma-1)\sup_{x\in K}\frac{1}{n}\sum_{j=1}^n \phi(\rho(x, X_j))\right)$$

$$\overset{\text{as}}{\leqslant} \delta_1 \left(\gamma\phi(1) + (\gamma-1)\alpha_K\right).$$

Where we applied 9 in the third inequality.

Thus by taking δ_1 sufficiently small, we get that $|F_n^\phi(x) - F_n^\phi(y)| \overset{as}{\leqslant} \varepsilon/3$ for any $x, y \in K$ with $\rho(x, y) \leqslant \delta_1$.

Furthermore, by Lemma 14, F_μ^ϕ is continuous, thus we can find some $\delta_2 > 0$ such that $|F_\mu^\phi(x) - F_\mu^\phi(y)| \leqslant \varepsilon/3$ for any $x, y \in K$ with $\rho(x, y) \leqslant \delta_2$. Now let $\delta = \min(\delta_1, \delta_2)$, by compactness of K, we can find a finite number of points $(z_i)_{i=1}^m$ such that $K \subseteq \bigcup_{i=1}^m B_\delta(z_i)$.

For every $i \in \{1, \ldots, m\}$, the SLLN guarantees the existence of an almost surely finite variable N_i such that

$$\forall n \geqslant N_i, \quad |F_n^\phi(z_i) - F^\phi(z_i)| \overset{\text{as}}{\leqslant} \varepsilon/3$$

Now, by finiteness of m, the random variable $N' := \max(N, \max_{i\in\{1,\ldots,m\}} N_i)$, is almost surely finite, and for all $n \geqslant N'$,

$$\begin{aligned}\sup_{x\in K} |F_n^\phi(x) - F^\phi(x)| \leqslant & \sup_{x\in K} |F_n^\phi(x) - F_n^\phi(z_{j_x})| + \\ & + \sup_{x\in K} |F_h^\phi(z_{j_x}) - F^\phi(z_{j_x})| + \\ & + \sup_{x\in K} |F^\phi(z_{j_x}) - F^\phi(x)| \\ \leqslant & \varepsilon/3 + \varepsilon/3 + \varepsilon/3 = \varepsilon.\end{aligned}$$

where we denoted $j_x := \min\{j \in \{1, \ldots, m\} : \rho(z_j, x) < \delta\}$. □

Lemma 18. *Let $m = \inf_{x\in K} F^\phi(x)$. There is an almost surely finite random variable N and a compact set A including $\mathfrak{m}^\phi$ such that for all $n \geqslant N$,*

$$\inf_{y\notin A} \min\left(F^\phi(y), F_n^\phi(y)\right) \geqslant m + 1$$

Proof. If $\mathcal{X}$ is compact, then we can just take $A = \mathcal{X}$ and we are done. Instead, assume $\mathcal{X}$ to be unbounded, then we can find a compact set B such that $\int_B \mu(\mathrm{d}x) > 3/4$ and then we consider $A_k = \{x \in \mathcal{X} : \rho(x, B) \leqslant k\}$. This is compact, and moreover,

$$\forall y \notin A_k, \quad F_\mu^\phi(y) \geqslant \int_B \phi(\rho(y, x))\mu(\mathrm{d}x) \geqslant \phi(k)\mu(B) = \frac{3}{4}\phi(k)$$

And since $\lim_{k\to\infty} \phi(k) = \infty$, we have that this quantity is larger than $m + 1$ for some k. As for the sample case, we have that

$$\forall y \notin A_k, \quad F_n^\phi(y) \geqslant \sum_{j:X_j\in B} \phi(\rho(y, X_j)) \geqslant \frac{|\{j : X_j \in B\}|}{n}\phi(k)$$

But by the SLLN, we have that $|\{j : X_j \in B\}|/n$ converges almost surely to $\mu(B) = 3/4$ so that, there is an almost sure finite random variable N such that if $n \geqslant N$, then $|\{j : X_j \in B\}|/n > 1/2$. Thus, we can take $k^* := \inf\{k : \phi(k)/2 > m+1\}$ and define $A := A_{k^*}$ and we are done. Indeed, it is easy to see that $\mathfrak{m}^\phi \subseteq A$, because, by definition of $\mathfrak{m}^\phi$, we have that $F^\phi(\mathfrak{m}^\phi) = m < m+1$. $\square$

Now we are ready to proceed with the proof of Theorem 12.

Proof of Theorem 12. Let $m = \inf_{x\in\mathcal{X}} F^\phi(x)$, and let $\varepsilon > 0$, we need to prove that there is a constant $\delta > 0$ and an almost-surely finite random variable N such that for any $n \geqslant N$,

$$\sup_{x\in\mathfrak{m}^\phi} F_n^\phi(x) \leqslant m+\delta, \qquad \inf_{x:\rho(x,\mathfrak{m}^\phi)>\varepsilon} F_n^\phi(x) \geqslant m+3\delta.$$

To this end, let A and $N_3 := N$ be as in the proof of Lemma 18, and define

$$K := \{x \in A : \rho(x,\mathfrak{m}^\phi) \geqslant \varepsilon\}.$$

This is compact and $m' := \min\{F^\phi(x) : x \in K\} > m$. Then set $\delta := (m'-m)/4$ so that $m' > m+4\delta$. Now we can apply the Lemma 17 to the compact sets $\mathfrak{m}^\phi$ and K and find almost surely finite random variables N_1, N_2 so that if $n > \max(N_1, N_2, N_3)$, we have

$$\begin{aligned}
&\forall x \in \mathfrak{m}^\phi, && F_n^\phi(x) \leqslant m+\delta\\
&\forall x \in K, && F_n^\phi(x) \geqslant m'-\delta = m+3\delta\\
&\forall x \notin A, && F_n^\phi(x) \geqslant m+1
\end{aligned}$$

This forces the sample ϕ-mean set $\mathfrak{m}_n^\phi$ to be contained in an ϵ-engrossing of $\mathfrak{m}^\phi$. $\square$

3.3.3 Proof of Uniform Consistency

The proof of this result, despite being more technical, follows the same steps of the proof of Theorem 12 and uses the following lemmas. For the relevant concepts and tools of this section see Chapter 4 of Wainwright, 2019, Chapter 7 of Vaart and Wellner, 1996 and Chapter 3 Sen, 2022.

Lemma 19. *Let $\Psi \subseteq \Phi_{ex}$ be dominated, $\mu \in \mathcal{P}(\mathcal{X})$ be such that F^{ϕ}_{μ} is finite for all $\phi \in \Psi$, then as $n \to \infty$, for any compact set K,*

$$\sup_{\phi \in \Psi} \sup_{x \in K} |F^{\phi}_{n}(x) - F^{\phi}(x)| \overset{\mu-\text{as}}{\to} 0.$$

Proof. Under the hypotheses and by Lemma 8 there is some $\psi \in \Phi_{ex}$ such that

$$\sup_{\phi \in \Psi} \phi(x) \leqslant \psi(x)$$

for every $x \geqslant 0$. Thus, for any $z \in \mathcal{X}$, $\sup_{\phi \in \Psi} F^{\phi}(z) \leqslant F^{\psi}(z)$. Now we fix an arbitrary $\varepsilon > 0$.

Firstly, by inspecting the proof of Lemma 2.3, we see that, for any $\phi \in \Psi$, the Lipschitz constants of $F^{\phi}_{\mu}|_{K}$ are bounded by a common constant not depending on ϕ. Thus, for any $\varepsilon > 0$ there is a $\delta_1 > 0$ such that

$$\sup_{\psi \in \Psi} \sup_{x,y \in K : \rho(x,y) < \delta_1} |F^{\phi}(x) - F^{\phi}(y)| < \varepsilon/3$$

Secondly, let $\gamma := \sup_{\phi \in \Psi} \gamma_{\phi}$, $z \in K$ be an arbitrary point and $k := \lceil \mathrm{diam}(K) \rceil$, then

$$\begin{aligned}
\sup_{\phi \in \Psi} \sup_{x \in K} F^{\phi}_{n}(x) &\leqslant \sup_{x \in K} F^{\psi}_{n}(x) = \sup_{x \in K} \frac{1}{n} \sum_{j=1}^{n} \psi(\rho(x, X_j)) \\
&\leqslant \sup_{x \in K} \frac{1}{n} \sum_{j=1}^{n} \psi(\rho(x,z) + \rho(z, X_j)) \leqslant \frac{1}{n} \sum_{j=1}^{n} \psi(k + \rho(z, X_j)) \\
&\leqslant \gamma^{k} \frac{1}{n} \sum_{j=1}^{n} \psi(k + \rho(z, X_j)) = \gamma^{k} F^{\psi}_{n}(z)
\end{aligned}$$

Where we applied 9 in the last inequality. But $\mathbb{E}[F^{\psi}_{n}(z)] = F^{\psi}(z) < \infty$, thus by SLLN, there is an almost surely finite integer-valued random variable N such that if $n \geqslant N$, $F^{\psi}_{n}(z) \leqslant F^{\psi}(z) + 1 < \infty$ and thus, if $n > N$, we also have that

$$\sup_{\phi \in \Psi} \sup_{x \in K} F^{\phi}_{n}(x) < \gamma^{k}(F^{\psi}(z) + 1) =: \alpha_K < \infty.$$

Now, if δ is sufficiently small, we have that

$$\begin{aligned}
&\sup_{\phi\in\Psi}\sup_{x,y\in K:\rho(x,y)\leqslant\delta}|F_n^{\phi}(x)-F_n^{\phi}(y)| \\
&\leqslant \sup_{\phi\in\Psi}\sup_{x,y\in K:\rho(x,y)\leqslant\delta}\frac{1}{n}\sum_{j=1}^{n}|\phi(\rho(x,X_j))-\phi(\rho(y,X_j))| \\
&\leqslant \sup_{\phi\in\Psi}\sup_{x,y\in K:\rho(x,y)\leqslant\delta}\frac{1}{n}\sum_{j=1}^{n}|\rho(x,X_j)-\rho(y,X_j)|\left(\gamma_{\phi}\phi(1)+(\gamma_{\phi}-1)\phi(\rho(y,X_j))\right) \\
&\leqslant \delta\left(\gamma\psi(1)+(\gamma-1)\sup_{x\in K}\frac{1}{n}\sum_{j=1}^{n}\phi(\rho(x,X_j))\right) \\
&\overset{\text{if } n>N}{\leqslant} \delta\left(\gamma\phi(1)+(\gamma-1)\alpha_K\right).
\end{aligned}$$

Where we applied 9 in the second inequality. Thus, we can find some δ_2 such that for any $n>N$,

$$\sup_{\phi\in\Psi}\sup_{x,y\in K:\rho(x,y)\leqslant\delta_2}|F_n^{\phi}(x)-F_n^{\phi}(y)|<\varepsilon/3.$$

Thirdly, let $\delta_3=\min(\delta_1,\delta_2)$, by compactness of K, we can find finitely many balls $(B_{\delta_3}(z_i))_{i\in\{1,\dots,m\}}$ so that $K\subseteq\bigcup_{i=1}^{m}B_{\delta_3}(z_i)$. For each z_i and $\phi\in\Psi$ we have that $\mathbb{E}[F_n^{\phi}(z_i)]=F^{\phi}(z_i)$, thus by SLLN, $|F_n^{\psi}(z_i)-F^{\psi}(z_i)|\overset{\text{as}}{\to}0$. However we need now to show the stronger statement that at each z_i, $\sup_{\phi\in\Psi}|F_n^{\phi}(z_i)-F^{\phi}(z_i)|\overset{\text{as}}{\to}0$. This is the same as asking the collection of functions $\mathcal{F}:=\{f_{\phi}=\phi\circ\rho(z_i,\cdot):\mathcal{X}\to\mathbf{R}_+:\phi\in\Psi\}$ to have the Glivenko-Cantelli property. This is the case because of Lemma 20 below. Thus, for any $i\in\{1,\dots,m\}$ there is an N_i almost surely finite such that for any $n>N_i$, $\sup_{\phi\in\Psi}|F_n^{\phi}(z_i)-F^{\phi}(z_i)|<\varepsilon/3$ and, if $n>N':=\max_{i\in\{1,\dots,m\}}(N_i)\overset{as}{<}\infty$,

$$\max_{i\in\{1,\dots,m\}}\sup_{\phi\in\Psi}|F_n^{\phi}(z_i)-F^{\phi}(z_i)|\leqslant\varepsilon/3.$$

Finally, if $n > \max(N, N')$, by combining the previous three results, we conclude that

$$\begin{aligned}\sup_{\phi\in\Psi}\sup_{x\in K}|F_n^{\phi}(x) - F^{\phi}(x)| \leqslant &\sup_{\phi\in\Psi}\sup_{x\in K}|F_n^{\phi}(x) - F_n^{\phi}(z_{j_x})| \\ &+ \sup_{\phi\in\Psi}\sup_{x\in K}|F_n^{\phi}(z_{j_x}) - F^{\phi}(z_{j_x})| \\ &+ \sup_{\phi\in\Psi}\sup_{x\in K}|F^{\phi}(z_{j_x}) - F^{\phi}(x)| \leqslant \varepsilon\end{aligned}$$

□

Lemma 20. *Let $\mu \in \mathcal{P}(\mathcal{X})$, $\Psi \subseteq \Phi_{ex}$ be dominated and such that F_{μ}^{ψ} is finite. Then, for any $z \in \mathcal{X}$ the class of functions $\mathcal{F} := \{f_{\phi} : x \mapsto \phi(\rho(z,x)) \ : \ \phi \in \Psi\}$ is μ-Glivenko Cantelli.*

Proof. Given a class of measurable functions $\mathcal{F}$, Theorem 3.2 of Sen, 2022 guarantees that a sufficient condition for $\mathcal{F}$ to be μ-Glivenko-Cantelli is that for every $\varepsilon > 0$,

$$N_{[]}(\varepsilon, \mathcal{F}, L^1(\mu)) < \infty.$$

By $N_{[]}$ we denote the bracketing number (see Chapter 3 of Sen, 2022 and Chapter 7 of Vaart and Wellner, 1996 for definition and properties). In our case

$$\mathcal{F} := \{\phi \circ \rho(z, \cdot) : \phi \in \Psi\}.$$

However, for any $\mu \in \mathcal{P}(\mathcal{X})$, we can find some $\nu \in \mathcal{P}(\mathbf{R}_+)$ such that

$$\|\phi \circ \rho(z,\cdot)\|_{L^1(\mu)} = \int_{\mathcal{X}} \phi(\rho(z,x))\mu(\mathrm{d}x) = \int_{\mathbf{R}_+} \phi(r)\nu(\mathrm{d}r).$$

Thus, any bracketing of $(\mathcal{F}, L^1(\mu))$ corresponds to a bracketing of $(\Psi, L^1(\nu))$, and thus $N_{[]}(\varepsilon, \mathcal{F}, L^1(\mu)) = N_{[]}(\varepsilon, \Psi, L^1(\nu))$. Let $\psi \in \Phi_{ex}$ be such that $\phi \leqslant \psi$ for every $\phi \in \Psi$. Now, since $\|\psi\|_{L^1(\nu)} < \infty$, there must be some $a > 0$ such that $\int_{(a,\infty)} \psi(r)\nu(\mathrm{d}r) < \varepsilon/2$. Also let $b := \psi(a)$. Let now f_1, f_2 be two increasing functions defined from $[0,a]$ to $[0,b]$ such that $f_1 < f_2$ and $\|f_2 - f_1\|_{L^1(\nu|_{[0,a]})} < \varepsilon/2$, then by extending f_1 to be 0 and f_2 to coincide with ψ on (a,∞), we see that $\|f_2 - f_1\|_{L^1(\nu)} < \varepsilon$. This implies that

$$N_{[]}(\varepsilon, \Psi, L^1(\nu)) \leqslant N_{[]}(\varepsilon/2, \mathcal{L}_{a,b}, L^1(\nu|_{[0,a]})),$$

where $\mathcal{L}_{a,b}$ is the set of increasing functions from $[0, a]$ to $[0, b]$. But, by a simple grid argument exposed in Theorem 2.7.5 of Vaart and Wellner, 1996, we know that

$$N_{[]}(\varepsilon/2, \mathcal{L}_{a,b}, L^1(\nu|_{[0,a]})) < \infty$$

for any ε, and we are done. $\square$

Lemma 21. *Let $\lambda = \inf_{x \in K} F^{\psi}(x)$. There is an almost surely finite random variable N and a compact set A containing $\bigcup_{\phi \in \Psi} \mathfrak{m}^{\phi}$ such that for all $n \geqslant N$,*

$$\inf_{\phi \in \Psi} \inf_{y \notin A} \min\left(F^{\phi}(y), F_n^{\phi}(y)\right) > \lambda + 1$$

Proof. If $\mathcal{X}$ is compact, then we can just take $A = \mathcal{X}$ and we are done. Instead, assume $\mathcal{X}$ to be unbounded, we can find a compact set B such that $\int_B \mu(\mathrm{d}x) > 3/4$ and then we consider $A_k = \{x \in \mathcal{X} : \rho(x, B) \leqslant k\}$. This is compact, and moreover, by denoting $\varphi : t \mapsto \inf_{\phi \in \Psi} \phi(t)$,

$$\forall \phi \in \Psi, \forall y \notin A_k, \quad F_{\mu}^{\phi}(y) \geqslant \int_A \varphi(\rho(y, x))\mu(\mathrm{d}x) \geqslant \varphi(k)\mu(A) = \frac{3}{4}\varphi(k)$$

Since $\lim_{k\to\infty} \varphi(k) \geqslant \lim_{k\to\infty} k \inf_{\phi \in \Psi} \phi(1) = \infty$, we have that this quantity is larger than $\lambda + 1$ for some k. As for the sample case, we have that

$$\forall \phi \in \Psi, \forall y \notin A_k, \quad F_n^{\phi}(y) \geqslant \sum_{j : X_j \in A} \varphi(\rho(y, X_j)) \geqslant \frac{|\{j \leqslant n : X_j \in A\}|}{n} \varphi(k)$$

But by the SLLN, we have that $|\{j : X_j \in A\}|/n$ converges almost surely to $\mu(A) = 3/4$ so that, there is an almost sure finite random variable N such that if $n \geqslant N$, then $|\{j \leqslant n : X_j \in A\}|/n > 1/2$. Thus, we can take $k^* := \inf\{k : \varphi(k)/2 > \lambda + 1\}$ and define $K := A_{k*}$ and we are done. Indeed it is now trivial to see that $\bigcup_{\phi \in \Psi} \mathfrak{m}^{\phi}$ is contained in B, because for each $\phi \in \Psi$, we have that $F^{\phi}(\mathfrak{m}^{\phi}) = \inf_{x \in K} F^{\phi}(x) \leqslant \inf_{x \in K} F^{\psi}(\mathfrak{m}^{\phi}) = \lambda$. $\square$

Notice that Lemma 21 also proves that $\bigcup_{\phi \in \Psi} \mathfrak{m}^{\phi}$ is precompact. Now we are ready to proceed with the proof of Theorem 13.

Proof of Theorem 13. Let $m_\phi = \inf_{x \in \mathcal{X}} F^\phi(x)$, and let $\varepsilon > 0$, we would like to prove that there is $\delta > 0$ and an as finite random variable N such that for any $\phi \in \Psi$, $n \geqslant N$,

$$\sup_{x \in \mathfrak{m}^\phi} F_n^\phi(x) \leqslant m_\phi + \delta, \qquad \inf_{x:\rho(x,\mathfrak{m}^\phi)>\varepsilon} F_n^\phi(x) \geqslant m_\phi + 3\delta$$

This will conclude the proof. Let A and N be as in the proof of Lemma 19, and define

$$K := \{x \in A : \rho(x, \mathfrak{m}^\phi) \geqslant \varepsilon\}.$$

This is compact and $m'_\phi := \min\{F^\phi(x) : x \in K\} > m_\phi$. Then set $\delta := \inf_{\phi\in\Psi}(m'_\phi - m_\phi)/4$ which can be shown to be strictly positive because of the hypothesis of Φ being compact Indeed, this is a consequence of Lemmas 22 and 23). Then, for any $\phi \in \Psi$, $m'_\phi > \lambda_\phi + 4\delta$. Now we can apply the Lemma 18 to the compact set A and find almost surely finite random variables N' so that if $n > \max(N, N')$, we have at the same time

$$\begin{aligned}
&\forall x \in \mathfrak{m}^\phi, && F_n^\phi(x) \leqslant m_\phi + \delta \\
&\forall x \in K, && F_n^\phi(x) \geqslant m'_\phi - \delta = m_\phi + 3\delta \\
&\forall x \notin A, && F_n^\phi(x) \geqslant m + 1
\end{aligned}$$

This forces each of the sample ϕ-mean sets $\mathfrak{m}_n^\phi$ to be contained in an ε-engrossing of the corresponding $\mathfrak{m}^\phi$. $\square$

This lemma is a consequence of Berge's maximum theorem Berge, 2010, all the relevant notions are presented in Chapter 7 of Rockafellar and Wets, 1998.

Lemma 22. *Let $\mathcal{F}$ be a closed subset of continuous functions from a compact set A to $\mathbf{R}$ such that for each $F \in \mathcal{F}$, F ha a unique minimizer (that we call x_F) contained in $K_- := \{x \in A : \rho(x, \partial A) > \eta\}$ for some $\eta > 0$. Then, for any $\varepsilon < \eta$, the map $\delta : F \mapsto \min_{x\in A:\rho(x,x_F)>\varepsilon} F(x) - F(x_F)$ is continuous. Thus, $\delta := \inf_{F\in\mathcal{F}} \delta_\varepsilon(F) > 0$.*

Proof. By Theorem 7.33 of Rockafellar and Wets, 1998, since $\mathcal{F}$ is closed in the sup-norm and it contains only continuous functions, the map $F \mapsto \arg\min_{x\in K} F(x)$ (from $\mathcal{F}$ to 2^K) is epigraphically continuous. Moreover, since by assumption, each F has a unique minimum, this implies that the map $F \mapsto x_F$ is continuous from $\mathcal{F}$ to K. By the closed graph theorem (Theorem 1.4.8 of Aubin and Frankowska, 2009), we also have that the map $C : F \mapsto \{x \in K : \rho(x, x_F) > \varepsilon\} = \mathrm{cl}(K \backslash B_\varepsilon(x_F))$ is upper and lower epicontinous and compact valued, thus we can apply Berge's maximum theorem (see Berge, 2010) and deduce that the map : $F \mapsto \min_{x \in C(F)} F(x)$ is continuous too. $\square$

Lemma 23. *If Ψ is a "compact" collection of maps in Φ_{ex} such that $\psi \geqslant \phi$ for any $\phi \in \Psi$ and F_μ^ψ is finite, then for any compact set $A \in \mathcal{X}$, the collection $\mathcal{F} := \{F_\mu^\phi|_A : \phi \in \Psi\}$ is compact.*

Proof. We can define a norm on Ψ as $\|\cdot\|_\Psi : \phi \mapsto \|\frac{\phi}{1+\psi}\|_\infty$. This is well defined and makes Ψ compact because each function of the form $\phi/(1+\psi)$ is continuous and bounded by one. Now we show that the map sending $\phi \in \Psi$ to $F_\mu^\phi|_A \in L_\infty(A)$ is continuous.

$$
\begin{aligned}
\|F_\mu^\phi|_A\|_\infty &= \sup_{x \in A} \int_{\mathcal{X}} \phi(\rho(x,y)) \mathrm{d}\mu(y) \\
&= \sup_{x \in A} \int_{\mathcal{X}} \frac{\phi(\rho(x,y))}{1 + \psi(\rho(x,y))} (1 + \psi(\rho(x,y))) \mathrm{d}\mu(y) \\
&\leqslant \|\phi\|_\Psi \left(1 + \sup_{x \in A} \int_{\mathcal{X}} \psi(\rho(x,y)) \mathrm{d}\mu(y)\right) \\
&\leqslant \|\phi\|_\Psi \left(1 + \sup_{x \in A} F_\mu^\psi(x)\right).
\end{aligned}
$$

$\square$

3.4 An Examples

Here we show an example of application of our results. We drewn $n = 100$ samples from a bimodal distribution on the sphere S^3. In particular, the generating process is as follows

$$\begin{aligned} s_i &\overset{iid}{\sim} (\delta_1 + \delta_{-1})/2, \\ y_i &\overset{iid}{\sim} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right), \\ x_i &= y_i / \|y_i\|_2. \end{aligned}$$

This is effectively a mixture of two isotropic distributions centered respectively on the north pole and the south pole of the sphere. As a consequence, the resulting population ϕ-mean (for any $\phi \in \Phi$) is given by all the points on the equator, that is points with null third component. However, when drawing a sample from such a distribution, the resulting ϕ-mean $\mathfrak{m}^\phi$ will almost surely be given by just a single point, as the empirical distribution is almost surely not symmetric with respect to the equator. However, according to the consistency result, the one-sided Hausdorff distance between $\mathfrak{m}_n^\phi$ and $\mathfrak{m}^\phi$ will converge to zero almost surely. This case makes it clear why it is essential to use the one-sided Hausdorff distance instead of the symmetric Hausdorff distance. The reason is that using the symmetric version we would get that the distance between a single point and the equator can never go below π. Instead, using the one-sided version, accurately assign a value of zero for the distance between a point on the equator and the equator itself.

Below is an illustration of a simulation with $\phi : t \mapsto t^p$ for $p \in [1, 3]$.

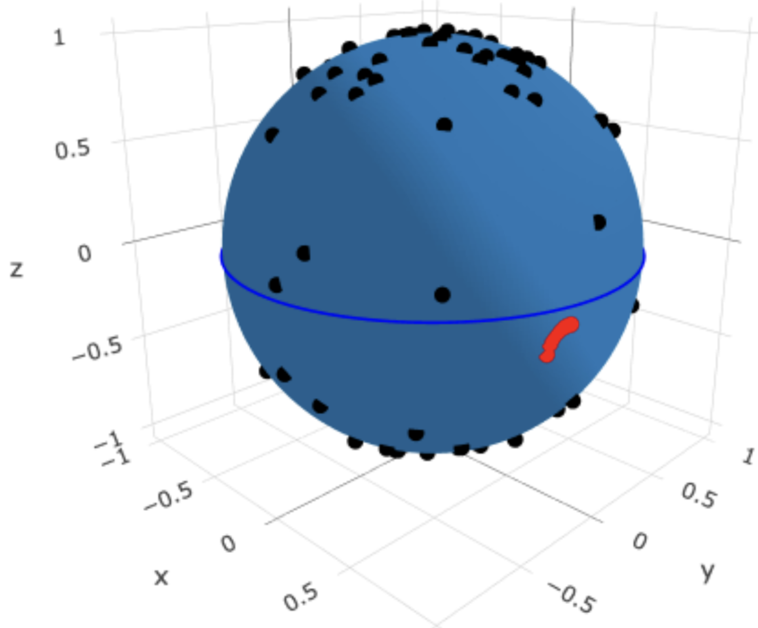

FIGURE 3.1: In black are the 100 samples, in blue is the population ϕ-mean (which does not depend on ϕ) and in red are the sample ϕ-means. As we can see they are not spread all over the equator, but they are all close to it.

4. Central Limit Theorems

In the previous Chapter and in the article Aveni and Mukherjee, 2024 we established a Consistency result for ϕ-means which shows that, under mild hypothesis, the sample ϕ-mean $\mathfrak{m}_n^\phi$ converges almost surely to its population counterpart $\mathfrak{m}^\phi$.

In this section, we better quantify the speed of this convergence by providing a CLT for generalized Fréchet means. However, in order to state such a result we need to restrict ourselves to a more restrictive setting than just metric spaces with the Heine Borel property. We need to be able to locally approximate our space $\mathcal{X}$ with a linear space. For this reason we assume from now on $\mathcal{X}$ to be a Riemannian manifold equipped with its geodesic distance ρ, even though more general settings have been recently explored by Mattingly et al., 2024.

In the first section, we will briefly recall some relevant notions from Riemannian geometry and then we will provide and prove our CLT. Finally we will illustrate these results with examples on common Riemannian manifolds like spheres and the Lie group of spatial rotations SO_3.

4.1 Notation and results on the cut locus

The basic notions of Riemannian geometry that we need are covered, for instance, in Chapter 6 of Lee, 2021, Chapter 5 of do Carmo, 2013, Chapter 5.4 of Jost, 2011 or Chapter 3.2 of Chavel, 2006. For the more specialized results on the cut locus, essential references are Eltzner et al., 2024, Le and Barden, 2014, Barden and Le, 1997 and most importantly for our work Hotz et al., 2022. The reason why we are interested in the cut locus is that this is a set of points where the distance function is not smooth and, despite the cut locus having volume measure zero, this can slow down the convergence of the sample ϕ-means. In particular, the contribution of the cut locus to the asymptotic variance of the sample ϕ-means is known in the literature as smeariness and it was first described in Eltzner, 2020.

We denote by $\mathrm{T}_x\mathcal{X}$ the tangent space of $\mathcal{X}$ at x and by $\exp_x : \mathrm{T}_x\mathcal{X} \to \mathcal{X}$ the exponential map, which maps a tangent vector $v \in \mathrm{T}_x\mathcal{X}$ to the point reached at $t = 1$ by the unique geodesic $\gamma : \mathbf{R} \to \mathcal{X}$ with $\gamma(0) = x$ and $\dot{\gamma}(0) = v$.

We denote by V_x and N_x the largest neighborhood of $0 \in \mathrm{T}_x\mathcal{X}$ and of $x \in \mathcal{X}$ for which $\exp_x$ is a diffeomorphism from V_x to N_x. The set N_x has full volume measure in $\mathcal{X}$ and its complement C_x is the cut locus of x which is known to have at least Hausdorff codimension one (see Barden and Le, 1997). We have that $\mathrm{C}_x = \exp_x(\partial \mathrm{V}_x)$. For this reason the tangential preimage of C_x is sometimes called the tangential cut locus (see for instance Jin-ichi and Minoru, 1998).

On N_x we can define the log map as the inverse of the exponential map and we denote it by $\exp_x^{-1} : \mathrm{N}_x \to \mathrm{T}_x\mathcal{X}$. This is by definition of N_x a diffeomorphism. When $y \in \mathrm{C}_x$, by $\exp^{-1}(y)$ we mean the preimage of y in the tangential cut locus ∂V_x through $\exp_x$. The failure of $\exp_x$ to be a diffeomorphism can be due to a lack of injectivity or to a degeneracy in its derivative. This allows to partition C_x into two (or indeed countably many) disjoint parts, in particular, let H_x be the set of points $y \in \mathrm{C}_x$ for which there are exactly two minimal geodesics connecting x to y. This set is known (Barden and Le, 1997) to be a (possibly empty) disjoint union of hypersurfaces (codimension one properly embedded submanifolds). Moreover, its complement in C_x has at least codimension two and it contains the set of first conjugate loci (i.e. the points in C_x at which the exponential map has a degenerate derivative). Notice that H_x can be also written as $\{y \in \mathrm{C}_x : |\exp_x^{-1}(y)| = 2\}$.

4.2 Central Limit Theorem

We now consider the regularity of the map $\rho_x : \mathcal{X} \to \mathbf{R}_+$ sending $y \mapsto \rho(x,y)$. This function is smooth on $\mathrm{N}_x \backslash \{x\}$ with unitary gradient and it is just continuous on $\mathrm{C}_x \sqcup \{x\}$. Despite this, ρ_x is relatively well behaved on H_x (and at x where it is completely analogous to the cone $y \mapsto \|y\|_2$ in $\mathbf{R}^k$). In particular, we recall the following facts proved in Hotz et al., 2022.

Lemma 24. *Let $S \subseteq \mathrm{H}_x$ be a sufficiently small subset of a connected component of H_x, then $\exp_x^{-1}(\mathrm{H}_x)$ will be made by two disconnected regions of C_x. There are two disjoint neighborhoods $V^1, V^2 \subseteq \mathrm{T}_x\mathcal{X}\backslash\{0\}$ containing these two regions such that $(\exp_x|_{V^i})_{i=1,2}$ are two diffeomorphism onto the same set $U = \exp_x(V^1) = \exp_x(V^2)$. We can now define $\rho_x^i : U \to \mathbf{R}_+$ as the map* *sending $y \mapsto \|(\exp_x|_{V^i})^{-1}(y)\|$. We have that these two maps are smooth and that they agree only on $U \cap \mathrm{H}_x$. Moreover,*

$$\forall y \in U, \quad \rho_x(y) = \min(\rho_x^1(y), \rho_x^2(y)).$$

In addition for $i \in \{1,2\}$, if $y \in S$ and v^i is the only element in $\exp_x^{-1}(y) \cap V^i$, then

$$\nabla \rho_x^i(y) = v^i / \|v^i\|$$

and the geodesics γ^1, γ^2 with $\gamma^i(0) = x$ and $\dot{\gamma}^i(0) = v^i$ are the two minimal geodesics connecting x to y in a unit of time.

Now, for each connected component of H_x, following the constructions in the previous lemma (and carefully patching together the different maps constructed at each connected component), we define ρ_x^i on a neighborhood of all these connected components and by making arbitrary choices of signs, we define $\chi_x : y \mapsto \rho_x^1(y) - \rho_x^2(y)$. This map is well defined on some neighborhood of H_x. In [] it is proved that $\mathrm{H}_x \ni y \mapsto \nabla \chi(y|x)$ is a non-vanishing vector field normal to H_x.

From this we define for every $y \in \mathrm{H}_x$,

$$\kappa_x : y \mapsto \|\nabla \chi_x(y)\| \quad \mathbf{n}_x : y \mapsto \nabla \chi_x(y)/\kappa_x(y).$$

We have that, if $\phi : \mathbf{R}_+ \to \mathbf{R}_+$ is a C^1 function, then for $y \in \mathrm{H}_x$,

$$\begin{aligned} \left\| \nabla \left(\phi(\rho_x^1(y)) - \phi(\rho_x^2(y)) \right) \right\| &= \left\| \phi'(\rho_x^1(y)) \nabla \rho_x^1(y) - \phi'(\rho_x^2(y)) \nabla \rho_x^2(y) \right\| \\ &= \phi'(\rho(x,y)) \|\nabla \chi_x(y)\| \\ &= \phi'(\rho(x,y)) \kappa_x(y) \end{aligned}$$

We also have that if $y \in \mathrm{H}_x$,

$$\nabla \rho_x^i(y) = \lim_{z \to y, \langle \exp_y^{-1}(z), \nabla \rho_x^i(y) \rangle < 0} \nabla \rho_x^i(z).$$

For $y \in \mathrm{H}_x$ we define the 1-form $\mathrm{d}y^x : v \mapsto \langle \mathbf{n}_x(y), v \rangle$ and the $(0,2)$-tensor $J_x(y) : (u,v) \mapsto \phi'(\rho(x,y)) \kappa_x(y) \langle \mathbf{n}_x(y), u \rangle \langle \mathbf{n}_x(y), v \rangle$. Now impossible definition of $\tau_{x|y}(t)$.

Now, for any C^2 function $\phi \in \Phi$, we define $\lambda^\phi : t \mapsto \phi'(t)/t$ this is well defined and $\lambda^\phi(0) = \phi''(0)$. If μ is a probability measure on $\mathcal{X}$, we define the following vector field

Definition 25.

$$G_\mu^\phi(x) = \int_{\mathrm{N}_x} \lambda^\phi(\|\exp_x^{-1}(y)\|)\exp_x^{-1}(y)\mathrm{d}\mu(y)$$

Now, if $\mu(\mathrm{C}_x) = 0$ which is always the case if μ is absolutely continuous with respect to the volume measure in a neighborhood of C_x, then

$$\begin{aligned}
\nabla F_\mu^\phi(x) &= \int_{\mathcal{X}} \nabla\phi(\rho(x,y))\mathrm{d}\mu(y) \\
&= -\int_{\mathcal{X}} \frac{\phi'(\|\exp_x^{-1}(y)\|)}{\|\exp_x^{-1}(y)\|}\exp_x^{-1}(y)\mathrm{d}\mu(y) \\
&= -\int_{N_x} \lambda^\phi(\|\exp_x^{-1}(y)\|)\exp_x^{-1}(y)\mathrm{d}\mu(y) = -G_\mu(x).
\end{aligned}$$

Indeed, as can be seen from an application of the chain rule, we have that the gradient of $f_x : y \mapsto \phi(\rho_x(y))$ is the vector field

$$\nabla\phi \circ \rho_y(x) = -\frac{\phi'(\rho_y(x)}{\rho_y(x)}\exp_x^{-1}(y)$$

Moreover, the Hessian of f_x is the symmetric (0,2)-tensor

$$\mathrm{Hess} f_x(y) : (u,v) \mapsto \langle D_v(\nabla f_x)(y)u, v\rangle_y$$

By calling $A_x(y)$ the (1,1)-tensor sending $v \in \mathrm{T}_x\mathcal{X}$ to

$$D_v\left(\lambda^\phi(\|\exp_x^{-1}(y)\|)\exp_x^{-1}(y)\right) \in \mathrm{T}_x\mathcal{X},$$

we have that $\mathrm{Hess} f_x(y)(u,v) = -\langle A_x(y)(u), v\rangle$. In what follows we will need to consider the expected value of A_x which is the map $\Lambda_x^1 : \mathrm{T}_x\mathcal{X} \to \mathrm{T}_x\mathcal{X}$ sending

$$\Lambda_x^1 : v \mapsto \int_{\mathrm{N}_x} A_x(y)(v)\mathrm{d}\mu(y) \overset{\text{if } \mu(\mathrm{C}_x)=0}{=} \int_{\mathcal{X}} A_x(y)(v)\mathrm{d}\mu(y) \tag{4.1}$$

Given a probability measure μ absolutely continuous around H_x with density $\tilde{\mu} := \frac{\mathrm{d}\mu}{\mathrm{dvol}}$, we also define the $(1,1)$-tensor field J_μ as

$$\Lambda_x^2 : v \mapsto \int_{\mathrm{H}_x} \phi'(\rho(x,y))\kappa_y(x)\langle \mathbf{n}_y(x), v\rangle \tau'_{y|x}(0)\mathbf{n}_y(x)\tilde{\mu}(y)\mathrm{dvol}_{\mathrm{H}_x}(y) \tag{4.2}$$

Clearly, if H_x is empty (that is if C_x has codimension $\geqslant 2$) or if μ has vanishing density over C_x, then Λ_x^2 is zero. Finally, putting together these two terms $\Lambda_x^\phi = \Lambda_x^1 + \Lambda_x^2$ we get the essential term for the next linearization result, which will allow us to compute the Hessian of F_μ^ϕ.

Finally, before stating our we recall the notions of metric continuity. We denote by

$$\mathcal{A}_\delta(x) := \mathrm{C}_{B_\delta(x)} = \bigcup_{y \in B_\delta(x)} \mathrm{C}_y, \quad \mathcal{B}_\delta(x) := B_\delta(\mathrm{C}_x) = \bigcup_{y \in \mathrm{C}_x} B_\delta(y)$$

Since $x \in \mathrm{C}_y$ iff $y \in \mathrm{C}_x$, we can rewrite the first set as

$$\mathcal{A}_\delta(x) = \{y \in \mathcal{X} : \mathrm{C}_y \cap B_\delta(x) \neq \varnothing\}.$$

Metric continuity, which holds for every Riemannian manifold, proven in Jin-ichi and Minoru, 1998 for compact manifolds and in Le and Barden, 2014 for general ones, says that

Lemma 26. *For any Riemannian manifold $\mathcal{X}$ and for any $x \in \mathcal{X}$,*

$$\forall \varepsilon > 0 \exists \delta > 0 : \mathcal{A}_\delta(x) \subseteq \mathcal{B}_\varepsilon(x).$$

Now we present the following CLT, which is a generalization of the one presented in Hotz et al., 2022. The assumptions are extremely weak, in particular, it is conjectured that assumption (ii) is always satisfied on any Riemannian manifold. In particular it is known to be satisfied on any symmetric space.

4.3 Central Limit Theorem for Smooth Functions

Theorem 27. *Let $\mathcal{X}$ be a connected Riemannian manifold, μ a Borel probability measure on $\mathcal{X}$ with a unique ϕ-mean $\mathfrak{m}^\phi$. Under the following assumptions*

- *There is some $\delta > 0$ such that μ has a continuous density with respect to the volume measure on $\mathcal{B}_\delta(\mathfrak{m}^\phi)$.*
- *As $\delta \downarrow 0$, $\mathrm{vol}(\mathcal{A}_\delta(\mathfrak{m}^\phi)) = O(\delta)$*
- *The terms $\Lambda^1_{\mathfrak{m}^\phi}, \Lambda^2_{\mathfrak{m}^\phi}$ are well defined and finite.*

Then

$$\sqrt{n}\exp^{-1}_{\mathfrak{m}^\phi}(\mathfrak{m}^\phi_n) \xrightarrow{w} \mathcal{N}_d\left(0, (\Lambda^\phi_{\mathfrak{m}^\phi})^{-1}\Sigma^\phi(\Lambda^\phi_{\mathfrak{m}^\phi})^{-T}\right)$$

The proof of this theorem will result from the following lemmas.

Lemma 28. *Under the hypothesis of Theorem 27, the vector field G^ϕ_μ admits the linearization*

$$\Pi_{x,\mathfrak{m}^\phi} G^\phi_\mu(x) = -\Lambda^\phi_{\mathfrak{m}^\phi}(\exp^{-1}_{\mathfrak{m}^\phi}(x)) + o(\rho(x,\mathfrak{m}^\phi)).$$

Proof. Let $z \in \mathcal{X}$ be fixed, we need to linearize the vector field g^ϕ_z sending

$$\mathrm{N}_z \ni x \mapsto \lambda^\phi(\|\exp^{-1}_x(z)\|)\exp^{-1}_x(z) \in \mathrm{T}_x\mathcal{X}$$

In particular, let $x, y \in \mathrm{N}_z$ be contained in a geodesically convex neighborhood, so that there is a unique minimal geodesic γ with $\gamma(0) = x$ and $\gamma(\rho(x,y)) = y$. We now have two possibilities. We could have that $\gamma([0,\rho(x,y)]) \subseteq \mathrm{N}_z$, or that $\gamma([0,\rho(x,y)]) \cup \mathrm{C}_z \neq \emptyset$. In the first case, we have that by definition of covariant derivative,

$$\Pi_{y,x}(g_z(y)) = g_z(x) + A_z(x)(\exp^{-1}_x(y)) + o(\rho(x,y)) \tag{4.3}$$

As for the second case, by taking x, y generic and sufficiently small, we will only consider the case in which $\gamma([0,\rho(x,y)])$ intersects C_z just once at some time $t^* \in (0,\rho(x,y))$ and that $\xi := \gamma(t^*) \in \mathrm{H}_z$. In this case, we have that

$$\Pi_{y,x}(g_z(y)) = g_z(x) + A_z(x)(\exp^{-1}_x(y)) + \phi'(\rho_z(\xi))\kappa_z(\xi)\Pi_{\xi,x}(\mathbf{n}_z(\xi)) + o(\rho(x,y)) \tag{4.4}$$

Indeed, if ρ^1_z and ρ^2_z are the functions defined in Lemma ? that we now label in such a way that $\rho_z(\gamma(t)) = \rho_z^1(\gamma(t))\mathbf{1}_{t\in[0,t^*]} + \rho_z^2(\gamma(t))\mathbf{1}_{t\in(t^*,\rho(x,y))]}$, then we can write

$$\begin{aligned}
\Pi_{y,x}(g_z(y)) - g_z(x) =& \Pi_{y,x}(g_z(y)) - \Pi_{\xi,x}\left(\nabla(\phi\circ\rho_z^1)(\xi)\right) \\
&+ \Pi_{\xi,x}\left(\nabla(\phi\circ\rho_z^1)(\xi)\right) - \Pi_{\xi,x}\left(\nabla(\phi\circ\rho_z^2)(\xi)\right) \\
&+ \Pi_{\xi,x}\left(\nabla(\phi\circ\rho_z^2)(\xi)\right) - g_z(x) \\
=& - \Pi_{y,x}(\Pi_{\xi,y}\left(\nabla(\phi\circ\rho_z^1)(\xi)\right) - g_z(y)) \\
&+ \Pi_{\xi,x}\left(\phi'(\rho_z(\xi))\kappa_z(\xi)\mathbf{n}_z(\xi)\right) \\
&+ A_z(x)(\exp_x^{-1}(\xi)) + o(\rho(x,\xi)) \\
=& - \Pi_{y,x}\left(A_z(y)(\exp_y^{-1}(\xi)) + o(\rho(\xi,y))\right) \\
&+ \phi'(\rho_z(\xi))\kappa_z(\xi)\Pi_{\xi,x}(\mathbf{n}_z(\xi)) \\
&+ A_z(x)(\exp_x^{-1}(\xi)) + o(\rho(x,\xi)).
\end{aligned}$$

Where we used the definitions of $\kappa, \mathbf{n}$, and the fact that, along a geodesic γ, $\Pi_{\gamma(a),\gamma(b)} \circ \Pi_{\gamma(b),\gamma(c)} = \Pi_{\gamma(a),\gamma(c)}$. Thus, we have that $\Pi_{y,x}(g_z(y)) - g_z(x)$ is given, up to an error $o(\rho(x,y))$, by

$$A_z(x)(\exp_x^{-1}(\xi)) + \Pi_{y,x}\left(A_z(y)(-\exp_y^{-1}(\xi))\right) + \phi'(\rho_z(\xi))\kappa_z(\xi)\Pi_{\xi,x}(\mathbf{n}_z(\xi)).$$

However, x, ξ, y are on a geodesic, and thus we have that if $\alpha := \rho(x,\xi)/\rho(x,y), \beta := \rho(\xi,y)/\rho(x,y) = 1-\alpha$, then

$$\begin{aligned}
&A_z(x)(\exp_x^{-1}(y)) - A_z(x)(\exp_x^{-1}(\xi)) - \Pi_{y,x}\left(A_z(y)(-\exp_y^{-1}(x)))\right) \\
&= A_z(x)(\exp_x^{-1}(y)) - \alpha A_z(x)(\exp_x^{-1}(y)) + \beta\Pi_{y,x}\left(A_z(y)(\exp_y^{-1}(x))\right) \\
&= \beta\left(A_z(x)(\exp_x^{-1}(y)) - \Pi_{y,x}\left(A_z(y)(\Pi_{x,y}(\exp_x^{-1}(y)))\right)\right) = o\left(\rho(x,y)\right)
\end{aligned}$$

Thus, we have indeed shown (4.4).

Now we can Let x be a point in a geodesically convex neighborhood of $\mathfrak{m}$ at distance

$\rho(x, \mathfrak{m}) = t$ and let γ be the unit speed geodesic from $\mathfrak{m}$ to x. For $k \geqslant 1$, we consider

$$L_x^k := \{y \in \mathcal{X} : |\gamma((0,t)) \cap \mathrm{C}_y| = |\gamma((0,t)) \cap \mathrm{H}_y| = k\}$$

$$L_x^\infty := \{y \in \mathcal{X} : |\gamma((0,t)) \cap \mathrm{H}_y| = \infty \text{ or } |\gamma((0,t)) \cap (\mathrm{C}_y \backslash \mathrm{H}_y)| > 0\}$$

$$L_x^{tot} := \{y \in \mathcal{X} : \gamma((0,t)) \cap \mathrm{C}_y \neq \varnothing\}.$$

Clearly, L_x^{tot} is the union of L_x^∞ and all the L_x^k, and $L_x^{tot} \subseteq \mathcal{A}_t(x)$, thus, by hypothesis $\mathrm{vol}(L_x^k) = O(t)$, moreover, if t is sufficiently small so that μ is absolutely continuous in $\mathcal{A}_t(x)$, we also have that $\mu(L_{x,y}) = O(t)$. Now, the ϕ-mean $\mathfrak{m}$ is such that $G_\mu^\phi(\mathfrak{m}) = 0$, thus, if x is in a geodesically convex neighborhood of $\mathfrak{m}$,

$$\begin{aligned}
&\Pi_{x,\mathfrak{m}}(G_\mu^\phi(x)) \\
&= \int_{\mathcal{X}} \Pi_{x,\mathfrak{m}}(g_y(x)) \mathrm{d}\mu(y) \\
&\overset{(4.3)}{=} \int_{\{y:\gamma((0,t))\cap \mathrm{C}_y = \varnothing\}} g_y(\mathfrak{m}) + A_y(\mathfrak{m})(\exp_{\mathfrak{m}}^{-1} x) + o(\rho(\mathfrak{m}, x)) \mathrm{d}\mu(y) \\
&\quad \overset{(4.4)}{+} \int_{L_x^1} g_y(\mathfrak{m}) + A_y(\mathfrak{m})(\exp_{\mathfrak{m}}^{-1} x) + \phi'(\rho_y(\xi))\kappa_y(\xi)\Pi_{\xi,\mathfrak{m}}(\mathbf{n}_y(\xi)) + o(\rho(\mathfrak{m}, x)) \mathrm{d}\mu(y) \\
&\quad + \int_{\bigcup_{k>1} L_x^k} \Pi_{x,\mathfrak{m}}(g_y(x)) \mathrm{d}\mu(y) \\
&\quad + \int_{L_x^\infty} \Pi_{x,\mathfrak{m}}(g_y(x)) \mathrm{d}\mu(y) \\
&= G_\mu^\phi(\mathfrak{m}) + \left(\int_{\mathcal{X}} A_y(\mathfrak{m}) \mathrm{d}\mu(y) \right) (\exp_{\mathfrak{m}}^{-1} x) \\
&\quad + \int_{L_x^1} \phi'(\rho_y(\xi))\kappa_y(\xi)\Pi_{\xi,\mathfrak{m}}(\mathbf{n}_y(\xi)) \mathrm{d}\mu(y) + o(\rho(\mathfrak{m}, x))
\end{aligned}$$

Now, since $\kappa, \mathbf{n}$ are defined and continuous in a neighborhood of each $\xi \in \mathrm{H}_y$ and since μ is absolutely continuous in L_x^1, with density $\tilde{\mu}$, we have that

$$\begin{aligned}
\int_{L_x^1} \phi'(\rho_y(\xi))\kappa_y(\xi)\Pi_{\xi,\mathfrak{m}}(\mathbf{n}_y(\xi)) \mathrm{d}\mu(y) = &\int_{L_x^1} \phi'(\rho_y(\mathfrak{m}))\kappa_y(\mathfrak{m})\mathbf{n}_y(\mathfrak{m})\tilde{\mu}(y) \mathrm{dvol}(y) \\
&+ O(\rho_{\mathfrak{m}}(x)^2)
\end{aligned}$$

Given this formula, we can now apply the same geometric construction in the proof of Theorem 1 of Hotz et al., 2022 and conclude the proof of the Lemma.

□

4.4 Uniform CLT

We are currently working toward the proof of a Uniform CLT. In particular,

Conjecture 29. *Under the same hypothesis of Therem 27 and when $\Psi \subseteq \Phi_{ex}$ satisfies the same hypothesis in Theorem 13 with dominating function $\hat{\psi}$, then*

$$\left(\phi \mapsto \sqrt{n}\exp^{-1}_{\mathfrak{m}^{\phi}_{\mu}}\left(\mathfrak{m}^{\phi}_{n}\right)\right) \overset{w,\, L^{\infty,\hat{\psi}}(\Psi)}{\to} \mathcal{GP}\left(0, [(\Lambda^{\phi})^{-1}\Sigma^{\phi,\psi}(\Lambda^{\psi})^{-T}]_{\phi,\psi\in\Psi}\right)$$

Where Λ is the same as in (4.1) *and* (4.2),

$$\Sigma^{\phi,\psi} = \mathbb{C}\left[\nabla_{v=0}\left(h^{\phi}_{X}(v)\right), \nabla_{v=0}\left(h^{\psi}_{X}(v)\right)\right],$$

and by $L^{\infty,\hat{\psi}}(\Psi)$ we mean the space Ψ equipped with the norm $\|\phi\|_{\infty,\hat{\psi}} := \|\phi/\hat{\psi}\|_{\infty}$.

We provide here an illustration of the validity of this statement in a particular case.

Let $\mathcal{X} = [-1,1]/1 \sim -1$ be a circle. We consider the isotropic measure with density given by $\mu(\mathrm{d}x) = \left(\frac{1}{2} + \frac{1}{4}\cos(\pi x)\right)\mathrm{d}x$. Since the circle is compact, symmetric and μ is isotropic, we have that $\mathfrak{m}^{\phi}_{\mu} = \{0\}$ for any ϕ. We now consider the collection of functions $\Psi = \{t \mapsto t^{p} : p \in [1,3]\}$. In this case, we have that according to our conjecture, the covariance of the limiting Gaussian process should be given by

$$2\frac{2^{-p-q}/(p+q-1) + f(p+q)}{(p-1)(q-1)f(p)f(q)} \tag{4.5}$$

where $f : p \mapsto \int_0^{1/2} x^{p-2}\cos^2(\pi x)\mathrm{d}x$. The proof can be found in the next chapter in 5.1.2. In the next figure, we compare the theoretical covariance (4.5) with the empirical covariance we obtained from simulations.

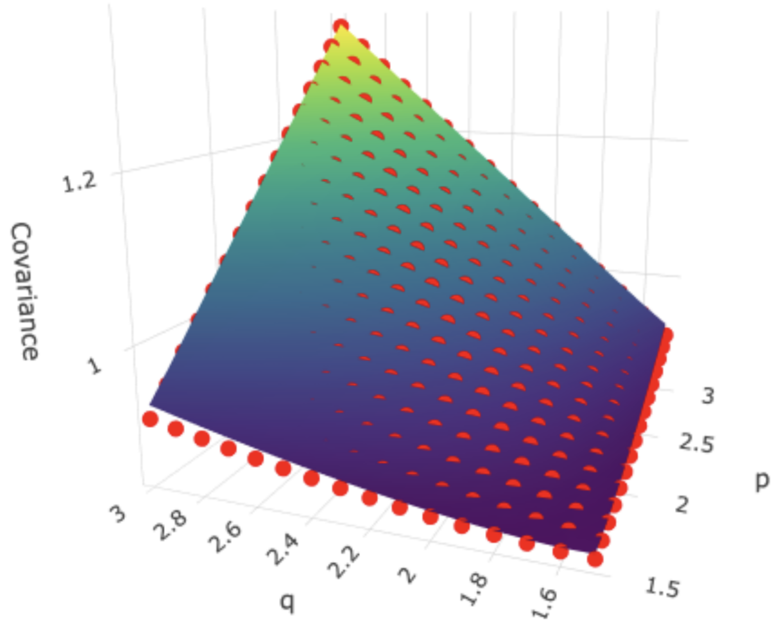

FIGURE 4.1: The covariance of $\sqrt{n}\exp^{-1}_{\mathfrak{m}^{\phi}_{\mu}}\left(\mathfrak{m}^{\phi}_{n}\right)$ for $\phi : t \mapsto t^p$ and $p \in [1.5, 3]$. The same quantity, both according to our conjecture and (in red) according to simulations.

5. Examples

5.1 The simplest case: The real line

Let $\phi \in \Phi_{ex}$ and let μ be a probability measure with a unique ϕ-mean that we can assume to be in zero. In this case, $\exp_0 = \mathrm{id}$ and $C_0 = \varnothing$ so that Λ^2 vanishes. As for the other terms we have

$$\Sigma^{\phi} = \mathbb{V}[\phi'^2(X)] = \int_{\mathbf{R}} \phi'^2(x)\mathrm{d}\mu(x), \qquad \Lambda^{\phi} = \Lambda^1 = \mathbb{E}[\phi''(X)] = \int_{\mathbf{R}} \phi''(x)\mathrm{d}\mu(x)$$

Assuming these two quantities to be well defined and non-zero, we will get the following asymptotic distribution for the sample ϕ-mean

$$\mathfrak{m}_n\sqrt{n} \xrightarrow{w} \mathcal{N}\left(0, \Sigma/(\Lambda)^2\right) = \mathcal{N}\left(0, \frac{\int_{\mathbf{R}} \phi'^2(x)\mathrm{d}\mu(x)}{\left(\int_{\mathbf{R}} \phi''(x)\mathrm{d}\mu(x)\right)^2}\right)$$

For instance, if $\phi : t \mapsto |t|^p$ for some $p > 1$, then the asymptotic variance will be

$$\frac{\int_{\mathbf{R}} \phi'^2(x)\mathrm{d}\mu(x)}{\left(\int_{\mathbf{R}} \phi''(x)\mathrm{d}\mu(x)\right)^2} = \frac{\int_{\mathbf{R}} p^2|x|^{2(p-1)}\mathrm{d}\mu(x)}{\left(\int_{\mathbf{R}} p(p-1)|x|^{p-2}\mathrm{d}\mu(x)\right)^2} = \frac{1}{(p-1)^2}\frac{\int_{\mathbf{R}} (x^2)^{p-1}\mathrm{d}\mu(x)}{\left(\int_{\mathbf{R}} |x|^{p-2}\mathrm{d}\mu(x)\right)^2}.$$

While, if $p = 1$ and μ has a continuous density $\tilde{\mu}$ in a neighborhood of zero, then

$$\frac{\int_{\mathbf{R}} \phi'^2(x)\mathrm{d}\mu(x)}{\left(\int_{\mathbf{R}} \phi''(x)\mathrm{d}\mu(x)\right)^2} = \frac{\int_{\mathbf{R}} 1\mathrm{d}\mu(x)}{\left(\int_{\mathbf{R}} 2\delta_0(x)\mathrm{d}\mu(x)\right)^2} = \frac{1}{4\tilde{\mu}^2(0)}$$

For instance, if μ is a Normal with mean zero and variance σ^2 and $\phi : t \mapsto |t|^p$ (for $p \geqslant 1$), then we get asymptotic variance

$$\frac{\int_{\mathbf{R}} \phi'^2(x)\mathrm{d}\mu(x)}{\left(\int_{\mathbf{R}} \phi''(x)\mathrm{d}\mu(x)\right)^2} = \frac{\sqrt{\pi}}{2}\frac{\Gamma(p-1/2)}{\Gamma^2((p+1)/2)}\sigma^2$$

Which is a well defined convex function in p, taking its minimum at $p = 2$, where it is just σ^2. Indeed, the sample mean (p = 2) is the UMVU estimator in this case. Moreover, we see that the asymptotic variance is diverging when $p \to \infty$ and indeed, it is well known that the variance of the mid-range of a Normal population converge to zero at the slow rate of $1/\ln(n)$.

If instead μ has a uniform distribution on $(-a, a)$, then we get asymptotic variance

$$\frac{\int_{\mathbf{R}} \phi'^2(x)\mathrm{d}\mu(x)}{\left(\int_{\mathbf{R}} \phi''(x)\mathrm{d}\mu(x)\right)^2} = \frac{a^2}{2p-1}.$$

Which is a well defined convex function decreasing to zero as $p \to \infty$. This is in line with the known fact that the mid-range ($p = \infty$) of a uniform distribution is the UMVU estimator of the mean and it converges to a Cauchy distribution when multiplied by the sample size n (and thus it will go to zero when scaled by $\sqrt{n}$ instead).

5.1.1 A Variational Problem

When μ is fixed and it has full support and a non-vanishing continuous density $\tilde{\mu}$ it is possible to ask which ϕ makes the asymptotic variance of the sample ϕ-mean minimal. Given formula (5.1), to solve this variational problem we can use the Lagrangian

$$\mathcal{L}(\phi) = \int_{\mathbf{R}} \phi'^2(x)\tilde{\mu}\mathrm{d}x - \lambda \left(\int_{\mathbf{R}} \phi''(x)\tilde{\mu}\mathrm{d}x\right)^2$$

Under regularity conditions, the functional derivative of this Lagrangian is (twice)

$$\phi'(x)\tilde{\mu}(x) - \alpha\tilde{\mu}'(x)$$

So that, we obtain the solution $\phi(x) = a + b\ln(\tilde{\mu}(x))$ for some constants $a, b \in \mathbf{R}$. In the case of μ being the normal distribution we get as a solution $\phi(t) = t^2$.

5.1.2 Non trivial cut locus: The Circle

Let $\mathcal{X}$ be a circle $[-1/2, 1/2]/\sim$ where $-1/2 \sim 1/2$ and let μ be the probability measure with density

$$\tilde{\mu}(x) = \frac{1}{2} + \cos^2(\pi x)$$

According to Theorem 31, we have that the ϕ-mean of μ is 0 for any $\phi \in \Phi$. And, clearly, $\exp_0(x) = x$ for any $x \in V_0 = (-1/2, 1/2)$. The cut locus at zero is given by the singleton

$C_0 = H_0 = \{\pm 1/2\}$. We now consider $\phi_p : t \mapsto |t|^p$.

$$\begin{aligned}\Sigma_p &= \mathbb{V}_{X\sim\mu}[\lambda^{\phi_p}(\|\exp_0^{-1}(X)\|)\exp_0^{-1}(X)] \\ &= \mathbb{V}_{X\sim\mu}[p|X|^{p-1}\mathrm{sign}(X)] = p^2\mathbb{E}[|X|^{2(p-1)}] \\ &= 2p^2\int_0^{1/2} x^{2p-2}\left(\frac{1}{2}+\cos^2(\pi x)\right)\mathrm{d}x\end{aligned}$$

Thus, if for instance $p = 1, 2, 3, 4$ we get respectively

$$\Sigma_1 = 1,\ \ \Sigma_2 = \frac{1}{3} - \frac{1}{\pi^2},\ \ \Sigma_3 = \frac{9}{80} - \frac{9(\pi^2-6)}{8\pi^4},\ \ \Sigma_4 = \frac{1}{28} - \frac{3(120-20\pi^2+\pi^4)}{4\pi^6}.$$

As for Λ we have that

$$A_0(x) = p\partial_x(|x|^{p-1}\mathrm{sign}(x)) = \begin{cases} p(p-1)|x|^{p-2} & \text{if } p > 1 \\ 2\delta_0(x) & \text{if } p = 1\end{cases}$$

so that if $p = 1$, we get $\Lambda^1 = 2\tilde{\mu}(0) = 3$ while if $p > 1$,

$$\Lambda^1 = 2p(p-1)\int_0^{1/2} x^{p-2}\left(\frac{1}{2}+\cos^2(\pi x)\right)\mathrm{d}x$$

So that $\Lambda^1 = 2, 3/2 - 3/\pi^2, 1 - 3/\pi^2$ when $p = 2, 3, 4$ respectively. Finally, for Λ^2, we have that $\tau'(0) \equiv 1$, $\mathbf{n}$ can be taken as 1, while $\kappa \equiv 2$, thus we have

$$\Lambda^2 = -2\phi'(1/2)\tilde{\mu}(1/2) = -p/2^{p-1}.$$

Thus if $p = 1$, $\Lambda^\phi = 2$, otherwise

$$\Lambda^\phi = 2p(p-1)\int_0^{1/2} x^{p-2}\cos^2(\pi x)\mathrm{d}x.$$

Overall, the asymptotic distribution of $\sqrt{n}\mathfrak{m}_n$ is a normal with variance $\Sigma^\phi/(\Lambda^\phi)^2$ which for $p = 1, 2, 3, 4$ is respectively equal to

$$\frac{1}{4},\ \frac{1}{3} - \frac{1}{\pi^2},\ \frac{60-10\pi^2+\pi^4}{5(\pi^2-4)^2},\ \frac{\pi^6-21\pi^4+420\pi^2-2520}{7\pi^2(\pi^2-6)^2}.$$

In the following plot we illustrate how the asymptotic variance varies with p.

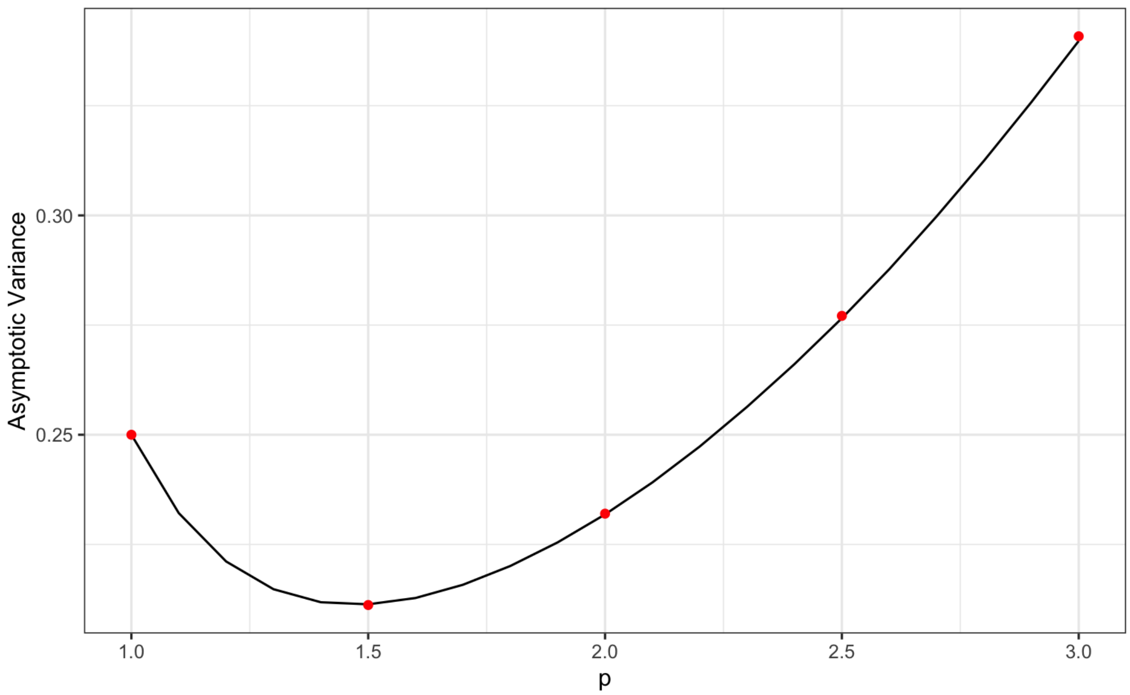

FIGURE 5.1: A numerical computation of the asymptotic variance Λ^{ϕ} for the circle when $\phi : t \mapsto t^p$ and $p \in [1, 3]$. In red, the exact values for $p = 1, 3/2, 2, 5/2, 3$. In these cases, we have $\frac{1}{4}$, $\frac{\pi^2-2}{\pi^2(\sqrt{2}+C(\sqrt{2})^2}$, $\frac{1}{3} - \frac{1}{\pi^2}$, $\frac{4(\pi^4-6\pi^2+24)}{\pi^2(4\pi-3\sqrt{2}S(\sqrt{2}))^2}$, $\frac{60-10\pi^2+\pi^4}{5(\pi^2-4)^2}$. where C and S are the Fresnel integral functions.

5.1.3 Spheres

We consider $S^n = \{x \in \mathbf{R}^{n+1} : \|x\|_2 = 1\}$ and a probability measure absolutely continuous with respect to the volume measure with isotropic monotone density, more precisely, we assume that $\tilde{\mu}(x) = f(\rho(x,e))$ with $f : [0,\pi] \to \mathbf{R}$ smooth monotonically decreasing with $f'(0) = f'(\pi) = 0$. Let now $\sigma_n := 2\pi^{(n+1)/2}/\Gamma((n+1)/2)$ be the surface volume of the unit n-sphere and $\bar{\sigma}_n = \pi^{(n+1)/2}/\Gamma((n+3)/2)$ the volume of the unit ball. Then, by 31, $\mathfrak{m}^\phi = e$ and

$$
\begin{aligned}
\Sigma^\phi &= \mathbb{V}[\lambda^\phi(|\exp_e^{-1}(X)|)\exp_e^{-1}(X)] \\
&= \mathbb{E}[(\lambda^\phi(|\exp_e^{-1}(X)|))^2(\exp_e^{-1}(X))^T\exp_e^{-1}(X)] \\
&= \int_0^\pi \sigma_{n-1}\sin^{n-1}(r)f(r)\phi'^2(r)\fint_{S^{n-1}} y^T y \mathrm{d}y\mathrm{d}r \\
&= \int_0^\pi \sigma_{n-1}\sin^{n-1}(r)f(r)\phi'^2(r)I\fint_{S^{n-1}} y_1^2 \mathrm{d}y\mathrm{d}r \\
&= \int_0^\pi \sigma_{n-1}\sin^{n-1}(r)f(r)\phi'^2(r)I\int_0^{\pi/2} 2\sigma_{n-2}\cos^{n-2}(t)\sin^2(t)\mathrm{d}t\mathrm{d}r \\
&= \int_0^\pi \sigma_{n-1}\sin^{n-1}(r)f(r)\phi'^2(r)\frac{1}{n}I\mathrm{d}r = \frac{1}{n}\mathbb{E}[\phi'^2(\rho(e,X))]I.
\end{aligned}
$$

Where $\fint f(x)\mathrm{d}(x) := \int f(x)\mathrm{d}(x)/\int \mathrm{d}(x)$. In spheres the cut locus is always just a single point, and H_e is empty for $n > 1$ and it is equal to $-e$ when $n = 1$. Thus, $\Lambda^2 \equiv 0$ if $n > 1$ and when $n = 1$, $\kappa = 2$, $\tau' \equiv 1$ and thus we have $\Lambda^2 = -2\phi'(\pi)f(\pi)$. As for Λ^1, we can caluclate in general that (if $r = \rho(e,X)$)

$$
A_e(X) : v \mapsto \phi'(r)\cot(r)v + \phi''(r)(1-\phi'(r)\cot(r))\langle v, \frac{\exp_e^{-1}(X)}{r}\rangle\frac{\exp_e^{-1}(X)}{r}
$$

Thus, assuming f to be differentiable,

$$\begin{aligned}
\Lambda^1 : v \mapsto & \mathbb{E}[A_e(X)(v)] \\
=& \mathbb{E}\left[\phi'(r)\cot(r)v + \frac{(1-\phi'(r)\cot(r))\phi''(r)}{r^2}\langle v, \exp_e^{-1}(X)\rangle \exp_e^{-1}(X)\right] \\
=& v\int_0^{\pi} \sigma_{n-1}\phi'(r)\cos(r)\sin^{n-2}(r)\mathrm{d}r \\
& + \int_0^{\pi} \sigma_{n-1}\phi''(r)(1-\phi'(r)\cot(r))\sin^{n-1}(r) \fint_{S^{n-1}} \langle x, v\rangle x \mathrm{d}x\mathrm{d}r \\
&= v\mathbb{E}\left[\phi'(r)\cot(r) + \frac{1}{n}\phi''(r)(1-\phi'(r)\cot(r))\right] \\
&= v\frac{\sigma_{n-1}}{n}\int_0^{\pi} \left(\phi''(r)\sin^{n-1}(r) + \phi'(r)(n-1)\cos(r)\sin^{n-2}(r)\right) f(r)\mathrm{d}r \\
&= -v\frac{\sigma_{n-1}}{n}\int_0^{\pi} \phi'(r)\sin^{n-1}(r)f'(r)\mathrm{d}r
\end{aligned}$$

Which is finite and strictly positive as long as f is continuous, non-increasing and non-constant. Thus, if $n > 1$ we conclude that on S^n,

$$\sqrt{n}\exp_e^{-1}(\mathfrak{m}_n) \xrightarrow{w} \mathcal{N}_n\left(0, \frac{n}{\sigma_{n-1}}\frac{\int_0^{\pi}\phi'^2(r)\sin^{n-1}(r)f(r)\mathrm{d}r}{\left(\int_0^{\pi}\phi'(r)\sin^{n-1}(r)f'(r)\mathrm{d}r\right)^2}\right)$$

And finally, if $n = 1$, we get the following CLT on the circle S^1

$$\sqrt{n}\exp_e^{-1}(\mathfrak{m}_n) \xrightarrow{w} \mathcal{N}\left(0, \frac{1}{2}\frac{\int_0^{\pi}\phi'^2(r)f(r)\mathrm{d}r}{\left(\int_0^{\pi}\phi'(r)f'(r)\mathrm{d}r + \phi'(\pi)f(\pi)\right)^2}\right).$$

Notice that the term $2\phi'(\pi)f(\pi)$ is the dampening effect due to the presence of some mass on $\mathrm{H}_{\mathfrak{m}}$. This phenomenon has been described for the first time (in the case of $\phi(t) = t^2$) in Eltzner, 2020 and named smeariness.

5.1.4 Real Projective Spaces

The case of Real projective spaces is similar to spheres as these are quotients of spheres by the antipodal map. In particular,

$$\Sigma^{\phi} = I\frac{1}{n}\int_0^{\pi/2} \sigma_{n-1}\sin^{n-1}(r)f(r)\phi'^2(r)\mathrm{d}r.$$

Similarly,

$$\Lambda^1 = -I\frac{1}{n}\int_0^{\pi/2} \sigma_{n-1}\sin^{n-1}(r)f'(r)\phi'(r)\mathrm{d}r.$$

In this case, however, the cut locus has codimension 1 and Λ^2 does not vanish. Indeed,

$$\Lambda^2 = -\int_{S_+^{n-1}} 2\phi'(\pi/2)f(\pi/2)y^T y\mathrm{d}y = -\bar{\sigma}_{n-1}\phi'(\pi/2)f(\pi/2) = -\frac{\sigma_{n-1}}{n}\phi'(\pi/2)f(\pi/2).$$

Thus, in this case we obtain the following CLT,

$$\sqrt{n}\exp_e^{-1}(\mathfrak{m}_n) \xrightarrow{w} \mathcal{N}\left(0, \frac{n}{\sigma_{n-1}}\frac{\int_0^{\pi/2}\phi'^2(r)\sin^{n-1}(r)f(r)\mathrm{d}r}{\left(\int_0^{\pi/2}\phi'(r)\sin^{n-1}(r)f'(r)\mathrm{d}r + \phi'(\pi/2)f(\pi/2)\right)^2}\right)$$

5.1.5 A useful example: Rotations in 3 dimensions

We consider SO_3 as the Lie group of tridimensional special orthogonal matrices (i.e. 3×3 matrices x with positive determinant such that $x^T x = xx^T = I$) equipped with its bi-invariant Riemannian metric (corresponding, up to a constant, to the Frobenius inner product on $\mathfrak{so}_3$) which we normalize so that $\mathrm{diam}(\mathrm{SO}_3) = \pi$. Notice that under this normalization we have that $\rho(x,y) = \|\exp_I^{-1}(x^{-1}y)\|_F/\sqrt{2}$. This manifold is isometric to the Real Projective Space $\mathbb{RP}^3$. In particular the cut locus C_I of the identity is the geodesic sphere $\{x \in \mathrm{SO}_3 : \rho(I,x) = \pi\}$ which is the same as the set of all rotations by π around any axis. This is a two dimensional embedded submanifold isometric to $\mathbb{RP}^2$. Let μ be a probability measure with density $\tilde{\mu}(x) = f(\rho(I,x))$. In order to integrate such probability measures we need to know the area of the geodesic spheres $S^2_{\mathrm{SO}_3}(r) = \{x \in \mathrm{SO}_3 : \rho(I,x) = r\}$ which is given by $A_r = 4\pi\sin^2(r)$. Given the formula for the area of geodesic spheres in SO_3 we can deduce the radial part of the Laplacian from which we recover the heat equation. In general, if A_r is the area of a geodesic sphere at distance r, then

$$\Delta_r = \partial_r^2 + \frac{\partial_r A_r}{A_r}\partial_r \overset{\text{in }\mathrm{SO}_3}{=} \partial_r^2 + 2\cot(r)\partial_r$$

So that the heat equation in SO_3 for a radially symmetric function is $\partial_t = \partial_r^2 + 2\cot(r)\partial_r$, whose solution is given by

$$f(r,t) = \frac{1}{2\pi^2}\sum_{k=1}^{\infty} k\frac{\sin(kx)}{\sin(x)}e^{-t(k^2-1)}$$

This can also be derived (with some effort) from Proposition 3.6 of Avramidi, 2023. This function reaches it maximum at $r = 0$ and its minimum at $r = \pi$, where

$$f(0,t) = \frac{1}{2\pi^2}\sum_{k=1}^{\infty} k^2 e^{-t(k^2-1)}, \qquad f(\pi,t) = \frac{1}{2\pi^2}\sum_{k=1}^{\infty}(-1)^{k+1}k^2 e^{-t(k^2-1)}.$$

This is properly normalized, indeed

$$\begin{aligned}\int_{\mathsf{SO}_3} \tilde{\mu}_t(x)\mathrm{dvol}(x) &= \int_0^{\pi} 4\pi\sin^2(x)\tilde{\mu}_t(x)\mathrm{d}x \\ &= \frac{2}{\pi}\sum_{k=1}^{\infty} ke^{-t(k^2-1)}\int_0^{\pi}\sin(kx)\sin(x)\mathrm{d}x \\ &= \sum_{k=1}^{\infty} ke^{-t(k^2-1)}\delta_{k,1} = 1.\end{aligned}$$

Moreover,

$$
\begin{aligned}
\mathbb{E}[|x|] = \int_{\mathrm{SO}_3} |x|\tilde{\mu}_t(x)\mathrm{dvol}(x) &= \int_0^\pi 4\pi x \sin^2(x)\tilde{\mu}_t(x)\mathrm{d}x \\
&= \frac{2}{\pi}\sum_{k=1}^{\infty} k e^{-t(k^2-1)} \int_0^\pi x\sin(kx)\sin(x)\mathrm{d}x \\
&= \frac{2}{\pi}\sum_{k=1}^{\infty} k e^{-t(k^2-1)} \left(\frac{\pi^2}{4}\delta_{k,1} - \frac{4k}{(k^2-1)^2}\delta_{keven}\right) \\
&= \frac{\pi}{2} - \frac{2}{\pi}\sum_{k=1}^{\infty} \frac{16k^2}{(4k^2-1)^2} e^{-t(4k^2-1)}, \\
\mathbb{E}[x^2] = \int_{\mathrm{SO}_3} x^2\tilde{\mu}_t(x)\mathrm{dvol}(x) &= \int_0^\pi 4\pi x^2 \sin^2(x)\tilde{\mu}_t(x)\mathrm{d}x \\
&= \frac{2}{\pi}\sum_{k=1}^{\infty} k e^{-t(k^2-1)} \int_0^\pi x^2\sin(kx)\sin(x)\mathrm{d}x \\
&= \frac{2}{\pi}\sum_{k=1}^{\infty} k e^{-t(k^2-1)} \left(\frac{2\pi^3-3\pi}{12}\delta_{k,1} - \frac{4(-1)^k\pi k}{(k^2-1)^2}\delta_{keven}\right) \\
&= \frac{2\pi^2-3}{6} - \sum_{k=2}^{\infty} \frac{8(-1)^k k^2}{(k^2-1)^2} e^{-t(k^2-1)}.
\end{aligned}
$$

We can compute Λ^2 in general on $\mathbb{RP}^n$. We have that

$$
\Lambda^2 = -\pi^{3n/2-1}\phi'(\pi)\frac{n-1}{\Gamma(n/2+1)} \overset{n=3}{=} -\frac{8}{3}\pi^3\phi'(\pi)\tilde{\mu}_t(\pi) \overset{\phi'(r)=1}{=} -\frac{4}{3}\pi\sum_{k=1}^{\infty}(-1)^{k+1}k^2 e^{-t(k^2-1)}.
$$

As for Λ^1, we have that if $v \in \mathrm{T}_I\mathbb{RP}^n$ and $y \neq I$, then v can be decomposed into a component parallel to $\exp_I^{-1}(y)$ and one orthogonal to it. Then we find easily that

$$
A_I(y) = \phi'(r)\frac{\cos(r)}{\sin(r)} + \phi''(r)\frac{}{r^2}\langle v, \rangle
$$

When integrating this with a radially symmetric density, the second term vanishes and we are left with

$$\Lambda^1 = 4\pi \int_0^\pi \phi'(r)\cos(r)\sin(r)\tilde{\mu}(r)\mathrm{d}r$$

$$= \frac{2}{\pi}\sum_{k=1}^{\infty} ke^{-t(k^2-1)} \int_0^\pi \phi'(r)\sin(kr)\cos(r)\mathrm{d}r$$

$$\overset{\phi'(r)=1}{=} \frac{2}{\pi}\sum_{k=1}^{\infty} \frac{8k^2}{4k^2-1} e^{-t(4k^2-1)}$$

5.2 Computational Results

We start this section with a powerful result that removes the need to perform any computations in a vast class of interesting cases. In particular, we will now prove that on any symmetric space if the probability measure is isotropic around a point z and decreasing, then all the well defined ϕ-means coincide with this unique point $\hat{x}$. This result is quite intuitive, but perhaps surprisingly it fails completely when the hypothesis of symmetry is removed.

5.2.1 Isotropic Probabilities on Symmetric Spaces

In this section, we will restrict ourselves to symmetric spaces. We also consider absolute continuous measures with isotropic decreasing density, that is measures of the form $\mu_{f,\hat{x}}(\mathrm{d}x) = f(\rho(x,\hat{x}))\mathrm{vol}(\mathrm{d}x)$ for some point $\hat{x} \in \mathcal{X}$ and f, is any non-constant, decreasing functions from $[0,\infty)$ to itself.

Definition 30. *A complete Riemannian Manifold $\mathcal{X}$ is said to be* ***symmetric*** *if for any two distinct points $x, y \in \mathcal{X}$, there is an isometry $R_{x,y}$ exchanging x and y of period two (meaning that $R_{x,y}^2 = \mathrm{id}$).*

Theorem 31. *If f is non-negative, decreasing and non constant and $\mathcal{X}$ is symmetric, then for every $\hat{x} \in X$, the measure $\mu(\mathrm{d}x) \propto f(\rho(x,\hat{x}))\mathrm{vol}(\mathrm{d}x)$ has ϕ-Fréchet mean $\hat{x}$ for every ϕ for which F_μ^ϕ is finite.*

Proof. Let $y \in X \backslash \{\hat{x}\}$, then we have that $X^+ := \{x \in X : \rho(x,\hat{x}) > \rho(x,y)\}$ is congruent with $X^- := \{x \in X : \rho(x,\hat{x}) < \rho(x,y)\}$ because the isometry R exchanging $\hat{x}$ and y also exchanges these two sets. In particular, any point $x \in X^+$ with $\phi(\rho(x,\hat{x})) - \phi(\rho(x,y)) > 0$ is sent to a point Rx with $\phi(\rho(Rx,\hat{x})) - \phi(\rho(Rx,y)) = -[\phi(\rho(x,\hat{x})) - \phi(\rho(x,y))] < 0$. Moreover, we have that for any $x \in X^+$, $\rho(x,\hat{x}) < \rho(Rx,\hat{x})$, thus

$$\begin{aligned}
F_\mu^\phi(\hat{x}) - F_\mu^\phi(y) &= \int_X (\phi(\rho(x,\hat{x})) - \phi(\rho(x,y)))f(\rho(x,\hat{x}))\mathrm{vol}(\mathrm{d}x) \\
&= \int_{X^+} (\phi(\rho(x,\hat{x})) - \phi(\rho(x,y)))f(\rho(x,\hat{x}))\mathrm{vol}(\mathrm{d}x) + \\
&\quad + \int_{X^-} (\phi(\rho(x',\hat{x})) - \phi(\rho(x',y)))f(\rho(x',\hat{x}))\mathrm{vol}(\mathrm{d}x') \\
&= \int_{X^+} (\phi(\rho(x,\hat{x})) - \phi(\rho(x,y)))f(\rho(x,\hat{x}))\mathrm{vol}(\mathrm{d}x) + \\
&\quad + \int_{X^+} (\phi(\rho(Rx,\hat{x})) - \phi(\rho(Rx,y)))f(\rho(Rx,\hat{x}))\mathrm{vol}(\mathrm{d}x) \\
&= \int_{X^+} [\phi(\rho(x,\hat{x})) - \phi(\rho(x,y))][f(\rho(x,\hat{x})) - f(\rho(Rx,\hat{x}))]\mathrm{vol}(\mathrm{d}x) < 0
\end{aligned}$$

Where the last inequality holds because of the fact that $\mathrm{vol}(X^+) = \mathrm{vol}(X)/2 > 0$. This proves that $\hat{x}$ is the global minimum of F^ϕ. □

Symmetric spaces have been completely classified in Cartan, 1926, Cartan, 1927, and they include the spheres, the real, complex, and quaternionic projective spaces, and all the compact semi-simple Lie Groups. Moreover, the family of symmetric spaces is closed under Cartesian products. Thus, for instance, the torus T^n belongs to the category.

Thus our result is useful in dealing with the population ϕ-means but not with the sample ϕ-mean. For this we considered the algorithms described in the next subsections.

5.2.2 Newton Method on nested grids

This method is extremely fast and precise on low dimensional manifolds but becomes unpractical when the dimension increases. We used when dealing with the line, the circle, the plane, and the sphere.

We illustrate it in the case of the plane. In the following code, X is a matrix with the samples.

```
fre=function(X){
```

```
  n=nrow(X); h=1/10000
  loss=function(x){
    o=0
    for(j in 1:n){
      o=o+phi(sqrt((x[1]-X[j,1])^2+(x[2]-X[j,2])^2))/n
    }
    return(o)
  }
  x0=y0=0
  for(s in 1:4){
    L=xx=yy=c()
    for(i in 1:100){
      x=x0+(i-50)/100^s
      for(j in 1:100){
        y=y0+(j-50)/100^s
        xx=c(xx,x); yy=c(yy,y)
        L=c(L,loss(c(x,y)))
      }
    }
    x0=xx[which.min(L)]; y0=yy[which.min(L)]; L=xx=yy=c()
  }
  return(c(x0,y0))
}
```

This algorithm at each step produces two additional digits of the ϕ-mean by exploring the loss function on an increasingly small grid in the neighborhood of the previously found minimum.

5.2.3 Gradient descent

This is a standard algorithm and its performance is widely studied. In particular, its convergence to the global minimum is not certain and in any case, it is not able to provide the entire set of minima in the case in which there are multiple. In our context the performance of this algorithm is optimal in the case of moderately convex functions like $\phi : t \mapsto t^2$, but it is not usable when ϕ is flat (like $\phi : t \mapsto |t|$) or when it is too convex (like when $\phi : t \mapsto t^{10}$). In the first case, the gradient of F^ϕ is constant in magnitude and thus the estimates constantly orbit around the global minimum without ever reaching it; in the second case, the convergence eventually occurs but becomes exceedingly slow at points close to the minimum. These problems are known, and there are some solutions in the literature.

5.2.4 Tangent plane flipping

This is the most promising algorithm that we considered and we used it on spheres and Projective Planes. It relies on the Lemma we prove below which guarantees that at a local minimum x of F^{ϕ}, the pushforward of μ on $\mathrm{T}_x\mathcal{X}$ though the exponential map has ϕ-mean equal to zero. There are two main drawback of this algorithm. Firstly it requires to compute the log and exponential maps at several points on the manifold which is doable only when these maps are known in closed form (which they are on symmetric spaces). Secondly, it requires to compute the ϕ-mean on the tangent plane, which is straightforward only when $\phi : t \mapsto t^2$ (in which case $\mathfrak{m}_n^{\phi}$ is the classical mean) but in general requires the appeal to one of the previous algorithms.

Inputs: The sample $(X_j)_{j=1}^n$ and an initial guess m_1
Set $k = 1$ and $d = 1$
while $d > \varepsilon$ **do**
  $v_k \leftarrow \arg\min_{v \in T_{m_k}\mathcal{X}} \frac{1}{n}\sum_{j=1}^n \phi(\|v - \exp_{m_k}^{-1}(X_j)\|_{m_k})$
  $d \leftarrow \|v_k\|_{m_k}$
  $m_{k+1} \leftarrow \exp_{m_k}(v_k)$
  $k = k + 1$
end

Algorithm 1: Tangent flipping algorithm

Lemma 32. *Let μ be a probability measure supported on $\mathcal{X}$ such that $\mu(\bigcup_{q \in \mathfrak{m}^{\phi}(\mu)} C_q) = 0$, then, if $\tilde{\mu}_p$ denotes the pushfoward of μ through the map $\exp_p$ in $V_p \subseteq T_p\mathcal{X}$, we have that,*

$$p \in \mathfrak{m}^{\phi}(\mu) \qquad \textit{only if} \qquad 0 \in \arg\min_{v \in T_p\mathcal{X}} \int_{T_p\mathcal{X}} \phi(\|v - w\|_g)\tilde{\mu}_p(\mathrm{d}w)$$

Proof. By definition a point $p \in \mathfrak{m}^{\phi}(\mu)$ iff $p \in \arg\min_{q \in \mathcal{X}} F_{\mu}^{\phi}(q)$. If this is the case, then we must have that moving away from p in any possible direction does not decrease F_{μ}^{ϕ}. That is, if $p \in \mathfrak{m}^{\phi}(\mu)$, we must have that

$$\forall v \in T_p\mathcal{X}, \quad \partial_t^+\big|_0 F_{\mu}^{\phi}(\exp_p(tv)) \geqslant 0$$

Where by $\partial_t^+\big|_0 f(t)$, we mean $\lim_{h \downarrow 0}(f(h) - f(0))/h$. Notice that this quantity is well defined and finite because of lemma ?. But we, remembering that that the exponential map is a radial isometry (so that $\rho(p, \exp_p(w)) = \|w\|$) we get that

$$
\begin{aligned}
0 \leqslant \partial_t^+|_0 F_\mu^\phi(\exp_p(tv)) &= \partial_t^+\big|_0 \int_{N_p} \phi(\rho(x, \exp_p(tv)))\mu(\mathrm{d}x) \\
&= \int_{N_p} \partial_t^+\big|_0 \, \phi(\rho(x, \exp_p(tv)))\mu(\mathrm{d}x) \\
&= \int_{N_p} \phi'_{\mathrm{sign}(\,\partial_t^+|_0 \rho(x, \exp_p(tv)))}(\rho(x,p)) \ \partial_t|_0 \, \rho(x, \exp_p(tv))\mu(\mathrm{d}x) \\
&= \|v\| \int_{N_p} \phi'_{\mathrm{sign}(\cos(\theta_{v, \exp_p^{-1}(x)}))}(\rho(x,p)) \cos(\theta_{v, \exp_p^{-1}(x)})\mu(\mathrm{d}x) \\
&= \|v\| \int_{V_p} \phi'_{\mathrm{sign}(\cos(\theta_{v,w}))}(\rho(\exp_p(w), \exp_p(0))) \cos(\theta_{v,w})\tilde{\mu}_p(\mathrm{d}w) \\
&= \int_{V_p} \phi'_{\mathrm{sign}(\langle v, w \rangle)}(\|w\|) \frac{\langle v, w \rangle}{\|w\|} \tilde{\mu}_p(\mathrm{d}w) \\
&= \int_{V_p} \phi'_{\mathrm{sign}(\langle v, w \rangle)}(\|w\|) \ \partial_t|_0 \, (\|tv - w\|)\tilde{\mu}_p(\mathrm{d}w) \\
&= \int_{V_p} \partial_t^+\big|_0 \, \phi(\|tv - w\|)\tilde{\mu}_p(\mathrm{d}w) = \partial_t^+\big|_0 \, \tilde{F}_{\tilde{\mu}_p}^\phi(0 + tv)
\end{aligned}
$$

But now notice that $\tilde{F}_{\tilde{\mu}_p}^\phi : v \mapsto \int_{T_p\mathcal{X}} \phi(\|v - w\|_g)\tilde{\mu}_p(\mathrm{d}w)$ is a convex function defined on a linear space. Thus $0 \in \arg\min \tilde{F}_{\tilde{\mu}_p}^\phi$ precisely iff $\partial_t^+\big|_0 \, \tilde{F}_{\tilde{\mu}_p}^\phi(0 + tv) \geqslant 0$ $\square$

6. A discrete application

6.1 Introduction

Analyzing how individual preferences differ across a set of alternatives is crucial for many fields, including economics, political science, and computer science. For instance, researchers often run experiments to elicit participants' preferences and then assess the dissimilarity in their choices to understand the variability or polarizations of preferences in the aggregate. Similarly, in modern recommender systems, assessing the difference in users' past clicking behaviors on movies or shopping items is a common practice to improve future recommendations.

In all these contexts it is often needed to find ways to aggregate preferences and possibly to cluster them around some "central" rankings. This is particularly useful for profiling users.

Once a distance d is defined on the space of rankings over a finite number of objects $C = [n]$, a natural way to produce a "central ranking" is to consider a ϕ-mean. In particular, in the literature, the median (corresponding to $\phi : t \mapsto t$) is frequently considered.

In what follows given a permutation σ, the k-th element $\sigma_k \in C$ is interpreted as the k-th preferred option by σ. If A is some menu of alternatives (that is, a subset of C), we denote by $M_A(\sigma)$ the most preferred element by σ over A.

The pioneering works Kendall, 1938 and Kemeny, 1959, proposed and axiomatized a distance function, known as the Kendall metric, defined over the space of permutations (equivalently, rankings). In particular, for some

The Kendall metric assesses the dissimilarity between two rankings by counting the number of times they have different favorite elements over all possible menus of size two. Formally,

$$\mathrm{d}_{\mathrm{K}}^{|\cdot|}(\sigma, \pi) := \sum_{\{i,j\} \subseteq C} \mathbf{1}(M_{\{i,j\}}(\sigma) \neq M_{\{i,j\}}(\pi)). \tag{6.1}$$

This metric has been employed successfully in several contexts. For instance, when used to retrieve a median (*consensus*) ranking from a collection of individual preferences, it has been shown to induce desirable welfare properties, such as Pareto optimality and the Majority principle.

However, despite its popularity, by focusing solely on the total number of disagreements over menus of size two, the Kendall metric has several shortcomings. First, it gives the same importance to disagreements over top-ranked alternatives as those over lower-ranked alternatives. Moreover, it does not account for the identity of the alternatives involved. Consider the following rankings,

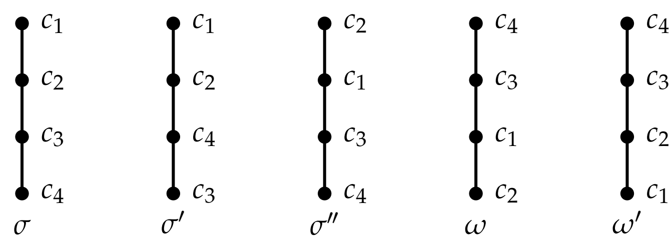

FIGURE 6.1: Rankings σ, σ', σ'', ω and ω'. Notice that from σ to σ' (and from ω to ω') the last two items are exchanged, while when going from σ to σ'' the first two items are exchanged.

6.1.0.0.1 Top-bottom trade-off. In Figure 6.1, observe that σ' is derived from σ by swapping the positions of its two least preferred elements, c_3 and c_4. Similarly, σ'' is obtained from σ by reversing the order of its two most preferred elements, c_1 and c_2. As such, σ and σ' differ only when faced with the menu of alternatives $\{c_3, c_4\}$, whereas σ and σ'' disagree on all four menus that contain both c_1 and c_2.[1] The greater number of disagreements in the second case naturally suggests that $\mathrm{d}(\sigma, \sigma') < \mathrm{d}(\sigma, \sigma'')$. However, this is not captured by the Kendall metric, which disregards the choices over menus with sizes larger than two and concludes that the two distances are identical.

6.1.0.0.2 Heterogeneous alternatives. Even if two couples of rankings disagree over the same number of menus, their respective choices may differ. For example, in Figure 6.1, σ and σ' differ only over the menu $\{c_3, c_4\}$, just as ω and ω' differ solely over $\{c_1, c_2\}$. Therefore,

[1] Specifically, they disagree over the menus $\{c_1, c_2\}$, $\{c_1, c_2, c_3\}$, $\{c_1, c_2, c_4\}$, and $\{c_1, c_2, c_3, c_4\}$.

their Kendall distances are equal. Yet, if the options in C have distinct characteristics—such as movies from different genres or products with different prices—it can be inappropriate to see these rankings as equally distant.

These limitations recently led to the development of the *top-difference* distance in Nishimura and Ok, 2023, where the specular point of view is taken. While the Kendall metric focuses exclusively on menus of size two, the metric of Nishimura and Ok, 2023 counts the total number of disagreements between two preferences across *all* possible menus of alternatives.[2] Formally,

$$\mathbf{d_1}(\sigma, \pi) := \sum_{A \subseteq C} | \triangle_A (\sigma, \pi)|, \tag{6.2}$$

where $\triangle_A(\sigma, \pi) := M_A(\sigma) \triangle M_A(\pi)$ and $\triangle$ denotes the symmetric difference between sets. By taking into account menus larger than two, the top-difference distance agrees with our expectations in the example of Figure 6.1, concluding that $\mathbf{d_1}(\sigma, \sigma') < \mathbf{d_1}(\sigma, \sigma'')$. Moreover, when the cardinality $|\cdot|$ in (6.2) is replaced by a general measure μ over C, the resulting distance, denoted by $\mathbf{d}_\mathbf{1}^\mu$, introduces some degree of heterogeneity among the alternatives, since in these cases their identity matters for computing $\mathbf{d}_\mathbf{1}^\mu$.

While providing some remedy to the main limitations of the classical Kendall metric, some aspects of the top-difference idea require further scrutiny. First, the underlying assumption behind the top-difference metric is that all menus are *equally important*. As long as the total number of menus with disagreements remains the same, the size of these menus is irrelevant.

[2] More precisely, it returns twice the total number of disagreements over all menus.

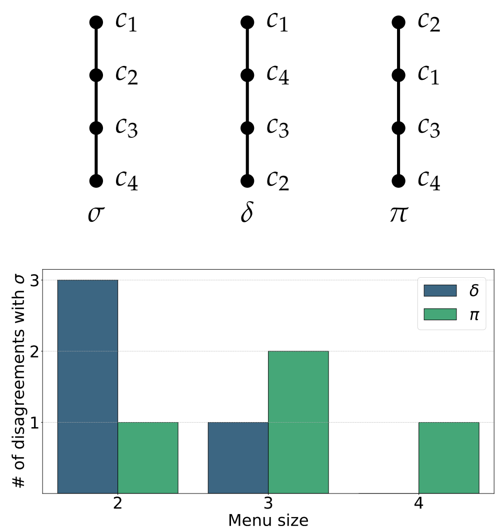

FIGURE 6.2: Rankings σ, δ, and π, together with a plot summarizing their number of disagreements relative to σ.

In Figure 6.2, the top-difference metric determines that $\mathbf{d_1}(\sigma, \pi) = \mathbf{d_1}(\sigma, \delta)$, as both rankings π and δ disagree with σ over four menus. However, the right plot illustrates that the distribution of these disagreements differs significantly. Specifically, while the disagreements between σ and π are more evenly distributed across menu sizes, those between σ and δ are mainly concentrated on smaller menus of size two, without involving the full set of alternatives.

Failing to account for the menus' size is undesirable in many applications, particularly in settings where the primary focus is on top-ranked candidates. Consider, for example, the task of aggregating user preferences over a subset of available movies of size n to generate a (ranked) recommendation list of length $k < n$. In this case, a natural approach is to focus solely on the top k elements of each ranking. This can be implemented by restricting attention to menus of size at least $n - k + 1$, as no candidate ranked below k can ever be the favorite element over these menus. The rigidity of the top-difference metric prevents such a specification and could bias our assessment by giving too much weight to small menus. Moving beyond the example, assigning greater weights to larger menus reflects the idea that the importance of disagreements grows as the menu size increases.

Second, while replacing the cardinality $|\cdot|$ in (6.2) with a general measure μ provides some degree of variability, we next argue that this modification does not capture the whole extent of heterogeneity, as it does not account for the relation between the alternatives (clustering). To illustrate this, consider a simple scenario where four candidates are running for an election: two from the Republican party and two from the Democratic party.

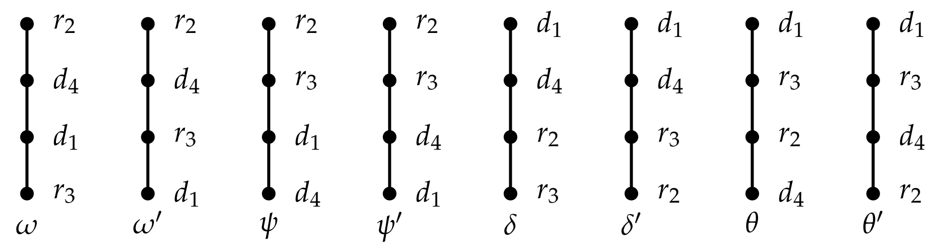

FIGURE 6.3: Voters' preferences over $C = \{d_1, r_2, r_2, d_4\}$: r_1, r_2 are Republican while d_1, d_4 are Democrats.

Observe that each ranking with a superscript is obtained from the same ranking without a superscript, by exchanging the order of the last two entries. Consider the comparison between the distances $\mathrm{d}(\omega, \omega')$ and $\mathrm{d}(\psi, \psi')$. It is reasonable to expect that $\mathrm{d}(\omega, \omega') > \mathrm{d}(\psi, \psi')$ since ψ and ψ' only differ for the ranking of the two democratic candidates, whereas ω and ω' disagree in their preference between a democratic and a republican. Similarly, when comparing $\mathrm{d}(\delta, \delta')$ with $\mathrm{d}(\theta, \theta')$, we would expect $\mathrm{d}(\delta, \delta') < \mathrm{d}(\theta, \theta')$, as r_2 and r_3 are both Republicans, while d_4 is a Democrat. However, satisfying these intuitive requirements using any μ-top-difference distance leads to a contradiction. Indeed, we would need

$$\mathrm{d}_{\mathbf{1}}^{\mu}(\omega, \omega') = \mu(d_1) + \mu(r_3) > \mu(d_1) + \mu(d_4) = \mathrm{d}_{\mathbf{1}}^{\mu}(\psi, \psi')$$
$$\mathrm{d}_{\mathbf{1}}^{\mu}(\delta, \delta') = \mu(r_2) + \mu(r_3) < \mu(r_2) + \mu(d_4) = \mathrm{d}_{\mathbf{1}}^{\mu}(\theta, \theta').$$

Motivated by the considerations above, in Aveni et al., 2024, we propose a new class of distances over rankings that can deal with the top-bottom trade-off and items' heterogeneity while maintaining desirable computational properties. This class extends and unifies several existing semimetrics, including the Kendall metric and the top-difference semi-metrics, within a single formulation (see §6.4).

6.2 Mathematical notation and preliminaries

Rankings.

We consider a set C of $n \geqslant 3$ elements which we interpret as items or candidates. Our investigation revolves around rankings over C represented as lists that arrange candidates based on preference, with the foremost candidate being the most favored and the ultimate candidate the least favored. We arbitrarily label the candidates with the numbers $[n] := \{1, \ldots, n\}$, so that a ranking is represented by a *permutation*; we use the terms permutations and rankings interchangeably. By a permutation we mean a bijective function $\sigma : [n] \to [n]$, with σ_i denoting the candidate in the i^{th}-position according to σ. Therefore, the set of all rankings over $[n]$, is the symmetric group denoted by $\mathbb{S}_n$. Often, we write permutations as ordered lists, e.g., $\sigma = (\sigma_1, \ldots, \sigma_n)$. For all permutations $\sigma \in \mathbb{S}_n$ and alternatives $i, j \in [n]$, we write $i >_\sigma j$ if σ ranks item i before j, that is, $\sigma_i^{-1} < \sigma_j^{-1}$.[3] The identity permutation $(1, \ldots, n)$ is denoted by id. Often we employ the group structure of $\mathbb{S}_n$, and we adopt the convention that for two permutations $\sigma, \pi \in \mathbb{S}_n$, $\sigma\pi : i \mapsto \sigma_{\pi_i}$. Particularly relevant for our work are *transpositions*. For distinct $i, j \in [n]$, we denote by $t_{i,j}$ the i,j-transposition, that is, the permutation that agrees with the identity id, but swaps the items i and j; formally, $t_{i,j} = (1, \ldots, i-1, j, i+1, \ldots, j-1, i, j+1, \ldots, n)$. Whenever $|i-j| = 1$, we refer to $t_{i,j}$ as *adjacent*. Note that, in general, for $\sigma \in \mathbb{S}_n$, we have $\sigma t_{i,j} \neq t_{i,j}\sigma$. While $\sigma t_{i,j}$ is the permutation obtained from σ by exchanging σ_i with σ_j, $t_{i,j}\sigma$ is obtained from σ by exchanging the items i and j.

Given a ranking $\sigma \in \mathbb{S}_n$ and an alternative $j \in [n]$, we denote by $j^{\uparrow,\sigma}$ the collection of alternatives preferred to j by σ, and $j^{\downarrow,\sigma}$ is defined analogously. More precisely,

$$j^{\uparrow,\sigma} := \{i \in [n] : i >_\sigma j\}, \qquad j^{\downarrow,\sigma} := \{i \in [n] : i <_\sigma j\}.$$

We denote by $\mathcal{P}$ the collection of subsets of $[n]$ with cardinality larger or equal than two,

[3] The symbol $<_\sigma$ has the analogous meaning.

and occasionally we refer to members of $\mathcal{P}$ as *menus*. Given a set $S \in \mathcal{P}$ and a ranking $\sigma \in \mathbb{S}_n$, the best element in S with respect to σ is written as $M_S(\sigma)$.[4] We often refer to $M_S(\sigma)$ as the choice induced by σ over S, for all $S \in \mathcal{P}$ and $\sigma \in \mathbb{S}_n$. For any ranking σ, we denote by σ^* its antipodal permutation, that is, the ranking obtained by reversing the order of all items in σ.

Distances and (semi)metrics over rankings.

We study distances and (semi)metrics over the space of rankings $\mathbb{S}_n$. A function $\mathrm{d} : \mathbb{S}_n \times \mathbb{S}_n \to \mathbf{R}_+$ is said to be a *distance* if $\mathrm{d}(x,x) = 0$ and $\mathrm{d}(x,y) = \mathrm{d}(y,x)$ for all $x, y \in \mathbb{S}_n$. A distance d is a *semimetric* if it satisfies the triangle inequality, that is, $\mathrm{d}(x,z) \leqslant \mathrm{d}(x,y) + \mathrm{d}(y,z)$ for all x, y, z, while it is a *metric* if it is a semimetric satisfying $x = y$ whenever $\mathrm{d}(x,y) = 0$. For any couple of rankings σ and π, the classical *Kendall metric*, denoted by $\mathrm{d}_K(\sigma,\pi)$, counts the minimum number of adjacent transpositions necessary to transform σ into π. The set of all *paths* of adjacent transpositions that transform σ into π is denoted by

$$T(\sigma,\pi) := \left\{ (t_{a_k,a_k+1})_{k=1}^{\mathrm{d}_K(\sigma,\pi)} : \pi = \sigma \prod_{k=1}^{\mathrm{d}_K(\sigma,\pi)} t_{a_k,a_k+1} \right\}.$$

In particular, given a path, observe that its 1^{st} transposition t_{a_1,a_1+1} swaps in σ the items in positions a_1 and a_1+1, and, for $k > 1$, its k^{th} transposition swaps in $\sigma \prod_{j=1}^{k-1} t_{a_j,a_j+1}$ the items in positions a_k and a_k+1. Similarly, for each k, we write i_k and j_k to refer to the unique items that are exchanged by the k^{th} transposition along the chosen path. When we want to make explicit the dependence on the chosen path $\mathbf{t} \in T(\sigma,\pi)$, we write instead $a_k(\mathbf{t}), i_k(\mathbf{t})$, and $j_k(\mathbf{t})$.

Finally, a distance over rankings $\mathrm{d} : \mathbb{S}_n \times \mathbb{S}_n \to \mathbf{R}_+$ is said to be a *graphic distance* if for all σ, π, there exists a path from σ to π, i.e., $\mathbf{t} \in T(\sigma,\pi)$, such that $\mathrm{d}(\sigma,\pi)$ can be

[4] Formally, $M_S(\sigma)$ is the σ-maximum element over S, that is, $M_S(\sigma) >_\sigma i$ for all $i \in S \backslash M_S(\sigma)$ and $S \in \mathcal{P}$.

decomposed as

$$\mathrm{d}(\sigma,\pi)=\sum_{j=1}^{\mathrm{d}_K(\sigma,\pi)}\mathrm{d}\left(\sigma\prod_{k=1}^{j-1}t_{a_k(\mathbf{t}),a_k(\mathbf{t})+1},\sigma\prod_{k=1}^{j}t_{a_k(\mathbf{t}),a_k(\mathbf{t})+1}\right). \tag{6.3}$$

When d is a semimetric and a graphic distance, it is termed *graphic semimetric* and, by the triangle inequality, $\mathrm{d}(\sigma,\pi)$ equals the minimum value of the right-hand side of (6.3) taken over all $\mathbf{t}\in T(\sigma,\pi)$.

Parameters.

As illustrated in the introduction, the distances we study depend on two parameters: a vector of weights $\boldsymbol{\beta}=(\beta_2,\ldots,\beta_n)\in\mathbf{R}^{n-1}$ and a function $g:[n]\times[n]\to\mathbf{R}$. The functions we consider are symmetric, $g_{ij}=g_{ji}$, and vanish on the diagonal, $g_{ii}=0$. We denote the collection of such functions by $\mathcal{G}$ and refer to its elements as *dissimilarity functions*. We also denote by $\mathcal{G}^{\mathrm{s}}$ ($\mathcal{G}^{\mathrm{m}}$) the sub-collections of $\mathcal{G}$ that are non-trivially zero and positive (strictly positive) outside the diagonal. The following sub-collections are employed to determine when $\rho_{\boldsymbol{\beta}}^{g}$ is a (semi)metric,

$$\mathcal{G}^{\mathrm{s}}_{\rhd}=\{g\in\mathcal{G}^{\mathrm{s}}:g_{ij}+g_{jk}\geqslant g_{ik}\text{ for all distinct }i,j,k\}$$
$$\mathcal{G}^{\mathrm{m}}_{\rhd}=\{g\in\mathcal{G}^{\mathrm{m}}:g_{ij}+g_{jk}>g_{ik}\text{ for all distinct }i,j,k\}$$

Let $g\in\mathcal{G}^{\mathrm{s}}$, if there is some $\mu\in\mathbf{R}^n$ such that $g_{ij}=\mu_i+\mu_j$ for all distinct $i,j\in[n]$, we say that g is *additive*, and we refer to such μ as *representing function of* g. If $g\in\mathcal{G}^{\mathrm{s}}$ is additive, then the representing function μ is unique, and we denote by $\mathcal{M}^{\mathrm{s}}$ ($\mathcal{M}^{\mathrm{m}}$) the collection of all functions μ that represent the additive elements of $\mathcal{G}^{\mathrm{s}}$ ($\mathcal{G}^{\mathrm{m}}$). The collections $\mathcal{M}^{\mathrm{s}}_{\rhd}$ and $\mathcal{M}^{\mathrm{m}}_{\rhd}$ are defined analogously.[5]

[5] One can verify that

$$\mathcal{M}^{\mathrm{s}}=\{\mu\in\mathbf{R}^n\setminus\{\mathbf{0}\}:\mu_i+\mu_j\geqslant 0\text{ for all distinct }i,j\}\text{ and }\mathcal{M}^{\mathrm{s}}_{\rhd}=\mathbf{R}^n_{+}\setminus\{\mathbf{0}\}$$
$$\mathcal{M}^{\mathrm{m}}=\{\mu\in\mathbf{R}^n\setminus\{\mathbf{0}\}:\mu_i+\mu_j>0\text{ for all distinct }i,j\}\text{ and }\mathcal{M}^{\mathrm{m}}_{\rhd}=\mathbf{R}^n_{++}.$$

6.3 Definition, dual representation, and computability

Consider a dissimilarity function $g \in \mathcal{G}$ and some weights $\boldsymbol{\beta} = (\beta_i)_{i=2}^n \in \mathbf{R}^{n-1}$. If a distance d on $\mathbb{S}_n$ can be written as

$$\mathrm{d} : (\sigma, \pi) \mapsto \sum_{S \in \mathcal{P}} \beta_{|S|} g_{M_S(\sigma), M_S(\pi)}, \tag{6.4}$$

then it is termed $(\boldsymbol{\beta}, g)$-*distance*, and denoted by $\rho^g_{\boldsymbol{\beta}}$. By construction, both the size of each menu and the identity of the items matter to assess the difference between any couple of rankings. Different choices of the weights $\boldsymbol{\beta}$ and the dissimilarity function g induce a large class of distances, offering a unified framework to study distances over rankings.

For instance, by specifying $\boldsymbol{\beta} = (1, 0, \dots, 0)$, the comparison between rankings is restricted to binary menus, in the spirit of the classical Kendall distance. We denote such distances by d^g_K.[6] Contrarily, by choosing $\boldsymbol{\beta} = (1, 1, \dots, 1)$ we impose that each menu contributes equally to the overall disagreement. This approach aligns with the distances proposed by Nishimura and Ok, 2023, and we denote them by $\mathrm{d}^g_{\mathbf{1}}$.[7] By modulating the parameter $\boldsymbol{\beta}$, we can bridge these two cases allowing for scenarios where the menu sizes are weighted differently. By changing g one controls how the identity of the alternatives contributes to measuring the disagreement. If g is constant outside the diagonal, then all alternatives are treated symmetrically. In this case, the total disagreement reduces to a weighted count of the number of menus where the rankings disagree. To ease notation, in these cases, we suppress the dependence of $\rho^g_{\boldsymbol{\beta}}$ on g. When the function g is additive, the dissimilarity between two rankings is unaffected by the distinct groups to which the alternatives may belong, and only their individual characteristics are considered. In this case, we use as superscript the representing function μ, i.e., $\mathrm{d}^\mu_{\boldsymbol{\beta}}$. Instead, with a general g we can consider both the individual features of each alternative, as well as the clusters they belong to.

[6] When g is equal to one outside the main diagonal, $\mathrm{d}^g_K = 2\mathrm{d}_K$, but this is equivalent to d_K for the purposes of our analysis.

[7] The semimetric proposed by Nishimura and Ok, 2023 are obtained when $\boldsymbol{\beta} = (1, \dots, 1)$ and g is μ-additive with $\mu \in \mathbf{R}^n_+$.

We now illustrate how the degrees of freedom provided by the parameters $\boldsymbol{\beta}$ and g allow $\rho_{\boldsymbol{\beta}}^{g}$ to overcome the difficulties that previously proposed semimetrics face. In particular, revisiting the issues described in the introduction, we propose the following parametrizations.

Consider a context in which the analyst is interested in the difference two rankings express in their top k positions (e.g., Figure 6.2). A reasonable choice of the parameters would be to set $\beta_j = \mathbf{1}_{j \geqslant n-k+1}$. This ensures that swaps between items below position k are irrelevant. On the other hand, if the alternatives are clustered into groups (e.g., Figure 6.3), one may want to model that within-group disagreements are less important than across-group disagreements. A reasonable choice of g for such case is to select for some $h > \ell \geqslant 0$:

$$g_{ij} = \begin{cases} \ell & \text{if } i \text{ and } j \text{ belong to the same group} \\ h & \text{if } i \text{ and } j \text{ belong to different groups.} \end{cases}$$

A dual representation

According to their definition, our distances compare the choices each ranking makes over different menus. We next establish a dual representation that reveals the nature of $\rho_{\boldsymbol{\beta}}^{g}$ as graphical semimetrics, providing additional insights about their structure. Following Farnoud and Milenkovic, 2014, a possible way to quantify the disagreement between two rankings is as follows: assign weights (costs) to each adjacent transposition and take the shortest (i.e., minimum cost) path that transforms one ranking into the other. In particular, for some costs $\boldsymbol{\phi} \in \mathbf{R}_{+}^{n-1}$, the following semimetric is considered,

$$\mathrm{d}_{\boldsymbol{\phi}} : (\sigma, \pi) \mapsto \min_{\mathbf{t} \in T(\sigma,\pi)} \sum_{k=1}^{\mathrm{d}_K(\sigma,\pi)} \phi_{a_k(\mathbf{t})}, \tag{6.5}$$

where we recall that $T(\sigma, \pi)$ denotes the set of all *paths* $\mathbf{t}$ of adjacent transpositions transforming σ into π in $\mathrm{d}_K(\sigma, \pi)$ steps,[8] and $a_k(\mathbf{t})$ denotes the smallest position involved in the k^{th} transposition. Observe that if $\boldsymbol{\phi} \equiv 1$, then $\mathrm{d}_{\boldsymbol{\phi}}$ reduces to the classical Kendall metric

[8] It is known that $T(\sigma, \pi)$ is not empty and there is no path transforming σ into π in fewer than $\mathrm{d}_K(\sigma, \pi)$ transpositions.

d_K. A dual representation for the special case of $\boldsymbol{\beta} \equiv 1$ and μ constant is provided in Nishimura and Ok, 2023. In particular, the authors observed that $\mathrm{d}_{\mathbf{1}}$ is equivalent to $\mathrm{d}_{\boldsymbol{\phi}}$ with exponential costs $\phi_k = 2^{n-1-k}$. However, when μ is not constant, the same authors show that there is no $\boldsymbol{\phi}$ for which the equality $\mathrm{d}_{\boldsymbol{\phi}} = \mathrm{d}_{\mathbf{1}}^{\mu}$ can hold.

We next prove that *in fact* it is possible to obtain a dual representation for each member in the class of $(\boldsymbol{\beta}, g)$-distances by suitably generalizing (6.5). Given $\boldsymbol{\phi} \in \mathbf{R}_+^{n-1}$ and $g \in \mathcal{G}$, we define,

$$\rho_{\boldsymbol{\beta}}^{g} : (\sigma, \pi) \mapsto \min_{\mathbf{t} \in T(\sigma,\pi)} \sum_{k=1}^{\mathrm{d}_K(\sigma,\pi)} \phi_{a_k(\mathbf{t})} g_{i_k(\mathbf{t}), j_k(\mathbf{t})},$$

where we recall that $i_k(\mathbf{t})$ and $j_k(\mathbf{t})$ are the unique elements in $[n]$ that are exchanged by the k^{th} transposition of $\mathbf{t}$. Therefore, in $\rho_{\boldsymbol{\beta}}^{g}$ the cost of an adjacent transposition depends on both the *position* and the *identity* of the swapped elements.

We relate $\rho_{\boldsymbol{\beta}}^{g}$ and $\rho_{\boldsymbol{\beta}}^{g}$ by showing that when the weights $\boldsymbol{\phi}$ are decreasing, there is a unique vector of weights $\boldsymbol{\beta}$ (depending on $\boldsymbol{\phi}$) such that $\rho_{\boldsymbol{\beta}}^{g} = \rho_{\boldsymbol{\beta}}^{g}$. Specifically, consider the following linear functions

$$F : \boldsymbol{\beta} \mapsto \boldsymbol{\phi} = \left(\sum_{k=2}^{n} \binom{n-a-1}{k-2} \beta_k \right)_{a=1}^{n-1}, \; G : \boldsymbol{\phi} \mapsto \boldsymbol{\beta} = \left(\sum_{k=0}^{a-2} (-1)^{a+k} \binom{a-2}{k} \phi_{n-1-k} \right)_{a=2}^{n}. \tag{6.6}$$

One can show that F and G are bijective maps such that $F = G^{-1}$. Via these functions, we can finally establish a dual representation for $\rho_{\boldsymbol{\beta}}^{g}$. To this end, we define the set of decreasing weights $\boldsymbol{\phi}$ that are not identically equal to zero by

$$\Phi^{\mathrm{s}} := \{\boldsymbol{\phi} \in \mathbf{R}_+^{n-1} \backslash \{\mathbf{0}\} : (i \geqslant j) \Rightarrow (\phi_i \leqslant \phi_j)\}.$$

Proposition 33. *Let $g \in \mathcal{G}$ be non-zero. If $\rho_{\boldsymbol{\beta}}^{g} = \mathrm{d}_{\boldsymbol{\beta}}^{g}$, then $\boldsymbol{\beta} = G(\boldsymbol{\phi})$. Moreover, if $g \in \mathcal{G}_{\triangleright}^{\mathrm{s}}$ and $\boldsymbol{\phi} \in \mathbf{R}_+^{n-1}$, then $\rho_{\boldsymbol{\beta}}^{g} = \mathrm{d}_{G(\boldsymbol{\phi})}^{g}$ if and only if $\boldsymbol{\phi} \in \Phi^{\mathrm{s}}$.*

This proposition plays a central role in the rest of the paper. The dual representation $\rho_{\boldsymbol{\beta}}^{g}$ turns out to be useful for both axiomatization purposes and parameter identification.

Indeed, later in the paper, we provide an axiomatic characterization of $\rho^g_{\boldsymbol{\beta}}$ using the equivalence established in Proposition 33. We defer the reader to §6.4 for a broader discussion. Second, we observe that fitting from data the costs of swapping adjacent elements $\boldsymbol{\phi}$ might be easier than trying to select directly the weights $\boldsymbol{\beta}$. Through Proposition 33, calibrating $\boldsymbol{\phi}$ allows to identify $\boldsymbol{\beta}$ using the function G. Finally, although this paper mainly focuses on rankings, we highlight that $\rho^g_{\boldsymbol{\beta}}$ can be directly extended to larger domains, including preferences that can be incomplete, and exhibit ties and intransitivity. On the other hand, we observe that the dual semimetric $\rho^g_{\boldsymbol{\beta}}$ cannot be directly extended beyond $\mathbb{S}_n$.

Parameter space

Having outlined and explained the logic behind our distances, in this section, we answer a basic, yet fundamental, question: which choices of $\boldsymbol{\beta}$ and g make $\rho^g_{\boldsymbol{\beta}}$ a (semi)metric? We begin our analysis by focusing on the simpler case when g is constant outside the diagonal. To this end, we collect all the relevant parametrizations of $\boldsymbol{\beta}$ in the following sets

$$\mathrm{B}^{\mathrm{s}} := \{\boldsymbol{\beta} \in \mathbf{R}^{n-1} : \mathrm{d}_{\boldsymbol{\beta}} \text{ is a non-trivial semimetric}\} \text{ and } \mathrm{B}^{\mathrm{m}} := \{\boldsymbol{\beta} \in \mathbf{R}^{n-1} : \mathrm{d}_{\boldsymbol{\beta}} \text{ is a metric}\},$$

where we are using s and m, as mnemonics for semimetric and metric. Leveraging the dual representation established in Proposition 33, we characterize these sets in terms of the costs $\boldsymbol{\phi}$.

Lemma 34. *Recall G in* (6.6). *We have* $\mathrm{B}^{\mathrm{s}} = G(\Phi^{\mathrm{s}})$ and $\mathrm{B}^{\mathrm{m}} = G(\Phi^{\mathrm{m}})$, where $\Phi^{\mathrm{m}} := \Phi^{\mathrm{s}} \cap \mathbf{R}^{n-1}_{++}$.

Thus, whenever the costs $\boldsymbol{\phi}$ are decreasing and positive the corresponding distance $\mathrm{d}_{G(\boldsymbol{\phi})}$ is a semimetric. If, in addition, each component of $\boldsymbol{\phi}$ is strictly positive, then $\mathrm{d}_{G(\boldsymbol{\phi})}$ becomes a metric.

To tackle the general case $\rho^g_{\boldsymbol{\beta}}$ is convenient to partition the set B^{s} (B^{m}) into two subsets as follows. We define $\mathrm{B}^{\mathrm{s}1}$ ($\mathrm{B}^{\mathrm{m}1}$) as the collection of $\boldsymbol{\beta} \in \mathrm{B}^{\mathrm{s}}$ (B^{m}) such that only first component, β_2, is non-zero, and $\mathrm{B}^{\mathrm{s}2}$ ($\mathrm{B}^{\mathrm{m}2}$) as its complement in B^{s} (B^{m}).[9]

[9] Formally, $\mathrm{B}^{\mathrm{s}1} := \{\boldsymbol{\beta} \in \mathrm{B}^{\mathrm{s}} : \forall j > 2 : \beta_j = 0\}$, $\mathrm{B}^{\mathrm{m}1} := \{\boldsymbol{\beta} \in \mathrm{B}^{\mathrm{m}} : \forall j > 2 : \beta_j = 0\}$, $\mathrm{B}^{\mathrm{s}2} := \{\boldsymbol{\beta} \in \mathrm{B}^{\mathrm{s}} : \exists j > 2 : \beta_j \neq 0\}$, and $\mathrm{B}^{\mathrm{m}2} := \{\boldsymbol{\beta} \in \mathrm{B}^{\mathrm{m}} : \exists j > 2 : \beta_j \neq 0\}$. Notice that $\mathrm{B}^{\mathrm{s}} = \mathrm{B}^{\mathrm{s}1} \sqcup \mathrm{B}^{\mathrm{s}2}$, $\mathrm{B}^{\mathrm{s}1} = \mathrm{B}^{\mathrm{m}1} = \mathbf{R}_{++} \times \{0\}^{n-2}$.

First, observe that if $g \equiv 0$ or $\boldsymbol{\beta} = \mathbf{0}$, then $\rho^g_{\boldsymbol{\beta}}$ reduces to the *trivial* semimetric, which is constantly equal to zero. Proposition 35 characterizes the parameters $\boldsymbol{\beta}$ and g that make $\rho^g_{\boldsymbol{\beta}}$ a (semi)metric.

Proposition 35. *$\rho^g_{\boldsymbol{\beta}}$ is a non-trivial semimetric iff precisely one of the following conditions holds*

- *$\boldsymbol{\beta} \in \mathrm{B}^{\mathrm{s}1}$ and $g \in \mathcal{G}^{\mathrm{s}}$ (or, $\boldsymbol{\beta} \in -\mathrm{B}^{\mathrm{s}1}$ and $g \in -\mathcal{G}^{\mathrm{s}}$)*
- *$\boldsymbol{\beta} \in \mathrm{B}^{\mathrm{s}2}$ and $g \in \mathcal{G}^{\mathrm{s}}_{\triangleright}$ (or, $\boldsymbol{\beta} \in -\mathrm{B}^{\mathrm{s}2}$ and $g \in -\mathcal{G}^{\mathrm{s}}_{\triangleright}$)*

Similarly, $\rho^g_{\boldsymbol{\beta}}$ is a metric iff precisely one of the following conditions holds

- *$\boldsymbol{\beta} \in \mathrm{B}^{\mathrm{m}1}$ and $g \in \mathcal{G}^{\mathrm{m}}$ (or, $\boldsymbol{\beta} \in -\mathrm{B}^{\mathrm{m}1}$ and $g \in -\mathcal{G}^{\mathrm{m}}$)*
- *$\boldsymbol{\beta} \in \mathrm{B}^{\mathrm{m}2}$ and $g \in \mathcal{G}^{\mathrm{m}}_{\triangleright}$ (or, $\boldsymbol{\beta} \in -\mathrm{B}^{\mathrm{m}2}$ and $g \in -\mathcal{G}^{\mathrm{m}}_{\triangleright}$)*

The proof of this result is based on the restrictions imposed by two of the defining properties of semimetrics. In particular, apart from the trivial cases, the positivity of the semimetric imposes that g belongs to $\mathcal{G}^{\mathrm{s}}$. Further, if $\beta_k \neq 0$ for some $k > 2$, then the triangle inequality leads to the positivity of $g_{ij} + g_{jk} - g_{ik}$, that is $g \in \mathcal{G}^{\mathrm{s}}_{\triangleright}$. The inclusion of the cases within parentheses in Proposition 35 is unavoidable because we clearly have that $\mathrm{d}^{-g}_{-\boldsymbol{\beta}} = \rho^g_{\boldsymbol{\beta}}$. It is easy to check that $\rho^g_{\boldsymbol{\beta}}$ is neutral, i.e., it treats *symmetrically* all alternatives, if and only if g is constant outside the diagonal.

Computability

The computation of the distance $\rho^g_{\boldsymbol{\beta}}$ between any two rankings, according to its definition, is prohibitive for large n as it requires computing the most preferred elements for $O(2^n)$ menus. However, we provide a convenient reformulation of $\rho^g_{\boldsymbol{\beta}}$ computable in polynomial time.

Lemma 36. *Fix a dissimilarity function $g \in \mathcal{G}$ and some weights $\boldsymbol{\beta} \in \mathbf{R}^{n-1}$. Then,*

$$\rho^g_{\boldsymbol{\beta}}(\sigma, \pi) = \sum_{\substack{i,j: \\ i >_\sigma j, j >_\pi i}} g_{ij} \tilde{f}_{\boldsymbol{\beta}}(|i^{\downarrow,\sigma} \cap j^{\downarrow,\pi}|), \text{ where } \tilde{f}_{\boldsymbol{\beta}} : t \mapsto \sum_{k=2}^{n} \beta_k \binom{t}{k-2}. \tag{6.7}$$

This representation (6.7) significantly reduces the computational complexity of $\rho_{\boldsymbol{\beta}}^{g}$. Using Pascal's identity, we can pre-compute the weighted binomial sum $\tilde{f}_{\boldsymbol{\beta}}$ in $O(n^2)$ time. Consequently, the overall time complexity for computing $\rho_{\boldsymbol{\beta}}^{g}$ using (6.7) can be reduced from $O(2^n)$ to $O(n^3)$. This follows from the fact that there are $O(n^2)$ ordered couples (i, j), and, for each of them, the value $|i^{\downarrow,\sigma} \cap j^{\downarrow,\pi}|$ can be evaluated in $O(n)$ time.

In addition, when g is additive (6.7) can be simplified further:

$$\rho_{\boldsymbol{\beta}}^{\mu}(\sigma, \pi) = \sum_{i \in [n]} \left(f_{\boldsymbol{\beta}}(|i^{\downarrow,\pi}|) + f_{\boldsymbol{\beta}}(|i^{\downarrow,\sigma}|) - 2 f_{\boldsymbol{\beta}}(|i^{\downarrow,\sigma} \cap i^{\downarrow,\pi}|) \right) \mu_i, \tag{6.8}$$

where

$$f_{\boldsymbol{\beta}} : t \mapsto \sum_{k=2}^{n} \beta_k \binom{t}{k-1}.$$

The computational complexity of $\rho_{\boldsymbol{\beta}}^{\mu}$ reduces to $O(n^2)$.

6.4 Axiomatic characterization

In this section we establish our main result, providing an axiomatic foundation for the class $(\boldsymbol{\beta}, g)$-distances. Our postulates allow us to break down the fundamental structure behind how $\rho_{\boldsymbol{\beta}}^{g}$ measures the dissimilarity between any two rankings.

Our first postulate, *betweeness axiom*, describes how our distances deal with in-between rankings. In particular, a ranking ω is said to be *between* σ and π, denoted by $\sigma - \omega - \pi$, if for each pair of alternatives, ω agrees with σ or π or both. More formally, $\sigma - \omega - \pi$ if for all $i \neq j$

$$M_{\{i,j\}}(\sigma) = M_{\{i,j\}}(\pi) \implies M_{\{i,j\}}(\omega) = M_{\{i,j\}}(\sigma).$$

We say that ω is *strictly between* σ and π if $\sigma - \omega - \pi$ and it is distinct from both of them.

Axiom 37 (Betweenness). *For all $\sigma, \pi \in \mathbb{S}_n$ disagreeing on more than one pair of elements there exists $\omega \in \mathbb{S}_n$ strictly between σ and π such that $\mathrm{d}(\sigma, \pi) = \mathrm{d}(\sigma, \omega) + \mathrm{d}(\omega, \pi)$.*

Axiom 37 links the notion of betweenness for rankings with its geometric counterpart, as captured by the triangle *equality*. In particular, for any couple of non-adjacent rankings, there is a ranking strictly in-between the two of them for which the triangle inequality holds as equality. The main implication of Axiom 37 for semimetrics is established in the following proposition.

Proposition 38. *A semimetric* d *satisfies Axiom 37 if and only if* d *is a graphic semimetric, that is, for all* $\sigma, \pi \in \mathfrak{S}_n$*,*

$$\mathrm{d}(\sigma, \pi) = \min_{\mathbf{t} \in T(\sigma,\pi)} \sum_{j=1}^{\mathrm{d}_K(\sigma,\pi)} \mathrm{d}\left(\sigma \prod_{k=1}^{j-1} t_{a_k(\mathbf{t}),a_k(\mathbf{t})+1}, \sigma \prod_{k=1}^{j} t_{a_k(\mathbf{t}),a_k(\mathbf{t})+1}\right). \tag{6.9}$$

Thus, a semimetric d satisfies Axiom 37 if and only if, for any σ and π, $\mathrm{d}(\sigma, \pi)$ can be decomposed into a sum of distances over adjacent rankings, obtained by successively swapping two adjacent alternatives until σ is transformed into π. Graphically, this proposition links all semimetrics that satisfy Axiom 37 to the structure of the permutahedron, as shown in Figure 6.4 for the case $n = 3$.

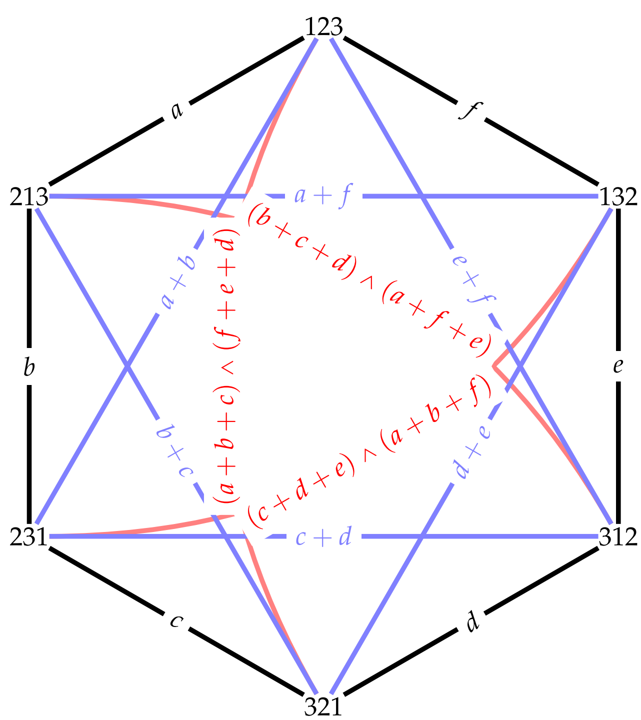

FIGURE 6.4: Permutahedron (black edges) together with its completion (colored edges) on [3].

The permutahedron is a graph whose vertices are the elements of $\mathbb{S}_n$, and its edges connect adjacent permutations. Each edge is assigned a weight corresponding to the distance between the connected rankings. In Figure 6.4, the edges of the permutahedron are represented in black and are arbitrarily weighted from a to f. In the same figure, we also link all the remaining couples of permutations with colored edges. Under Axiom 37, their weights are *uniquely* determined by the weights of the edges of the permutahedron (black). Therefore, to characterize a semimetric that satisfies Axiom 37, it is sufficient to describe how it measures the distance between *adjacent* rankings, i.e., rankings that differ *only* in the positioning of two consecutive items. Our second axiom is the following.

Axiom 39 (Monotonicity). *For all $\sigma, \pi \in \mathbb{S}_n$ and positions $a \leqslant b$,*

$$\{\sigma_a, \sigma_{a+1}\} = \{\pi_b, \pi_{b+1}\} \implies \mathrm{d}(\sigma, \sigma t_{a,a+1}) \geqslant \mathrm{d}(\pi, \pi t_{b,b+1}).$$

Axiom 39 captures the idea that disagreements over items placed higher in a ranking have at least as much importance as disagreements over *the same* items in lower positions. This axiom addresses the top-bottom issue discussed in Farnoud and Milenkovic, 2014 and Nishimura and Ok, 2023. When $a = b$, Axiom 39 also implies an invariance property. If we swap two alternatives placed in positions a and $a+1$, the order of the other items is irrelevant to assess the distance between any ranking σ and $\sigma t_{a,a+1}$. Mixing axioms 37 and 39 yields the next proposition.

Proposition 40. *If a semimetric* d *satisfies axioms 37 and 39, then there exists $g : [n-1] \times [n]^2 \to \mathbf{R}_+$ symmetric in the last two entries with $g_\cdot(i,i) = 0$ for all $i \in [n]$ such that for all $\sigma, \pi \in \mathbb{S}_n$,*

$$\mathrm{d}(\sigma, \pi) = \min_{\mathbf{t} \in T(\sigma,\pi)} \sum_{k=1}^{\mathrm{d}_K(\sigma,\pi)} g_{a_k(\mathbf{t})}(i_k(\mathbf{t}), j_k(\mathbf{t})). \tag{6.10}$$

This proposition is one of the building blocks of our main result. Its key implication is that now the distance between any couple of adjacent rankings σ and $\sigma t_{a,a+1} = t_{i,j}\sigma$ depends *only* on the interaction between the identity of the swapped items (i,j) and their positions $(a, a+1)$, *via* some function $g_a(i,j)$. Our next axiom models this interaction.

Axiom 41 (Proportionality). *For all $\sigma, \pi, \rho, \tau \in \mathbb{S}_n$ and distinct $a, b \in [n-1]$ with*

$$\{\sigma_a, \sigma_{a+1}\} = \{\pi_b, \pi_{b+1}\} \text{ and } \{\rho_a, \rho_{a+1}\} = \{\tau_b, \tau_{b+1}\}$$

we have that

$$\mathrm{d}(\sigma, \sigma t_{a,a+1})\mathrm{d}(\tau, \tau t_{b,b+1}) = \mathrm{d}(\pi, \pi t_{b,b+1})\mathrm{d}(\rho, \rho t_{a,a+1}).$$

Equivalently, Axiom 41 requires the existence of a function $c : [n] \times [n-1] \to \mathbf{R}_+$ such that for all rankings $\sigma, \pi \in \mathbb{S}_n$ and positions $a < b$

$$\{\sigma_a, \sigma_{a+1}\} = \{\pi_b, \pi_{b+1}\} \implies \mathrm{d}(\pi, \pi t_{b,b+1}) = c(b,a)\mathrm{d}(\sigma, \sigma t_{a,a+1}). \tag{6.11}$$

To evaluate the consequences of Axiom 41 we analyze its interplay with Axiom 39. Consider two rankings, σ and π, in which the items i and j are adjacent but occupy different positions. Suppose that in both rankings, i precedes j, and let the positions of i be a in σ and b in π. We aim to compare the distances between σ and $\sigma t_{a,a+1}$, and between π and $\pi t_{b,b+1}$. By Axiom 39, the order of the items not affected by the swaps plays no role in determining these distances. Therefore, any difference between these two distances can only be due to the interaction of the identity of the items being swapped and their positions. Axiom 41 imposes a (multiplicative) separation of this interaction. This is clear by looking at the reformulation (6.11), which imposes that the ratio between the corresponding distances depends only on the positions of the items, a and b.

These three axioms completely characterize the semimetrics $\rho^g_{\boldsymbol{\beta}}$ and $\rho^g_{\boldsymbol{\beta}}$, which coincide for a suitable choice of the parameters $\boldsymbol{\beta}, \boldsymbol{\phi}$, and g.

Theorem 42. *For a non-trivial semimetric* d*, the following are equivalent:*

1. d *satisfies axioms 37, 39, and 41.*
2. *There is $(\boldsymbol{\phi}, g) \in F(\mathrm{B}^{\mathrm{s}1}) \times \mathcal{G}^{\mathrm{s}} \cup F(\mathrm{B}^{\mathrm{s}2}) \times \mathcal{G}^{\mathrm{s}}_{\rhd}$ such that for all $\sigma, \pi \in \mathbb{S}_n$,*

$$\mathrm{d}(\sigma, \pi) = \min_{\mathbf{t} \in T(\sigma,\pi)} \sum_{k=1}^{\mathrm{d}_K(\sigma,\pi)} \phi_{a_k(\mathbf{t})} g_{i_k(\mathbf{t}), j_k(\mathbf{t})}$$

3. *There is* $(\boldsymbol{\beta}, g) \in \mathrm{B}^{\mathrm{s}1} \times \mathcal{G}^{\mathrm{s}} \cup \mathrm{B}^{\mathrm{s}2} \times \mathcal{G}^{\mathrm{s}}_{\triangleright}$ *such that for all* $\sigma, \pi \in \mathbb{S}_n$,

$$\mathrm{d}(\sigma, \pi) = \sum_{S \in \mathcal{P}} \beta_{|S|} g\left(M_S(\sigma), M_S(\pi)\right).$$

Moreover, we have the following remark

Remark 43. *Given any non-trivial semimetric satisfying axioms 37, 39, and 41, the choice of* $(\boldsymbol{\phi}, g)$ *or* $(\boldsymbol{\beta}, g)$ *is unique up to multiplicative constants. In particular,* g *can be chosen to be the same in the two representations, in which case* $\boldsymbol{\beta} = G(\boldsymbol{\phi})$ *and* $\boldsymbol{\phi} = F(\boldsymbol{\beta})$.

The proof of the implication $1 \Rightarrow 2$ is based on three steps. First of all, by Axiom 37, d must be a graphic semimetric (Proposition 38), and thus it is completely determined by the value it assigns to adjacent permutations. Secondly, adding Axiom 39 we retrieve (Proposition 40) that the distance between $\sigma, \sigma' = \sigma t_{a,a+1} = t_{i,j}\sigma$ depends only on three factors: the position a at which the transposition occurs and the swapped items i and j. Thus, we recover the expression $\mathrm{d}(\sigma, \sigma') = g_a(i, j)$. Finally, Axiom 41 allows to break $g_a(i, j)$ into a factor depending only on the position of the swap ϕ_a and another depending only on the identity of the swapped items g_{ij}, so that overall $g_a(i, j) = \phi_a g_{ij}$.
The proof of $2 \Leftrightarrow 3$ follows from the dual representation of $\rho^g_{\boldsymbol{\beta}}$ (Proposition 33). In conclusion, the necessity of the axioms is easily retrieved using the two equivalent representations $\rho^g_{\boldsymbol{\beta}}$ and $\mathrm{d}^g_{\boldsymbol{\phi}}$.

To conclude the section, we retrieve the axiomatic characterization of the additive version of $\rho^g_{\boldsymbol{\beta}}$. It is natural to imagine that, to characterize $\rho^g_{\boldsymbol{\beta}}$ with an additive dissimilarity function g, one needs to impose a further separability condition. This exact condition is provided by the following axiom.

Axiom 44 (Additivity). *For all* $\sigma, \pi, \rho, \tau \in \mathbb{S}_n$ *and* $a \in [n-1]$ *with*

$$\{\sigma_a, \sigma_{a+1}\} \sqcup \{\pi_a, \pi_{a+1}\} = \{\rho_a, \rho_{a+1}\} \sqcup \{\tau_a, \tau_{a+1}\}$$

we have that

$$\mathrm{d}(\sigma, \sigma t_{a,a+1}) + \mathrm{d}(\pi, \pi t_{a,a+1}) = \mathrm{d}(\rho, \rho t_{a,a+1}) + \mathrm{d}(\tau, \tau t_{a,a+1}).$$

Adding it to the mix of axioms in Theorem 42 yields the desired characterization.

Proposition 45. *For a non-trivial semimetric* d*, the following are equivalent:*

1. d *satisfies axioms 37, 39, 41, and 44.*
2. *There exists* $(\boldsymbol{\phi}, \mu) \in F(\mathrm{B}^{\mathrm{s}1}) \times \mathcal{M}^{\mathrm{s}} \cup F(\mathrm{B}^{\mathrm{s}2}) \times \mathcal{M}^{\mathrm{s}}_{\triangleright}$ *such that for all* $\sigma, \pi \in \mathbb{S}_n$*,*

$$\mathrm{d}(\sigma, \pi) = \min_{\mathbf{t} \in T(\sigma,\pi)} \sum_{k=1}^{\mathrm{d}_K(\sigma,\pi)} \phi_{a_k(\mathbf{t})} \left(\mu(i_k(\mathbf{t})) + \mu(j_k(\mathbf{t}))\right). \tag{6.12}$$

   Moreover, μ *can be chosen so that* $\sum_i \mu(i) = 1$*.*

3. *There exists* $(\boldsymbol{\beta}, \mu) \in \mathrm{B}^{\mathrm{s}1} \times \mathcal{M}^{\mathrm{s}} \cup \mathrm{B}^{\mathrm{s}2} \times \mathcal{M}^{\mathrm{s}}_{\triangleright}$ *such that for all* $\sigma, \pi \in \mathbb{S}_n$*,*

$$\mathrm{d}(\sigma, \pi) = \sum_{S \in \mathcal{P}} \beta_{|S|} \mu \left(\triangle_S(\sigma, \pi)\right).$$

   Moreover, μ *can be chosen so that* $\sum_i \mu(i) = 1$*.*

Parametrizations of $\boldsymbol{\beta}$ and g

In addition to the cases discussed in §6.3, our distance $\rho^g_{\boldsymbol{\beta}}$ relates to well-known dissimilarity functions established in the literature:

- If $\boldsymbol{\beta} = (1, \ldots, 1, 0, \ldots, 0)$ and g is μ-additive we get the metrics in Gilbert et al., 2022.
- If $\boldsymbol{\beta} = \left((\alpha-1)^{k-2}\right)_{k=2}^n$ for some $\alpha > 1$ and g is constant, then we recover the *unavailable candidate model* of Lu and Boutilier, 2010.
- If g is constant, then we recover the weighted Kendall distance (with decreasing weights $\boldsymbol{\phi}$) proposed and axiomatized by Farnoud and Milenkovic, 2014.
- If g is constant and the weights $\boldsymbol{\phi}$ are completely monotone and positive, we retrieve the distances of Baldiga and Green, 2013 with exchangeable distributions.

6.5 Generalized Diaconis-Graham inequalities

In this section, we restrict attention to μ-additive dissimilarity functions and provide a generalization of the celebrated Diaconis-Graham (DG) inequality Diaconis and Graham, 1977, which can be used to approximate our distances. The DG inequality establishes a connection between the Kendall metric and another measure of disagreement, known as Spearman distance:

$$\mathrm{s}(\sigma, \pi) := \sum_{x \in [n]} |\sigma_x^{-1} - \pi_x^{-1}| = \sum_{x \in [n]} ||x^{\downarrow,\sigma}| - |x^{\downarrow,\pi}||.$$

Thus, s quantifies the disagreement between rankings, σ and π, by summing, over all items, the absolute value of the difference between the position of each item with respect to σ and π.[10] It was established in Diaconis and Graham, 1977 that the Spearman distance is never too far from the Kendall metric. Specifically, they always stay within a factor of two, that is, for all σ and π,

$$\mathrm{s}(\sigma, \pi) \leqslant \mathrm{d}_K^{|\cdot|}(\sigma, \pi) \leqslant 2\mathrm{s}(\sigma, \pi).^{11} \tag{6.13}$$

This relation has found application on several fronts: for instance, it has been exploited by Dwork et al., 2001 to provide a 2-approximation algorithm to the problem of finding a median ranking (with respect to the Kendall metric) starting from a collection of individual rankings.

The main result of this section shows that, in fact, there exists a family of these inequalities, and that the factor two appearing in (6.13) is the *worst-possible* that can emerge inside the family.

For all $\boldsymbol{\beta} \in \mathbf{R}^{n-1}$ and $\mu \in \mathbf{R}_+^n$, we define the ($\boldsymbol{\beta}$,μ)-*Spearman's distance* as

$$\mathrm{s}_{\boldsymbol{\beta}}^{\mu} : (\sigma, \pi) \mapsto \sum_{x \in [n]} \left| f_{\boldsymbol{\beta}}(|x^{\downarrow,\sigma}|) - f_{\boldsymbol{\beta}}(|x^{\downarrow,\pi}|) \right| \mu(x), \tag{6.14}$$

where $f_{\boldsymbol{\beta}}$ is the function defined in (6.8). Differently from the classical Spearman's distance, the cardinalities of $x^{\downarrow,\sigma}$ and $x^{\downarrow,\pi}$ are inputs for the function $f_{\boldsymbol{\beta}}$ and each item x is weighted by μ_x.[12] We first study the case where μ is constant. The following is the main result of this section.

[10] Recall that in our notation σ_x^{-1} denotes the position of item x in the ranking σ.

[11] Recall that $\mathrm{d}_K^{|\cdot|}(\sigma, \pi) = 2\mathrm{d}_K(\sigma, \pi)$, where d_K is the Kendall metric.

[12] Observe that when $\boldsymbol{\beta} = (1, 0, \ldots, 0)$, then, for all t, $f_{\boldsymbol{\beta}}(t) = t$, so that if μ is constant to one, we are back to the classical Spearman's distance. Thus, the $(\boldsymbol{\beta}, \mu)$-Spearman distance is a proper generalization.

Theorem 46. *If $\boldsymbol{\beta} \in \mathrm{B}^{\mathrm{s}}$, then, for all $\sigma, \pi \in \mathbb{S}_n$, it holds*

$$\mathrm{s}_{\boldsymbol{\beta}}(\sigma, \pi) \leqslant \mathrm{d}_{\boldsymbol{\beta}}(\sigma, \pi) \leqslant \gamma_{\boldsymbol{\beta}} \mathrm{s}_{\boldsymbol{\beta}}(\sigma, \pi),$$

where the approximation factor $\gamma_{\boldsymbol{\beta}}$ takes the form

$$\gamma_{\boldsymbol{\beta}} := \max_{h \in \mathcal{H}_{\boldsymbol{\beta}}} \left(1 + \frac{f_{\boldsymbol{\beta}}(n-h)}{f_{\boldsymbol{\beta}}(n-h+1)}\right) \text{ with } \mathcal{H}_{\boldsymbol{\beta}} := \{h \in [2, n] : f_{\boldsymbol{\beta}}(n-h+1) > 0\}.$$

As the function $f_{\boldsymbol{\beta}}$ is increasing, Theorem 46 readily implies that for each choice of the weights $\boldsymbol{\beta}$, the $\boldsymbol{\beta}$-semimetric and the $\boldsymbol{\beta}$-Spearman distance remain within a factor of two, extending the conclusion of Diaconis and Graham, 1977 to the family of $\boldsymbol{\beta}$-distances. However, as we next illustrate, factor 2 is the worst that can be obtained within the family of generalized DG inequalities.

We then move to address the case when μ is not constant. In these instances, the approximation gets worse, since the approximation factor degrades with the support of μ. However, for a "small" amount of heterogeneity, the approximation can still be useful. In particular, we have

Corollary 47. *Let $\boldsymbol{\beta} \in \mathrm{B}^{\mathrm{s}}$ and $\mu \in \mathbf{R}^n_{++}$ be such that $u \leqslant \mu_i \leqslant U$, for some positive u and U. Then, for all $\sigma, \pi \in \mathbb{S}_n$, it holds*

$$\mathrm{s}^{\mu}_{\boldsymbol{\beta}}(\sigma, \pi) \leqslant \rho^{\mu}_{\boldsymbol{\beta}}(\sigma, \pi) \leqslant \frac{\gamma_{\boldsymbol{\beta}} U}{u} \mathrm{s}^{\mu}_{\boldsymbol{\beta}}(\sigma, \pi).$$

To illustrate the power behind these inequalities, we now adapt the arguments of Dwork et al., 2001 and show how Corollary 47 yields a polynomial-time $\gamma_{\boldsymbol{\beta}} U/u$-approximation algorithm for the median rank aggregation problem (with respect to $(\boldsymbol{\beta}, \mu)$-distances). The running time of the algorithm is $O(mn^3)$, where n is the number of alternatives and m is the number of rankings.

Dwork et al. Dwork et al., 2001 proposed to obtain a polynomial-time approximation algorithm for the Kendall rank aggregation problem by minimizing the sum of Spearman's distances via a minimum cost perfect matching algorithm.

Following a similar spirit, we leverage on Corollary 47 to provide an approximation algorithm for our rank aggregation problems. First, we show that, by minimizing the sum of $(\boldsymbol{\beta}, \mu)$-Spearman distances, we can obtain $\gamma_{\boldsymbol{\beta}} U/u$-approximations for the median rank aggregation problem induced by $\mathrm{d}^{\mu}_{\boldsymbol{\beta}}$. For any collection of input rankings $V = (v^j)_{j=1}^m$ defined over $[n]$, denote by

$$P^{\mathrm{apx}}_{\boldsymbol{\beta},\mu}(V) := \operatorname*{argmin}_{\sigma \in \mathbb{S}_n} \sum_{j=1}^{m} \mathrm{s}^{\mu}_{\boldsymbol{\beta}}(\sigma, v^j),$$

Then,

Corollary 48. *Let $\boldsymbol{\beta} \in \mathrm{B}^{\mathrm{s}}$, $\mu \in \mathbf{R}_{++}$ with $0 < u \leqslant u_i \leqslant U$, for all $i \in [n]$, $V = (v^j)_{j=1}^m$, and $\sigma^{\mathrm{apx}} \in P^{\mathrm{apx}}_{\boldsymbol{\beta},\mu}(V)$. We have,*

$$\min_{\sigma \in \mathbb{S}_n} \sum_{j=1}^{m} \mathrm{d}^{\mu}_{\boldsymbol{\beta}}(\sigma, v^j) \leqslant \mathrm{d}_{\boldsymbol{\beta}}(\sigma^{\mathrm{apx}}, V) \leqslant \gamma_{\boldsymbol{\beta}} \frac{U}{u} \min_{\sigma \in \mathbb{S}_n} \sum_{j=1}^{m} \mathrm{d}^{\mu}_{\boldsymbol{\beta}}(\sigma, v^j).$$

Proof of Corollary 48. Let $\sigma^{\mathrm{apx}} \in P^{\mathrm{apx}}_{\boldsymbol{\beta},\mu}(V)$ and σ^* be a minimizer of $\sum_{j=1}^m \mathrm{d}^{\mu}_{\boldsymbol{\beta}}(\sigma, v^j)$. The following inequalities hold

$$\begin{aligned}\min_{\sigma \in \mathbb{S}_n} \sum_{j=1}^{m} \mathrm{d}^{\mu}_{\boldsymbol{\beta}}(\sigma, v^j) \leqslant \sum_{j=1}^{m} \mathrm{d}_{\boldsymbol{\beta}}(\sigma^{\mathrm{apx}}, v^j) &\overset{(a)}{\leqslant} \gamma_{\boldsymbol{\beta}} \frac{U}{u} \sum_{j=1}^{m} \mathrm{s}^{\mu}_{\boldsymbol{\beta}}(\sigma^{\mathrm{apx}}, v^j) \\ &\overset{(b)}{\leqslant} \gamma_{\boldsymbol{\beta}} \frac{U}{u} \sum_{j=1}^{m} \mathrm{s}_{\boldsymbol{\beta}}(\sigma^*, v^j) \overset{(c)}{\leqslant} \gamma_{\boldsymbol{\beta}} \frac{U}{u} \sum_{j=1}^{m} \mathrm{d}_{\boldsymbol{\beta}}(\sigma^*, v^j) = \gamma_{\boldsymbol{\beta}} \frac{U}{u} \min_{\sigma \in \mathbb{S}_n} \sum_{j=1}^{m} \mathrm{d}^{\mu}_{\boldsymbol{\beta}}(\sigma, v^j),\end{aligned}$$

where (a) and (c) hold by using Corollary 47, and (b) holds since, by construction, σ^{apx} is a minimizer of $\sum_{j=1}^m \mathrm{s}_{\boldsymbol{\beta}}(\cdot, v^j)$. □

The Hungarian algorithm can be implemented to retrieve a permutation minimizing the quantity $\sum_{j=1}^m \mathrm{s}^{\mu}_{\boldsymbol{\beta}}(\cdot, v^j)$.

Lemma 49. *σ^{apx} is computable in $O(mn^3)$ time.*

Proof of Lemma 49. Let $(v^j)_{j=1}^m$ be m input rankings over $[n]$. Construct the following bipartite graph $G = ([n], [n], E)$, where each vertex $c \in [n]$ corresponds to an alternative in $[n]$, and each vertex $p \in [n]$ corresponds to a possible ranking. Edges $e = \{c, p\} \in E$ are interpreted as "alternative c is in position p". We weight each edge $e = \{c, p\} \in E$ by

$$w_{c,p} = \sum_{j=1}^{|V|} \left| f_{\boldsymbol{\beta}}(n-p) - f_{\boldsymbol{\beta}}(|c^{\downarrow,v^j}|) \right| \mu(c)$$

Observe that we can compute all weights $\{w_{c,p}\}_{c,p}$ in time $O(mn^3)$. Indeed, by using Pascal's formula, we can pre-compute $\{f_{\boldsymbol{\beta}}(0), \ldots, f_{\boldsymbol{\beta}}(n-1)\}$ in time $O(n^2)$. Clearly, the number of edges $\{c, p\}$ is n^2. Fix $\{c, p\}$ and recall that we have a profile of m preferences. For each preference, we need to compute $|c^{\downarrow,v^j}|$ and this can be done in linear time. Therefore, in total, it takes $O(mn^3)$ to compute all weights.

Further, note that each perfect matching is associated with a linear order σ, such that $|c^{\downarrow,\sigma}| = n - p$, whenever the edge $\{c, p\}$ belongs to the matching. Particularly, by solving for a minimum weight perfect matching, we obtain $\sigma^{\mathrm{apx}} \in \mathbb{S}_n$ that minimizes $\omega \mapsto \sum_{j=1}^{|V|} \mathrm{s}_{\boldsymbol{\beta}}(\omega, v^j)$. Finally, we can compute a minimum weight perfect matching by relying on the Hungarian algorithm whose running time is $O(n^3)$. Thus, the overall running time is $O(mn^3)$. $\square$

By combining Corollary 47 with Corollary 48 and Lemma 49, we obtain the following.

Theorem 50. *Let $\boldsymbol{\beta}$, μ and V be as in the statement of Corollary 48. Then, there exists a $\gamma_{\boldsymbol{\beta}} U/u$-approximation algorithm for the $(\boldsymbol{\beta}, \mu)$-rank aggregation problem running in time $O(mn^3)$.*

Proof of Theorem 50. Combine Corollary 47 with Corollary 48 and Lemma 49. $\square$

6.6 Applications to voting and statistical inference

Median preference correspondences and majority

Aggregating individuals' preferences into a consensus ranking is a classical problem in the theory of social choice. Suppose that there is an election with several candidates and an ensemble of voters. Each voter submits a ballot, formalized as a ranking over the candidates, and a *voting rule* is then implemented to determine the election's winners. A typical approach (see e.g., Young, 1977) to obtain a *consensus ranking* is by solving a *median rank aggregation problem* with respect to some given distance. In this subsection, we study the median voting rules induced by our distances $\rho^g_{\boldsymbol{\beta}}$.

Formally, an *electorate* $V = (v^1, \ldots, v^m)$ is an ordered list of rankings over the set of candidates $[n]$, and we denote by $\mathbf{V}_n := \bigcup_{k=1}^{\infty} \mathbb{S}_n^k$ the set of all possible electorates. The theory of voting rules revolves around the notions of preference and social choice correspondences. A *preference correspondence* $P : \mathbf{V}_n \rightrightarrows \mathbb{S}_n$ associates to each electorate V a set of rankings $P(V)$ called the set of *consensus preferences*. In particular, the *median preference correspondence* induced by a distance d is

$$P_{\mathrm{d}} : V = (v^j)_{j=1}^{|V|} \mapsto \operatorname*{arg\,min}_{\sigma \in \mathbb{S}_n} \sum_{j=1}^{|V|} \mathrm{d}(\sigma, v^j),$$

Each preference correspondence P naturally defines a *social choice correspondence* $W_P : \mathbf{V}_n \rightrightarrows C$, that gives the set of possible *winners* - candidates ranked first by some consensus preference. Formally,

$$W_P(V) := \bigcup_{v \in P(V)} M_{[n]}(v).$$

When $\mathrm{d} = \rho^g_{\boldsymbol{\beta}}$, we denote $P^g_{\boldsymbol{\beta}} := P_{\mathrm{d}}$, and similarly, when $P = P^g_{\boldsymbol{\beta}}$, we denote $W^g_{\boldsymbol{\beta}} := W_{P^g_{\boldsymbol{\beta}}}$.

Studying the properties of each P and W_P is fundamental to deciding which voting rule should be adopted in different contexts. In this regard, the notion of *majority* is central to the analysis of voting rules and thus, we devote the remaining part of the section to the study of this property.

Given an electorate V and a candidate c, we denote by $n(V, c)$ the number of voters having c as their most favorite alternative *minus* the number of voters who do not. A preference correspondence P satisfies the *majority principle* if for all $V \in \mathbf{V}_n$ and $c \in C$,

$$n(V, c) \geqslant 0 \Longrightarrow c \in W_P(V).$$

Similarly, a preference correspondence P satisfies *unanimity* if $n(V,c) = |V|$ implies $c \in W_P(V)$.

Proposition 51. *If $\rho^g_{\boldsymbol{\beta}}$ is a semimetric, then $P^g_{\boldsymbol{\beta}}$ satisfies the majority principle and unanimity.*

The generality behind Proposition 51 highlights that it is possible to maintain desirable welfare properties while allowing to treat asymmetrically the candidates. This is relevant for contexts where the status quo is prioritized or where the candidates are clustered in different political groups.

Dispersion of ranking data

Probabilistic models are important to understanding the generating process for ranking data and providing forecasts. A popular probabilistic model for rankings has been proposed by Mallows, 1957 and assumes the presence of a central (modal) ranking $\bar{\sigma}$. In particular, the probability of selecting a ranking σ decreases with the distance $\mathrm{d}(\bar{\sigma},\sigma)$. Formally,

$$\mathbb{P}[\sigma] \propto e^{-\theta \mathrm{d}(\sigma,\bar{\sigma})}$$

where $\theta > 0$ is the dispersion parameter, $\bar{\sigma}$ is the central ranking and d is a distance on $\mathbb{S}_n$. Originally Mallows, considered the Kendall distance, but more general versions have been studied. For instance, Marden, 1995 considers the weighted Kendall metric $\mathrm{d}_{\boldsymbol{\phi}}$, and Diaconis, 1989 focuses on general neutral distances. It is easy to see that, for any Mallows model, the MLE estimator of the central ranking is the median ranking. In particular, given a sample of rankings $\Sigma = (\sigma^j)_{j=1}^m$,

$$\hat{\sigma}_{MLE}(\Sigma) = \underset{\sigma \in \mathbb{S}_n}{\arg\min}\, L^{\Sigma}_{\mathrm{d}}(\sigma) = \underset{\sigma \in \mathbb{S}_n}{\arg\min}\, \frac{1}{m} \sum_{\sigma \in \Sigma} \mathrm{d}(\pi,\sigma),$$

where L^{Σ}_{d} is the loss function induced by the distance d given the data Σ. In this section, we analyze the Mallows model with $\rho^g_{\boldsymbol{\beta}}$. We simulate data involving rankings formulated over six items that belong to two groups of movies: sci-fi anime

$$a_1 = \text{Ghost in the shell}, a_2 = \text{Akira}, a_3 = \text{Evangelion}$$

and historical dramas

$$b_1 = \text{The Handmaiden}, b_2 = \text{Rashomon}, b_3 = \text{Silence}.$$

The data-generating process we use in our experiment is inspired by the ratings provided by the users of Letterboxd, where voters tend to rank first the items in one class and then rank the movies in the other class, e.g., $(a_2, a_1, a_3, b_2, b_1, b_3)$ or $(b_2, b_3, b_1, a_2, a_3, a_1)$. More precisely, We generated 1000 rankings with the following procedure. For each user, we decided independently and with probability $1/2$ if this user will prefer anime or historical movies. Moreover, we select randomly how to permute the movies in each category. In particular, independently of the genre, we pick the following permutations with the corresponding probabilities.

permutation	1,2,3	1,3,2	2,1,3	2,3,1	3,1,2	3,2,1
probability	6/21	5/21	4/21	3/21	2/21	1/21

So that, for instance, the probability that any user produces the ranking $(b_1, b_3, b_2, a_2, a_1, a_3)$ is given by $1/2 * 5/21 * 4/21$. Moreover, we are not particularly interested in how different movies' genres are ordered by the voters, but, rather, in how movies in each category are ranked. Thus, for ρ^g_β, we consider a dissimilarity function g that takes into account the clusters of the two movie genres, and gives low importance to swaps between movies of different groups:

$$g_{ij} = \begin{cases} \alpha & \text{if } i, j \text{ both belong to the same cluster} \\ \beta & \text{if } i, j \text{ both belong to different clusters} \end{cases}$$

with $\beta < \alpha$. Within our sample Σ, we show that using the distance d^g_K (with $\beta = \alpha/10$) we observe a much smaller dispersion in Σ, in contrast to the one observed using the Kendall metric d_K. By dispersion $V_{\mathrm{d}}(\Sigma)$, we mean the value of the loss function evaluated at its minimum, normalized by the maximum value this minimum can reach over all possible samples of data. Formally,

$$V_{\mathrm{d}}(\Sigma) := \frac{\min_{\sigma \in \mathbb{S}_n} L^{\Sigma}_{\mathrm{d}}(\sigma)}{\sup_{\Sigma'} \min_{\sigma \in \mathbb{S}_n} L^{\Sigma'}_{\mathrm{d}}(\sigma)} = \frac{L^{\Sigma}_{\mathrm{d}}(\hat{\sigma}_{MLE}(\Sigma))}{\sup_{\Sigma'} L^{\Sigma'}_{\mathrm{d}}(\hat{\sigma}_{MLE}(\Sigma'))}.$$

This normalization makes sure that $V_{\mathsf{d}}(\Sigma)$ lies between 0 (zero dispersion, i.e., degenerate data - all rankings equal) and 1 (maximal dispersion). For each generated Σ, we observe that $V_{\mathsf{d}_K^g}(\Sigma) < V_{\mathsf{d}_K}(\Sigma)$ and the difference between the two dispersions is at least $V_{\mathsf{d}_K}(\Sigma) - V_{\mathsf{d}_K^g}(\Sigma) > 0.174$. This suggests that d_K^g is more suitable than d_K to capture the actual dispersion within a sample of rankings with clustered alternatives.

7. Conclusion

We defined and studied a generalization of Fréchet means on metric spaces and manifolds. We provided Uniform consistency results, a CLT and finally, we explored various applications.

Bibliography

Aubin, J., & Frankowska, H. (2009). *Set-valued analysis*. Modern Birkhäuser Classics.

Aveni, A., Crippa, L., & Principi, G. (2024). Distance basted ranking aggregation: An axiomatic approach. *work in progress available on arxiv*.

Aveni, A., & Mukherjee, S. (2024). Uniform consistency of generalized fréchet means. *work in progress available on arxiv*.

Aveni, A., & Mukherjee, S. (2025). Uniform central limit theorems of generalized fréchet means. *work in progress*.

Avramidi, I. (2023). *Heat kernel on lie groups and maximally symmetric spaces*. Birkhäuser Frontiers in Mathematics.

Baldiga, K. A., & Green, J. R. (2013). Assent-maximizing social choice. *Social Choice and Welfare*, *40*, 439–460.

Barden, D., & Le, H. (1997). Some consequences of the nature of the distance function on the cut locus in a riemannian manifold. *Journal of the London Mathematical Society*, *56*(2), 369–383.

Berge, C. (2010). *Topological spaces: Including a treatment of multi-valued functions, vector spaces and convexity*. Dover Books on Mathematics.

Bhattacharya, A., & Bhattacharya, R. (2012). *Nonparametric inference on manifolds: With applications to shape spaces*. Cambridge University Press.

Bhattacharya, R., & Lin, L. (2017). Omnibus clts for fréchet means and nonparametric inference on non-euclidean spaces. *Proceedings of the American Mathematical Society*, *145*(1), 413–428.

Bhattacharya, R., & Patrangenaru, V. (2003). Large sample theory of intrinsic and extrinsic sample means on manifolds i. *The Annals of Statistics*, *31*, 1–29.

Bhattacharya, R., & Patrangenaru, V. (2005). Large sample theory of intrinsic and extrinsic sample means on manifolds II. *The Annals of Statistics*, *33*(3), 1225–1259.

Cartan, E. (1926). *Sur une classe remarquable d'espaces de Riemann, I* (Vol. 54). Société Mathématique de France.

Cartan, E. (1927). *]Sur une classe remarquable d'espaces de Riemann, II* (Vol. 55). Société Mathématique de France.

Charlier, B. (2012). Necessary and sufficient condition for the existence of a Fréchet mean on the circle.

Chavel, I. (2006). *Riemannian geometry a modern introduction*. Cambridge studied in advanced Mathematics.

de Finetti, B. (1931). Sul Concetto di Media. *Giornale dell'Istituto Italiano degli Atuari*, *3*, 369–396.

Diaconis, P. (1989). A Generalization of Spectral Analysis with Application to Ranked Data. *The Annals of Statistics*, *17*(3), 949–979.

Diaconis, P., & Graham, R. L. (1977). Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *39*(2), 262–268.

do Carmo, M. (2013). *Riemannian geometry and geometric analysis*. Birkhäuser Mathematics Theory; Applications.

Dwork, C., Kumar, R., Naor, M., & Sivakumar, D. (2001). Rank aggregation methods for the web. *Proceedings of the 10th international conference on World Wide Web*, 613–622.

Eltzner, B. (2020). Geometrical smeariness – a new phenomenon of fréchet means. *Bernoulli*.

Eltzner, B., Galaz-García, F., Huckemann, S. F., & Tuschmann, W. (2024). Stability of the cut locus and a central limit theorem for fréchet means of riemannian manifolds. *Proceedings of the American Mathematical Society*, *149*, 3947–3963.

Evans, S. N., & Jaffe, A. Q. (2020). Limit theorems for Fréchet mean sets. *Bernoulli*.

Farnoud, F., & Milenkovic, O. (2014). An axiomatic approach to constructing distances for rank comparison and aggregation. *IEEE Trans. Inform. Theory*, *60*(10), 6417–6439.

Fréchet, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l'institut Henri Poincaré*, *10.4*, 215–310.

Gilbert, H., Portoleau, T., & Spanjaard, O. (2022). Beyond pairwise comparisons in social choice: A setwise kemeny aggregation problem. *Theoretical Computer Science*, *904*, 27–47.

Gromov, M. (2007). *Metric structures for riemannian and non-riemannian spaces*. Modern Birkhäuser Classics.

Grove, K., & Karcher, M. (1973). How to conjugate C1-close group actions. *Math Z*, *132*, 11–20.

Hotz, T., & Huckemann, S. (2015). Intrinsic means on the circle: uniqueness, locus and asymptotics. *Ann Inst Stat Math*, *67*, 177–193.

Hotz, T., Le, H., & Wood, A. (2022). Central limit theorem for intrinsic frechet means in smooth compact riemannian manifolds. *to appear*.

Jin-ichi, I., & Minoru, T. (1998). The dimension of a cut locus on a smooth Riemannian manifold. *Tohoku Mathematical Journal*, *50*(4), 571–575.

Jost, J. (2011). *Riemannian geometry and geometric analysis*. Springer Universitext.

Kemeny, J. G. (1959). Mathematics without numbers. *Daedalus*, *88*(4), 577–591.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*(1-2), 81–93.

Kolmogorov, A. (1930). Sur la notion de la moyenne. *Atti Accad. Naz. Lincei*, *12*, 388–391.

Le, H., & Barden, D. (2014). On the measure of the cut locus of a fréchet mean. *Bulletin of the London Mathematical Society*, *46*.

Le, H. (2016). A diffusion process associated with Fréchet means. *Annals of Applied Probability*, *25(6)*, 3033–3046.

Lee, J. (2021). *Introduction to riemannian manifolds*. Springer.

Lu, T., & Boutilier, C. E. (2010). The unavailable candidate model: A decision-theoretic view of social choice. *Proceedings of the 11th ACM conference on Electronic commerce*, 263–274.

Mallows, C. L. (1957). Non-null ranking models. i. *Biometrika*, *44*(1-2), 114–130.

Marden, J. I. (1995). *Analyzing and modeling rank data*. Chapman; Hall/CRC.

Mattingly, J. C., Miller, E., & Tran, D. (2024). Central limit theorems for Fréchet means on stratified spaces. *to appear*.

Narula, S. C., & Wellington, J. F. (1982). The Minimum Sum of Absolute Errors Regression: A State of the Art Survey. *International Statistical Review*, *50 (3)*, 317–326.

Nishimura, H., & Ok, E. (2023). A class of dissimilarity metrics for acyclic preference relations. *Mathematics of Operations Research*.

Rockafellar, T., & Wets, R. (1998). *Variational analysis*. Springer Grundlehren der mathematischen Wissenschaften.

Romon, G., & Brunel, V.-E. (2023). Convex generalized Fréchet means in a metric tree. *to appear*.

Schötz, C. (2018). Convergence rates for the generalized Fréchet mean via the quadruple inequality. *Electronic Journal of Statistics*.

Sen, B. (2022). *A gentle introduction to empirical process theory and applications*. Stanford Lecture Notes.

Vaart, A., & Wellner, J. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics.

Wainwright, M. (2019). *High-dimensional statistics*. Cambridge Series in Statistical; Probabilistic Mathematics.

Young, H. P. (1977). Extending condorcet's rule. *Journal of Economic Theory*, *16*(2), 335–353.

Biography

Andrea Aveni got his Bachelor's and Master's degrees from Bocconi University, where he was advised by Simone Cerreia-Vioglio, Antonio Lijoi, and Igor Prünster. Then he moved to Duke University, earning a PhD in Statistical Sciences and a Concurrent Master's degree in Mathematics. In Duke, he was advised by Sayan Mukherjee, whom he also visited in the Max Plank Institute for Mathematics in the Sciences in Leipzig, Germany. Andrea is now in Italy with his two black cats, Lipschitz and Hölder, whom he found while he was in Duke. He is waiting to start a postdoc in the Department of Mathematics of the University of Copenhagen, where he will work with Martin Bladt.