INAUGURAL-DISSERTATION

zur
Erlangung der Doktorwürde
der
Naturwissenschaftlich-Mathematischen Gesamtfakultät
der Ruprecht-Karls-Universität
Heidelberg

vorgelegt von

Christof Schötz
M. Sc. Informatik

aus Gunzenhausen

Tag der mündlichen Prüfung:

The Fréchet Mean and Statistics in Non-Euclidean Spaces

Betreuer: Prof. Dr. Enno Mammen
Prof. Dr. Jan Johannes

Acknowledgments

I want to thank my supervisors Jan Johannes and Enno Mammen for their support, advice, the freedom to choose the topics that were most interesting to me, and all the scientific and nonscientific discussions. In particular, the many conversations over coffee after lunch are remembered fondly. I also thank Dagmar Neubauer and Jutta Wiech for taking care of all the administrative issues.

I gratefully acknowledge support by the German Research Foundation (DFG) through the Research Training Group (RTG) 1953. The RTG enabled me to look beyond me own plate, connect with other researchers, and participate in seminars, workshops, and retreats.

Representative for all colleagues and friends I am thankful to, I want to mention Alex I. for runs, Alex K. for his flat, Joseph M. for board games, Lena R. for organization, Marilena M. for walks and talks, Martin K. for the office, Mehmet M. for generosity, Moritz v. R. for hiking, Nathawut P. for apple news, Sandra S. for nut paste, Sergio B. M. for jokes, Stefan K. for riddles, Stefan R. for being Stefan, Stephan B. for discussing R, Simon W. for biking, and Xavier L. for crumble.

Over the last four years I had many fruitful discussions with scientists outside my closer work environment. For this I want to thank Alexander Aue, Benjamin Eltzner, Stefan Heyder, Thomas Hotz, Stephan Huckemann, Steve Marron, Quentin Paris, Wolfgang Polonik, and Suhasini Subba Rao.

Moreover, I am grateful to the administration of the *Combined Faculty of Natural Sciences and Mathematics* of the Ruprecht-Karls-Universität Heidelberg.

Lastly, I want to thank my parents and my brother, who have supported me my whole life, and made it possible for me to write this thesis.

Zusammenfassung

In dieser Dissertation werden statistische Eigenschaften des Fréchet-Mittelwertes und seiner Verallgemeinerungen in abstrakten Rahmen untersucht. Dadurch werden vielerlei verschiedene Anwendungen abgedeckt, welche insbesondere in der Untersuchung von Nicht-Standard-Daten von Interesse sind. Der Fokus der Arbeit liegt auf der Konvergenz und Konvergenzgeschwindigkeit des empirischen Fréchet-Mittelwertes unabhängiger Beobachtungen. Die abstrakten Ergebnisse werden beispielhaft in spezifischen Räumen angewendet.

Der Erwartungswert einer reellwertigen, quadratintegrierbaren Zufallsvariable kann dadurch charakterisiert werden, dass er den erwarteten quadratischen Abstand zu dieser Zufallsvariable eindeutig minimiert. Diese Eigenschaft kann benutzt werden, um den Begriff Mittelwert zu verallgemeinern. Ein Fréchet-Mittelwert einer Zufallsvariable mit Werten in einem metrischen Raum ist jeder Minimierer des erwarteten quadratischen Abstandes zu dieser Zufallsvariable. Durch diese Definition werden zwei Dinge erreicht: Erstens werden viele gebräuchliche Arten von Mittelwerten – etwa der Erwartungswert, der Median oder das geometrische Mittel – in einem Begriff umfasst. Zweitens wird ein Mittelwertsbegriff für nicht-euklidische Räume – wie etwa die Kugel, der Raum phylogenetischer Bäume oder die Wasserstein-Räume – definiert, wodurch diese der Anwendung von Wahrscheinlichkeitstheorie und Statistik zugänglich gemacht werden.

Wir zeigen starke Gesetze der großen Zahlen für Mengen von Fréchet-Mittelwerten mit zwei verschiedenen Begriffen der Konvergenz von Mengen. Dabei setzen wir nur ein endliches erstes Moment voraus. Als nächstes wenden wir uns der Geschwindigkeit dieser Konvergenz zu. Zuerst zeigen wir anhand des projizierten Mittelwertes – einer Instanz des Fréchet-Mittelwertes – dass hierbei sehr unterschiedliche Konvergenzraten zustande kommen können, abhängig von der Geometrie des zugrundeliegenden Raumes und einiger Eigenschaften der Verteilung der Daten. Danach beweisen wir Konvergenzraten in einem allgemeinen Rahmen. Eine der Bedingungen, die wir dafür aufstellen, ist die Quadrupelungleichung – eine Verallgemeinerung der Cauchy-Schwarz-Ungleichung. Diese und einige weitere der von uns aufgestellten abstrakten Bedingungen sind in Hadamard-Räumen – geodätische metrische Räume mit nicht-positiver Krümmung – erfüllt, sodass sie sich besonders zur Untersuchung im Kontext des Fréchet-Mittelwertes eignen. Wir zeigen eine Quadrupelungleichung für Potenzen von Hadamard-Metriken – ein rein geometrisches Resultat mit verblüffend komplexem Beweis. Zuletzt untersuchen wir Regressionsmodelle mit Zielwerten aus metrischen Räumen und einem bedingten Fréchet-Mittelwert als Regressionsfunktion. Wir vergleichen zwei Ansätze, wie bekannte Schätzer auf nicht-euklidische Szenarien angepasst werden können. Dabei zeigen wir Konvergenzraten für vier verschiedene Schätzer; zwei davon sind neue Methoden. Die Verfahren werden auf der Kugel angewendet und verglichen. Zu diesem Zweck wurde eigens ein R-Paket entwickelt.

Abstract

In this thesis, we study statistical properties of the Fréchet mean and its generalizations in abstract settings. These settings include large classes of scenarios, which may be of great interest in practice when dealing with nonstandard data. Our main focus is on the convergence of sample Fréchet means of independent observations to their population counterpart. The results are exemplarily applied to some specific spaces.

The expectation of a real-valued, square-integrable random variable is characterized by being the unique constant value that minimizes the expected squared difference to the random variable. One can use this property to generalize the notion of mean. A Fréchet mean of a metric space-valued random variable is any minimizer of the expected squared distance to that random variable. This definition achieves two important things: Firstly, it encompasses many commonly used types of mean – like the expectation, the median, or the geometric mean – allowing to state powerful, general, and far-reaching theorems about properties of means. Secondly, it defines a mean for non-Euclidean spaces – like the sphere, the space of phylogenetic trees, or Wasserstein spaces – opening up these spaces for profound applications of probability theory and statistics.

We show strong laws of large numbers of Fréchet mean sets with two different notions of convergence of sets assuming only a first moment condition. After having established consistency of the sample Fréchet mean, we investigate the rate of this convergence. We demonstrate, using projected means, an instance of the Fréchet mean, that Fréchet means may exhibit very different rates depending on the geometry of the metric space and properties of the distribution of the data. Then we prove rates of convergence in a general setting under some conditions. One of these is the quadruple inequality – a generalization of the Cauchy-Schwarz inequality. This and some other conditions are fulfilled in Hadamard spaces – geodesic metric spaces of nonpositive curvature – which makes them particularly interesting to study in the context of Fréchet means. We show a quadruple inequality for certain powers of Hadamard metrics – a purely geometric result with an intriguingly complex proof. Lastly, we examine regression models where responses live in a metric space and the regression function is a conditional Fréchet mean. We compare two approaches to transform known estimators to this non-Euclidean setting. In doing so, we establish rates of convergence for four different estimation procedures, two of which are new methods. To illustrate these regression estimators, an R-package was developed that allows their application and comparison on the sphere.

Contents

1 Introduction

Contents

1.1 The Fréchet Mean

To understand a collection of observations, the first statistic one may want to calculate is the mean, as it summarizes the data in one value. But what does *mean* mean? Depending on the kind of data and the goal of the statistician the most suitable notion of the concept *mean* may vary. For real-valued data, obvious candidates are the arithmetic mean and the median, but in some cases the geometric or harmonic mean might be preferable. If the data lives in a set without vector spaces structure, like a manifold or an abstract metric space, a different concept of *mean* is required.

In this thesis, we explore one type of general mean that can be defined with only little structure but encompasses the common notions of mean – the *Fréchet mean*. Based on Gauß' idea of least squares (*Methode der kleinsten Quadrate* [Gau09]), Fréchet defined the mean value of a collection of objects in a metric space as the minimizer of the summed squared distances to the objects [Fré48].

The Fréchet mean achieves two major things: It provides a common construction for many well-known notions of mean, so that theorems proven for the Fréchet mean imply properties of many interesting objects and statistics. Secondly, it provides a notion of mean in spaces with less or different structure than the Euclidean spaces, e.g., abstract metric spaces or Riemannian manifolds, and thus widens the possibilities of applying probability theory and statistics in these spaces.

1.1.1 Definitions

Before we formally define the Fréchet mean, we first inspect a characterizing property of the Euclidean mean or expectation.

Notation 1.1. Let $s \in \mathbb{N}$. Denote the Euclidean norm in $\mathbb{R}^s$ as $\|\cdot\|$. Denote the Euclidean metric in $\mathbb{R}^s$ as $d_{\mathbb{R}^s}$, i.e., $d_{\mathbb{R}^s}(q,p) := \|q-p\|$ for $q,p \in \mathbb{R}^s$.

Notation 1.2. For a set Q and a function $f\colon Q \to \mathbb{R} \cup \{\infty, -\infty\}$, denote

$$\operatorname*{arg\,min}_{q\in Q} f(q) := \left\{ q \in Q\colon f(q) \leq \inf_{p\in Q} f(p) \right\}.$$

If this set contains only on element m, we may write $m = \arg\min_{q\in Q} f(q)$ instead of $\{m\} = \arg\min_{q\in Q} f(q)$.

For a random variable Y with values in $\mathbb{R}^s$ and $\mathbb{E}[\|Y\|^2] < \infty$, it holds

$$\mathbb{E}[Y] = \operatorname*{arg\,min}_{q\in\mathbb{R}^s} \mathbb{E}[d_{\mathbb{R}^s}(Y,q)^2]\,.$$

Moreover, by adding a constant term, we can also write

$$\mathbb{E}[Y] = \operatorname*{arg\,min}_{q\in\mathbb{R}^s} \mathbb{E}[d_{\mathbb{R}^s}(Y,q)^2 - d_{\mathbb{R}^s}(Y,0)^2]\,.$$

In the latter equation, we only require Y to be once integrable, $\mathbb{E}[\|Y\|] < \infty$, as $|\|y-q\|^2 - \|y\|^2| \leq 2\|y\|\|q\| + \|q\|^2$ for $y,q \in \mathbb{R}^s$. A direct generalization of this characterizing property of the Euclidean mean to arbitrary metric spaces is the Fréchet mean.

Notation 1.3. For a metric space $(\mathcal{Q}, d)$, we may write $\overline{y,q} := d(y,q)$ for $y,q \in \mathcal{Q}$.

Remark 1.4. Measurability concerns are not the focus of this thesis. It is always silently assumed that functions are measurable if necessary, so that all objects are well-defined. For random variables, say X, with values in a set, say $\mathcal{S}$, we assume that there is a silently underlying probability space $(\Omega, \Sigma_\Omega, \mathbb{P})$ and a measurable space $(\mathcal{S}, \Sigma_\mathcal{S})$ such that $X\colon \Omega \to \mathcal{S}$ is measurable. If $\mathcal{S}$ is a metric space, a natural choice of σ-algebra is the Borel σ-algebra.

Definition 1.5. Let $(\mathcal{Q}, d)$ be a metric space. Let $o \in \mathcal{Q}$. Let μ be a probability measure on $\mathcal{Q}$ with $\int \overline{y,o}\,\mathrm{d}\mu(y) < \infty$. The **Fréchet mean set** of μ is defined as the set $\mathbb{M}(\mu) := \mathbb{M}(\mathcal{Q}, d; \mu) := \arg\min_{q\in\mathcal{Q}} \int \overline{y,q}^2 - \overline{y,o}^2\,\mathrm{d}\mu(y)$. An element of $\mathbb{M}(\mu)$ is referred to as **Fréchet mean**.

Remark 1.6. Definition 1.5 does not depend on o. The reason for subtracting $\overline{y,o}^2$ is the same as in the Euclidean case, i.e., we need to make less moment assumptions to obtain a meaningful value: The triangle inequality implies

$$\left|\overline{y,q}^2 - \overline{y,o}^2\right| = |\overline{y,q} - \overline{y,o}|\,(\overline{y,q} + \overline{y,o}) \leq \overline{o,q}\,(\overline{o,q} + 2\overline{y,o})$$

for all $y, q, o \in \mathcal{Q}$. Thus, if $\int \overline{y,o}\,\mathrm{d}\mu(y) < \infty$, then $\int |\overline{y,q}^2 - \overline{y,o}^2|\,\mathrm{d}\mu(y) < \infty$ for all $q \in \mathcal{Q}$. Furthermore, note that $\int \overline{y,o}\,\mathrm{d}\mu(y) < \infty$ if and only if $\int \overline{y,q}\,\mathrm{d}\mu(y) < \infty$ for all $q \in \mathcal{Q}$.

Notation 1.7. We abbreviate *Fréchet mean* as FM and *Fréchet mean set* as FMS.

From the definition of FM for probability measures, we can derive further objects and terminology: Let $(\mathcal{Q}, d)$ be a metric space and $o \in \mathcal{Q}$.

- Let $y_1, \ldots, y_n \in \mathcal{Q}$ and $w_1, \ldots, w_n \in [0, 1]$ with $\sum_{i=1}^n w_i = 1$. Set $\mu_n = \sum_{i=1}^n w_i \delta_{y_i}$, where δ_y denotes the Dirac-delta at point y. Then

$$\mathbb{M}(\mu_n) = \arg\min_{q\in\mathcal{Q}} \sum_{i=1}^n w_i(\overline{y_i,q}^2 - \overline{y_i,o}^2) \tag{1.1}$$

  is the **weighted FMS** of $(y_i)_{i=1,\ldots,n}$ with weights $(w_i)_{i=1,\ldots,n}$.

Notation 1.8. Let $\mathcal{S}, \tilde{\mathcal{S}}$ be sets. Let μ be a probability measure on $\mathcal{S}$. Let $f\colon \mathcal{S} \to \tilde{\mathcal{S}}$. Denote the pushforward as $f_*\mu$, i.e., $f_*\mu$ is a measure on $\tilde{\mathcal{S}}$ with $f_*\mu(B) = \mu(f^{-1}(B))$ for subsets $B \subseteq \tilde{\mathcal{S}}$. Recall that measurability of $\mathcal{S}, \tilde{\mathcal{S}}, f, B$ is silently assumed, see Remark 1.4.

- For a **random variable** Y with values in $\mathcal{Q}$ and distribution $Y_*\mathbb{P}$, its **FMS** is
$$\mathbb{M}[Y] := \mathbb{M}(Y_*\mathbb{P}) = \operatorname*{arg\,min}_{q\in\mathcal{Q}} \mathbb{E}[\overline{Y,q}^2 - \overline{Y,o}^2]\,.$$
- Let $Y_1, \dots, Y_n$ be random observations in $\mathcal{Q}$. Let $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$ be the empirical distribution. The **sample FMS** is
$$\mathbb{M}(\mu_n) = \operatorname*{arg\,min}_{q\in\mathcal{Q}} \frac{1}{n}\sum_{i=1}^n (\overline{Y_i,q}^2 - \overline{Y_i,o}^2)\,.$$
Accordingly, we may call $\mathbb{M}[Y_1]$ the **population FMS** (given that the observations have identical distribution).
- Let (X, Y) be a pair of random variables, where Y has values in $\mathcal{Q}$. For a function $h\colon \mathcal{Q} \to \mathbb{R}$, denote the conditional expectation of $h(Y)$ given X as $\mathbb{E}[h(Y) \mid X]$. Then the **conditional FMS** of Y given X is
$$\mathbb{M}[Y \mid X] := \operatorname*{arg\,min}_{q\in\mathcal{Q}} \mathbb{E}[\overline{Y,q}^2 - \overline{Y,o}^2 \mid X]\,.$$

Remark 1.9. Random arg min-sets, like the conditional FMS or the sample FMS, or random elements of these sets may not always be measurable. One might want to use measurable majorants and outer integrals to be able to derive upper bounds of certain nonmeasurable objects. For a detailed discussion of a technique to deal with such settings, see [VW96].

1.1.2 Sets of Means

Whenever the FMS is a singleton, i.e., it has exactly one element, we may refer to that element as *the* FM. In Euclidean spaces, the FM of a random variable is its expectation, as seen in section 1.1.1.

The FMS may be empty: Let Y be a $\mathbb{R}^s$-valued random variable with $\mathbb{E}[\|Y\|] < \infty$ and $\mathbb{P}(Y = \mathbb{E}[Y]) = 0$. Set $\mathcal{Q} := \mathbb{R}^s \setminus \{\mathbb{E}[Y]\}$. We can view Y as a $\mathcal{Q}$-valued random variable. Then its FMS in $\mathcal{Q}$ with the Euclidean metric is empty, $\mathbb{M}[\mathcal{Q}, d_{\mathbb{R}^s}; Y] = \emptyset$, as $\mathbb{E}[d_{\mathbb{R}^s}(Y,q)^2 - d_{\mathbb{R}^s}(Y,0)^2] = \|q\|^2 - 2\langle q, \mathbb{E}[Y]\rangle$ is always greater than $\mathbb{E}[d_{\mathbb{R}^s}(Y,\mathbb{E}[Y])^2 - d_{\mathbb{R}^s}(Y,0)^2] = -\|\mathbb{E}[Y]\|^2$ for $q \neq \mathbb{E}[Y]$.

Any set can be the FMS in some space: Let $\mathcal{S}$ be a nonempty set. Let $\mathcal{Q} := \mathcal{S} \cup \{\bullet, \star\}$, where $\bullet \neq \star$ and $\bullet, \star \notin \mathcal{S}$. Define the metric $d\colon \mathcal{Q} \times \mathcal{Q} \to [0, \infty)$ as
$$d(y,q) := \begin{cases} 0\,, & \text{if } y = q\,, \\ 1\,, & \text{if } y \neq q \text{ and } |\{y,q\} \cap \{\bullet,\star\}| < 2\,, \\ 2\,, & \text{if } |\{y,q\} \cap \{\bullet,\star\}| = 2\,. \end{cases}$$
Let $\mu := \frac{1}{2}\delta_\bullet + \frac{1}{2}\delta_\star$. Then $\mathbb{M}(\mathcal{Q}, d; \mu) = \mathcal{S}$, as $\int \overline{y,q}^2 \,\mathrm{d}\mu(y) = 1$ for $q \in \mathcal{S}$ and $\int \overline{y,q}^2 \,\mathrm{d}\mu(y) = 2$ for $q \in \{\bullet, \star\}$. We will later see examples of nonsingleton FMSs that are less technical.

1.1.3 Moments

Let $(\mathcal{Q}, d)$ be a metric space, μ be a probability measure on $\mathcal{Q}$, and $\alpha \in (0, \infty)$. We say that μ has a **finite α-moment** if $\int \overline{y,q}^\alpha \, \mathrm{d}\mu(y) < \infty$ for one (and thus, by the triangle inequality, for all) $q \in \mathcal{Q}$.

Notation 1.10. For a set $\mathcal{S}$, we denote by $\mathcal{P}(\mathcal{S})$ the set of all probability measures on $\mathcal{S}$ (measurable structure silently implied). For a metric space $(\mathcal{Q}, d)$ and $\alpha > 0$, denote $\mathcal{P}_\alpha(\mathcal{Q}, d)$ the set of all probability measures μ with $\int d(y, q)^\alpha \, \mathrm{d}\mu(y) < \infty$ for all $q \in \mathcal{Q}$. We may shorten $\mathcal{P}_\alpha(\mathbb{R}^s) := \mathcal{P}_\alpha(\mathbb{R}^s, d_{\mathbb{R}^s})$.

We call $\inf_{q \in \mathcal{Q}} \int \overline{y,q}^2 \, \mathrm{d}\mu(y)$ the **Fréchet variance**. If $m \in \mathbb{M}(\mu)$, then the Fréchet variance equals $\int \overline{y,m}^2 \, \mathrm{d}\mu(y)$. On the euclidean real line $(\mathbb{R}, |\cdot|)$, it is identical to the common notion of variance. [DM19a] present a central limit theorem for the Fréchet variance and use it to obtain an analysis of variance procedure for metric spaces. Building upon these results, a method for change point detection is proposed in [DM19b].

1.1.4 Terminology

The Fréchet mean is also called *barycenter*, *Karcher mean* (although Karcher objects to this name, see [Kar14]), or *center of mass*. One may argue that Fréchet might have had similar concerns as Karcher and conclude that *barycenter* is a better name. In this thesis, the term *Fréchet mean* is used, as this seems to be the term that is best recognized in the literature and statistics community.

1.1.5 Constructions

Given one or more metric spaces, we can construct new metric spaces. We briefly explore some of these constructions and try to describe the behavior of the FM in the new spaces.

Lemma 1.11 (Isometries)**.** Let $(\mathcal{Q}, d)$, $(\tilde{\mathcal{Q}}, \tilde{d})$ be metric spaces such that there is a bijective isometry $f \colon \mathcal{Q} \to \tilde{\mathcal{Q}}$. Let $\mu \in \mathcal{P}_1(\mathcal{Q}, d)$. Then $f_*\mu \in \mathcal{P}_1(\tilde{\mathcal{Q}}, \tilde{d})$ and $\mathbb{M}(\tilde{\mathcal{Q}}, \tilde{d}; f_*\mu) = f(\mathbb{M}(\mathcal{Q}, d; \mu))$.

Proof. We leave out the $d(y, o)^2$-term. The proof is the same when keeping that term. Let $q \in \mathcal{Q}$. Then, as f is an isometry,

$$\int \tilde{d}(z, f(q))^2 \, \mathrm{d}f_*\mu(z) = \int \tilde{d}(f(y), f(q))^2 \, \mathrm{d}\mu(y) = \int d(y, q)^2 \, \mathrm{d}\mu(y) \,.$$

Thus, as f is bijective,

$$\arg\min_{\tilde{q}\in\tilde{\mathcal{Q}}} \int \tilde{d}(z,\tilde{q})^2 \,\mathrm{d}f_*\mu(z) = \arg\min_{q\in\mathcal{Q}} \int d(y,q)^2 \,\mathrm{d}\mu(y)\,. \qquad \square$$

With the same reasoning, we can describe FMs in sets, where a metric is induced by mapping elements to a given metric space.

Lemma 1.12 (Images)**.** Let $(\mathcal{Q}, d)$ be a metric space and $\mathcal{S}$ be a set. Let $f\colon \mathcal{S} \to \mathcal{Q}$. Let $d_f\colon \mathcal{S}\times\mathcal{S} \to [0,\infty), (q,p) \mapsto d(f(q), f(p))$. Then $(\mathcal{S}, d_f)$ is a metric space. Let $\mu \in \mathcal{P}_1(\mathcal{S}, d_f)$. Then $f^{-1}(\mathbb{M}(\mathcal{Q}, d; f_*\mu)) \subseteq \mathbb{M}(\mathcal{S}, d_f; \mu)$ with equality if f is bijective.

We can create new metric spaces, by transforming the distances by a subadditive function. This is an important construction, e.g., for defining medians, but it is not easy to describe the resulting FMs.

Lemma 1.13 (Transformations)**.** Let $(\mathcal{Q}, d)$ be a metric space. Let $g\colon [0,\infty) \to [0,\infty)$ be nondecreasing with $g(0) = 0$ and subadditive, i.e., $g(a+b) \le g(a) + g(b)$ for all $a, b \in [0,\infty)$. Let $d^g\colon \mathcal{Q}\times\mathcal{Q} \to [0,\infty), (q,p) \mapsto g(d(q,p))$. Then $(\mathcal{Q}, d_g)$ is a pseudometric space. If $g(x) = 0$ implies $x = 0$, then $(\mathcal{Q}, d_g)$ is a metric space.

To prove the lemma, one can easily check all requirements of a (pseudo-)metric space. For this construction it may come in handy to know that a function $g\colon [0,\infty) \to [0,\infty)$ that is concave is also subadditive.

The expectation of a random vector is the vector of the expectations of its components. A generalization of this statement is true for FMs, as the following lemma shows.

Lemma 1.14 (Products)**.** Let $J \in \mathbb{N}$. Let $(\mathcal{Q}_1, d_1), \dots, (\mathcal{Q}_J, d_J)$ be metric spaces. Let $w_1, \dots, w_J \in (0,\infty)$. Let $\mathcal{Q} := \bigtimes_{j=1}^J \mathcal{Q}_j$ and $d\colon \mathcal{Q}\times\mathcal{Q} \to [0,\infty)$, $d(q,p) := (\sum_{j=1}^J w_j d_j(q_j,p_j)^2)^{\frac{1}{2}}$. Then $(\mathcal{Q}, d)$ is a metric space. Furthermore, let $\mu \in \mathcal{P}_1(\mathcal{Q}, d)$ with marginal distributions $\mu_1, \dots, \mu_J$. Then $\mu_j \in \mathcal{P}_1(\mathcal{Q}_j, d_j)$, $j = 1, \dots, J$ and $\mathbb{M}(\mathcal{Q}, d; \mu) = \bigtimes_{j=1}^J \mathbb{M}(\mathcal{Q}_j, d_j; \mu_j)$.

Proof. For $o, q \in \mathcal{Q}$,

$$\begin{aligned}\int d(y,q)^2 - d(y,o)^2 \,\mathrm{d}\mu(y) &= \int \sum_{j=1}^J w_j \left(d_j(y_j,q_j)^2 - d_j(y_j,o_j)^2\right) \mathrm{d}\mu(y) \\ &= \sum_{j=1}^J w_j \int \left(d_j(y_j,q_j)^2 - d_j(y_j,o_j)^2\right) \mathrm{d}\mu_j(y_j)\,.\end{aligned}$$

Thus,

$$\inf_{q\in\mathcal{Q}} \int d(y,q)^2 - d(y,o)^2 \,\mathrm{d}\mu(y) = \sum_{j=1}^{J} w_j \inf_{q\in\mathcal{Q}_j} \int d_j(y,q)^2 - d_j(y,o_j)^2 \,\mathrm{d}\mu_j(y)\,,$$

$$\operatorname*{arg\,min}_{q\in\mathcal{Q}} \int d(y,q)^2 - d(y,o)^2 \,\mathrm{d}\mu(y) = \bigtimes_{j=1}^{J} \operatorname*{arg\,min}_{q\in\mathcal{Q}_j} \int d_j(y,q)^2 - d_j(y,o_j)^2 \,\mathrm{d}\mu_j(y)\,. \quad \square$$

Note that in the setting of Lemma 1.14 one can combine the individual metrics d_j differently to a metric on $\mathcal{Q}$. If $\|\cdot\|_\bullet$ is a norm on $\mathbb{R}^J$, then $\|(d_j)_{j=1,\dots,J}\|_\bullet$ is a metric on $\mathcal{Q}$. But then the FM in $\mathcal{Q}$ may not be described as easily. We later introduce power Fréchet means, where we minimize d^α, instead of d^2. Then a similar results as Lemma 1.14 can be shown when replacing the Euclidean (ℓ^2-) norm by the ℓ^α-norm.

1.2 Examples

We present some examples of FMs including common notions of mean as well as nonstandard spaces where the FM can be applied. An overview over the FM and its applications (and its extensions, see section 1.3) is given by the *Map of Means*, Figure 1.1.

1.2.1 Standard Spaces

We call convex subsets of separable Hilbert spaces equipped with the metric induced by the inner product **standard spaces** (also it could be argued that not all of the examples below are commonly considered to be *standard* in an intuitive sense). We can take expectations in these spaces (Lebesgue or Bochner integral) and these expectations coincide with the (unique) FM. The standard spaces include, among others, the Euclidean spaces $\mathbb{R}^s$, the sequence space $\ell^2(\mathbb{R})$, and the 2-Wasserstein space of $\mathbb{R}$: For $\mu,\nu\in\mathcal{P}_2(\mathbb{R})$, define $W_2(\mu,\nu) := (\inf\int (x-y)^2 \,\mathrm{d}\gamma(x,y))^{\frac{1}{2}}$, where the infimum is taken over all probability measures γ on $\mathbb{R}^2$ with marginals μ and ν. The metric W_2 can be expressed as a Hilbert metric on quantile functions, $W_2(\mu,\nu)^2 = \int_0^1 (F_\mu^-(x) - F_\nu^-(x))^2 \,\mathrm{d}x$ [Vil03], where F_μ^-, F_ν^- are the quantile functions of μ and ν. Moreover, $(\mathcal{P}_2(\mathbb{R}), W_2)$ is a metric space that is isometrically isomorphic to a closed and convex subset of the separable Hilbert space $\mathbb{L}_2(\mathbb{R})$ of square integrable functions, see [Big+17].

1.2.2 Kolmogorov Means

Let $I\subseteq\mathbb{R}$ be convex. Let $f\colon I\to f(I)$ be a strictly monotone and continuous function with inverse $f^{-1}\colon f(I)\to I$. Define the metric d_f on I as

$$d_f(y,q) := |f(y) - f(q)|\,.$$

Let $\mu\in\mathcal{P}_1(I,d_f)$. The **Kolmogorov mean** (also *quasi-arithmetic mean* or *generalized f-mean*) is defined as

$$m_f(\mu) := f^{-1}\Big(\int f(y)\,\mathrm{d}\mu(y)\Big)\,.$$

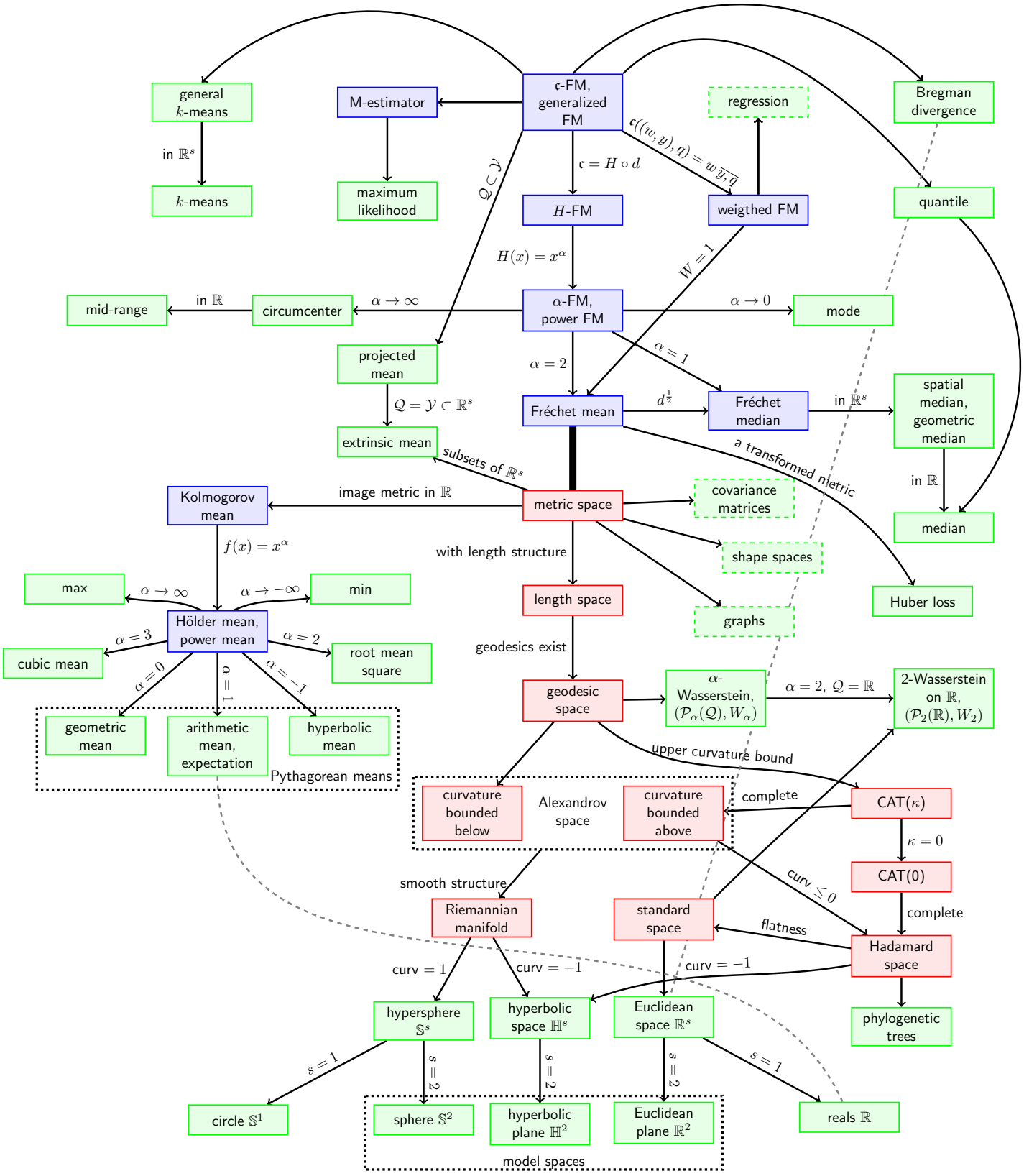

Figure 1.1: Map of Means. This graphic summarizes the instances and extensions of FMs in this chapter. Classes of means are blue, classes of spaces are red, and more specific instances of FMs in certain spaces are green. Dashed colored boxes mark topics that are broad and not described in detail here.

It is the unique Fréchet mean $M(I, d_f; \mu) = \{m_f(\mu)\}$, see Lemma 1.12. Instances of the Kolmogorov mean are the **Hölder means** $m_\alpha(\mu)$ (also *generalized means* or *power means*), where $f(x) = x^\alpha$ with $\alpha \neq 0$ and $I = \mathbb{R}$ for $\alpha \in \mathbb{Z}$ and $I = (0, \infty)$ otherwise. Notable instances of the Hölder means are the **harmonic mean** ($\alpha = -1$), the **arithmetic mean** (or *Euclidean mean*, $\alpha = 1$), the **root mean square** (also *quadratic mean*, $\alpha = 2$), and the **cubic mean** ($\alpha = 3$).

Notation 1.15. For a measure μ, denote the support of μ as $\overline{\mathsf{supp}}(\mu)$.

The definition of Hölder mean can be extended to $\alpha \in \{-\infty, 0, \infty\}$ by taking the respective limit:

$$\begin{aligned} m_{-\infty}(\mu) &:= \lim_{\alpha \to -\infty} m_\alpha(\mu) = \min \overline{\mathsf{supp}}(\mu)\,, \\ m_0(\mu) &:= \lim_{\alpha \to 0} m_\alpha(\mu) = \exp\biggl(\int \log(y)\,\mathrm{d}\mu(y)\biggr)\,, \\ m_\infty(\mu) &:= \lim_{\alpha \to \infty} m_\alpha(\mu) = \max \overline{\mathsf{supp}}(\mu)\,, \end{aligned}$$

where the necessary integrability conditions are assumed. The Hölder mean with $\alpha = 0$ is also called **geometric mean**.

The arithmetic, geometric, and harmonic mean collectively are known as the **Pythagorean means**. For $y_1, \dots, y_n \in (0, \infty)$ and the empirical measure $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$, the geometric mean is $m_0(\mu_n) = (\prod_{i=1}^n y_i)^{\frac{1}{n}}$. It is the side length of a n-dimensional hypercube with the same volume as the n-dimensional box with side lengths $y_1, \dots, y_n$. The harmonic mean can be illustrated as follows: If an athlete runs n rounds on a running track, each with velocity $y_1, \dots, y_n$, respectively, then another athlete, who always runs at constant speed, runs the same distance in the same time if their velocity is equal to the harmonic mean $m_{-1}(\mu_n) = n(\sum_{i=1}^n \frac{1}{y_i})$.

The Hölder means are ordered in the sense that $m_\alpha(\mu) \leq m_{\tilde\alpha}(\mu)$ for $\alpha \leq \tilde\alpha$, as the following lemma shows with $x \mapsto x^{\frac{\tilde\alpha}{\alpha}}$ convex.

Lemma 1.16. Let $I \subseteq \mathbb{R}$ be convex. Let $f, \tilde f \colon I \to \mathbb{R}$ be strictly monotone and continuous functions. Let μ be a probability measure on I with $\int |\tilde f(y)|\,\mathrm{d}\mu(y) < \infty$. Assume that $\tilde f \circ f^{-1}$ is convex. Then $m_f \leq m_{\tilde f}$.

Proof. Use Jensen's inequality to obtain

$$f^{-1}\biggl(\int f(y)\,\mathrm{d}\mu(y)\biggr) = \tilde f^{-1} \circ \tilde f \circ f^{-1}\biggl(\int f(y)\,\mathrm{d}\mu(y)\biggr) \leq \tilde f^{-1}\biggl(\int \tilde f(y)\,\mathrm{d}\mu(y)\biggr)\,. \qquad \square$$

All Kolmogorov means have properties that one naturally associates with the term *mean*. Kolmogorov even defined the mean by four axioms of desirable properties:

Definition 1.17 (Kolmogorov's axioms of means, [Kol30])**.** A collection of functions $(M_n)_{n\in\mathbb{N}}$, $M_n\colon \mathbb{R}^n \to \mathbb{R}$ is called **regular mean** if following four conditions are fulfilled:

1. All M_n are continuous and nondecreasing in each variable.
2. All M_n are invariant under permutation of their arguments.
3. $M_n(y,\dots,y) = y$ for $y \in \mathbb{R}$ and $n \in \mathbb{N}$.
4. Let $k, n \in \mathbb{N}$ with $k < n$, $y \in \mathbb{R}^n$ and $m = M_k(y_1,\dots,y_k)$. Then $M_n(y) = M_n(m,\dots,m,y_{k+1},\dots,y_n)$.

Kolmogorov showed that any regular mean is of the form $M_n(y) = f^{-1}\Big(\frac{1}{n}\sum_{i=1}^n f(y_i)\Big)$, where f is continuous and strictly monotone. For further elementary properties of these means, see [HLP52, chapter 3] and [AD89, chapter 17]. [Car16] shows a central limit theorem for Kolmogorov means.

1.2.3 Median and Huber Loss

A **median** of a probability distribution μ on $\mathbb{R}$ is any point m with $\int_{-\infty}^m y\,\mathrm{d}\mu(y) = \int_m^\infty y\,\mathrm{d}\mu(y)$. It is well-known that the set of all medians is the FMS of $(\mathbb{R}, d^{\frac{1}{2}})$, where $d^{\frac{1}{2}}(y,q) := \sqrt{|y-q|}$.

Notation 1.18. Let S, Q be sets with $S \subseteq Q$. Then $\mathbb{1}_S\colon Q \to \{0,1\}$ denotes the indicator function of S., i.e., $\mathbb{1}_S(x) = 1$ if and only of $x \in S$.

A mixture of median and expectation can be attained by the Huber loss: Let $\delta > 0$. Let $h_\delta\colon [0,\infty) \to [0,\infty), x \mapsto \frac{1}{\sqrt{2}}x\mathbb{1}_{[0,\delta]}(x) + \sqrt{\delta(|x| - \frac{1}{2}\delta)}\mathbb{1}_{(\delta,\infty)}(x)$. Then h^2 is called **Huber loss**, [Hub64]. Furthermore, $(\mathbb{R}, d^h)$ is a metric space, see Lemma 1.13, as h is concave. The FMs in this space are commonly used in robust statistics.

Now, let $(\mathcal{Q}, \|\cdot\|)$ be a Banach space with induced metric $d_{\|\cdot\|}(q,p) = \|q-p\|$. A generalization of the median to Banach spaces is the **geometric median** (or spatial median), see [Kem87]. It is the FM in $(\mathcal{Q}, d^{\frac{1}{2}}_{\|\cdot\|})$, i.e.,

$$\mathbb{M}(\mathcal{Q}, d^{\frac{1}{2}}_{\|\cdot\|}, \mu) = \operatorname*{arg\,min}_{q\in\mathcal{Q}} \int \|y-q\| - \|y\|\,\mathrm{d}\mu(y)\,.$$

The geometric median is unique if there is no one-dimensional subspace where μ has mass one [MD87].

1.2.4 Extrinsic and Projected Mean

If a Riemannian manifold $\mathcal{Q}$ is embedded in an ambient Euclidean space, $\mathcal{Q} \subseteq \mathbb{R}^s$, we can take the FM with respect to the extrinsic distance, i.e., the Euclidean distance $d_{\mathbb{R}^s}$.

More generally, let $\mathcal{Q} \subseteq \mathbb{R}^s$. Then $(\mathcal{Q}, d_{\mathbb{R}^s})$ is a metric space. The FM in this space is called **extrinsic mean**. We may allow the distribution μ to have mass in the ambient space $\mathbb{R}^s$, not only on $\mathcal{Q}$. Then the respective FMs are called **projected means**: Let $\mu \in \mathcal{P}_1(\mathbb{R}^s, d_{\mathbb{R}^s})$. Let $m = \int y \, \mathrm{d}\mu(y)$. Then

$$\operatorname*{arg\,min}_{q \in \mathcal{Q}} \int \overline{y,q}^2 - \overline{y,o}^2 \, \mathrm{d}\mu(y) = \operatorname*{arg\,min}_{q \in \mathcal{Q}} \|m - q\| \,,$$

i.e., the projected means really are the projections of the Euclidean mean m to $\mathcal{Q}$. If $\mathcal{Q}$ is convex and $\overline{\mathsf{supp}}(\mu) \subseteq \mathcal{Q}$, then $m \in \mathcal{Q}$ and $\mathbb{M}(\mathcal{Q}, d_{\mathbb{R}^s}, \mu) = \{m\}$. The projected mean can only be unique if m does not lie on the so-called medial axis of $\mathcal{Q}$. The **medial axis** $\mathcal{M}_\mathcal{Q}$ is the set of all points in $\mathbb{R}^s$ that do not have a unique projection:

$$\mathcal{M}_\mathcal{Q} := \left\{ z \in \mathbb{R}^s \;\middle|\; \exists p_1, p_2 \in \mathcal{Q}, p_1 \neq p_2 \colon \|p_1 - z\| = \|p_2 - z\| = \inf_{p \in \mathcal{Q}} \|p - z\| \right\} .$$

For geometric properties of the medial axis, see [BD17]. In chapter 3, we discuss the projected mean in more detail. In particular, we investigate the influence of the distance of m to $\mathcal{M}_\mathcal{Q}$ on the rate of convergence of sample projected means to their population counterpart.

1.2.5 Geodesic Spaces

[Kar77] developed theory on the FM for Riemannian manifolds $(\mathcal{Q}, g)$, where $\mathcal{Q}$ is a smooth manifold and g is a Riemannian metric. The tuple $(\mathcal{Q}, d_g)$ with intrinsic distance d_g (length of shortest path between two points) is a metric space. In this context the FM is often called (Riemannian) center of mass or intrinsic mean. In contrast to the extrinsic mean, it is not necessary to refer to an ambient Euclidean space.

There have been many contributions to the theory and applications of FMs on Riemannian manifolds since Karcher's original article. We only mention few and refer the reader to [HE20] for an overview. [BP03; BP05] show consistency results and a central limit theorem for FMs of random observations on Riemannian manifolds. [EH19] extend this central limit theorem to a more general setting, in which – due to the geometry of the space – rates of convergence slower than $n^{\frac{1}{2}}$ can be observed.

A generalization of Riemannian manifolds with a bound on the sectional curvature are geodesic spaces with curvature bounds. These include many interesting nonsmooth spaces and seem to be a more natural realm for the study of Fréchet means in general as their structure is based on a metric like the definition of the FM. We quickly introduce some necessary terms of metric geometry. See [BBI01] for a more detailed introduction.

Let $(\mathcal{Q}, d)$ be a metric space. For a continuous map $\gamma \colon [a, b] \to \mathcal{Q}$ define its **length** as

$$L(\gamma) := \sup \left\{ \sum_{i=1}^{n} d(\gamma(x_{i-1}), \gamma(x_i)) \;\middle|\; a = x_0 < x_1 < \cdots < x_n = b, n \in \mathbb{N} \right\} .$$

Define the **inner metric** (also *intrinsic metric*) of $(\mathcal{Q}, d)$ as $d_{\mathrm{i}}(q, p) := \inf L(\gamma)$, where the infimum is taken over all continuous maps $\gamma \colon [a, b] \to \mathcal{Q}$ with $\gamma(a) = q$ and $\gamma(b) = p$.

A **length space** is a metric space $(\mathcal{Q}, d)$ with $d = d_{\mathrm{i}}$. Now, let $(\mathcal{Q}, d)$ be a length space. A continuous map $\gamma\colon [a,b] \to \mathcal{Q}$ is called **shortest path** if $L(\gamma) \leq L(\tilde{\gamma})$ for all continuous maps $\tilde{\gamma}\colon [\tilde{a}, \tilde{b}] \to \mathcal{Q}$ with $\gamma(a) = \tilde{\gamma}(\tilde{a})$ and $\gamma(b) = \tilde{\gamma}(\tilde{b})$. A continuous map $\gamma\colon [a,b] \to \mathcal{Q}$ is **locally minimizing** if for every $t \in (a,b)$ there is $\epsilon > 0$ such that the restriction $\gamma_{|[t-\epsilon, t+\epsilon]}$ is a shortest path. A continuous map $\gamma\colon [a,b] \to \mathcal{Q}$ has **constant speed** if there is $v \geq 0$ such that $L(\gamma_{|[a',b']}) = v(b'-a')$ for all $a', b' \in [a,b]$ with $a' < b'$. A **geodesic** is a locally minimizing continuous map with constant speed. The tuple $(\mathcal{Q}, d)$ is a **geodesic space** if there is a geodesic for every pair of points.

Curvature bounds in geodesic spaces can be defined via comparison of triangles with model spaces of constant curvature. The **model spaces** M_κ are the unique complete simply connected real 2-dimensional Riemannian manifolds of constant sectional curvature κ, i.e.

- the Euclidean plane $\mathbb{R}^2$ with Euclidean metric $d_{\mathbb{R}^2}$, for $\kappa = 0$,
- the sphere $\mathbb{S}^2(\kappa^{-\frac{1}{2}})$ with radius $\kappa^{-\frac{1}{2}}$ with intrinsic metric, for $\kappa > 0$,
- the hyperbolic plane $\mathbb{H}^2$ with the standard metric multiplied by $(-\kappa)^{-\frac{1}{2}}$, for $\kappa < 0$.

See [BBI01, chapter 4 and 5] for a precise definition of the model spaces.

Notation 1.19. For a metric space $(\mathcal{Q}, d)$ and $B \subseteq \mathcal{Q}$ denote the diameter of b as $\mathsf{diam}(B) := \sup_{q,p \in B} d(q,p)$.

Let $D_\kappa := \mathsf{diam}(M_\kappa)$, i.e., $D_\kappa = \pi/\sqrt{\kappa}$ for $\kappa > 0$, otherwise $D_\kappa = \infty$. A geodesic space $(\mathcal{Q}, d)$ has **curvature bounded below** (respectively **above**) by κ if for any three points $y, q, p \in \mathcal{Q}$ with $d(q,p) + d(y,q) + d(y,p) < 2D_\kappa$ it holds

$$d(y, \gamma_t) \geq d_\kappa(\bar{y}, \bar{\gamma}_t) \qquad (\text{ respectively } \quad d(y, \gamma_t) \leq d_\kappa(\bar{y}, \bar{\gamma}_t) \ )$$

for $t \in [0,1]$, where $\gamma\colon [0,1] \to \mathcal{Q}$ is a geodesic with $\gamma_0 = q$, $\gamma_1 = p$, $(\bar{y}, \bar{q}, \bar{p})$ is a triple of points in M_κ that is isometric to (y, q, p), and $\bar{\gamma}\colon [0,1] \to M_\kappa$ is the geodesic connecting $\bar{\gamma}_0 = \bar{q}$ with $\bar{\gamma}_1 = \bar{p}$. Informally, triangles in negatively curved spaces are thinner than the Euclidean triangle and thicker in positively curved spaces.

Geodesic spaces with upper curvature bound κ are called **CAT(κ)**-spaces. Complete geodesic spaces with curvature bound are collectively called **Alexandrov spaces**. Complete CAT(0)-spaces are also called **Hadamard spaces** or **global NPC-spaces** (**n**on**p**ositive **c**urvature). Instances of these spaces are Riemannian manifolds with the respective upper or lower bound κ on the sectional curvature. Standard spaces are flat Hadamard spaces, i.e., spaces with constant curvature 0. The hyperbolic spaces $\mathbb{H}^s$ are CAT(-1), the Euclidean spaces $\mathbb{R}^s$ are CAT(0), and the unit hyperspheres $\mathbb{S}^s$ are CAT(1). All three are Alexandrov spaces.

Hadamard spaces have some desirable properties, e.g., unique FMs, that make them particularly interesting. We show rates of convergence of sample FMs in Hadamard spaces in 4.4.4 and 4.5, and throughout chapter 5, we present different results for regression estimators with responses in Hadamard spaces. Following property characterizes Hadamard spaces.

Lemma 1.20 ([Stu03])**.** A nonempty complete metric space $(\mathcal{Q}, d)$ is Hadamard if and only if for all $q, p \in \mathcal{Q}$, there is $m \in \mathcal{Q}$ such that $d(y,m)^2 \leq \frac{1}{2}d(y,q)^2 + \frac{1}{2}d(y,p)^2 - \frac{1}{4}d(q,p)^2$ for all $y \in \mathcal{Q}$.

In Hadamard spaces, all geodesics are minimizing. [Stu03] shows how in these spaces some classical results of probability theory in Euclidean spaces (e.g., strong law of large numbers, Jensen's inequality) can be transferred to the Fréchet mean setting. An algorithm for calculating Fréchet means in Hadamard spaces is described in [Bač14a].

One important application of statistics in Hadamard spaces is the space phylogenetic trees. A phylogenetic tree represents the genetic relatedness of biological species, including bacteria and viruses. The geometry of the space of phylogenetic trees T_m with m leaves is studied in [BHV01]. In particular, it is shown that T_m is a Hadamard space. There has been a lot of recent interest in statistics on T_m. E.g., [BLO18] show a central limit theorem for the Fréchet mean in T_m and [Nye11] apply principal component analysis in that space.

Hadamard spaces are complete CAT(0)-spaces. In CAT(κ)-spaces with $\kappa \in \mathbb{R}$, there are simple conditions for the FM to be unique.

Theorem 1.21 ([Stu03; Yok17])**.** Let $(\mathcal{Q}, d)$ be a complete CAT(κ)-space. Let $\mu \in \mathcal{P}_1(\mathcal{Q}, d)$.

1. Assume $\kappa \leq 0$. Then the FM of μ is unique.
2. Assume $\kappa > 0$. Assume $\sqrt{\kappa}\ \mathsf{diam}(\overline{\mathsf{supp}}(\mu)) < \pi$. Then the FM of μ is unique.

In both cases the FM lies in the convex hull of $\overline{\mathsf{supp}}(\mu)$.

Similar results hold for power FMs, which are introduced in section 1.3.2.

1.2.6 Wasserstein Spaces

Let $(\mathcal{Q}, d)$ be a metric space and $\alpha \geq 1$. Let $\mu, \nu \in \mathcal{P}_\alpha(\mathcal{Q}, d)$. The α-Wasserstein distance [Vil09, Definition 6.1] is

$$W_\alpha(\mu, \nu) := \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int d(x,y)^\alpha \, \mathrm{d}\gamma(x,y) \right)^{\frac{1}{\alpha}}.$$

where $\Gamma(\mu, \nu)$ is the set of all probability measures on $\mathcal{Q} \times \mathcal{Q}$ with marginals μ and ν. $(\mathcal{P}_\alpha(\mathcal{Q}, d), W_\alpha)$ is a metric space. If $(\mathcal{Q}, d)$ is complete and separable, then so is $(\mathcal{P}_\alpha(\mathcal{Q}, d), W_\alpha)$ [Vil09, Theorem 6.18]. If $\mathcal{Q}$ is compact, then so is $\mathcal{P}_\alpha(\mathcal{Q}, d)$, [Vil09, Remark 6.18]. If $\mathcal{Q}$ is a locally compact geodesic space, then $\mathcal{P}_\alpha(\mathcal{Q}, d)$ is a geodesic space, [LV09, Lemma 2.4 and Proposition 2.6]. The Wasserstein distance metrizises weak convergence: convergence in the Wasserstein space is weak convergence of probability measures plus convergence of the α-moment [Vil09, Theorem 6.9].

In Wasserstein spaces, FMs are usually called (Wasserstein) barycenters. [AC11] discuss existence, uniqueness, and characterizations of the 2-Wasserstein barycenters in $\mathbb{R}^s$. [LL15] show existence and consistency of α-Wasserstein barycenters in geodesic spaces. [KP17] discuss 2-Wasserstein barycenters on Riemannian manifolds. [ZP19] explore the link between Fréchet means in the Wasserstein space and Procrustes analysis. The Wasserstein covariance is introduced in [PM19b] to analyze the dependence between multiple random densities. A regression framework, where predictors and responses are distributions in developed in [CLM20].

1.2.7 Further applications

The FM has been applied to find a mean of graphs, e.g., [GSK12; GGR18]. In [WM07], it is used in tree spaces to analyze blood vessel data. The FM is used in Kendall's shape spaces and Procrustes analysis [Gow75; Ken84; DM16]. Covariance matrices can be averaged using FMs of different metric spaces that induce certain desirable properties, e.g., [DKZ09; PDM19]. For more applications and an overview over further methods for nonstandard spaces, see [MA14] and [HE20].

1.3 Extensions

Aside from application of Fréchet means on different spaces, there are also many modifications and generalizations of the concept that encompass even more interesting objects in one description.

1.3.1 Restriction of the Descriptor Set

Let $(\mathcal{Q}, d)$ be a metric space. Let $o \in \mathcal{Q}$. Let μ be a probability distribution on $\mathcal{Q}$. Instead of minimizing over the whole set $\mathcal{Q}$ one might search for a minimizer only in a subset of $\mathcal{Q}$, see also section 1.2.4. The subset might represent known theoretical or computational constraints. One specific instance of interest is the **support-restricted FMS**,

$$\operatorname*{arg\,min}_{q \in \overline{\mathsf{supp}}(\mu)} \int \overline{y,q}^2 - \overline{y,o}^2 \,\mathrm{d}\mu(y)\,.$$

It can be useful for computation of an approximate FM. [Sve81; EJ20] show that elements of $\mathbb{M}(\overline{\mathsf{supp}}(\mu_n), d, \mu_n)$ converge to elements of $\mathbb{M}(\overline{\mathsf{supp}}(\mu), d, \mu)$ for $\mu \in \mathcal{P}_2(\mathcal{Q}, d)$ and the empirical measure $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$ of independent and identically distributed random variables Y_i with distribution μ.

1.3.2 Power Fréchet Means

Let $(\mathcal{Q}, d)$ be a metric space. Let $o \in \mathcal{Q}$. Let μ be a probability distribution on $\mathcal{Q}$. Let $\alpha > 0$. Let $\mu \in \mathcal{P}(\mathcal{Q})$. For $\alpha > 1$, assume $\mu \in \mathcal{P}_{\alpha-1}(\mathcal{Q}, d)$. The **power Fréchet mean**

set or α-**Fréchet mean set** of μ is

$$\mathbb{M}^\alpha(\mu) := \mathbb{M}^\alpha(\mathcal{Q}, d; \mu) := \operatorname*{arg\,min}_{q \in \mathcal{Q}} \int \overline{y,q}^\alpha - \overline{y,o}^\alpha \, \mathrm{d}\mu(y) \,.$$

In chapter 2, we show strong laws of large numbers for α-FMSs. Note that the power Fréchet means are not directly connected to the power means of section 1.2.2.

For $\alpha = 2$ the α-FMS is the usual FMS, $\mathbb{M}^2(\mu) = \mathbb{M}(\mu)$. For $\alpha \in (0, 2]$, $d^{\frac{\alpha}{2}}$ is a metric. Thus, $\mathbb{M}(\mathcal{Q}, d^{\frac{\alpha}{2}}; \mu) = \mathbb{M}^\alpha(\mathcal{Q}, d; \mu)$. For $\alpha > 2$, d^α may not be a metric. Hence, the power FMs are more general than the FM.

For $\alpha = 1$ in a Banach space $\mathcal{Q}$, we again obtain the (geometric) median, see section 1.2.3. In general metric spaces, the elements of $\mathbb{M}^1(\mu)$ are called **Fréchet median**, e.g., [ABY13]. As mentioned before, the Fréchet median for a metric space is the Fréchet mean of another metric space. In [FVJ09] the Fréchet median on Riemannian manifolds is discussed. [Yok17, Corollary 41] presents conditions for uniqueness of Fréchet medians in CAT(κ)-spaces. A Nadaraya-Watson-type nonparametric regression procedure on tree spaces is presented in [Wan+12], where the target function is described by a conditional Fréchet median.

In $(\mathbb{R}, d_\mathbb{R})$, the limit cases $\alpha \to 0$ and $\alpha \to \infty$ are the familiar **mode** and **mid-rage** of a distribution. They generalize to metric spaces, by defining

$$\begin{aligned} \mathsf{mode}(\mu) &:= \left\{ m \in \mathcal{Q} \colon \limsup_{r \to 0} \frac{\mu(B_r(q))}{\mu(B_r(m))} \leq 1 \text{ for all } q \in \mathcal{Q} \right\} , \\ \mathsf{circumcenter}(\mu) &:= \operatorname*{arg\,min}_{q \in \mathcal{Q}} \sup_{y \in \overline{\mathsf{supp}}(\mu)} d(y, q) \,. \end{aligned}$$

The limit cases of the power-FM can be interpreted as mode ($\alpha \to 0$) and circumcenter ($\alpha \to \infty$), see [Mac67]. The precise relationship between the family $(\mathbb{M}^\alpha(\mu))_{\alpha \in (0,\infty)}$, $\mathsf{mode}(\mu)$, and $\mathsf{circumcenter}(\mu)$ may be complex, in particular if the sets are not singletons.

Slightly more general than power FMs are H-FMs: Let $H \colon [0, \infty) \to [0, \infty)$ be a convex and nondecreasing function. The set of minimizers

$$\mathbb{M}^H(\mu) := \mathbb{M}^H(\mathcal{Q}, d; \mu) := \operatorname*{arg\,min}_{q \in \mathcal{Q}} \int H(\overline{y,q}) - H(\overline{y,o}) \, \mathrm{d}\mu(y)$$

is called H-**Fréchet mean set** or **convex Fréchet mean set**. For $H(x) = x^\alpha$ with $\alpha \geq 1$, we obtain the power FMs. [Yok17, Theorem 40] shows conditions for uniqueness of H-FMs in CAT(κ)-spaces.

The assumption that H is convex is not restrictive: For a concave, nondecreasing function G with $G^{-1}(\{0\}) = \{0\}$, we may interpret $G \circ d$ as a metric. Thus, this case is covered by the original definition of FMS.

1.3.3 Generally Weighted Fréchet Mean

In section 1.1.1, we introduced the weighted FM of points in a metric space, which is nothing but the usual FM of a certain probability measure. Only positive weights can

be treated like that. In some settings, it might be of use to allow negative weights, e.g., linear regression can be viewed as a weighted mean with not necessarily positive weights as we will see shortly. A negatively weighted FM can be justified as a generalization of a Euclidean setting:

Lemma 1.22. Let (W, Y) be an pair of random variables with values in $\mathbb{R} \times \mathbb{R}^s$ such that $\mathbb{E}[\|WY\|] < \infty$. Define $a := \mathbb{E}[W]$. Assume $a > 0$. Then it holds

$$\operatorname*{arg\,min}_{q \in \mathbb{R}^s} \mathbb{E}\Big[W\left(d_{\mathbb{R}^s}(Y,q)^2 - d_{\mathbb{R}^s}(Y,0)^2\right)\Big] = \frac{1}{a}\mathbb{E}[WY]\,.$$

Proof. Set $m := \mathbb{E}[WY]$. For every $q \in \mathbb{R}^s$, it holds

$$\mathbb{E}\Big[W\left(d_{\mathbb{R}^s}(Y,q)^2 - d_{\mathbb{R}^s}(Y,0)^2\right)\Big] = -2q^\top m + a\|q\|^2\,.$$

Hence,

$$\mathbb{E}\left[W d_{\mathbb{R}^s}(Y,q)^2 - W d_{\mathbb{R}^s}\left(Y, \frac{1}{a}m\right)^2\right] = \frac{1}{a}\|m\|^2 - 2q^\top m + a\|q\|^2 = \left\|\frac{1}{\sqrt{a}}m - \sqrt{a}q\right\|^2 \geq 0,$$

which shows that $\frac{1}{a}m$ minimizes $q \mapsto \mathbb{E}[W\left(d_{\mathbb{R}^s}(Y,q)^2 - d_{\mathbb{R}^s}(Y,0)^2\right)]$. □

The lemma shows that $\mathbb{E}[W] = 1$ is required to make the weighing meaningful for minimizers and leads to following definition: Let $(\mathcal{Q}, d)$ be a metric space. Let $o \in \mathcal{Q}$. Let W be a real valued random variable with $\mathbb{E}[W] = 1$. Let Y be a $\mathcal{Q}$-valued random variable such that $\mathbb{E}[|W|\,\overline{Y,o}] < \infty$. Then the **generally weighed FMS** of (W, Y) is

$$\operatorname*{arg\,min}_{q \in \mathcal{Q}} \mathbb{E}\Big[W\left(\overline{Y,q}^2 - \overline{Y,o}^2\right)\Big]\,.$$

An analogous definition for probability measures is possible.

Remark 1.23. We can view a weighted mean $\mathbb{E}[WY] = \int W(\omega)Y(\omega)\,\mathrm{d}\mathbb{P}(\omega)$ as the usual mean with respect to the transformed measure $\nu = W \cdot \mathbb{P}$. If $\mathbb{E}[W] = 1$ and $W(\omega) \geq 0$ $\mathbb{P}$-almost everywhere, W is just the density of the probability measure ν with respect to $\mathbb{P}$. In the upcoming section we usually have to deal with the case $\mathbb{E}[W] = 1$, but $\mathbb{P}(W < 0) > 0$. Then ν is a signed measure.

We can apply the generally weighted FM to least squares regression estimators, where target values live in a metric space. But, first consider a general regression setting with real-valued targets, which encompasses different common regression scenarios. Let (X, Y) be a pair of random variables with values in $(\mathcal{X} \times \mathbb{R})$. Let $\Psi\colon \mathcal{X} \to \mathbb{R}^r$ be a feature function, $K\colon \mathcal{X} \to \mathbb{R}$ a localizing function. We are interested in the value

$$m(t) := \Psi(t)^\top \theta_0$$

with the least squares parameters

$$\theta_0 \in \operatorname*{arg\,min}_{\theta\in\mathbb{R}^r} \mathbb{E}\Big[\Big(Y - \theta^\top\Psi(X)\Big)^2 K(X)\Big] = \operatorname*{arg\,min}_{\theta\in\mathbb{R}^r}\Big(-2\theta^\top a + \theta^\top B\theta\Big)$$

where $a = \mathbb{E}[\Psi(X)K(X)Y]$ and $B = \mathbb{E}[\Psi(X)\Psi(X)^\top K(X)]$. The matrix B is symmetric. Assume that B is also positive definite. Then $\theta_0 = B^{-1}a$ and

$$m(t) = \Psi(t)^\top B^{-1}a = \mathbb{E}\Big[\Psi(t)^\top B^{-1}\Psi(X)K(X)Y\Big] = \mathbb{E}[w(t,X)Y]$$

for $w(t,x) = \Psi(t)^\top B^{-1}\Psi(x)K(x)$. This general setting includes many well-known regression estimators.

Example 1.24.

(i) **Expectation:**
Let $r = 1$, $K \equiv \Psi \equiv 1$, $\mathcal{X} = \{\star\}$. Then $m(t) = m = \mathbb{E}[Y]$.

(ii) **linear regression:**
Let $r \in \mathbb{N}$, $K \equiv 1$, $\Psi = \mathrm{id}$, $\mathcal{X} = \mathbb{R}^r$. Then $m(t) = t^\top\theta_0$, where

$$\theta_0 \in \operatorname*{arg\,min}_{\theta\in\mathbb{R}^r} \mathbb{E}\Big[\Big(Y - \theta^\top X\Big)^2\Big].$$

(iii) **Projection estimator:**
Let $\psi_k\colon \mathcal{X} \to \mathbb{R}$ for $k \in \mathbb{N}$, $K \equiv 1$, $\Psi(x) = \Psi_r(x) = (\psi_k(x))_{k=1,\dots,r}$. Then $m(t) = m_r(t)$ is the expectation of the projection estimator with projection to the space spanned by $\psi_1, \dots, \psi_r$. An example is trigonometric projection, where $\mathcal{X} = [0,1]$ and $(\psi_k)_{k\in\mathbb{N}}$ is the trigonometric basis of $\mathbb{L}_2([0,1])$. See, e.g., [Tsy08, section 1.7].

(iv) **Local polynomial estimator:**
Let $\mathcal{X} = \mathbb{R}$, $\kappa\colon \mathbb{R} \to \mathbb{R}$ be a kernel, $h > 0$, and

$$K(x) = K_{t,h}(x) = \frac{1}{h}\kappa\Big(\frac{x-t}{h}\Big).$$

Let $N \in \mathbb{N}_0$ specify the polynomial degree, set $r = N + 1$. Denote

$$\psi(x) = \left(\frac{x^k}{k!}\right)_{k=0,\dots,N}, \quad \Psi(x) = \Psi_{t,h}(x) = \psi\Big(\frac{x-t}{h}\Big).$$

Then $m(t) = m_h(t)$ is the expectation of the local polynomial estimator of degree N with bandwidth h. See, e.g., [Tsy08, section 1.6].

The regression function $m(t) = \mathbb{E}[w(t,X)Y]$ has the form of a weighted mean, where the weight may take negative values. We can easily generalize this setting to cases where Y

lives in a metric space $(\mathcal{Q}, d)$ by using the generally weighted FM:

$$\mathbb{M}(t) := \operatorname*{arg\,min}_{q \in \mathcal{Q}} \mathbb{E}\Big[w(t, X)\left(d(Y, q)^2 - d(Y, o)^2\right)\Big],$$

where $o \in \mathcal{Q}$ is an arbitrary element. To justify this transfer, we need $\mathbb{E}[w(t, X)] = 1$, see Lemma 1.22, which is easy to obtain as the next lemma shows.

Lemma 1.25. With the above definitions, if there is a vector $R \in \mathbb{R}^r$ with $\Psi(x)^\top R = 1$ for all $x \in \mathcal{X}$, then $\mathbb{E}[w(t, X)] = 1$.

Proof. As $\Psi(x)^\top R = 1$, we can write the weight function as

$$w(t, x) = \Psi(t)^\top B^{-1} \Psi(x) K(x) = \Psi(t)^\top B^{-1} \Psi(x) \Psi(x)^\top R K(x)\,.$$

Taking the expectation then yields the desired result,

$$\mathbb{E}[w(t, X)] = \Psi(t)^\top B^{-1} B R = \Psi(t)^\top R = 1\,. \qquad \square$$

The condition of the lemma is fulfilled if the first entry of Ψ is constant 1, which is common. E.g., for linear regression, this only requires the inclusion of an intercept in the model.

The regression function may be estimated via a plug-in estimator, replacing the expectation by a sum over observations. This scenario is called **Fréchet regression** and is investigated in [PM19a] and chapter 5 of this thesis.

1.3.4 Generalized Fréchet Mean

Let $\mathfrak{c}\colon \mathcal{Y} \times \mathcal{Q} \to \mathbb{R}$ be a function. Let Y be a random variable with values in $\mathcal{Y}$. Let $M := \arg\min_{q \in \mathcal{Q}} \mathbb{E}[\mathfrak{c}(Y, q)]$ assuming the expectations exist. In this context, $\mathfrak{c}$ is called *cost function*, $\mathcal{Y}$ is called *data space*, $\mathcal{Q}$ is called *descriptor space*, $q \mapsto \mathbb{E}[\mathfrak{c}(Y, q)]$ is called *objective function* (or Fréchet function), and M is called *generalized Fréchet mean set* or $\mathfrak{c}$*-Fréchet mean set.*

This is setting encompasses the previous generalizations of FM. Set $\mathfrak{c}(y, q) = \overline{y,q}^\alpha - \overline{y,o}^\alpha$ to obtain power Fréchet means. Set $\mathfrak{c}((y, w), q) = w(\overline{y,q}^2 - \overline{q,o}^2)$ to obtain weighted Fréchet means. Set $\mathfrak{c}((x, y), \beta) = (y - \beta^\top x)^2$ for linear regression.

This general scenario contains the setting of general M-estimation. It includes many important statistical frameworks like maximum likelihood estimation, where $\mathcal{Q} = \Theta$ parameterizes a family of densities $(f_\vartheta)_{\vartheta \in \Theta}$ on $\mathcal{Y} = \mathbb{R}^s$ and $\mathfrak{c}(x, \vartheta) = -\log f_\vartheta(x)$, or linear regression, where $\mathcal{Q} = \mathbb{R}^{s+1}$, $\mathcal{Y} = (\{1\} \times \mathbb{R}^s) \times \mathbb{R}$, $\mathfrak{c}((x, y), \beta) = (y - \beta^\top x)^2$. It also includes nonstandard settings, e.g., [Huc11], where geodesics in $\mathcal{Q}$ are fitted to points in $\mathcal{Y}$.

We use this general setting throughout this thesis to present theorems on consistency (chapter 2) and rate of convergence of $\mathfrak{c}$-FMs (chapter 4) as well as regression for metric spaces-valued functions (chapter 5). These results encompass large classes of instances as is visualized in the *Map of Means*, Figure 1.1.

1.3.5 Further Examples

There is a vast number of objects that can be defined as $\mathfrak{c}$-FMs. We only show a few more and again refer the reader to [HE20] for more applications.

1.3.5.1 Bregman Divergence

The Bregman divergence is an example of a large class of cost functions with identical FMs. Let $\mathcal{Q} \subseteq \mathbb{R}^s$ be a closed convex set. Let $\psi\colon \mathcal{Q} \to \mathbb{R}$ be a continuously differentiable and strictly convex function. The Bregman divergence $D_\psi\colon \mathcal{Q} \times \mathcal{Q} \to [0, \infty)$ associated with ψ for points $y, q \in \mathcal{Q}$ is defined as $D_\psi(y, q) := \psi(y) - \psi(q) - \langle \nabla\psi(q)\,,\, y - q\rangle$. It is the difference between the value of ψ at point y and the value of the first-order Taylor expansion of ψ around point q evaluated at point y. It is well-known, that the $\mathfrak{c}$-FM with $\mathfrak{c} = D_\psi$ is the expectation, see [BGW05, Theorem 1]. If we set $\psi(q) = \|q\|^2$, the cost function becomes the squared Euclidean metric. The Bregman Divergence yields an example of the quadruple inequality, see section 4.3.2.2, which is a condition introduced in this thesis to obtain rates of convergence for FMs.

1.3.5.2 k-Means

The k-means algorithm is a clustering method [Ste56], usually applied in Euclidean spaces. It can be formulated for general metric spaces as a $\mathfrak{c}$-FM: Let $(\mathcal{Q}, d)$ be a metric space. Let $\mu \in \mathcal{P}_2(\mathcal{Q}, d)$. Let $k \in \mathbb{N}$ be the number of means. Define

$$\mathfrak{c}\colon \mathcal{Q} \times \mathcal{Q}^k \to \mathbb{R}\,, (y, q) \mapsto \min_{j=1,\dots,k} d(y, q_j)^2\,.$$

A $\mathfrak{c}$-FM of μ is a tuple of k points that are centers of clusters with minimal within-cluster Fréchet variance with respect to μ. The common k-means algorithm is an instance of the general setting with $(\mathcal{Q}, d) = (\mathbb{R}^s, d_{\mathbb{R}^s})$ and $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$ for observations $y_1, \dots, y_n \in \mathbb{R}^s$, i.e., one tries to find an element of $\arg\min_{q\in(\mathbb{R}^s)^k} \sum_{i=1}^n \min_{j=1,\dots,k} \|y_i - q_j\|^2$. The classical algorithm [Llo82] is an iterative approach, which always converges after a finite number of steps but the limit may be a local minimizer of the objective function.

1.3.5.3 Quantiles

Let $\tau \in (0, 1)$. For $y, q \in \mathbb{R}$, define $\rho_\tau(y) = y(\tau - \mathbb{1}_{(-\infty,0)}(y))$ and set $\mathfrak{c}(y, q) = \rho_\tau(y - q)$. Then the $\mathfrak{c}$-FM of $\mu \in \mathcal{P}_1(\mathbb{R})$ is the τth quantile,

$$\inf\{m \in \mathbb{R}\colon F_\mu(m) \geq \tau\} \in \operatorname*{arg\,min}_{q\in\mathbb{R}} \int \rho_\tau(y - q)\,\mathrm{d}\mu(y)\,,$$

where F_μ is the distribution function of μ. This property of quantiles is used in quantile regression, see, e.g., [Koe05]. As the median, quantiles may not be unique.

1.3.6 Elicitability – Is Everything a Fréchet Mean?

After the overwhelming amount of objects that can be written as a (generalized) FM, one may ask whether the notion is trivial and everything is in fact a form of FM. The property of being a generalized FM for some cost function may be called elicitability – a term that is usually studied in the financial mathematics literature, e.g., [LPS08; Gne11; FZ16; FHR21].

Definition 1.26. A functional $S\colon \mathcal{P}(\mathcal{Y}) \to 2^{\mathcal{Q}}$ is called **elicitable** if there is a cost function $\mathfrak{c}$ such that $S(\mu) = \arg\min_{q\in\mathcal{Q}} \int \mathfrak{c}(y,q)\,\mathrm{d}\mu(y)$.

We have seen that expectation, median, maximum likelihood statistics, and more are all elicitable. It may seem, that the definition of $\mathfrak{c}$-FM is so general that every property of a distribution is elicitable. Unfortunately, there are counterexamples. It can be shown that the variance of a real-valued random variable is not elicitable, [Gne11]. The mode, too, is not elicitable, [Hei14].

Interestingly, the vector of expectation and variance is elicitable: Let Y be a real-valued random variable with $\mathbb{E}[Y^4] < \infty$. Then

$$(\mathbb{E}[Y], \mathbb{V}[Y]) = \operatorname*{arg\,min}_{q\in\mathbb{R}, s\in[0,\infty)} \mathbb{E}\Big[(Y-q)^2 + \Big((Y-q)^2 - s\Big)^2\Big]\,.$$

By subtracting $(Y-q)^4$ on the right hand side, we can reduce the moment requirement to $\mathbb{E}[Y^2]$.

In a way, every property of a distribution can be part of a tuple that is a $\mathfrak{c}$-FM of that distribution. To show that, we elicit the distribution itself. Let $\mathcal{F}([a,b])$ be the set of distribution functions of distributions with support in $[a,b]$. Let Θ be a set and let $\theta\colon \mathcal{F}([a,b]) \to \Theta$ be any property of a distribution in $\mathcal{F}([a,b])$. Assume for a distribution $F_* \in \mathcal{F}([a,b])$, we want to find $\vartheta_* := \theta(F_*)$. Let $\mathcal{F}([a,b],\theta) := \{(F, \theta(F)) \in \mathcal{F}([a,b]) \times \Theta\}$. Then

$$(F_*, \vartheta_*) = \operatorname*{arg\,min}_{(F,\vartheta)\in\mathcal{F}([a,b],\theta)} \int\int_a^b \Big(\mathbb{1}_{(-\infty,x]}(y) - F(x)\Big)^2 \,\mathrm{d}x\,\mathrm{d}F_*(y)\,,$$

i.e., (F_*, ϑ_*) is the $\mathfrak{c}$-FM of the distribution induced by F_*, where $\mathfrak{c}\colon \mathcal{Y}\times\mathcal{Q} \to \mathbb{R}$ with $\mathcal{Y} := [a,b]$ and $\mathcal{Q} := \mathcal{F}([a,b],\theta)$ is defined by

$$\mathfrak{c}(y,(F,\vartheta)) := \int_a^b \Big(\mathbb{1}_{(-\infty,x]}(y) - F(x)\Big)^2 \,\mathrm{d}x\,.$$

Although this construction makes every property part of an elicitable vector, it seems very unlikely that one can show nontrivial statements because of it.

1.4 Contributions of this Thesis

The major concern of this thesis is the theory of convergence of the sample Fréchet mean to its population counterpart in general settings. Here, we summarize our main results.

A more detailed description of the results and their relation to existing literature can be found at the beginnings of each upcoming chapter.

- **Strong Laws of Large Numbers – Chapter 2**
  — Available as preprint [Sch20b].

  After a general result on the convergence of sets of minimizers, Theorem 2.6, we show strong laws of large numbers for $\mathfrak{c}$-FMS (Theorem 2.9 and Theorem 2.10), H-FMS (Corollary 2.12 and Corollary 2.13), and α-FMS (Corollary 2.14 and Corollary 2.15). All results are given in outer limit and in one-sided Hausdorff distance, which are the types of convergence usually considered in this context.

  Over the years there have been several statements of strong laws of large numbers for FMS, see [Zie77; Sve81; KW01; BP03; Huc11]. The novelty of our results is the weakness of their assumptions. In particular, we only require an $(\alpha - 1)$-moment for the α-FMS with $\alpha > 1$ to converge. Moreover, the α-FMS with $\alpha \in (0, 1]$ converges without a requirement on the moment.

- **Rates of Convergence and the Projected Mean – Chapter 3**
  — Available as preprint [Sch19a].

  We consider the projected mean for nonconvex subsets $\mathcal{Q}$ of the Euclidean plane $\mathbb{R}^2$. For a wide range of rate sequences $(a_n)_{n\in\mathbb{N}} \subseteq (0, \infty)$, $a_n \to 0$, we construct $\mathcal{Q}$ such that a sample projected mean m_n converges to a population projected mean m with rate a_n; to be more precise, $a_n^{-1}(m_n - m) \xrightarrow{n\to\infty} \nu$ in distribution for some nondegenerate distribution ν. Corollary 3.9 presents sets $\mathcal{Q}$ for polynomial, logarithmic, and exponential rates of convergence. This result is a consequence of Theorem 3.5, which generally describes a central limit theorem with distortion due to projection.

  Conditions for parametric rates of convergence of the projected mean have been known for some time, e.g., [HL98]. How breaking these assumptions yields different rates is the major contribution in this chapter.

- **Rates of Convergence via the Quadruple Inequality – Chapter 4**
  — Published in [Sch19b].

  In this chapter, we find conditions to obtain a rate of convergence for the generalized FM. To that end, we use empirical process theory and the generic chaining. We require an entropy condition on $\mathcal{Q}$, which is typical when using these tools. Moreover, to be applicable in descriptor spaces $\mathcal{Q}$ with infinite diameter, a moment condition and a quadruple inequality, which is a generalization of the Cauchy–Schwarz inequality, are assumed. We prove nonasymptotic bounds in probability, Theorem 4.1 and two ways of obtaining rates in expectation: In Theorem 4.5 we assume a stronger version quadruple inequality (and a weak entropy condition) to obtain finite sample bounds. Theorem 4.7 yields asymptotic rates of convergence, requiring only a weak quadruple inequality but a stronger entropy condition. It is known that in Hadamard spaces an instance of the quadruple inequality holds.

This enables us to show rates of convergence in expectation for the FM in these spaces, see Corollary 4.9. Furthermore, we show that we also obtain a quadruple inequality – called power inequality – for certain powers of Hadamard metrics:

$$d(y,q)^\alpha - d(y,p)^\alpha - d(z,q)^\alpha + d(z,p)^\alpha \leq 4\alpha 2^{-\alpha} d(y,z)^{\alpha-1} d(q,p)\,,$$

for $\alpha \in [1,2]$ and y, z, q, p in a Hadamard space, see Theorem 4.10, which leads to finite sample bounds and rates of convergence for the respective power FMs, see Corollary 4.11.

[PM19a] and [ALP20] show rates of convergence for FMs in metric spaces with finite diameter. In Alexandrov spaces, [Gou+19] present conditions for a parametric rates of convergence. The contents of this chapter are set apart from these results by being far more general – we use $\mathfrak{c}$-FMs instead of 2-FMs – and being applicable to descriptor spaces with infinite diameter via introduction of the quadruple inequality. Moreover, we contribute to the study of Hadamard spaces via rates of convergence of the FM and the power inequality with the resulting rates of convergence of power FMs.

- **Regression in Non-Euclidean Spaces – Chapter 5**
  — Available as preprint [Sch20a].

  We compare two approaches – geodesic and Fréchet – to regression with responses in metric spaces, where the regression function is modeled as a conditional FM and covariates are assumed to be deterministic, equidistant points in the unit interval. Both approaches are applied to three estimators – linear, local linear, and trigonometric projection. We show finite sample bounds for linear geodesic regression (Theorem 5.2), both localized estimators (Theorem 5.12 and Theorem 5.17), and the trigonometric projection Fréchet estimator (Theorem 5.23). The obtained rates reflect the typical parametric and nonparametric rates of convergence. Linear Fréchet regression is shown to be inconsistent (Theorem 5.8) in non-Euclidean spaces. For specific spaces, namely the hyperspheres and hyperbolic spaces, we introduce a parametric alternative, which we call cosine regression. Furthermore, it is argued that a geodesic trigonometric projection estimator is suboptimal in non-Euclidean spaces.

  These general results are applied to the sphere to underline their relevance. Moreover, we compare all estimators in a simulation study. To that end, they have been implemented using the statistical programming language R [R D08]. The code is freely available at `https://github.com/ChristofSch/spheregr`.

  The geodesic approach builds on [Fle13], which introduces linear geodesic regression; the Fréchet approach is based on [PM19a], which introduces linear and local linear Fréchet regression. Local geodesic regression and trigonometric Fréchet regression are new methods, making the results on the rate of convergence for these two also a new contribution. For linear geodesic regression and local linear Fréchet regression, rates have been establish in the literature. But our statements here

require weaker assumptions and hold in greater generality. Furthermore, we introduce cosine regression, a new method for regression on the sphere.

2 Strong Laws of Large Numbers

Contents

2.1 Introduction

We begin the study of Fréchet means by establishing strong laws of large numbers. In contrast to upcoming chapters, we do not assume that the FMS is a singleton. As discussed in section 1.1.2, the FMS a can be an arbitrary set. For conditions of uniqueness, see, e.g., section Theorem 1.21. Prominent examples of nonunique FMs are the median (section 1.2.3) and FMs on hyperspheres like the circle, see [HH15] and section 3.4.4.

We will show strong laws of large numbers for $\mathfrak{c}$-FMSs. Recall the setting of section 1.3.4: Let $(\mathcal{Q}, d)$ be a metric space, the descriptor space. Let $\mathcal{Y}$ the data space. Let $\mathfrak{c}\colon \mathcal{Y} \times \mathcal{Q} \to \mathbb{R}$ be the cost function. Let Y be a random variable with values in $\mathcal{Y}$. Let $M := \arg\min_{q\in\mathcal{Q}} \mathbb{E}[\mathfrak{c}(Y, q)]$ be the $\mathfrak{c}$-FMS assuming the expectations exist. This general scenario contains many interesting notions of means and other statistics, including M-estimators and regression settings, see section 1.3.4

We will apply our results on $\mathfrak{c}$-FMSs to H- and α-FMSs, see section 1.3.2: Fix an arbitrary element $o \in \mathcal{Q}$. We will set $\mathfrak{c}(y, q) = H(\overline{y,q}) - H(\overline{y,o})$, where $H(x) = \int_0^x h(t)\mathrm{d}t$ for a nondecreasing function h, or $\mathfrak{c}(y, q) = \overline{y,q}^\alpha - \overline{y,o}^\alpha$ with $\alpha > 0$. In both cases the set of minimizers does not depend on o. The H-Fréchet means serve as a generalization of α-Fréchet means for $\alpha > 1$ as well as an intermediate result for proving strong laws of large numbers for α-Fréchet mean sets with $\alpha \in (0, 1]$.

We have described the population FMS M; next we define its empirical version M_n. For a function $f\colon \mathcal{Q} \to \mathbb{R}$ and $\epsilon \geq 0$, define $\epsilon\text{-}\arg\min_{q\in\mathcal{Q}} f(q) := \{q \in \mathcal{Q} \mid f(q) \leq \epsilon + \inf_{p\in\mathcal{Q}} f(p)\}$. Let $Y_1, \ldots, Y_n$ be independent random variables with the same distribution as Y. Choose $(\epsilon_n)_{n\in\mathbb{N}} \subseteq [0, \infty)$ with $\epsilon_n \xrightarrow{n\to\infty} 0$. Let $M_n := \epsilon_n\text{-}\arg\min_{q\in\mathcal{Q}} \frac{1}{n}\sum_{i=1}^n \mathfrak{c}(Y_i, q)$. Our goal is to show almost sure convergence of elements in M_n to elements in M. To this end, there are different possibilities of how this convergence of sets can be described.

Definition 2.1. Let $(\mathcal{Q}, d)$ be a metric space.

(i) Let $(B_n)_{n\in\mathbb{N}}$ with $B_n \subseteq \mathcal{Q}$ for all $n \in \mathbb{N}$. Then the *outer limit* of $(B_n)_{n\in\mathbb{N}}$ is

$$\lim \overline{\sup}_{n\to\infty} B_n := \bigcap_{n\in\mathbb{N}} \overline{\bigcup_{k\geq n} B_k}\,,$$

where $\overline{B}$ denotes the closure of the set B.

(ii) The *one-side Hausdorff distance* between $B, B' \subseteq \mathcal{Q}$ is

$$d_{\subseteq}(B, B') := \sup_{x\in B} \inf_{x'\in B'} d(x, x')\,.$$

(iii) The *Hausdorff distance* between $B, B' \subseteq \mathcal{Q}$ is

$$d_{\mathsf{H}}(B, B') := \max(d_{\subseteq}(B, B'), d_{\subseteq}(B', B))\,.$$

Remark 2.2.

(a) The outer limit is the set of all points of accumulation of all sequences $(x_n)_{n\in\mathbb{N}}$ with $x_n \in B_n$.

(b) It holds $d_{\subseteq}(B, B') = 0$ if and only if $B \subseteq \overline{B'}$, but $d_{\mathsf{H}}(B, B') = 0$ if and only if $\overline{B} = \overline{B}'$. The function d_{H} is a metric on the set of closed and bounded subsets of $\mathcal{Q}$.

(c) Elements from a sequence of sets might have sub-sequences that have no point of accumulation and are bounded away from the outer limit of the sequence of sets. That cannot happen with the one-sided Hausdorff limit. Here, every sub-sequence is eventually arbitrarily close to the limiting set. As an example, the outer limit of the sequence of sets $\{0, n\}$, $n \in \mathbb{N}$ on the Euclidean real line is $\{0\}$, but $d_{\subseteq}(\{0, n\}, \{0\}) \xrightarrow{n\to\infty} \infty$.

We will give conditions so that $\lim \overline{\sup}_{n\to\infty} M_n \subseteq M$ almost surely or $d_{\subseteq}(M, M_n) \xrightarrow{n\to\infty}_{\mathsf{a.s.}} 0$, where the index $\mathsf{a.s.}$ indicates almost sure convergence. It is not easily possible to show $d_{\mathsf{H}}(M, M_n) \xrightarrow{n\to\infty}_{\mathsf{a.s.}} 0$ if M is not a singleton. These results may be called strong laws of large numbers of the Fréchet mean set or (strong) consistency of the empirical Fréchet mean set.

[Zie77] shows a strong law for the outer limit of Fréchet mean sets with a second moment condition. [Sve81] shows a strong law in outer limit for power Fréchet mean sets in compact spaces. [BP03] shows convergence of Fréchet mean sets in one-sided Hausdorff distance with a second moment condition. In contrast, we show strong laws of large numbers for power Fréchet mean sets in outer limit and in one-sided Hausdorff distance with less moment assumptions: For $\alpha > 1$, we require $\mathbb{E}[\overline{Y,o}^{\alpha-1}] < \infty$, and

for $\alpha \in (0,1]$ no moment assumption is made, see Corollary 2.14 and Corollary 2.15. Thus, α-Fréchet means may be of interest in robust statistics. [Huc11] shows almost sure convergence in one-side Hausdorff distance as well as in outer limit for generalized Fréchet means. Our results for $\mathfrak{c}$-Fréchet means in one-side Hausdorff distance require different assumptions, which make them applicable in a larger class of settings, see Theorem 2.9 and Remark 2.8. Results by [AW95; KW01; CHS03] imply strong laws and ergodic theorems in outer limit for generalized Fréchet means. We recite parts of these results to state Theorem 2.10. Furthermore, we show strong laws of large numbers for H-Fréchet means sets in outer limit, Corollary 2.13, and one-sided Hausdorff distance, Corollary 2.12.

Before we consider the probabilistic setting, we present theory on convergence of minimizing sets for deterministic functions in section 2.2, where we partially follow [RW98]. Thereafter, we derive strong laws of large numbers for $\mathfrak{c}$-Fréchet mean sets in section 2.3, for H-Fréchet mean sets in section 2.4, and for α-Fréchet mean sets in section 2.5.

2.2 Convergence of Minimizer Sets of Deterministic Functions

Let $(\mathcal{Q}, d)$ be a metric space. We use two notions of convergence of functions, which will lead to different convergence results of their minimizers.

Definition 2.3. Let $f, f_n \colon \mathcal{Q} \to \mathbb{R}$, $n \in \mathbb{N}$.

(i) The sequence $(f_n)_{n\in\mathbb{N}}$ *epi-converges* to f *at* $x \in \mathcal{Q}$ if and only if

$$\forall (x_n)_{n\in\mathbb{N}} \subseteq \mathcal{Q}, x_n \to x \colon \liminf_{n\to\infty} f_n(x_n) \geq f(x) \qquad \text{and}$$
$$\exists (y_n)_{n\in\mathbb{N}} \subseteq \mathcal{Q}, y_n \to x \colon \limsup_{n\to\infty} f_n(y_n) \leq f(x)\,.$$

The sequence $(f_n)_{n\in\mathbb{N}}$ *epi-converges* to f if and only if it epi-converges at all $x \in \mathcal{Q}$. We then write $f_n \xrightarrow{n\to\infty}_{\mathsf{epi}} f$.

(ii) The sequence $(f_n)_{n\in\mathbb{N}}$ *converges* to f *uniformly on bounded sets* if and only if for every $B \subseteq \mathcal{Q}$ with $\mathsf{diam}(B) < \infty$,

$$\lim_{n\to\infty} \sup_{x\in B} |f_n(x) - f(x)| = 0\,.$$

We then write $f_n \xrightarrow{n\to\infty}_{\mathsf{ubs}} f$.

We introduce some short notation. Let $f \colon \mathcal{Q} \to \mathbb{R}$ and $\epsilon \geq 0$. Denote $\inf f := \inf_{x\in\mathcal{Q}} f(x)$, $\arg\min f := \{x \in \mathcal{Q} \,|\, f(x) = \inf f\}$, $\epsilon\text{-}\arg\min f := \{x \in \mathcal{Q} \,|\, f(x) \leq \epsilon + \inf f\}$. Let $\delta > 0$, $x_0 \in \mathcal{Q}$, and $A \subseteq \mathcal{Q}$. Denote $\mathrm{B}_\delta(x_0) := \{x \in \mathcal{Q} \,|\, d(x, x_0) < \delta\}$ and $\mathrm{B}_\delta(A) := \bigcup_{x\in A} \mathrm{B}_\delta(x)$. Furthermore, f is called *lower semi-continuous* if and only if $\liminf_{x\to x_0} f(x) \geq f(x_0)$ for all $x_0 \in \mathcal{Q}$.

Definition 2.4. A function f has *approachable minimizers* if and only if for all $\epsilon > 0$ there is a $\delta > 0$ such that $\delta\text{-}\arg\min f \subseteq B_\epsilon(\arg\min f)$.

The definition directly implies that $d_\subseteq(\delta\text{-}\arg\min f, \arg\min f) \xrightarrow{\delta\to 0} 0$ is equivalent to f having approachable minimizers. Furthermore, if f has approachable minimizers, then $\arg\min f \neq \emptyset$, as for every $\delta > 0$ the set $\delta\text{-}\arg\min f$ is nonempty, but $\mathrm{B}_\epsilon(\emptyset) = \emptyset$.

To state convergence results for minimizing sets of deterministic functions, we need one final definition.

Definition 2.5. A sequence $(B_n)_{n\in\mathbb{N}}$ of sets $B_n \subseteq \mathcal{Q}$ is called *eventually bounded* if and only if

$$\limsup_{n\to\infty} \mathsf{diam}\left(\bigcup_{k=n}^{\infty} B_k\right) < \infty\,.$$

The main theorem of this section states conditions for convergence of sets of minimizers in outer limit and in one-sided Hausdorff distance.

Theorem 2.6. Let $f, f_n\colon \mathcal{Q} \to \mathbb{R}$. Let $(\epsilon_n)_{n\in\mathbb{N}} \subseteq [0,\infty)$ with $\epsilon_n \xrightarrow{n\to\infty} 0$.

(i) Assume $f_n \xrightarrow{n\to\infty}_{\mathsf{epi}} f$. Then

$$\limsup_{n\to\infty} \epsilon_n\text{-}\arg\min f_n \subseteq \arg\min f$$

and

$$\limsup_{n\to\infty} \inf f_n \leq \inf f\,.$$

(ii) Assume f has approachable minimizers, $f_n \xrightarrow{n\to\infty}_{\mathsf{ubs}} f$, and $(\epsilon_n\text{-}\arg\min f_n)_{n\in\mathbb{N}}$ is eventually bounded. Then

$$d_\subseteq(\epsilon_n\text{-}\arg\min f_n, \arg\min f) \xrightarrow{n\to\infty} 0$$

and

$$\inf f_n \xrightarrow{n\to\infty} \inf f\,.$$

Large parts of this theorem can be found e.g., in [RW98, chapter 7]. To make this thesis more self-contained, we give a proof here.

Proof.

(i) Let $x \in \limsup_{n\to\infty} \epsilon_n\text{-}\arg\min f_n$. Then there is a sequence $x_n \in \epsilon_n\text{-}\arg\min f_n$ with a subsequence converging to x, i.e., $x_{n_i} \xrightarrow{i\to\infty} x$, where $n_i \xrightarrow{i\to\infty} \infty$. Let $y \in \mathcal{Q}$ be arbitrary. As $f_n \xrightarrow{n\to\infty}_{\mathsf{epi}} f$, there is a sequence $(y_n)_{n\in\mathbb{N}} \subseteq \mathcal{Q}$ with $y_n \xrightarrow{n\to\infty} y$ and $\limsup_{n\to\infty} f_n(y_n) \leq f(y)$. It holds $f_{n_i}(x_{n_i}) \leq \epsilon_{n_i} + \inf f_{n_i} \leq \epsilon_{n_i} + f_{n_i}(y_{n_i})$.

Thus, by the definition of epi-convergence and $\epsilon_n \xrightarrow{n\to\infty} 0$, we obtain

$$f(x) \leq \liminf_{i\to\infty} f_{n_i}(x_{n_i}) \leq \liminf_{i\to\infty} (\epsilon_{n_i} + f_{n_i}(y_{n_i})) \leq \limsup_{i\to\infty} f_{n_i}(y_{n_i}) \leq f(y)\,.$$

Thus, $x \in \arg\min f$. Next, we turn to the inequality of the infima. For $\epsilon > 0$ choose an arbitrary $x \in \epsilon\text{-}\arg\min f$. There is a sequence $(y_n)_{n\in\mathbb{N}} \subseteq \mathcal{Q}$ with $y_n \xrightarrow{n\to\infty} x$ and $f_n(y_n) \xrightarrow{n\to\infty} f(x)$. Thus,

$$\limsup_{n\to\infty} \inf f_n \leq \limsup_{n\to\infty} f_n(y_n) \leq \inf f + \epsilon\,.$$

(ii) Let $\epsilon > 0$. As f has approachable minimizers, there is $\delta > 0$ such that

$$(3\delta)\text{-}\arg\min f \subseteq \mathrm{B}_\epsilon(\arg\min f)\,.$$

Furthermore, $\arg\min f \neq \emptyset$. Let $y \in \arg\min f$. As $f_n(y) \xrightarrow{n\to\infty} f(y)$, there is $n_1 \in \mathbb{N}$ such that $\inf f_n \leq \inf f + \delta$ for all $n \geq n_1$. As $\epsilon_n \xrightarrow{n\to\infty} 0$, there is $n_2 \in \mathbb{N}$ such that $\epsilon_n \leq \delta$ for all $n \geq n_2$. As $(\epsilon_n\text{-}\arg\min f_n)_{n\in\mathbb{N}}$ is eventually bounded, there is $n_3 \in \mathbb{N}$ such that $\mathsf{diam}(B) < \infty$ for $B = \bigcup_{n\geq n_3} \epsilon_n\text{-}\arg\min f_n$. As $f_n \xrightarrow{n\to\infty}_{\mathsf{ubs}} f$ there is n_4 such that $\sup_{x\in B} |f_n(x) - f(x)| \leq \delta$. Let $n \geq \max(n_1, n_2, n_3, n_4)$ and $x \in \epsilon_n\text{-}\arg\min f_n$. Then

$$f(x) \leq f_n(x) + \delta \leq \inf f_n + 2\delta \leq \inf f + 3\delta\,.$$

Thus, $x \in (3\delta)\text{-}\arg\min f$. By the choice of ϵ and δ, we obtain $\epsilon_n\text{-}\arg\min f_n \subseteq \mathrm{B}_\epsilon(\arg\min f)$ or equivalently $d_\subseteq(\epsilon_n\text{-}\arg\min f_n, \arg\min f) \leq \epsilon$.

Finally, we show the convergence of the infima. We already know $\inf f_n \leq \inf f + \epsilon$ for all $\epsilon > 0$ and n large enough. If $\inf f_n \xrightarrow{n\to\infty} \inf f$ does not hold, there is a sequence $x_n \in \epsilon_n\text{-}\arg\min f_n$ and $\epsilon > 0$ such that $f_n(x_n) < \inf f - \epsilon$ for all n large enough. As before, because of eventual boundedness and uniform convergence on bounded sets, we have $\sup_{k\in\mathbb{N}} |f_n(x_k) - f(x_k)| \xrightarrow{n\to\infty} 0$. Therefore, for all $\epsilon > 0$ we have $f(x_n) \leq f_n(x_n) + \epsilon$ for n large enough, which contradicts $f_n(x_n) < \inf f - \epsilon$.

□

The conditions for subset convergence in outer limit are minimal. In the following, we construct examples to show that none of the conditions for one-sided Hausdorff convergence can be dropped.

Example 2.7.

(i) Let $f, f_n \colon \mathbb{N}_0 \to \mathbb{R}$, $f_n := 1 - \mathbb{1}_{\{0,n\}}$, $f := 1 - \mathbb{1}_{\{0\}}$, $d(i,j) := 1$ for $i \neq j$. It holds that f is continuous and has approachable minimizers, and the sequence of nonempty sets $\arg\min f_n = \{0, n\}$ is eventually bounded, as $\mathsf{diam}(A) \leq 1$ for every $A \subseteq \mathbb{N}_0$. Furthermore, f_n converges to f uniformly on compact sets, which are exactly the finite subsets of $\mathbb{N}_0$, but not uniformly on bounded sets

like $\mathbb{N}_0$ itself. There is a subsequence of minimizers $x_n = n \in \arg\min f_n$ that is always bounded away from 0, the minimizer of f. This shows that uniform convergence on compact sets (instead of bounded sets) is not enough.

(ii) As above, let $f, f_n \colon \mathbb{N}_0 \to \mathbb{R}$, $f_n := 1 - \mathbb{1}_{\{0,n\}}$, $f := 1 - \mathbb{1}_{\{0\}}$, but define $d(i,j) := |i-j|$. It holds that f is continuous and has approachable minimizers, and $f_n \xrightarrow{n\to\infty}_{\mathsf{ubs}} f$, but the sequence of nonempty sets $\arg\min f_n = \{0, n\}$ is not eventually bounded. Again, there is a subsequence of minimizers $x_n = n \in \arg\min f_n$ that is always bounded away from 0, the minimizer of f. This shows that eventual boundedness of minimizer sets cannot be dropped.

(iii) Let $f, f_n \colon \mathbb{N}_0 \to \mathbb{R}$, $f(0) := 0$, $f(i) := \frac{1}{i}$, $f_n(i) := f(i)\mathbb{1}_{\{i<n\}}$, and set $d(i,j) := 1$ for $i \neq j$. It holds that f is continuous, but f does not have approachable minimizers. The sequence of nonempty sets $\arg\min f_n = \{0, n, n+1, \ldots\}$ is eventually bounded and $f_n \xrightarrow{n\to\infty}_{\mathsf{ubs}} f$. There is a subsequence of minimizers $x_n = n \in \arg\min f_n$ that is always bounded away from 0, the minimizer of f. This shows that approachability of minimizers of f cannot be dropped.

In the setting of the second part of the theorem, for an arbitrary sequence $\epsilon_n \xrightarrow{n\to\infty} 0$, $d_{\mathsf{H}}(\epsilon_n\text{-}\arg\min f_n, \arg\min f)$ does not necessarily vanish unless $\arg\min f$ is a singleton. For a result of full set convergence, see [RW98, Theorem 7.31 (c)].

2.3 Strong Laws for $\mathfrak{c}$-Fréchet Mean Sets

Let $(\mathcal{Q}, d)$ be a metric space, the descriptor space. Let $\mathcal{Y}$ be a set, the data space. Let $\mathfrak{c} \colon \mathcal{Y} \times \mathcal{Q} \to \mathbb{R}$ be a function, the cost function. Let Y be a random variable with values in $\mathcal{Y}$. Denote the $\mathfrak{c}$-Fréchet mean set of Y as $M := \arg\min_{q\in\mathcal{Q}} \mathbb{E}[\mathfrak{c}(Y,q)]$. Let $Y_1, \ldots, Y_n$ be independent random variables with the same distribution as Y. Choose $(\epsilon_n)_{n\in\mathbb{N}} \subseteq [0,\infty)$ with $\epsilon_n \xrightarrow{n\to\infty} 0$. Set $M_n := \epsilon_n\text{-}\arg\min_{q\in\mathcal{Q}} \frac{1}{n}\sum_{i=1}^n \mathfrak{c}(Y_i, q)$.

Assumptions.

- Heine–Borel: Every closed bounded set in $\mathcal{Q}$ is compact.
- Continuity: The function $q \mapsto \mathfrak{c}(Y,q)$ is almost surely continuous.
- UpperBound: $\mathbb{E}[\sup_{q\in B} |\mathfrak{c}(Y,q)|] < \infty$ for all bounded sets $B \subseteq \mathcal{Q}$.
- LowerBound: There are $o \in \mathcal{Q}$, $\psi^+, \psi^- \colon [0,\infty) \to [0,\infty)$, $\mathfrak{a}^+, \mathfrak{a}^- \in (0,\infty)$, and $\mathfrak{a}_n^+, \mathfrak{a}_n^- \in [0,\infty)$ depending on $(Y_i)_{i=1,\ldots,n}$ such that

$$\mathfrak{a}^+\psi^+(\overline{q,o}) - \mathfrak{a}^-\psi^-(\overline{q,o}) \leq \mathbb{E}[\mathfrak{c}(Y,q)]\,,$$
$$\mathfrak{a}_n^+\psi^+(\overline{q,o}) - \mathfrak{a}_n^-\psi^-(\overline{q,o}) \leq \frac{1}{n}\sum_{i=1}^n \mathfrak{c}(Y_i,q)$$

for all $q \in \mathcal{Q}$. Furthermore, $\mathfrak{a}_n^+ \xrightarrow{n\to\infty}_{\mathsf{a.s.}} \mathfrak{a}^+$ and $\mathfrak{a}_n^- \xrightarrow{n\to\infty}_{\mathsf{a.s.}} \mathfrak{a}^-$. Lastly, $\psi^+(\delta)/\psi^-(\delta) \xrightarrow{\delta\to\infty} \infty$.

Remark 2.8.

- On Heine–Borel: A space enjoying this property is also called *boundedly compact* or *proper* metric space. The Euclidean spaces $\mathbb{R}^s$, finite dimensional Riemannian manifolds, as well as $\mathcal{C}^\infty(\Omega)$ for open subsets $\Omega \subseteq \mathbb{R}^s$ fulfill Heine–Borel [Edw95, section 8.4.7]. See [WJ87] for a construction of further spaces where Heine–Borel is true.
- On LowerBound: We illustrate this condition in the linear regression setting with $\mathcal{Q} := \mathbb{R}^{s+1}$, $\mathcal{Y} := (\{1\} \times \mathbb{R}^s) \times \mathbb{R}$, $\mathfrak{c}((x,y),\beta) := (y - \beta^\top x)^2 - y^2 = -2\beta^\top xy + \beta^\top x x^\top \beta$. Let (X,Y) be random variables with values in $\mathcal{Y}$. Let $(X_1,Y_1),\dots,(X_n,Y_n)$ be independent with the same distribution as (X,Y). We can set $o := 0 \in \mathbb{R}^{s+1}$, $\mathfrak{a}^+ := \lambda_{\mathsf{min}}(\mathbb{E}[XX^\top])$, where λ_{min} denotes the smallest eigenvalue, $\mathfrak{a}^- := 2\|\mathbb{E}[XY]\|$, $\mathfrak{a}_n^+ := \lambda_{\mathsf{min}}(\frac{1}{n}\sum_{i=1}^n X_i X_i^\top)$, $\mathfrak{a}_n^- := 2\|\frac{1}{n}\sum_{i=1}^n X_i Y_i\|$, $\psi^+(\delta) := \delta^2$ and $\psi^-(\delta) := \delta$. If $\lambda_{\mathsf{min}}(\mathbb{E}[XX^\top]) > 0$, the largest eigenvalue $\lambda_{\mathsf{max}}(\mathbb{E}[XX^\top]) < \infty$, and $\mathbb{E}[\|XY\|] < \infty$, all conditions are fulfilled.

  For a further application of LowerBound, see the proof of Corollary 2.12 in the next section.

Theorem 2.9. Assume Heine–Borel, Continuity, UpperBound, and LowerBound. Then

$$d_\subseteq(M_n, M) \xrightarrow{n\to\infty}_{\mathsf{a.s.}} 0\,.$$

Proof. Define $F(q) := \mathbb{E}[\mathfrak{c}(Y,q)]$, $F_n(q) := \frac{1}{n}\sum_{i=1}^n \mathfrak{c}(Y_i,q)$. The proof consists of following steps:

1. Show that $F_n \xrightarrow{n\to\infty}_{\mathsf{ubs}} F$ almost surely.
2. Reduction to a bounded set.
3. Show that F has approachable minimizers.
4. Show that M_n is eventually bounded.
5. Apply Theorem 2.6.

Step 1. To show uniform convergence on bounded sets, we will use the uniform law of large numbers, Theorem 2.16 (appendix). Let $B \subseteq \mathcal{Q}$ be a bounded set. By Heine–Borel, $\overline{B}$ is compact. By Continuity, $q \mapsto \mathfrak{c}(Y,q)$ is almost surely continuous. By

UpperBound, $\mathbb{E}[\sup_{q\in B}|\mathfrak{c}(Y,q)|]<\infty$. Thus, Theorem 2.16 (appendix) implies that $q\mapsto F(q)$ is continuous and

$$\sup_{q\in B}|F_n(q)-F(q)| \xrightarrow{n\to\infty}_{\mathsf{a.s.}} 0\,.$$

Fix an arbitrary element $o\in\mathcal{Q}$. For all bounded sets B, there is $\delta\in\mathbb{N}$ such that $B\subseteq \mathrm{B}_\delta(o)$. By the previous considerations, uniform convergence holds almost surely for all $(\mathrm{B}_\delta(o))_{\delta\in\mathbb{N}}$. Thus, $F_n \xrightarrow{n\to\infty}_{\mathsf{ubs}} F$ almost surely.
Step 2. Next, we want to show that there is a bounded set $B\subseteq\mathcal{Q}$ such that $F(q)\geq F(m)+1$ and $F_n(q)\geq F_n(m)+1$ for all $q\in\mathcal{Q}\setminus B$ and $m\in M$. If $\mathcal{Q}$ is bounded, we can take $B=\mathcal{Q}$. Assume $\mathcal{Q}$ is not bounded.

Let $m\in M$. By UpperBound, $F(m)<\infty$. Let $o\in\mathcal{Q}$ from LowerBound. Due to LowerBound, $F(q)\geq \mathfrak{a}^+\psi^+(\delta)-\mathfrak{a}^-\psi^-(\delta)\geq F(m)+2$ for all $q\in\mathcal{Q}\setminus\mathrm{B}_\delta(o)$ and δ large enough. This holds for all $m\in M$ as $F(m)$ does not change with m. We set $B=\mathrm{B}_\delta(o)$. For F_n, it holds $F_n(m)\xrightarrow{n\to\infty}_{\mathsf{a.s.}} F(m)$ and $\inf_{q\in\mathcal{Q}\setminus B}F_n(q)\geq \mathfrak{a}_n^+\psi^+(\delta)-\mathfrak{a}_n^-\psi^-(\delta)$ with $\mathfrak{a}_n^+\xrightarrow{n\to\infty}_{\mathsf{a.s.}}\mathfrak{a}^+$ and $\mathfrak{a}_n^-\xrightarrow{n\to\infty}_{\mathsf{a.s.}}\mathfrak{a}^-$. Thus, there is a random variable N_0 such that almost surely $F_n(q)\geq F_n(m)+1$ for all $n\geq N_0$, $q\in\mathcal{Q}\setminus B$, and $m\in M$.
Step 3. Clearly, $M\subseteq B$ is bounded. Furthermore, for all $\epsilon>0$ small enough the set $D_\epsilon:=\overline{B\setminus\mathrm{B}_\epsilon(M)}$ is not empty (if it is, increase δ), does not contain any element of M and, by Heine–Borel, is compact. Thus, the continuous function $q\mapsto F(q)$ attains its infimum on D_ϵ where $\inf_{q\in D_\epsilon}F(q)>\inf_{q\in\mathcal{Q}}F(q)$. Take $\zeta:=\min(1,\frac{1}{2}(\inf_{q\in D_\epsilon}F(q)-\inf_{q\in\mathcal{Q}}F(q)))$. Then $\zeta\text{-}\arg\min_{q\in\mathcal{Q}}F(q)\subseteq\mathrm{B}_\epsilon(M)$, i.e., F has approachable minimizers.
Step 4. For $\epsilon_n<1$ and $n\geq N_0$, it holds $M_n\subseteq B$. Thus, $(M_n)_{n\in\mathbb{N}}$ is eventually bounded almost surely.
Step 5. Finally, Theorem 2.6 implies $d_\subseteq(M_n,M)\xrightarrow{n\to\infty}_{\mathsf{a.s.}}0$. □

Assumptions.

- Polish: $(\mathcal{Q},d)$ is separable and complete.
- LowerSemiContinuity: $q\mapsto\mathfrak{c}(y,q)$ is lower semi-continuous.
- Integrable: $\mathbb{E}[|\mathfrak{c}(Y,q)|]<\infty$ for all $q\in\mathcal{Q}$.
- IntegrableInf: $\mathbb{E}[\inf_{q\in\mathcal{Q}}\mathfrak{c}(Y,q)]>-\infty$.

Theorem 2.10. Assume Polish, LowerSemiContinuity, Integrable, and IntegrableInf. Then, almost surely,

$$\limsup_{n\to\infty} M_n\subseteq M\,.$$

Proof. Define $F(q) := \mathbb{E}[\mathfrak{c}(Y,q)]$, $F_n(q) := \frac{1}{n}\sum_{i=1}^n \mathfrak{c}(Y_i,q)$. By Integrable, $F(q) < \infty$. [KW01, Theorem 1.1] states that $F_n \xrightarrow{n\to\infty}_{\mathsf{epi}} F$ almost surely if Polish, LowerSemi-Continuity, and IntegrableInf are true. Theorem 2.6 then implies $\limsup_{n\to\infty} M_n \subseteq M$ almost surely. □

2.4 Strong Laws for H-Fréchet Mean Sets

Let $(\mathcal{Q}, d)$ be a metric space. Let Y be a random variable with values in $\mathcal{Q}$. Let $h\colon [0,\infty) \to [0,\infty)$ be a nondecreasing function. Define $H\colon [0,\infty) \to [0,\infty), x \mapsto \int_0^x h(t)\mathrm{d}t$. Fix an arbitrary element $o \in \mathcal{Q}$. Denote the H-Fréchet mean set of Y as $M := \arg\min_{q\in\mathcal{Q}} \mathbb{E}[H(\overline{Y,q}) - H(\overline{Y,o})]$. Let $Y_1, \dots, Y_n$ be independent random variables with the same distribution as Y. Choose $(\epsilon_n)_{n\in\mathbb{N}} \subseteq [0,\infty)$ with $\epsilon_n \xrightarrow{n\to\infty} 0$. Set $M_n := \epsilon_n\text{-}\arg\min_{q\in\mathcal{Q}} \frac{1}{n}\sum_{i=1}^n (H(\overline{Y_i,q}) - H(\overline{Y_i,o}))$.

Assumptions.

- InfiniteIncrease: $h(x) \xrightarrow{x\to\infty} \infty$.
- Additivity: There is $b \in [1,\infty)$ such that $h(2x) \le bh(x)$ for all $x \ge 0$.
- h-Moment: $\mathbb{E}[h(\overline{Y,o})] < \infty$.

Remark 2.11.

- On Additivity: This implies $h(x+y) \le b(h(x)+h(y))$ for all $x, y \ge 0$, see Lemma 2.17 (appendix). If h is concave, Additivity holds with $b = 2$ and we even have $h(x+y) \le h(x) + h(y)$. This condition is not very restrictive, but it excludes functions that grow exponentially.

Corollary 2.12. Assume Heine–Borel, Additivity, InfiniteIncrease, and h-Moment. Then

$$d_{\subseteq}(M_n, M) \xrightarrow{n\to\infty}_{\mathsf{a.s.}} 0\,.$$

Proof. We have to check the conditions of Theorem 2.9. Heine–Borel is an assumption of the corollary. Continuity is fulfilled as $(q,p) \mapsto d(q,p)$ and $x \mapsto H(x)$ are continuous.

For UpperBound, let $y, q, o \in \mathcal{Q}$ and apply Lemma 2.17 (i) (appendix),

$$\begin{aligned} |H(\overline{y,q}) - H(\overline{y,o})| &\le |\overline{y,q} - \overline{y,o}|\, h(\max(\overline{y,q}, \overline{y,o})) \\ &\le \overline{q,o}\, h(\overline{q,o} + \overline{y,o}) \\ &\le b\,\overline{q,o}\,(h(\overline{q,o}) + h(\overline{y,o}))\,, \end{aligned}$$

where the last inequality follows from Lemma 2.17 (ii) (appendix). Thus, h-Moment

implies UpperBound.

To show LowerBound, we note that H is nondecreasing and apply Lemma 2.17 (iii) (appendix),

$$\begin{aligned} H(\overline{y,q}) - H(\overline{y,o}) &\geq H(|\overline{y,o} - \overline{q,o}|) - H(\overline{y,o}) \\ &\geq b^{-1} H(\overline{q,o}) - 2\,\overline{q,o}\,h(\overline{y,o})\,. \end{aligned}$$

Thus, we can set $\psi^+(\delta) := b^{-1}H(\delta)$, $\psi^-(\delta) := 2\delta$, $\mathfrak{a}^+ := \mathfrak{a}_n^+ := 1$, $\mathfrak{a}^- := \mathbb{E}[h(\overline{Y,o})]$, and $\mathfrak{a}_n^- := \frac{1}{n}\sum_{i=1}^n h(\overline{Y_i,o})$ with $\mathfrak{a}_n^- \xrightarrow{n\to\infty}_{\mathsf{a.s.}} \mathfrak{a}^-$ because of h-Moment. As $H(\delta) = \int_0^\delta h(x)\mathrm{d}x \geq \frac{1}{2}\delta h(\frac{1}{2}\delta)$, $\psi^+(\delta)/\psi^-(\delta) \geq \frac{1}{4}b^{-1}h(\frac{1}{2}\delta) \xrightarrow{\delta\to\infty} \infty$ by InfiniteIncrease. □

Corollary 2.13. Assume Polish, Additivity, and h-Moment. Then, almost surely,

$$\limsup_{n\to\infty} M_n \subseteq M\,.$$

Proof. As in Corollary 2.12, $H \circ d$ is continuous. In particular, LowerSemiContinuity holds. Following the proof of Corollary 2.12 shows

$$\begin{aligned} |H(\overline{y,q}) - H(\overline{y,o})| &\leq b\,\overline{q,o}\,(h(\overline{q,o}) + h(\overline{y,o})) \\ H(\overline{y,q}) - H(\overline{y,o}) &\geq b^{-1}H(\overline{q,o}) - 2\,\overline{q,o}\,h(\overline{y,o}) \end{aligned}$$

due to Additivity with $H(\delta)/\delta \xrightarrow{\delta\to\infty} \infty$. With that, h-Moment implies IntegrableInf and Integrable. Thus, Theorem 2.10 can be applied using Polish. □

2.5 Strong Laws for α-Fréchet Mean Sets

Let $(\mathcal{Q}, d)$ be a metric space. Let Y be a random variable with values in $\mathcal{Y}$. Let $\alpha > 0$. Fix an arbitrary element $o \in \mathcal{Q}$. Denote the α-Fréchet mean set of Y as $M := \arg\min_{q\in\mathcal{Q}} \mathbb{E}[\overline{Y,q}^\alpha - \overline{Y,o}^\alpha]$. Let $Y_1, \ldots, Y_n$ be independent random variables with the same distribution as Y. Choose $(\epsilon_n)_{n\in\mathbb{N}} \subseteq [0,\infty)$ with $\epsilon_n \xrightarrow{n\to\infty} 0$. Set $M_n := \epsilon_n\text{-}\arg\min_{q\in\mathcal{Q}} \frac{1}{n}\sum_{i=1}^n (\overline{Y_i,q}^\alpha - \overline{Y_i,o}^\alpha)$.

Corollary 2.14. Let $\alpha > 1$. Assume $\mathbb{E}[\overline{Y,o}^{\alpha-1}] < \infty$.

(i) Assume Heine–Borel. Then $d_\subseteq(M_n, M) \xrightarrow{n\to\infty}_{\mathsf{a.s.}} 0$.

(ii) Assume Polish. Then $\limsup_{n\to\infty} M_n \subseteq M$ almost surely.

Proof. Set $h(x) := \alpha^{-1}x^{\alpha-1}$. This function is nondecreasing, fulfills Additivity with $b = 2^{\alpha-1}$ and InfiniteIncrease. Due to $\mathbb{E}[\overline{Y,o}^{\alpha-1}] < \infty$, h-Moment is fulfilled.

Furthermore, $H(x) = x^\alpha$. Thus, Corollary 2.12 and Corollary 2.13 imply the claims. □

Corollary 2.15. Let $\alpha \in (0, 1]$.

(i) Assume Heine–Borel. Then $d_\subseteq(M_n, M) \xrightarrow{n\to\infty}_{\text{a.s.}} 0$.

(ii) Assume Polish. Then $\limsup_{n\to\infty} M_n \subseteq M$ almost surely.

Proof. First, consider the case $\alpha = 1$. Apply Lemma 2.18 (appendix) on $\overline{Y,o}$ to obtain a function $h\colon [0,\infty) \to [0,\infty)$ which is strictly increasing, continuous, concave, fulfills InfiniteIncrease, and $\mathbb{E}[h(\overline{Y,o})] < \infty$. Concavity implies Additivity with $b = 1$. As its derivative is strictly increasing, $H(x) = \int_0^x h(t)\mathrm{d}t$ is convex and strictly increasing. Thus, H has an inverse H^{-1} and H^{-1} is concave. This implies that $d_H(q,p) = H^{-1}(\overline{q,p})$ is a metric.

As H^{-1} is concave, there are $u_0, u_1 \in [0,\infty)$ such that $H^{-1}(x) \le u_0 + u_1 x$ for all $x \ge 0$. As h is concave, there are $v_0, v_1 \in [0,\infty)$ such that $h(u_0 + u_1 x) \le v_0 + v_1 h(x)$ for all $x \ge 0$. Thus, $\mathbb{E}[h(d_H(Y,o))] = \mathbb{E}[h(H^{-1}(\overline{Y,o}))] \le v_0 + v_1\mathbb{E}[h(\overline{Y,o})] < \infty$. Hence, h-Moment is true for the metric d_H.

Moreover, Heine–Borel and Polish of $(\mathcal{Q}, d)$ are preserved in $(\mathcal{Q}, d_H)$, as H^{-1} is strictly increasing, concave, and continuous, with $H^{-1}(0) = 0$ and $H^{-1}(\delta) \xrightarrow{\delta\to\infty} \infty$, and thus, the properties boundedness, compactness, separability, and completeness coincide for d and d_H. Applying Corollary 2.12 and Corollary 2.13 on the minimizers of $\mathbb{E}[H(d_H(Y,q)) - H(d_H(Y,o))] = \mathbb{E}[\overline{Y,q} - \overline{Y,o}]$ now yields the claims for $\alpha = 1$.

For $\alpha \in (0,1)$ just note, that $\tilde{d}(q,p) = d(q,p)^\alpha$ is a metric, which preserves Heine–Borel and Polish, and apply the result for $\alpha = 1$ on $\tilde{d}$. □

Appendix of Chapter 2

2.A Auxiliary Results

There are many versions of uniform laws of large numbers in the literature. We state and prove one version that is tailored to our needs.

Theorem 2.16. Let $(\mathcal{Y}, \Sigma_{\mathcal{Y}})$ be a measurable space and Y be a random variable with values in $\mathcal{Y}$. Let $Y_1, \dots, Y_n$ be independent and have the same distribution as Y. Let $(\mathcal{Q}, d)$ be a metric space and $B \subseteq \mathcal{Q}$ compact. Let $f\colon \mathcal{Y} \times B \to \mathbb{R}$ be such that $q \mapsto f(Y, q)$ is almost surely continuous. Assume there is a random variable Z such that $|f(Y, q)| \leq Z$ for all $q \in B$ with $\mathbb{E}[Z] < \infty$. Then $q \mapsto \mathbb{E}[f(Y, q)]$ is continuous and

$$\sup_{q \in B} \left| \frac{1}{n} \sum_{i=1}^{n} f(Y_i, q) - \mathbb{E}[f(Y, q)] \right| \xrightarrow{n \to \infty}_{\text{a.s.}} 0\,.$$

Proof. Let $\epsilon > 0$. As B is compact, there is a finite set $\{q_1, \dots, q_k\} \subseteq \mathcal{Q}$ such that $B \subseteq \bigcup_{\ell=1}^{k} \mathrm{B}_\epsilon(q_\ell)$. We split the supremum,

$$\begin{aligned}
&\sup_{q \in B} \left| \frac{1}{n} \sum_{i=1}^{n} f(Y_i, q) - \mathbb{E}[f(Y, q)] \right| \\
&\leq \sup_{\ell \in \{1, \dots, k\}} \sup_{q \in \mathrm{B}_\epsilon(q_\ell)} \left| \frac{1}{n} \sum_{i=1}^{n} (f(Y_i, q) - f(Y_i, q_\ell)) - \mathbb{E}[f(Y, q) - f(Y, q_\ell)] \right| \\
&\quad + \sup_{\ell \in \{1, \dots, k\}} \left| \frac{1}{n} \sum_{i=1}^{n} f(Y_i, q_\ell) - \mathbb{E}[f(Y, q_\ell)] \right| .
\end{aligned}$$

For the second summand, by the strong law of large numbers with $\mathbb{E}[Z] < \infty$,

$$\sup_{\ell \in \{1, \dots, k\}} \left| \frac{1}{n} \sum_{i=1}^{n} f(Y_i, q_\ell) - \mathbb{E}[f(Y, q_\ell)] \right| \xrightarrow{n \to \infty}_{\text{a.s.}} 0\,.$$

For the first summand,

$$\sup_{\ell\in\{1,\dots,k\}}\sup_{q\in \mathrm{B}_\epsilon(q_\ell)}\left|\frac{1}{n}\sum_{i=1}^n (f(Y_i,q)-f(Y_i,q_\ell)) - \mathbb{E}[f(Y,q)-f(Y,q_\ell)]\right|$$
$$\le \frac{1}{n}\sum_{i=1}^n \sup_{q,p\in B,\,\overline{qp}\le\epsilon} |f(Y_i,q)-f(Y_i,p)| + \mathbb{E}\left[\sup_{q,p\in B,\,\overline{qp}\le\epsilon} |f(Y,q)-f(Y,p)|\right].$$

By the strong law of large numbers with $\mathbb{E}[Z]<\infty$,

$$\frac{1}{n}\sum_{i=1}^n \sup_{q,p\in B,\,\overline{qp}\le\epsilon} |f(Y_i,q)-f(Y_i,p)| \xrightarrow{n\to\infty}_{\text{a.s.}} \mathbb{E}\left[\sup_{q,p\in B,\,\overline{qp}\le\epsilon} |f(Y,q)-f(Y,p)|\right].$$

Thus,

$$\mathbb{P}\left(\limsup_{n\to\infty}\sup_{q\in B}\left|\frac{1}{n}\sum_{i=1}^n f(Y_i,q) - \mathbb{E}[f(Y,q)]\right| \le a_\epsilon\right) = 1\,, \tag{2.1}$$

where $a_\epsilon = 2\mathbb{E}\Big[\sup_{q,p\in B,\,\overline{qp}\le\epsilon}|f(Y,q)-f(Y,p)|\Big]$. As $q\mapsto f(Y,q)$ is almost surely continuous and B is compact, $q\mapsto f(Y,q)$ is almost surely uniformly continuous, i.e., for all $\delta>0$ there is $\epsilon>0$ such that $|f(Y,q)-f(Y,p)|\le\delta$ for all $\overline{q,p}\le\epsilon$. As $\mathbb{E}[Z]<\infty$, we can use dominated convergence to obtain

$$\lim_{\epsilon\searrow 0}\mathbb{E}\left[\sup_{q,p\in B,\,\overline{qp}\le\epsilon}|f(Y,p)-f(Y,p)|\right] = \mathbb{E}\left[\lim_{\epsilon\searrow 0}\sup_{q,p\in B,\,\overline{qp}\le\epsilon}|f(Y,p)-f(Y,p)|\right] = 0\,.$$

Thus, $a_\epsilon \xrightarrow{\epsilon\searrow 0} 0$. Together with (2.1), this implies

$$\sup_{q\in B}\left|\frac{1}{n}\sum_{i=1}^n f(Y_i,q) - \mathbb{E}[f(Y,q)]\right| \xrightarrow{n\to\infty}_{\text{a.s.}} 0\,.$$

We have also shown that $q\mapsto\mathbb{E}[f(Y,q)]$ is continuous, as $|\mathbb{E}[f(Y,q)]-\mathbb{E}[f(Y,p)]|\le a_{\overline{qp}}$. □

Lemma 2.17. Let $h\colon[0,\infty)\to[0,\infty)$ be a nondecreasing function. Define its integral function as $H\colon[0,\infty)\to[0,\infty), x\mapsto\int_0^x h(t)\mathrm{d}t$. Let $x,y\ge 0$. Then

(i) $|H(x)-H(y)|\le|x-y|\,h(\max(x,y))$.

Assume, there is $b\in[1,\infty)$ such that $h(2u)\le bh(u)$ for all $u\ge 0$. Then

(ii) $\frac{1}{2}h(x)+\frac{1}{2}h(y)\le h(x+y)\le b\,(h(x)+h(y))$,

(iii) $H(|x-y|)-H(x)\ge b^{-1}H(y)-2yh(x)$.

Proof. (i) This is a direct consequence of the mean value theorem.

(ii) As h is nondecreasing, $\max(h(x), h(y)) \leq h(x+y) \leq \max(h(2x), h(2y))$. By the definition of b and with $\frac{1}{2}(u+v) \leq \max(u,v) \leq u+v$ for $u, v \geq 0$ the claim follows.

(iii) First, consider the case $x \geq y$. Define $f(x,y) = H(x-y) - H(x) - b^{-1}H(y) + 2yh(x)$. We want to show $f(x,y) \geq 0$. The derivative of f with respect to y is

$$\partial_y f(x,y) = -h(x-y) - b^{-1}h(y) + 2h(x)\,.$$

By applying the first inequality of (ii) to $h(x) = h((x-y)+y)$, we obtain $\partial_y f(x,y) \geq 0$ as $b^{-1} \leq 1$. Hence, $f(x,y) \geq f(x,0) = 0$, as $H(y) = 0$.

Now, consider the case $x \leq y$. Set $g(x,y) = H(y-x) - H(x) - b^{-1}H(y) + 2yh(x)$, which yields

$$\partial_y g(x,y) = h(y-x) - b^{-1}h(y) + 2h(x)\,.$$

By applying the second inequality of (ii) to $h(y) = h((y-x)+x)$, we obtain $\partial_y g(x,y) \geq 0$ as $b^{-1} \leq 1$. Thus, $g(x,y) \geq g(x,x) = -(1+b^{-1})H(x) + 2xh(x)$ as $H(0) = 0$. By the definition of H, as h is nondecreasing, $H(x) \leq xh(x)$. Hence, $g(x,y) \geq 0$ as $1 + b^{-1} \leq 2$.

Together, we have shown $H(|x-y|) - H(x) - b^{-1}H(y) + 2yh(x) \geq 0$ for all $x, y \geq 0$. □

Lemma 2.18. Let X be a random variable with values in $[0,\infty)$. Then there is a strictly increasing, continuous, and concave function $h\colon [0,\infty) \to [0,\infty)$ with $h(\delta) \xrightarrow{\delta\to\infty} \infty$ such that $\mathbb{E}[h(X)] < \infty$.

Proof. If there is $K > 0$ such that $\mathbb{P}(X < K) = 1$ take $h(x) = x$. Now, assume that X is not almost surely bounded. We first construct a nondecreasing function $\tilde{h}\colon [0,\infty) \to [0,\infty)$ such that $\tilde{h}(x) \xrightarrow{x\to\infty} \infty$ with $\mathbb{E}[\tilde{h}(X)] < \infty$. Then we construct a function h from $\tilde{h}$ with all desired properties.

Let F be the distribution function of X, $F(x) = \mathbb{P}(X \leq x)$. Let $z_1 := 0$ and $z_{n+1} := \inf\{x \geq z_n + 1 \,|\, 1 - F(x) \leq \frac{1}{n}\}$. As $F(x) \xrightarrow{x\to\infty} 1$, $z_n < \infty$. Furthermore, $z_{n+1} - z_n \geq 1$. Moreover, as X is not almost surely bounded, $1 - F(x) > 0$ for all $x \geq 0$. Set

$$g(x) := \sum_{n=1}^{\infty} (z_{n+1} - z_n)^{-1} n^{-2} \mathbb{1}_{[z_n, z_{n+1})}(x)\,,$$
$$\tilde{h}(x) := \int_0^x \frac{g(t)}{1 - F(t)} \mathrm{d}t\,.$$

Then

$$\lim_{x\to\infty} \tilde{h}(x) = \int_0^\infty \frac{g(t)}{1-F(t)}\mathrm{d}t \geq \sum_{n=1}^{\infty} n^{-1} = \infty\,.$$

Moreover, $\tilde{h}(x)$ is strictly increasing, as $g(t) \geq 0$ and $1 - F(t) \geq 0$. The function $\tilde{h}$ is continuously differentiable everywhere except at point z_n, $n \in \mathbb{N}$. Thus,

$$\begin{aligned}
\mathbb{E}[\tilde{h}(X)] &= \int_0^\infty \mathbb{P}\Big(\tilde{h}(X) > t\Big)\mathrm{d}t \\
&= \int_0^\infty \mathbb{P}\Big(X > \tilde{h}^{-1}(t)\Big)\mathrm{d}t \\
&= \int_0^\infty \tilde{h}'(t)\mathbb{P}(X > t)\mathrm{d}t \\
&= \int_0^\infty g(t)\mathrm{d}t \\
&= \sum_{n=1}^{\infty} n^{-2} < \infty\,.
\end{aligned}$$

Let $a_0 := 1$, $x_0 := 0$, $x_{n+1} := \inf\{x \geq x_n + a_n^{-1} \,|\, \tilde{h}(x) \geq n+1\}$ and $a_{n+1} := (x_{n+1} - x_n)^{-1}$. Let $h\colon [0,\infty) \to [0,\infty)$ be the linear interpolation of $(x_n, n)_{n\in\mathbb{N}_0}$. As $\tilde{h}(x) \xrightarrow{x\to\infty} \infty$, all x_n are finite. Hence, $h(x) \xrightarrow{x\to\infty} \infty$. Because of $a_n > 0$, h is strictly increasing. Furthermore, $a_{n+1} \leq a_n$ as $x_{n+1} \geq x_n + a_n^{-1}$. As h is continuous and a_n is the derivative of h in the interval (x_n, x_{n+1}), h is concave. Lastly, $h(x) \leq \tilde{h}(x) + 1$. Thus, $\mathbb{E}[h(X)] < \infty$. □

3 Rates of Convergence and the Projected Mean

Contents

3.1 Introduction

We continue our journey through the realm of Fréchet means with one of its simplest nonstandard instances: the projected mean. After showing in chapter 2 that the sample FM does converge, we now want to know how fast it converges. To be more precise, we ask: What is the rate of convergence of a sample projected mean of independent and identically distributed data to the respective population projected mean? We will learn that although the projected mean often has a similar behavior as the Euclidean mean, in some instances it might act very differently. In fact, for any given target rate, we can construct cases where the sample projected mean converges with that rate to its population counterpart. This demonstrates the subtleties involved in finding rates of

convergence for FMs and empathizes the importance of the upcoming chapters, where rats for FMs are established in a general setting.

We recapitulate the definitions of projected and extrinsic mean from 1.2.4, here only for the Euclidean plane: Let Z be a random variable with values in $\mathbb{R}^2$ and finite second moment. Let $\mathcal{Q} \subseteq \mathbb{R}^2$ be a subset of the Euclidean plane. Assume $m := \arg\min_{p\in\mathcal{Q}} \|\mathbb{E}[Z] - p\|$ exists and is unique. We call m the *(population) projected mean* of Z in $\mathcal{Q}$. It is the FM of $(\mathbb{R}^2, d_{\mathbb{R}^2})$. Let $Z_1, \dots, Z_n$ be independent and identically distributed copies of Z. We estimate m by a *sample projected mean* $m_n \in \arg\min_{p\in\mathcal{Q}} \|\bar{Z}_n - p\|$, $\bar{Z}_n := \frac{1}{n}\sum_{i=1}^n Z_i$. If Z takes values only in $\mathcal{Q}$, then m and m_n are called *(population) extrinsic mean* and *sample extrinsic mean*, respectively, see, e.g., [BP03]. In [HL98], the extrinsic mean is called *mean location*.

For a given rate sequence $(a_n)_{n\in\mathbb{N}} \subseteq (0,\infty)$, $a_n \to 0$ our goal is to find a set $\mathcal{Q}$ such that for a large class of distributions of Z a central limit theorem of the form $a_n^{-1}(m_n - m) \xrightarrow{n\to\infty} \nu$ holds for some nondegenerate distribution ν. Then m_n converges to m in probability at rate a_n.

Asymptotics of extrinsic sample means in cases with parametric rate of convergence, i.e., $a_n = n^{-\frac{1}{2}}$, are well-studied in [HL98; Pat98; BP03; BP05] among others. This line of work is mostly concerned with finite dimensional manifolds, but results for infinite dimensional Hilbert manifolds are available [EPR13]. Slower rates for *intrinsic* sample means, i.e., minimizers of $p \mapsto \sum_{i=1}^n d_{\mathcal{Q}}(Z_i, p)^2$ with the intrinsic metric $d_{\mathcal{Q}}$, have been observed on the circle [HH15] and more general manifolds [EH19]. In some cases intrinsic and extrinsic means coincide [BP03, Theorem 3.3]. But this is not true in general.

The occurrence of a rate of convergence slower than the parametric one is called *smeariness*. If, in contrast, the sample mean is equal to its population counterpart with high probability, the behavior is called *stickiness*, which is observed for intrinsic means in certain negatively curved spaces [Hot+13; Huc+15].

3.1.1 Medial Axis and Reach

To be unique, m must not be located on the *medial axis* $\mathcal{M}_{\mathcal{Q}}$ of the set $\mathcal{Q}$, which is the set of all points that have more than one closest point in $\mathcal{Q}$. Recall from section 1.2.4,

$$\mathcal{M}_{\mathcal{Q}} = \left\{ z \in \mathbb{R}^2 \;\middle|\; \exists p_1, p_2 \in \mathcal{Q}, p_1 \neq p_2 \colon \|p_1 - z\| = \|p_2 - z\| = \inf_{p\in\mathcal{Q}} \|p - z\| \right\}.$$

The medial axis has been analyzed from a purely geometric perspective [BD17]. By the definition of medial axis $\mathcal{M}_{\mathcal{Q}}$ as it it is used here, it need not be a closed set, as the example $\mathcal{Q} = \{y = x^2\} \subseteq \mathbb{R}^2$, $\mathcal{M}_{\mathcal{Q}} = (1/2,\infty) \times \{0\}$ shows. Note that this contrasts some mentions of the term in the literature, e.g., in the context of [BP03, Theorem 3.2]. See [HHM10, Theorem A.5] for a sufficient condition for a closed medial axis.

The *reach* [Fed59] $\tau_{\mathcal{Q}} := \inf_{m\in\mathcal{M}_{\mathcal{Q}}, p\in\mathcal{Q}} \|m - p\|$ of a set $\mathcal{Q} \subseteq \mathbb{R}^2$ is the largest nonnegative real value such that any point in $\mathbb{R}^2$ with distance to $\mathcal{Q}$ less than $\tau_{\mathcal{Q}}$ has a unique closest point in $\mathcal{Q}$. If $\mathcal{Q}$ is a $\mathcal{C}^2$-manifold, the projection map $z \mapsto \Pi_{\mathcal{Q}}(z) = \arg\min_{p\in\mathcal{Q}} \|z - p\|$ is continuously differentiable on $\mathbb{R}^2 \backslash \overline{\mathcal{M}_{\mathcal{Q}}}$ with $\|\nabla \Pi_{\mathcal{Q}}(z)\| > 0$ [Aba78]. If additionally the reach $\tau_{\mathcal{Q}}$ is greater than the distance of $\mathbb{E}[Z]$ to $\mathcal{Q}$, the projected sample mean attains a

parametric rate of convergence [HL98; BP05]: The delta-method yields $\sqrt{n}\,(m_n - m) = \sqrt{n}\left(\Pi_{\mathcal{Q}}(\bar{Z}_n) - \Pi_{\mathcal{Q}}(\mathbb{E}[Z])\right) \xrightarrow{d} \mathcal{N}(0, \tilde{\Sigma})$ where $\tilde{\Sigma} = \nabla\Pi_{\mathcal{Q}}(\mathbb{E}[Z])' \cdot \mathbb{COV}(Z) \cdot \nabla\Pi_{\mathcal{Q}}(\mathbb{E}[Z])$. As convergence is a local phenomenon, we can replace the condition on the reach by the requirement that $\mathbb{E}[Z]$ is bounded away from the medial axis $\mathcal{M}_{\mathcal{Q}}$.

We construct sets $\mathcal{Q}$ with faster and slower rates of convergence than $1/\sqrt{n}$. In our examples, the sets $\mathcal{Q}$ for slow rates are $\mathcal{C}^2$-smooth, but $\mathbb{E}[Z]$ is too close to the medial axis, i.e., $\mathbb{E}[Z] \in \overline{\mathcal{M}_{\mathcal{Q}}} \setminus \mathcal{M}_{\mathcal{Q}}$. Sets $\mathcal{Q}$ with fast rates have reach $\tau_{\mathcal{Q}} > \inf_{p\in\mathcal{Q}} \|\mathbb{E}[Z] - p\|$ but are only $\mathcal{C}^1$- but not $\mathcal{C}^2$-manifolds.

3.1.2 Our Construction

For a continuous function f with $f(0) = 0$, we construct $\mathcal{Q} = \mathcal{Q}_f$ such that the projection of a point $(x, y)' \in \mathbb{R}^2$ to $\mathcal{Q}$ is roughly $(1, f(y))'$ for $|x|, |y|$ small enough. Assuming $\mathbb{E}[Z] = 0 \in \mathbb{R}^2$, the arithmetic mean $\bar{Z}_n = (\bar{X}_n, \bar{Y}_n)'$ concentrates at 0 with rate $1/\sqrt{n}$. Thus, $m_n = \Pi_{\mathcal{Q}}(\bar{Z}_n) \approx (1, f(\bar{Y}_n))'$ concentrates at $(1, 0)'$ with a rate depending on f. For a wide range of rates $(a_n)_{n\in\mathbb{N}} \subseteq (0, \infty)$, $a_n \to 0$, we can find a function f with corresponding set $\mathcal{Q}$ such that $a_n^{-1}(m_n - m) \xrightarrow{n\to\infty} \nu$ in distribution for some nondegenerate distribution ν. As an example, $f(y) = |y|^\gamma$, $\gamma > 0$ yields $a_n = n^{-\frac{\gamma}{2}}$, see Corollary 3.9. Examples of constructed sets for qualitatively different rates can be found in Figure 3.1.

3.1.3 Outline

In section 3.2, we present our main theoretical results. We state the requirements on the function f and describe how the set $\mathcal{Q}$ is constructed from f. Proposition 3.3 states a result on deterministic projection to $\mathcal{Q}$, while Theorem 3.5 describes how the projected sample mean converges to the projected population mean. The goal of section 3.3 is to illustrate the general statement of Theorem 3.5. We derive Corollary 3.9, which presents explicit functions f and sets $\mathcal{Q}$ for certain target rates a_n. In particular, we give examples where projected means attain polynomial, logarithmic, or exponential rates of convergence. Moreover, the results are visualized. All proofs are given in section 3.A. Lastly, to place these results in a larger context, we discuss some further circumstances where parametric means diverge from the parametric rate of convergence, in section 3.4.

3.2 Results

The possible choices of the function f, which determines the set $\mathcal{Q} = \mathcal{Q}_f$ and, thus, the rate of convergence, are not restricted very much.

(A0): Let $b > 0$. Let $f \in \mathcal{C}^0([0, b])$ be strictly increasing with $f(0) = 0$.

Under the assumption **(A0)**, we construct the set $\mathcal{Q}$ as follows. Set $B := f(b)$. We denote the inverse function of $f\colon [0, b] \to [0, B]$ by $g\colon [0, B] \to [0, b]$, i.e., $g(x) := f^{-1}(x)$.

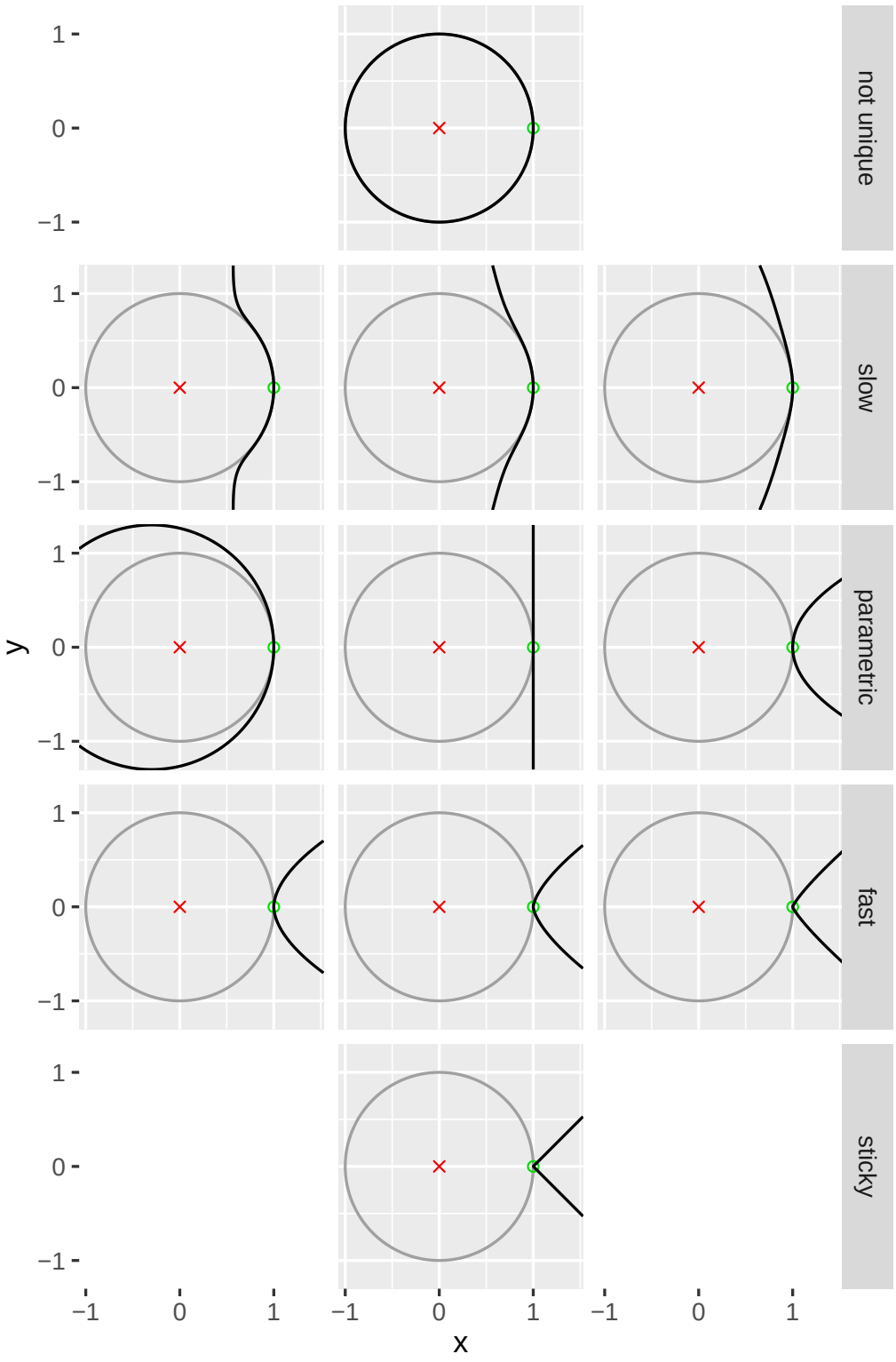

Figure 3.1: The images show the transition of the set $\mathcal{Q}$ (black) from nonunique projections, to slow, parametric, and fast rates, and sticky behavior of m_n. For reference, a circle (gray) with radius 1 around the origin is drawn. The expectation of Z and its projection to $\mathcal{Q}$ are marked in red and green, respectively.

For $t \in [0, B]$, define $r(t) := r_f(t) := 1 + \int_0^t g(x)\mathrm{d}x$. Finally, define

$$\begin{aligned} q(t) := q_f(t) := r(|t|) \begin{pmatrix} \cos(t) \\ \sin(t) \end{pmatrix} \text{ for } t \in [-B, B]\,, \\ \mathcal{Q} := \mathcal{Q}_f := \{q(t) \colon t \in [-B, B]\}\ . \end{aligned} \tag{3.1}$$

Our main results are based on the observation that the projection of a point $(x, y)'$ to $\mathcal{Q}$ for x, y small enough is essentially $(1, f(y))'$.

We denote the projection of $z \in \mathbb{R}^2$ to $\mathcal{Q}$ as $\Pi_{\mathcal{Q}}(z)$, i.e., $\Pi_{\mathcal{Q}}(z) := \arg\min_{p \in \mathcal{Q}} \|z - p\|$. If the argmin is not unique, we assume that one element of the argmin–set is chosen by a fixed arbitrary mechanism, e.g., smallest lexicographic order. The argmin–set cannot be empty as $\mathcal{Q}$ is compact by construction.

Lemma 3.1. Assume **(A0)** with $\mathcal{Q}$ from (3.1). For $y \in \mathbb{R}$, we consider $y \to 0$. Let $x = \mathbf{O}(y)$, and $t_y \in [-B, B]$ such that $\Pi_{\mathcal{Q}}((x, y)') = q(t_y)$. Then

$$g(t_y) = y + \mathbf{o}(y)\,.$$

Remark 3.2 (Simpler construction)**.** As can be seen from the proof of Lemma 3.1, a simpler construction in the case of $f(t) = \mathbf{o}(t)$, where $t \to 0$, is replacing $q(t)$ by

$$\tilde{q}(t) := \begin{pmatrix} 1 + tg(t) \\ t \end{pmatrix}\,.$$

This yields the same results, but does not include $g(t) = \mathbf{o}(t)$.

The projection of a point close to the origin is represented by t_y. Lemma 3.1 describes t_y in an indirect way, i.e., after applying the function g. To have a direct statement, we need to make additional assumptions.

(A1): Assume

$$\lim_{y \searrow 0} \frac{f(y + cy(y + f(y)))}{f(y)} = 1$$

for all $c \in \mathbb{R}$.

(A1)': Assume

$$\lim_{y \searrow 0} \frac{f(y + cyf(y)(y + f(y)))}{f(y)} = 1$$

for all $c \in \mathbb{R}$.

Proposition 3.3. Assume **(A0)** and **(A1)** with $\mathcal{Q}$ from (3.1). For $y \in \mathbb{R}$, consider $y \to 0$. Let $x = \mathbf{O}(y)$ and $t_y \in [-B, B]$ such that $\Pi_{\mathcal{Q}}((x, y)') = q(t_y)$. Then

$$t_y = f(y) + \mathbf{o}(f(y)) \qquad \text{and} \qquad \Pi_{\mathcal{Q}}\left(\begin{pmatrix} x \\ y \end{pmatrix}\right) = \begin{pmatrix} 1 \\ f(y) \end{pmatrix} + \mathbf{o}(f(y))\,.$$

Furthermore, if $x = 0$, we can replace the assumption **(A1)** by **(A1)'**.

This shows that a point close to the origin essentially is projected to $(1, f(y))'$. Let us shortly discuss the conditions for this result.

Remark 3.4 (On the assumptions **(A1)** and **(A1)'**)**.**

(a) We have

$$\lim_{y \searrow 0} \frac{f(y + \mathbf{o}(y))}{f(y)} = 1 \tag{3.2}$$

for any function of the form $f(y) = ay^\gamma$, with $a, \gamma > 0$. Furthermore, (3.2) implies **(A1)**, and **(A1)** implies **(A1)'**.

(b) It is unclear to the author, whether there is a function that fulfills **(A0)** but not **(A1')**.

(c) The function $f(y) = \exp(-1/y)$ fulfills **(A0)** and **(A1)'**, but does not fulfill **(A1)**. If we set $x = y$, we obtain

$$t_y = f\left(\frac{y}{1-y}\right) + \mathbf{o}\left(f\left(\frac{y}{1-y}\right)\right) = \exp(1) f(y) + \mathbf{o}(f(y)) \neq f(y) + \mathbf{o}(f(y))\,.$$

If we set $\tilde{f}(y) = \exp(-\exp(1/y))$, we even have $\tilde{f}(y) = \mathbf{o}(\tilde{t}_y)$.

Note that $x \mapsto \exp(-1/x)\mathbb{1}_{(0,\infty)}(x)$ is a classical example of a function that is infinitly often differentiable but not analytic: for every $k \in \mathbb{N}_0$ the k-th derivative at 0 is 0, $f^{(k)}(0) = 0$.

(d) If $f \in \mathcal{C}^k$, i.e., f is k-times continuously differentiable, $k \in \mathbb{N}$, and there is an $\ell \in \{1, \dots, k\}$ such that $f^{(\ell)}(0) \neq 0$, we set $\ell_0 = \min\{\ell \in \{1, \dots, k\} : f^{(\ell)}(0) \neq 0\}$. Then, by Taylor's theorem, $f(z) = \frac{f^{(\ell_0)}(0)}{\ell_0!} z^{\ell_0} + \mathbf{o}(z^{\ell_0})$. Thus, (3.2), **(A1)**, and **(A1)'** hold.

As taking the projected mean is projecting the Euclidean mean in $\mathbb{R}^2$ to $\mathcal{Q}$, Lemma 3.1 induces a central limit theorem for projected means. This is the main result of this chapter. It is illustrated in the next section.

Theorem 3.5. Assume **(A0)** with $\mathcal{Q}$ from (3.1). Let $Z = (X, Y)'$ be a random variable in $\mathbb{R}^2$ with finite second moment, $\mathbb{E}[Z] = 0 \in \mathbb{R}^2$, and $\mathbb{V}[Y] = \sigma^2 > 0$. Let $Z_1, \dots, Z_n$ be independent copies of Z. Then the projected population mean $m \in \mathcal{Q}$ exists, is unique, and

$$m = \Pi_{\mathcal{Q}}(\mathbb{E}[Z]) = \operatorname*{arg\,min}_{p \in \mathcal{Q}} \mathbb{E}[\|Z - p\|^2] = q(0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Let $(m_{n,1}, m_{n,2})' := m_n := \Pi_{\mathcal{Q}}\left(\bar{Z}_n\right)$, $\bar{Z}_n := \frac{1}{n}\sum_{i=1}^n Z_i$. Then m_n is a projected sample mean. Let $t_n \in [-B, B]$ such that $m_n = q(t_n)$. Then, for $s \geq 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(t_n \leq f\left(\frac{s}{\sqrt{n}}\right)\right) = \lim_{n\to\infty} \mathbb{P}\left(-t_n \leq f\left(\frac{s}{\sqrt{n}}\right)\right) =$$
$$\lim_{n\to\infty} \mathbb{P}\left(m_{n,2} \leq f\left(\frac{s}{\sqrt{n}}\right)\right) = \lim_{n\to\infty} \mathbb{P}\left(-m_{n,2} \leq f\left(\frac{s}{\sqrt{n}}\right)\right) = \Phi\left(\frac{s}{\sigma}\right),$$

where Φ denotes the distribution function of a standard normal random variable. Moreover,

$$\mathbb{P}\left(|m_{n,1} - 1| \geq f\left(\frac{s}{\sqrt{n}}\right)\right) \xrightarrow{n\to\infty} 0\,.$$

Before applying the theorem to better understand its implications, we quickly remark on some theoretical properties.

Remark 3.6 (Arc length)**.** The curve $q(t)$ in (3.1) is not necessarily parameterized by arc length. But $q \in \mathcal{C}^1((-B, B))$ and $\|\dot{q}(t)\| = 1 + \mathbf{o}(1)$ for $|t| \to 0$ as

$$\begin{aligned}\|\dot{q}(t)\|^2 &= (g(t)\cos(t) - r(t)\sin(t))^2 + (g(t)\sin(t) + r(t)\cos(t))^2 \\ &= 1 + (g(t) - t)^2 + \mathbf{O}(tg(t) + t^2)\,.\end{aligned}$$

Thus, the results on t_n in Theorem 3.5 also hold if t_n is replaced by an arc length parametrization.

Remark 3.7 (Why Theorem 3.5 does not require **(A1)**)**.** In contrast to Proposition 3.3, we do not require **(A1)** or **(A1')** in Theorem 3.5. In particular, in the setting of Remark 3.4 (c), $f(y) = \exp(-1/y)$, we have

$$\mathbb{P}\left(t_n \leq f\left(\frac{s}{\sqrt{n}}\right)\right) \xrightarrow{n\to\infty} \Phi\left(\frac{s}{\sigma}\right),$$

for $s \geq 0$ even though $t_n \neq f(\bar{Y}_n) + \mathbf{o}(f(\bar{Y}_n))$. The reason is that the difference between t_n and $f(\bar{Y}_n)$ is negligible in the scale that is used in Theorem 3.5. The

right scale for a central limit theorem of t_n is the one of $\bar{Y}_n$ (multiplied by $\sqrt{n}$), i.e., $g(t_n)$. The factor e in $t_n \approx \mathrm{e} f(\bar{Y}_n)$, see Remark 3.4 (c), is nonnegligible on the scale of t_n, but on the scale of $\bar{Y}_n$ it becomes

$$\frac{g(\mathrm{e} f(\bar{Y}_n))}{\bar{Y}_n} = \frac{\log\Big(\mathrm{e}^{-1} \cdot \exp\Big(\bar{Y}_n^{-1}\Big)\Big)^{-1}}{\bar{Y}_n} = 1/(1-\bar{Y}_n) \xrightarrow{n\to\infty} 1$$

almost surely, i.e., negligible.

Remark 3.8 (Nonuniqueness)**.** Nonunique closest points of $\bar{Z}_n$ are not a problem in Theorem 3.5 as $\mathbb{P}(\bar{Z}_n \in \mathcal{M}_{\mathcal{Q}}) \to 0$ by $\mathbb{V}[Y] = \sigma^2 > 0$. See also Remark 3.15.

3.3 Illustration

To illustrate Theorem 3.5, we apply it to explicit functions f, which yield polynomial, logarithmic, and exponential rates of convergence for $m_n \to m$, respectively.

Corollary 3.9. Use the setting of Theorem 3.5.

(i) Let $f(y) = y^\gamma$ with $\gamma > 0$. Then $r(t) = 1 + \frac{\gamma}{1+\gamma} t^{\frac{1+\gamma}{\gamma}}$ and

$$n^{\frac{\gamma}{2}} t_n \to T$$

in distribution, where $\mathbb{P}(T \le s) = \Phi\left(\frac{\mathsf{sgn}(s)|s|^{\frac{1}{\gamma}}}{\sigma}\right)$ for all $s \in \mathbb{R}$.

(ii) Let $f(y) = (-\log(y))^{-\gamma}$ with $\gamma > 0$. Then $r(t) = 1 + \int_0^t \exp\Big(-x^{-\frac{1}{\gamma}}\Big)\,\mathrm{d}x$ and

$$\left(\frac{1}{2}\log(n)\right)^\gamma t_n \to T$$

in distribution, where $\mathbb{P}(T = 1) = \mathbb{P}(T = -1) = \frac{1}{2}$.

(iii) Let $f(y) = \exp(-y^{-\gamma})$ with $\gamma > 0$. Then $r(t) = 1 + \int_0^t \log(x^{-1})^{-\frac{1}{\gamma}}\,\mathrm{d}x$. For $c > 0$, define $U_{n,c} := \exp((\sqrt{n}/c)^\gamma)\, t_n$ and $p_c := \Phi(\frac{c}{\sigma})$. Then, for all $u \in (0,\infty)$, $\mathbb{P}(U_{n,c} \ge u) \xrightarrow{n\to\infty} 1 - p_c$, $\mathbb{P}(U_{n,c} \le -u) \xrightarrow{n\to\infty} 1 - p_c$, and $\mathbb{P}(|U_{n,c}| \le u) \xrightarrow{n\to\infty} 2p_c - 1$.

The results also hold when t_n is replaced by $m_{n,2}$.

The results of Corollary 3.9 are also true in arc length, see Remark 3.6.

Remark 3.10 (On Corollary 3.9)**.**

(i) For any polynomial scale n^γ, part (i) of Corollary 3.9 gives an example of a central limit theorem with that scale.

(ii) In part (ii) we obtain a central limit theorem with logarithmic scale and a Bernoulli-type limiting distribution that does not depend on σ. This seems quite remarkable and can be explained as follows:

Scaling our observations Z_i by σ^{-1}, is roughly like scaling n by σ^2 as the variance is $\mathbb{V}[\sigma^{-1}\bar{Y}_n] = n^{-1} \approx \mathbb{V}[\bar{Y}_{[n\sigma^2]}]$, where $[n\sigma^2]$ denotes the closest integer to $n\sigma^2$. The scaling factor $\log(n)^\gamma$ is asymptotically equivalent to $\log(n\sigma^2)^\gamma$. Thus, constant factors like σ cannot influence the asymptotic distribution on the scale $\log(n)^\gamma$.

Densities of t_n in the case of normally distributed observations are plotted in Figure 3.2.

(iii) The statement of part (iii) of Corollary 3.9 can be summarized informally by

$$\exp\big((\sqrt{n}/c)^\gamma\big)\, t_n \to T_c\,,$$

where $\mathbb{P}(T_c = \infty) = \mathbb{P}(T_c = -\infty) = 1 - p_c$ and $\mathbb{P}(T_c = 0) = 2p_c - 1$. The limiting distribution has mass only at 0 and $\pm\infty$. If the scale is changed such that the limit does not have a point mass at 0, all mass escapes to $\pm\infty$. If the scale is such that no mass escapes to $\pm\infty$, then in the limit all mass is at 0.

Densities of t_n in the case of normally distributed observations are plotted in Figure 3.3 on a log-log-scale. Only the positive axis of the symmetric densities is displayed. The plot shows that the densities have nonnegligible mass at all small orders of magnitude. Thus, choosing one specific order of magnitude by a specific scale makes all mass on larger orders of magnitude escape to infinity and all mass at smaller orders of magnitude go to 0.

Remark 3.11 (Extrinsic mean)**.** For the sets $\mathcal{Q}$ constructed in Corollary 3.9, there might not be a distribution with support in $\mathcal{Q}$ that has expectation 0. In particular, they might not directly yield examples of extrinsic means with the described asymptotic behavior. This is but a technical inconvenience. We can extend $\mathcal{Q}$ with an arbitrary set of points which have a distance to the origin that is bounded away from 1, and the result does not change. By doing so, we can also construct 2-dimensional manifolds with boundary which induce the same convergence results as the 1-dimensional structures in Corollary 3.9.

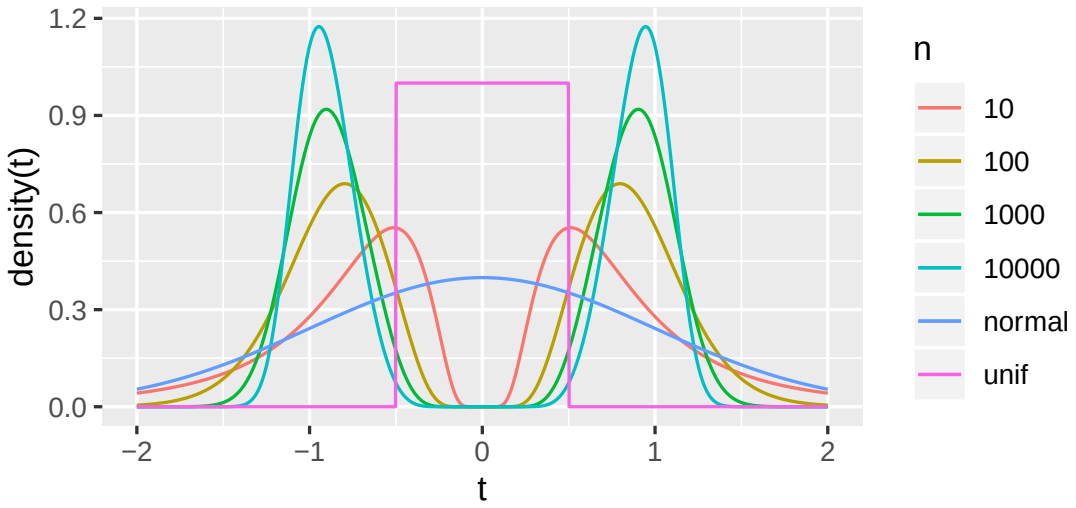

Figure 3.2: Plot of densities of $\frac{1}{2}\log(n)t_n$ for $f(y) = -\log(y)^{-1}$, $Z = (0, Y)'$ and $Y \sim \mathcal{N}(0, 1)$, with standard normal and uniform densities for reference.

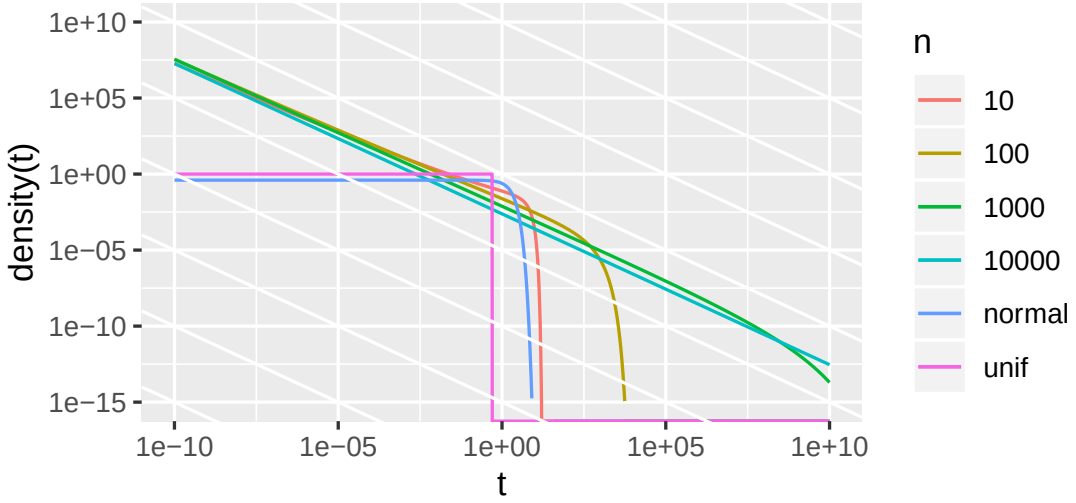

Figure 3.3: Log-log-plot of densities of $\exp(\sqrt{n})t_n$ for $f(y) = \exp(-y^{-1})$, $Z = (0, Y)'$ and $Y \sim \mathcal{N}(0, 1)$, with standard normal and uniform densities for reference.

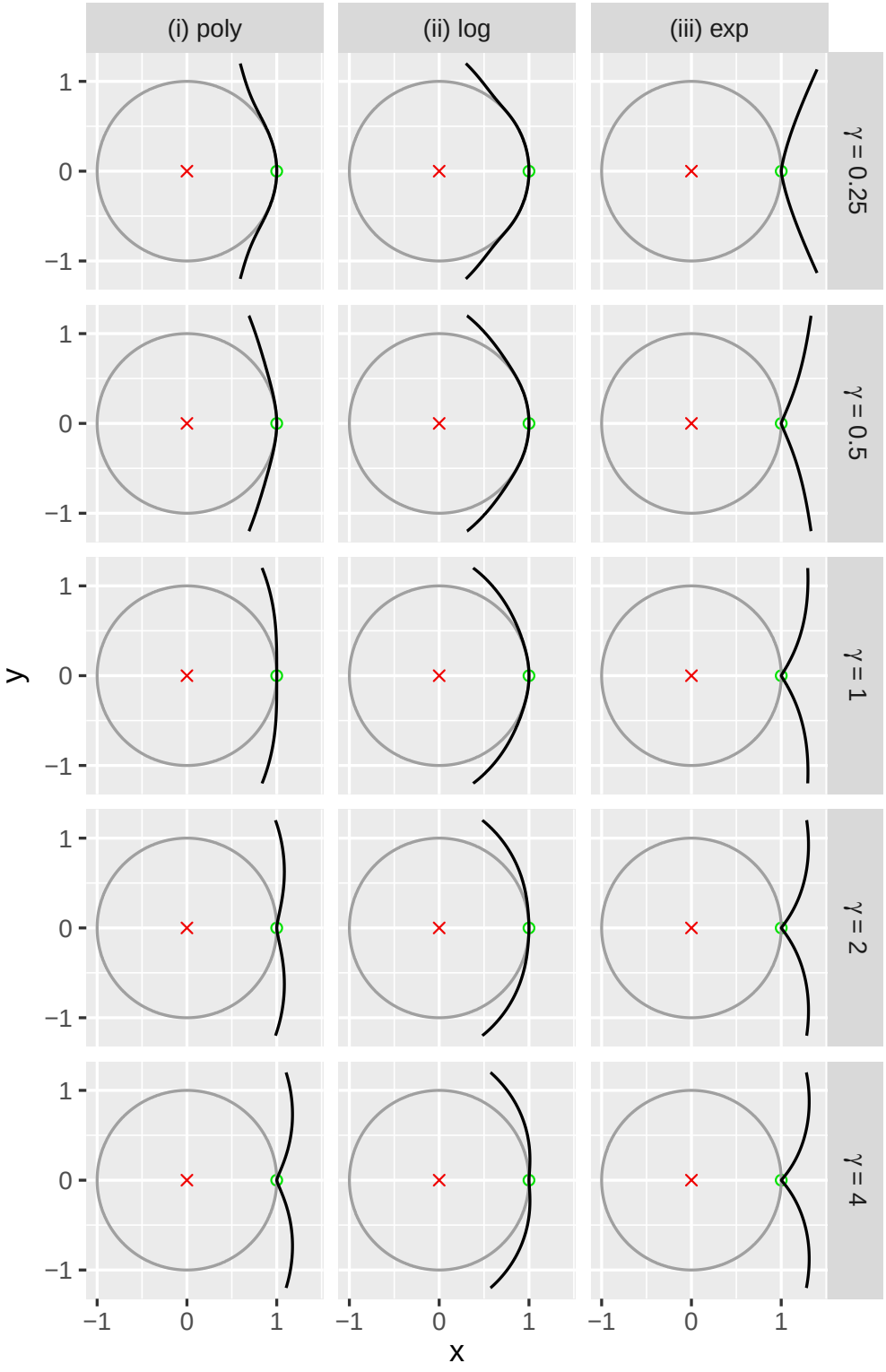

Figure 3.4: The images show the set $\mathcal{Q}$ (black) for different curves q, which are chosen as described in Corollary 3.9. For reference, a circle (gray) with radius 1 around the origin is drawn. The expectation of Z and its projection to $\mathcal{Q}$ are marked in red and green, respectively.

Remark 3.12 (Application of Proposition 3.3)**.** For the functions f in (i) and (ii), **(A1)** holds, see section 3.A.5. Thus, Proposition 3.3 implies

$$m_n = \Pi_{\mathcal{Q}}(\bar{Z}_n) \approx \begin{pmatrix} 1 \\ \mathsf{sgn}(\bar{Y}_n) f(|\bar{Y}_n|) \end{pmatrix},$$

meaning $\left| m_{n,2} - \mathsf{sgn}(\bar{Y}_n) f(|\bar{Y}_n|) \right| / f(|\bar{Y}_n|) \to 0$ and $|m_{n,1} - 1| / f(|\bar{Y}_n|) \to 0$ in probability. In (iii) only **(A1')** is true. Thus, the equation above is true for (iii) if $X = 0$ almost surely.

Remark 3.13 (Delta method)**.** In light of the delta method, note that, in the cases above, $f'(0)$ is 0 or ∞, except when f is equal to the identity in (i). This is the only case of Corollary 3.9 that yields the usual parametric rate and the conditions of the δ-method are fulfilled.

Figure 3.4 illustrates the sets $\mathcal{Q}$ constructed according to the functions f from Corollary 3.9. The results on the convergence rate described in Theorem 3.5 and Corollary 3.9 depend only on the form of the curve close to the point $(1, 0)'$. Even so the curve [*(ii) log*, $\gamma = 4$] looks like it is growing faster away from the circle than [*(i) poly*, $\gamma = 0.25$], the opposite is true when observing a neighborhood of $(1, 0)'$ that is small enough.

There is a smooth transition of the set $\mathcal{Q}$ between slow and fast rates, see Figure 3.1. A circle with radius 1 centered at the origin can be seen as one extreme case, in the sense that an arbitrarily small change of a point at the origin can change its projection by a large amount. If $\mathcal{Q}$ almost looks like this circle, but increases its radius $r(t)$ slow enough, i.e, $r(t) \lesssim 1 + t^2$, we still have large changes in the projection, but not arbitrarily large. For a larger circle with center $(-\delta, 0)'$ and radius $1 + \delta$ or a straight vertical line through $(1, 0)'$ the changes of point and projection are proportional, i.e, $r(t) \approx 1 + t^2$. Changes in the point effect the projection only little if $q(t)$ grows to the right quickly when moving away from $(1, 0)'$, i.e, $r(t) \gtrsim 1 + t^2$. For $\mathcal{Q} = \{(1 + |y|, y)' : y \in \mathbb{R}\}$ certain changes do not change the projection at all. In particular, $\mathbb{P}(m_n = m) \to 1$ (stickiness).

Remark 3.14 (Larger circles)**.** A circle with center at $(-\delta, 0)'$, $\delta > 0$, and radius $1 + \delta$, see Figure 3.5, can be described by our construction with

$$r(t) = \sqrt{\cos(t)^2 \delta^2 + 2\delta + 1} - \cos(t)\delta\,,$$

$t \in [-\pi, \pi]$. Thus,

$$g(t) = \dot{r}(t) = \delta \sin(t) - \frac{\cos(t)\sin(t)\delta^2}{\sqrt{\cos(t)^2\delta^2 + 2\delta + 1}} = \frac{\delta}{\delta + 1} t + \mathbf{O}(t^2)\,.$$

Hence, the projection $\Pi_{\mathcal{Q}}(\bar{Z}_n)$ scales the y-direction only by a constant factor with-

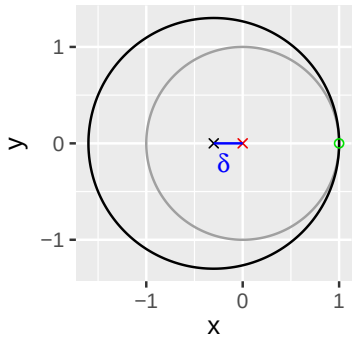

Figure 3.5: The black curve shows the set $\mathcal{Q}$ as described in Remark 3.14 with $\delta = 0.3$. For reference, a circle (gray) with radius 1 around the origin is drawn. The expectation of Z and its projection to $\mathcal{Q}$ are marked in red and green, respectively.

out affecting the rate of convergence. In particular we have a parametric rate of convergence. This can also be inferred by noting that $\mathcal{Q}$ is $\mathcal{C}^2$-smooth and has a reach larger than 1 as described in the introduction.

Remark 3.15 (Reach and Medial Axis)**.** A set $\mathcal{Q}$ of our construction has reach at most 1 if $g(t) = \mathbf{o}(t)$ for $t \searrow 0$. This can be seen from Remark 3.14: If every circle with center at $(-\delta, 0)'$ and radius $1 + \delta$ for $\delta \in (0, \delta_0)$, $\delta_0 > 0$ intersects $\mathcal{Q}$ at more than one point the reach can be at most 1. Moreover, such a circle is constructed with $g_{\mathsf{circle},\delta}(t)$ of order t, i.e., $g(t) = \mathbf{o}(g_{\mathsf{circle},\delta}(t))$ and $r(t) = \mathbf{o}(r_{\mathsf{circle},\delta}(t))$. Thus, $\{(-\delta, 0)' \colon \delta \in (0, \delta_0)\} \subseteq \mathcal{M}_{\mathcal{Q}}$ and $0 \in \partial\mathcal{M}_{\mathcal{Q}}$.

3.4 Digression: Nonstandard Rates of Convergence

In parametric statistics, one often observes the *parametric rate of convergence*, i.e., $d(\hat{\theta}_n, \theta)$ is of order $n^{-\frac{1}{2}}$ for an estimator $\hat{\theta}_n$ of θ in a metric space $(\mathcal{Q}, d)$, where usually $\mathcal{Q} \subseteq \mathbb{R}^s$ is convex and d is the Euclidean distance. In one way or another, the appearance of that rate is often connected to the central limit theorem (CLT): Let Z be a real-valued random variable with variance $\sigma^2 \in (0, \infty)$ and $(Z_i)_{i\in\mathbb{N}}$ be independent and identically distributed copies of Z. Then $n^{\frac{1}{2}}(\bar{Z}_n - \mathbb{E}[Z]) \xrightarrow{n\to\infty}_d \mathcal{N}(0, \sigma^2)$, where $\bar{Z}_n := \frac{1}{n}\sum_{i=1}^n Z_i$. Similarly the rate appears in the $\mathbb{L}^2$ distance, $\mathbb{E}[(\bar{Z}_n - \mathbb{E}[Z])^2]^{\frac{1}{2}} = \sigma n^{-\frac{1}{2}}$.

Knowing this, natural questions to ask could be: What other rates of convergence can occur, in which setting, and why? To get some satisfying answers, we need to make further restrictions: We assume that, as above $(Z_i)_{i\in\mathbb{N}}$ be independent and identically distributed copies of a random variable Z with distribution $Z_*\mathbb{P}$ and live in $\mathbb{R}$ or $\mathbb{R}^2$,

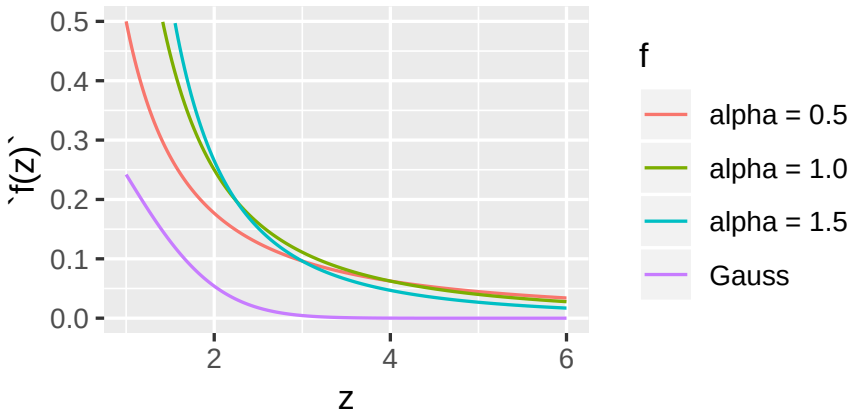

Figure 3.6: Heavy-tailed densities.

and $\hat{\theta}_n(Z_1, \dots, Z_n)$ is an estimator of $\theta(Z_*\mathbb{P})$. Furthermore, θ should be a Fréchet mean with respect to some metric (or a generalization of it).

We exclude nonparametric settings, where it is well-known that the *nonparametric rate of convergence* $n^{-\frac{\beta}{2\beta+s}}$ for a smoothness parameter $\beta > 0$ and a dimension parameter $s \in \mathbb{N}$ occurs when estimating β-smooth functions $\mathbb{R}^s \to \mathbb{R}$. Roughly, these slower rates are a consequence of estimating an object inside an infinite dimensional space from finite dimensional observations.

3.4.1 Levy α-stable distributions

The CLT for Levy α-stable distributions P_∞^α gives us a first answer, how we can create slower rates of convergence: by violating the second moment condition of the classical CLT.

Proposition 3.16 ([Kle14, chapter 16.2])**.** Let Z have a density $f\colon \mathbb{R} \to [0, \infty)$ that is symmetric, $f(z) = f(-z)$, with tail behavior $f(z) \sim |z|^{-\alpha-1}$ for $\alpha \in (0, 2]$. Then $n^{\frac{\alpha-1}{\alpha}} \bar{Z}_n \xrightarrow{d} P_\infty^\alpha$.

For $\alpha = 2$ this is the classical CLT. We obtain slow rates $n^{\frac{\alpha-1}{\alpha}} < n^{\frac{1}{2}}$ for $\alpha \in (0, 2)$. The rates for $\alpha < 1$ are even too slow for convergence in the sense of a law of large numbers. The second moment condition is violated as $\mathbb{E}[|Z|^\beta] = \infty$ for $\beta > \alpha$, due to the heavy tails of f, see Figure 3.6, and the strong dependence of the Euclidean mean on extreme observations. Note that nontrivial results for faster rates are not possible with this construction.

3.4.2 Max-stable distributions

Related to Levy α-stable distributions are max-stable distributions, which occur in extreme value theory. Extreme value theory studies the statistic $\theta_n := \max(Z_1, \dots, Z_n)$. For the maximum, different rates of convergence can be observed:

Proposition 3.17 ([HF06, chapter 1])**.** Let $\alpha > 0$.

(i) Let $Z \sim \mathsf{Pareto}(\alpha)$, i.e., $\mathbb{P}(Z \leq t) = 1 - t^{-\alpha}$ for $t \geq 1$. Then $n^{-\frac{1}{\alpha}}\theta_n \xrightarrow{d} Z_\infty \sim \mathsf{Frechet}(\alpha)$, i.e., $\mathbb{P}(Z_\infty \leq t) = \exp(-t^{-\alpha})$ for $t > 0$.

(ii) Let $Z + 1 \sim \mathsf{Beta}(1, \alpha)$, i.e., $\mathbb{P}(Z \leq t) = 1 - (-t)^\alpha$ for $t \in [-1, 0]$. Then $n^{\frac{1}{\alpha}}\theta_n \xrightarrow{d} Z_\infty \sim \mathsf{Weibull}(\alpha)$, i.e., $\mathbb{P}(Z_\infty \leq t) = \exp(-(-t)^\alpha)$ for $t \leq 0$.

Obviously, the empirical maximum is determined by values close to the right endpoint $z_{\mathsf{max}} := \sup \overline{\mathsf{supp}}(Z_*\mathbb{P}) \in \mathbb{R} \cup \{\infty\}$ of the support of the distribution of Z. Thus, we observe slow rates for densities f of Z with $f(z) \to 0$ as $z \to z_{\mathsf{max}}$ (i) and fast rates for $f(z) \to \infty$ as $z \to z_{\mathsf{max}}$ (ii).

The maximum is the limit case of Hölder means with power $\alpha \to \infty$, but it may not be a Fréchet mean, which are the objects we want to study. Fortunately, we can use the same idea and reasoning as above to obtain different rates for the median, as is shown in the next section.

3.4.3 Median

Let $\theta_n := \mathsf{median}(Z_1, \dots, Z_n)$, $\theta := \mathsf{median}(Z_*\mathbb{P})$. The empirical median θ_n is not influenced by extreme values, but depends strongly on the mass of $Z_*\mathbb{P}$ close to θ. This is opposite to the arithmetic mean. Assume the density f of Z is symmetric $f(z) = f(-z)$. Thus 0 is a median of Z. On one hand, if $f(z) = 0$ for all $z \in [-a, a]$, $a > 0$, then clearly the interval $[-a, a]$ is in the set of medians. On the other hand, if Z is equal to 0 with positive probability, e.g., $Z_*\mathbb{P} = p\delta_0 + (1-p)g \cdot \mathrm{d}z$ with $p \in (0, 1]$ and a probability density g, the median is sticky, i.e., after finitely many samples it holds $\theta_n = \theta$ with high probability. Furthermore, if the density f of Z is continuous and $0 < f(0) < \infty$, it is well-known that $n^{\frac{1}{2}}\theta_n \xrightarrow{d} \mathcal{N}(0, (2f(0))^{-2})$, see, e.g., [DN03, Theorem 10.3]. Now we interpolate between those cases such that we violate this bounded density condition but still have a unique but not sticky median. Set $Z \sim \mathsf{Beta}(1, \alpha)^{\mathsf{sym}}$, $\alpha > 0$, where $\mathsf{Beta}(1, \alpha)^{\mathsf{sym}}$ is the distribution with the density of a $\mathsf{Beta}(1, \alpha)$-distribution mirrored at 0 (and scaled by $\frac{1}{2}$), see Figure 3.8. Then $n^{\frac{1}{2\alpha}}|\theta_n|$ does not go to zero, but $n^{\frac{1}{2\alpha}}|\theta_n| = \mathbf{O}_{\mathbb{P}}(1)$, i.e., the median converges at rate $n^{-\frac{1}{2\alpha}}$. We do not prove this claim. Instead, we show a simulation that strongly affirms the claim, see Figure 3.9.

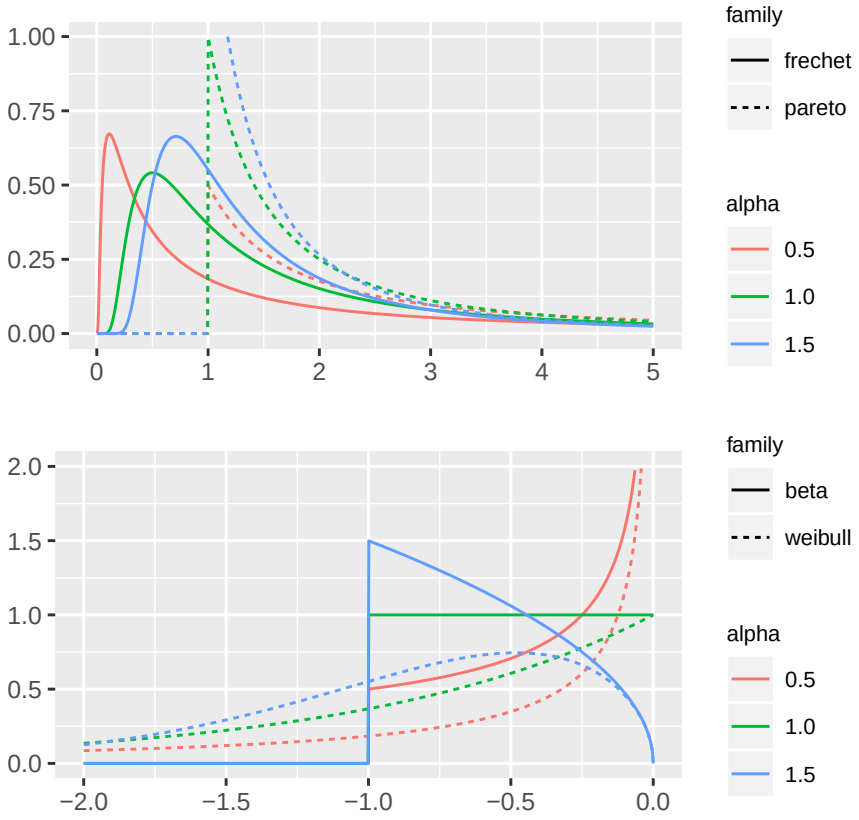

Figure 3.7: The densities of Pareto-, Fréchet-, Beta-, and Weibull-distribution for different parameters.

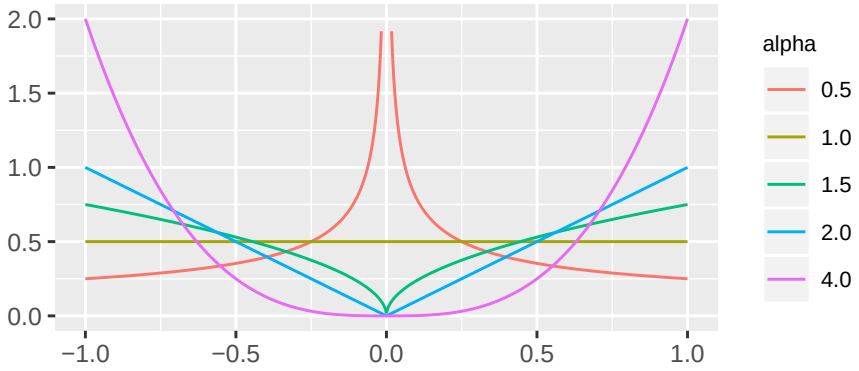

Figure 3.8: The densities of $\mathsf{Beta}(1,\alpha)^{\mathsf{sym}}$ are mirrored densities of $\mathsf{Beta}(1,\alpha)$. Their support is $[-1,1]$.

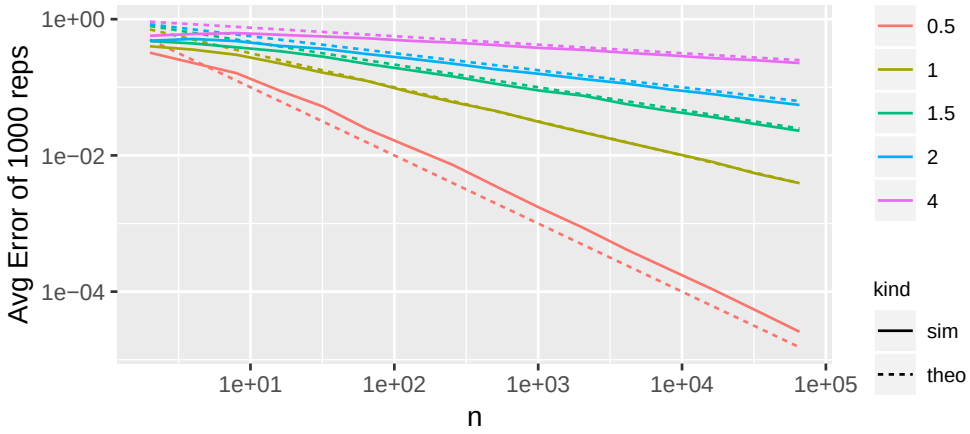

Figure 3.9: The results of a simulation study on a log-log-scale. The rate of convergence of the median behaves as expected. The constant difference of the red lines indicates a constant ratio between the mean error (solid) and the value n^{-1} (dashed).

3.4.4 Intrinsic Mean

So far, we have changed the behavior of probability densities on $\mathbb{R}$ to change the rate of convergence of certain statistics. Now, we will change the geometry of the underlying space to obtain a similar effect. We start out with the Fréchet mean with respect to the inner metric on the circle, also called intrinsic mean.

We shortly present here some results of [HH15]. Let Z have values on the unit circle, parameterized as $(-\pi, \pi]$. Let $d\colon (-\pi,\pi]^2 \to [0,\infty)$ be the arc-length metric on the circle. The population intrinsic mean is $\theta := \arg\min_{q\in(-\pi,\pi]} \mathbb{E}[d(Z,q)^2]$ with its sample counterpart $\theta_n := \arg\min_{q\in(-\pi,\pi]} \frac{1}{n}\sum_{i=1}^n d(Z_i,q)^2$.

To get a feeling of how the intrinsic mean behaves, we first take a look at some illustrative examples. In the left image of Figure 3.10, the location of the mean is intuitively meaningful. It also may be no surprise that the FMS of the uniform distribution on the circle (right image) is the whole circle. Slightly less simple, but still quite understandable is the setting with mass only at two andipodal points, where the FMS consists of two points, see Figure 3.11. Maybe surprisingly, we can actually construct distributions where whole segments of the circle make up the FMS, see Figure 3.12. Now, similar to the construction for the median, we interpolate between the settings with clearly unique means and with nonunique means, see Figure 3.13, and obtain scenarios with $n^{\frac{1}{2(\alpha+1)}}|\theta_n - \theta| = \mathbf{O}_{\mathbb{P}}(1)$ for $\alpha > 0$ ($\alpha \in \mathbb{N}$ is shown in the next proposition, $\alpha \notin \mathbb{N}$ is a conjecture).

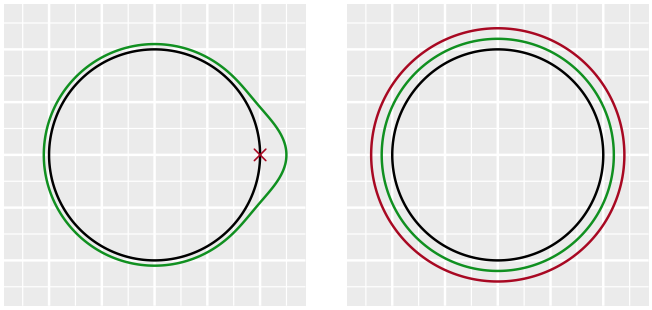

Figure 3.10: The intrinsic mean (red) can easily be located if the density (green) has a clear spike. For a uniform distribution of Z, every value $q \in (-\pi, \pi]$ minimizes $\mathbb{E}[d(Z, q)^2]$.

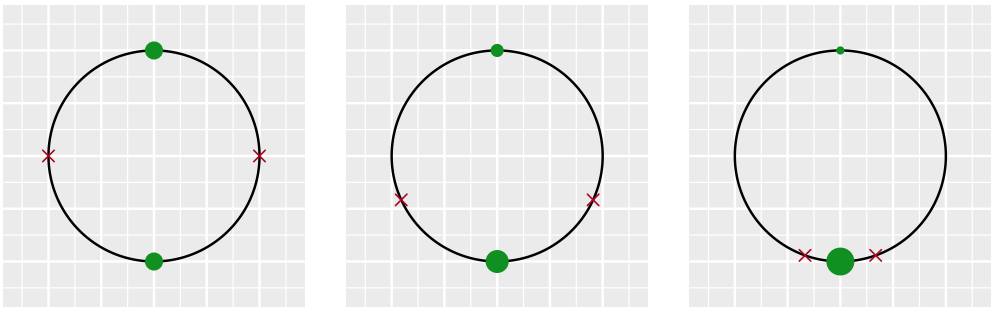

Figure 3.11: The distribution $\mathbb{P}_Z = (1-p)\delta_\pi + p\delta_0$, $p \in (0, 1)$ has two intrinsic means.

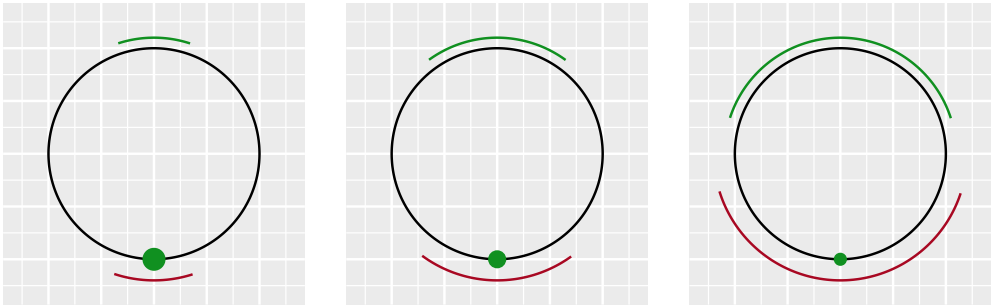

Figure 3.12: For the distribution $\mathbb{P}_Z = (1-p)\delta_\pi + p\mathsf{Unif}([-\frac{1}{2}p, \frac{1}{2}p])$, all θ with $d(\theta, \pi) \leq \frac{1}{2}p$ are intrinsic means.

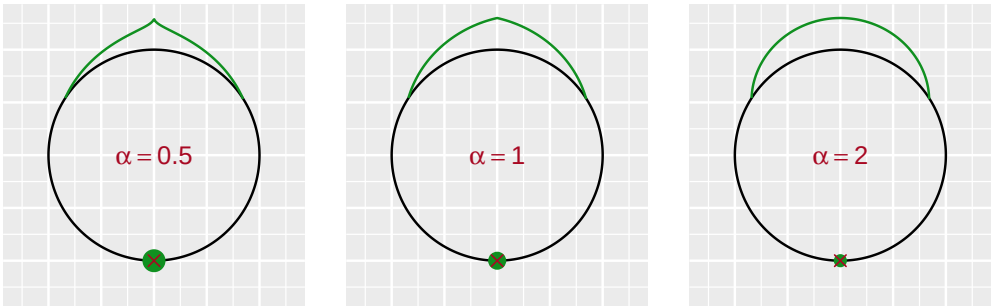

Figure 3.13: The distributions $\mathbb{P}_Z = (1-p)\delta_\pi + f(z)\mathrm{d}z$, $f(z) = (2\pi)^{-1}(1-|z|^\alpha)\mathbb{1}_{[-1,1]}(z)$, $\alpha > 0$, $p = \frac{1}{\pi}\frac{\alpha}{\alpha+1}$ have one single intrinsic mean $\theta = \pi$.

Proposition 3.18 ([HH15])**.** Assume $\theta = \pi$ is the unique intrinsic mean of Z and that the density f of Z is continuous on $\mathbb{S}^1$.

- If $f(0) < (2\pi)^{-1}$, then $n^{\frac{1}{2}}\theta_n \xrightarrow{d} \mathcal{N}(0, \tilde{\sigma}^2)$.
- If $f(0) = (2\pi)^{-1}$, and f is k-times continuously differentiable with $f^{(\ell)}(0) = 0$ for $\ell < k$ and $0 \neq |f^{(k)}(0)| < \infty$, then $n^{\frac{1}{2(k+1)}}|\theta_n - \theta| = \mathbf{O}_{\mathbb{P}}(1)$.

The proposition shows that the behavior of the density f of Z in neighborhood of 0, the antipodal point of the unique intrinsic mean π, is critical. These slower rates are also observed on more general manifolds, see [EH19].

3.4.5 Projected Mean and Conclusion

Finally, let us quickly summarize the results of section 3.2 and 3.3, and put them into the context of the previous rates of convergence.

With the projected mean, we interpolate between standard settings (Figure 3.14) and extreme cases (Figure 3.15) of nonunique and sticky means to obtain slow rates and fast rates of convergence, see Figure 3.16. This is exactly the same idea that drives all examples above to some extend.

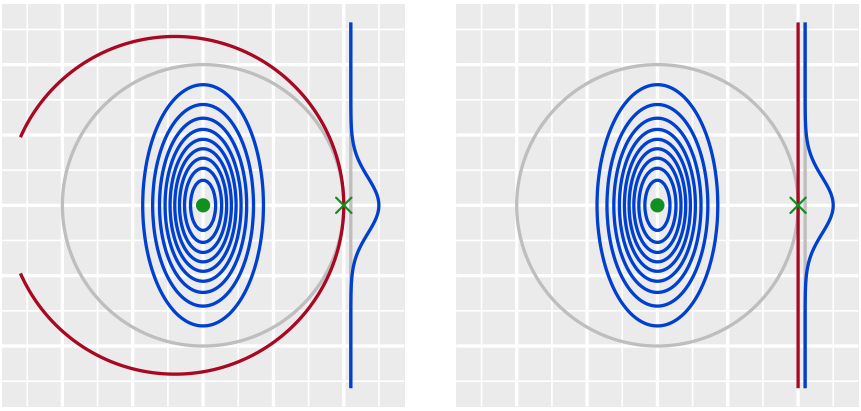

Figure 3.14: If $Q \subseteq \mathbb{R}^2$ (red) is $\mathcal{C}^2$ and points in neighborhood of origin have unique projection, then $n^{\frac{1}{2}}(\theta_n - \theta) \to \mathcal{N}(0, \tilde{\Sigma})$. The green dot is the Euclidean population mean. The distribution of the Euclidean sample mean is indicated by blue ellipses. The green cross is the projected population mean and the distribution of its sample version is indicated by the density on the right.

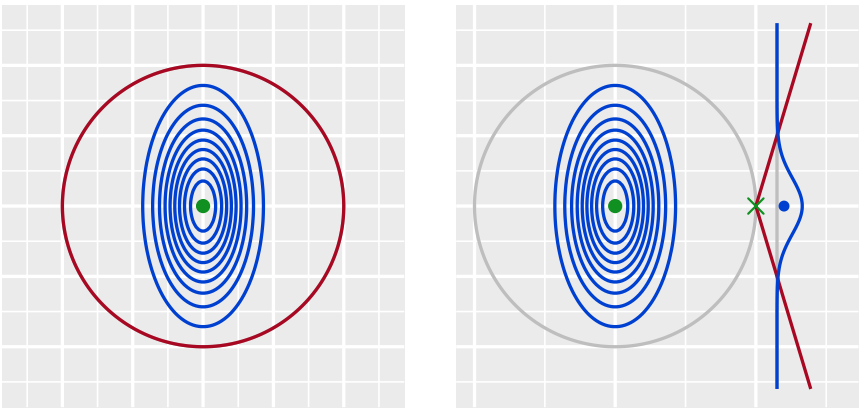

Figure 3.15: The extreme cases. On a circle (left), the projection of its center is not unique. For $a > 0$, $Q = \{(1 + a|y|, y) \colon y \in \mathbb{R}\}$ (right) the projected mean is sticky.

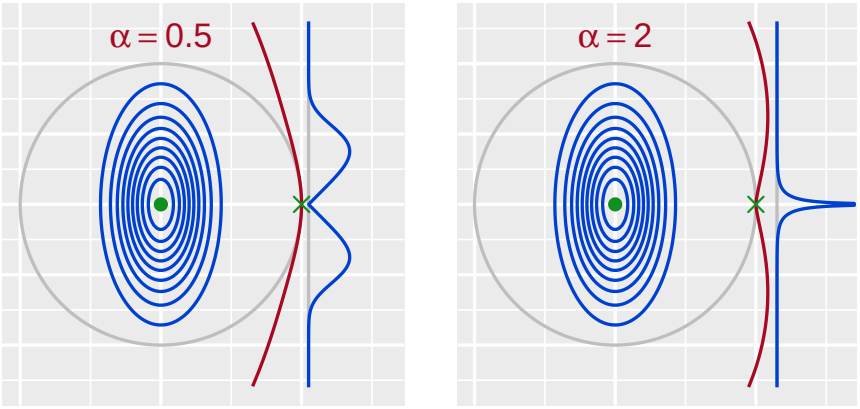

Figure 3.16: For $\alpha > 0$, $r(t) = 1 + |t|^{\frac{1+\alpha}{\alpha}}$, $Q = \{r(t)\exp(it) \colon t \in [-\pi, \pi]\}$ we get $n^{\frac{\alpha}{2}}(\theta_n - \theta) \xrightarrow{d} (0, Y_\infty^\alpha)$ where Y_∞^α is nondegenerated.

Appendix of Chapter 3

3.A Proofs

3.A.1 Lemma 3.1

Due to symmetry, we can restrict our analysis to $y \geq 0$ and $t \geq 0$ without loss of generality. To find the projection point, we have to minimize the squared distance $\ell \in \mathcal{C}^1([-B, B])$,

$$\ell(t) := \left\| q(t) - \begin{pmatrix} x \\ y \end{pmatrix} \right\|^2 .$$

For its derivative, we have

$$\frac{1}{2}\dot{\ell}(t) = r(t)\dot{r}(t) - x\left(\cos(t)\dot{r}(t) - \sin(t)r(t)\right) - y\left(\sin(t)\dot{r}(t) + \cos(t)r(t)\right) .$$

For $t \to 0$,

$$\begin{aligned} r(t) &= 1 + \mathbf{O}(tg(t)) , \\ \dot{r}(t) &= g(t) , \\ \sin(t) &= t + \mathbf{O}(t^3) , \\ \cos(t) &= 1 + \mathbf{O}(t^2) . \end{aligned}$$

Thus,

$$\begin{aligned} \cos(t)\dot{r}(t) - \sin(t)r(t) &= \mathbf{O}(g(t) + t) , \\ \sin(t)\dot{r}(t) + \cos(t)r(t) &= 1 + \mathbf{O}\Big(tg(t) + t^2\Big) , \\ r(t)\dot{r}(t) &= g(t) + \mathbf{O}\Big(tg(t)^2\Big) . \end{aligned}$$

Denote by t_y a global minimizer of $\ell(t)$. As $r(t)$ is strictly increasing for $t \geq 0$, we have $t_y \to 0$ as $x, y \to 0$.

Let $y \searrow 0$. From $\dot{\ell}(t_y) = 0$ with $x = \mathbf{O}(y)$, we obtain

$$0 = g(t_y) + \mathbf{O}(t_y g(t_y)^2) - y(1 + \mathbf{O}(g(t_y) + t_y)) ,$$

and in the setting of $x = 0$, we have

$$0 = g(t_y) + \mathbf{O}(t_y g(t_y)^2) - y(1 + \mathbf{O}(t_y g(t_y) + t_y^2)) .$$

For $a, b, u \in \mathbb{R}$ with $|b| \leq \frac{1}{2}$, it holds

$$\left| \frac{u+a}{1+b} - u \right| \leq 2\,|a| + 2\,|ub| \ .$$

Applied to the equations above with $u = g(t_y)$, $a = \mathbf{O}(t_y g(t_y)^2)$, and $b = \mathbf{O}(g(t_y)+t_y) = \mathbf{o}(1)$, this yields

$$y = g(t_y) + \mathbf{O}(g(t_y)^2 + t_y g(t_y))$$

for $x = \mathbf{O}(y)$, and for $x = 0$ with $b = \mathbf{O}(t_y g(t_y) + t_y^2)$,

$$y = g(t_y) + \mathbf{O}(t_y g(t_y)^2 + t_y^2 g(t_y))\,.$$

In particular, we always have

$$y = g(t_y) + \mathbf{o}(g(t_y))\,,$$

which implies

$$g(t_y) = y + \mathbf{o}(y)\,.$$

3.A.2 Proposition 3.3

Because of symmetry we can restrict our analysis to $y \geq 0$ and $t \geq 0$ without loss of generality. In the proof of Lemma 3.1, we have shown

$$y = g(t_y) + \mathbf{O}(g(t_y)^2 + t_y g(t_y))$$

for $x = \mathbf{O}(y)$, and for $x = 0$,

$$y = g(t_y) + \mathbf{O}(t_y g(t_y)^2 + t_y^2 g(t_y))\,.$$

Then, with $s := g(t_y)$ and $t_y = f(s)$, we have

$$\frac{t_y - f(y)}{f(y)} = \frac{f(s)}{f(s + \mathbf{O}(s^2 + s f(s)))} - 1 = \mathbf{o}(1)$$

by **(A1)** in the case of $x = \mathbf{O}(y)$, and by **(A1)'** in the case of $x = 0$,

$$\frac{t_y - f(y)}{f(y)} = \frac{f(s)}{f(s + \mathbf{O}(s^2 f(s) + s f(s)^2))} - 1 = \mathbf{o}(1)\,.$$

Hence, in both cases we get

$$t_y = f(y) + \mathbf{o}(f(y))\,.$$

Furthermore, for $t \searrow 0$,

$$q(t) = \begin{pmatrix} 1 \\ t \end{pmatrix} + \mathbf{o}(t)$$

and, thus,

$$\Pi_{\mathcal{Q}}\left(\begin{pmatrix} x \\ y \end{pmatrix} \right) = q(t_y) = \begin{pmatrix} 1 \\ f(y) \end{pmatrix} + \mathbf{o}(f(y))\,.$$

3.A.3 Theorem 3.5

Note that $\arg\min_{p\in\mathcal{Q}} \mathbb{E}[\|Z-p\|^2] = \arg\min_{p\in\mathcal{Q}} \|\mathbb{E}[Z]-p\|$, as $\mathbb{E}[\|Z-p\|^2] = \|\mathbb{E}[Z]-p\|^2 - \|\mathbb{E}[Z]\|^2 + \mathbb{E}[\|Z\|^2]$. As $\mathbb{E}[Z]=0$, $r(0)=1$, and $r(t)>1$ for $t>0$, the projected mean m of Z is unique and equal to $q(0)$.

Let $(\bar{X}_n, \bar{Y}_n)' := \bar{Z}_n = \frac{1}{n}\sum_{i=1}^n Z_i$. Fix $s \geq 0$. Our goal is to show

$$\mathbb{P}\Big(t_n \leq f\Big(\frac{s}{\sqrt{n}}\Big)\Big) \to \Phi\Big(\frac{s}{\sigma}\Big)\,. \tag{3.3}$$

For $L, \delta > 0$ define the following events,

$$\begin{aligned} A_{n,L} &:= \Big\{|\bar{X}_n| \leq L|\bar{Y}_n|\Big\}\,,\\ B_{n,s} &:= \Big\{t_n \leq f\Big(\frac{s}{\sqrt{n}}\Big)\Big\}\,,\\ C_{n,s} &:= \Big\{\sqrt{n}\bar{Y}_n + \Delta_n \leq s\Big\}\,,\\ D_{n,s,\delta} &:= \Big\{\sqrt{n}\bar{Y}_n \leq s(1+\delta)\Big\}\,, \end{aligned}$$

where $\Delta_n := \sqrt{n}\Big(g(t_n) - \bar{Y}_n\Big)$. Fix $\epsilon > 0$. We show (3.3) by proving $\big|\mathbb{P}(B_{n,s}) - \Phi(\frac{s}{\sigma})\big| < 5\epsilon$ for n large enough. We achieve this by splitting the left hand side into five parts by means of the triangle inequality and bound each summand by ϵ:

(i) By the central limit theorem for $(\bar{X}_n, \bar{Y}_n)'$, with $\mathbb{V}[Y] = \sigma^2 > 0$, there is $L > 0$ and $n_1 \in \mathbb{N}$ such that $\mathbb{P}(A_{n,L}^{\mathsf{c}}) < \epsilon$ for all $n > n_1$. Thus, $|\mathbb{P}(B_{n,s}) - \mathbb{P}(B_{n,s} \cap A_{n,L})| < \epsilon$.

(ii) Choose $\delta > 0$ such that $\Big|\Phi\Big(\frac{s}{\sigma(1+\delta)}\Big) - \Phi\big(\frac{s}{\sigma}\big)\Big| + \Big|\Phi\Big(\frac{s}{\sigma(1-\delta)}\Big) - \Phi\big(\frac{s}{\sigma}\big)\Big| < \epsilon$.

(iii) By Lemma 3.1, on the event $A_{n,L}$ for $\bar{Y}_n$ small enough, $g(t_n) = \bar{Y}_n + \mathbf{o}(\bar{Y}_n)$. Thus, there is $n_2 \in \mathbb{N}$ such that $\mathbb{P}(\{|\Delta_n| > \sqrt{n}\delta|\bar{Y}_n|\} \cap A_{n,L}) \leq \epsilon$ for all $n > n_2$. Therefore, $\mathbb{P}(D_{n,s,-\delta} \cap A_{n,L}) - \epsilon < \mathbb{P}(C_{n,s} \cap A_{n,L}) < \mathbb{P}(D_{n,s,\delta} \cap A_{n,L}) + \epsilon$.

(iv) As in (i), $|\mathbb{P}(D_{n,s,\pm\delta}) - \mathbb{P}(D_{n,s,\pm\delta} \cap A_{n,L})| < \epsilon$ for all $n > n_1$.

(v) By the central limit theorem, there is $n_3 \in \mathbb{N}$ such that $\Big|\mathbb{P}(D_{n,s,\pm\delta}) - \Phi\Big(\frac{s}{\sigma(1\pm\delta)}\Big)\Big| < \epsilon$ for all $n > n_3$.

As $B_{n,s} = C_{n,s}$, trivially $\mathbb{P}(B_{n,s} \cap A_{n,L}) = \mathbb{P}(C_{n,s} \cap A_{n,L})$. All points above together yield $\big|\mathbb{P}(B_{n,s}) - \Phi(\frac{s}{\sigma})\big| < 5\epsilon$ for all $n > \max(n_1, n_2, n_3)$. Hence, we have shown (3.3). As

$$\begin{pmatrix} m_{n,1} \\ m_{n,2} \end{pmatrix} = m_n = q(t_n) = \begin{pmatrix} 1 \\ t_n \end{pmatrix} + \mathbf{o}(t_n)\,,$$

equation (3.3) implies

$$\begin{aligned} \mathbb{P}\Big(m_{n,2} \leq f\Big(\frac{s}{\sqrt{n}}\Big)\Big) &\to \Phi\Big(\frac{s}{\sigma}\Big)\,,\\ \mathbb{P}\Big(|m_{n,1} - 1| \geq f\Big(\frac{s}{\sqrt{n}}\Big)\Big) &\to 0\,. \end{aligned}$$

The results for $-t_n$ and $-m_{n,2}$ are due to symmetry.

3.A.4 Corollary 3.9

We only show the statements for t_n as the results for y_n, $-t_n$, $-y_n$ follow similarly. Denote $F(s) := \Phi(\frac{s}{\sigma})$ and let $s \geq 0$.

(i) It is easy to see that **(A0)** holds for $f(y) = y^\gamma$. Thus, by Theorem 3.5,

$$\mathbb{P}\Big(n^{\frac{\gamma}{2}} t_n \leq s\Big) = \mathbb{P}\left(t_n \leq f\left(\frac{s^{\frac{1}{\gamma}}}{\sqrt{n}}\right)\right) \xrightarrow{n\to\infty} F\Big(s^{\frac{1}{\gamma}}\Big) \,.$$

Furthermore, $r(t) = 1 + \int_0^t x^{\frac{1}{\gamma}} \mathrm{d}x = 1 + \frac{\gamma}{1+\gamma} t^{\frac{1+\gamma}{\gamma}}$.

(ii) It is easy to check **(A0)** for $f(y) = (-\log(y))^{-\gamma}$.

The inverse function of f is $g(x) = \exp\Big(-x^{-\frac{1}{\gamma}}\Big)$, which yields the expression for $r(t)$.

For $x \in \mathbb{R}$, $s \geq 0$, $n \in \mathbb{N}$, set

$$G_n(x) = \mathbb{P}\left(\left(\frac{1}{2}\log(n)\right)^\gamma t_n \leq x\right) \qquad \text{and} \qquad b_{s,n} = \left(\frac{\log(\sqrt{n})}{\log\left(\frac{\sqrt{n}}{s}\right)}\right)^\gamma .$$

Let $\epsilon > 0$. Let $s > 1$ large enough such that $F(s) > 1 - \epsilon$. As $b_{n,s} \xrightarrow{n\to\infty} 1$ from above and G_n is right-continuous, there is n_0 such that $|G(b_{n,s}) - G(1)| < \epsilon$ for all $n \geq n_0$. Furthermore, by Theorem 3.5, there is $n_1 \in \mathbb{N}$ such that

$$G_n(b_{s,n}) = \mathbb{P}\left(t_n \leq f\left(\frac{s}{\sqrt{n}}\right)\right) \geq F(s) - \epsilon$$

for all $n \geq n_1$. Thus,

$$G_n(1) \geq G_n(b_{s,n}) - \epsilon \geq F(s) - 2\epsilon \geq 1 - 3\epsilon$$

for all $n \geq \max(n_0, n_1)$. We can argue similarly for $\lim_{t\nearrow 1} G_n(t)$ and obtain the final result

$$G_n(t) \xrightarrow{n\to\infty} \begin{cases} F(0) = \frac{1}{2} & \text{for } 0 < t < 1\,, \\ \lim_{s\to\infty} F(s) = 1 & \text{for } t \geq 1\,. \end{cases}$$

Together with the symmetry of the distribution, this shows the convergence of $(\frac{1}{2}\log n)^\gamma t_n$ in distribution to a uniform distribution on $\{-1, 1\}$.

(iii) It is easy to check **(A0)** for $f(y) = \exp(-y^{-\gamma})$. The inverse function of f is $g(x) = (-\log(x))^{-\frac{1}{\gamma}}$, which yields the expression for $r(t)$.

Let $c, u > 0$. For $s \in (1, \infty)$ and n large enough, it holds $u \exp(-(\sqrt{n}/c)^\gamma) \leq \exp(-(\sqrt{n}/(cs))^\gamma) = f(csn^{-\frac{1}{2}})$. Thus, with $U_{n,c} := \exp((\sqrt{n}/c)^\gamma)\, t_n$,

$$\mathbb{P}(U_{n,c} \leq u) \leq \mathbb{P}\left(t_n \leq f\left(\frac{cs}{\sqrt{n}}\right)\right) \xrightarrow{n\to\infty} F(cs)$$

by Theorem 3.5. Similarly, for $s \in (0,1)$,

$$\mathbb{P}(U_{n,c} \leq u) \geq \mathbb{P}\Big(t_n \leq f\Big(\frac{cs}{\sqrt{n}}\Big)\Big) \xrightarrow{n\to\infty} F(cs)\,.$$

Thus,

$$\mathbb{P}(U_{n,c} \leq u) \xrightarrow{n\to\infty} F(c) =: p_c\,,$$

which implies $\mathbb{P}(U_{n,c} \geq u) \xrightarrow{n\to\infty} 1 - p_c$. As t_n is symmetric, $\mathbb{P}(U_{n,c} \leq -u) \xrightarrow{n\to\infty} 1 - p_c$, which leaves $\mathbb{P}(|U_{n,c}| < u) \xrightarrow{n\to\infty} 2p_c - 1$.

3.A.5 Remark 3.12

(i) It is easy to see that **(A1)** hold for $f(y) = y^\gamma$.

(ii) To verify **(A1)** for $f(y) = (-\log(y))^{-\gamma}$, note

$$\lim_{y\to 0} \frac{\log(y)}{\log(y + h(y))} = \lim_{y\to 0} \frac{y + yh'(y)}{y + h(y)} = 1$$

by L'Hôpital's rule for a continuously differentiable function h with $h(y) = \mathbf{o}(y)$ and $h'(y) = \mathbf{o}(y)$. Here we use $h(y) = cy\left(y + \log(\frac{1}{y})^{-\gamma}\right)$.

(iii) To verify **(A1)'** for $f(y) = \exp(-y^{-\gamma})$, note

$$\frac{\exp(-(y + h(y))^{-a})}{\exp(-y^{-a})} = \exp\big(y^{-a} - (y + h(y))^{-a}\big) \xrightarrow{y\to 0} 1$$

for $a > 0$ and $h(y) = \mathbf{o}(y^2)$, as

$$y^{-a} - (y + h(y))^{-a} \to 0\,.$$

Here, we use $h(y) = c\left(y^2 \exp(-y^{-\gamma}) + y \exp(-2y^{-\gamma})\right)$.

4 Rates of Convergence via the Quadruple Inequality

Contents

4.1 Introduction

After having established consistency in a general setting in chapter 2 and rates of convergence in a specific setting in chapter 3, we now want to investigate how rates of convergence for FMs can be established in a general setting.

Recall the setting of the generalized FM from section 1.3.4: Let $\mathcal{Q}, \mathcal{Y}$ be sets, Y a $\mathcal{Y}$-valued random variable, and $\mathfrak{c}\colon \mathcal{Y} \times \mathcal{Q} \to \mathbb{R}$ a cost function. Every element m of the set $\arg\min_{q \in \mathcal{Q}} \mathbb{E}[\mathfrak{c}(Y, q)]$ is a generalized FM or $\mathfrak{c}$-FM. Given independent copies $Y_1, \ldots, Y_n$ of Y, natural estimators of the generalized FM are elements m_n of the set $\arg\min_{q \in \mathcal{Q}} \frac{1}{n} \sum_{i=1}^{n} \mathfrak{c}(Y_i, q)$. Our goal is to find suitable conditions for establishing finite sample bounds and convergence rates for such plug-in estimators.

We are particularly interested in finite sample bounds in expectation, i.e., bounds on $\mathbb{E}[\mathfrak{l}(m, m_n)]$, where $\mathfrak{l}$ is a loss function, e.g., $\mathfrak{l} = d^2$ if $(\mathcal{Q}, d)$ is a metric space, as these are stronger statements than bounds in probability and seem very natural for Euclidean means:

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - \mathbb{E}[Y]\right)^2\right] = \frac{1}{n}\mathbb{E}\Big[(Y - \mathbb{E}[Y])^2\Big]$$

for $Y, Y_1, \ldots, Y_n$ independent and identically distributed real-valued random variables with $\mathbb{E}[Y^2] < \infty$. Results on convergence rates in expectation seem to be rare in the literature on the FM. Common are convergence rates in probability or exponential concentration. The latter also implies rates in expectation, but under rather strong assumptions. One publication that establishes rates in expectation more directly, for general cost functions in Euclidean spaces is [BFW19].

The FM estimator is a *M-estimator*. Thus, we can build upon many classical and deep results from the M-estimation literature, see, e.g., [VW96; Gee00; Tal14].

4.1.1 Our Contribution

Our contribution consists of three parts:

(a) We introduce a condition, which we call *quadruple inequality*, that is used to establish finite sample bounds and convergence rates in probability and expectation for spaces with infinite diameter, see Theorem 4.1, Theorem 4.5, and Theorem 4.7.

(b) We formulate our results in the setting of the generalized FM with a cost-function $\mathfrak{c}$ that is not restricted to being the square of a metric.

(c) We prove a quadruple inequality for powers of metrics of Hadamard spaces, Theorem 4.10. We apply it to obtain finite sample bounds and rates of convergence for estimators of power FMs.

[PM19a] and [ALP20] show rates of convergence and finite sample bounds for metric spaces which have a finite diameter (or at least the support of the distribution of observations must be bounded). The proofs in both papers rely on *empirical process theory*. In particular, they make use of *symmetrization* and the *generic chaining* to bound the supremum of an empirical process. But where [ALP20] use that bound to be able to apply *Talagrand's inequality* [Bou02], [PM19a] employ a *peeling device* (also called *slicing*; see, e.g., [Gee00]) to obtain rates. As a consequence, [ALP20] achieve stronger results (nonasymptotic exponential concentration instead of $\mathbf{O}_{\mathbb{P}}$-statements), but they rely more heavily on the boundedness of the metric. As our goal is to obtain results for spaces with infinite diameter, our proof technique is closer to [PM19a], i.e., we also apply a peeling device.

A law of large numbers (see chapter 2), such that the estimator of the Fréchet mean converges in probability to the true value, implies that the estimator eventually is in a subset with finite diameter. Thus, for asymptotic rates in probability as in [PM19a], it is not very restrictive to assume a finite diameter. Our motivation to directly deal with infinite diameter comes from our interest in nonasymptotic results and in bounds in expectation.

Similar to [PM19a] and [ALP20], we use the *generic chaining*. Therefore we have entropy bounds as conditions of our theorems. These entropy bounds can be stated by requiring a bound on the *covering numbers*

$$N(Q,d,r) := \min\left\{ k \in \mathbb{N} \,\middle|\, \exists q_1,\dots,q_k \in \mathcal{Q}\colon Q \subseteq \bigcup_{j=1}^{k} \mathrm{B}_r(q_j) \right\},$$

where $(\mathcal{Q},d)$ is a metric space, $Q \subseteq \mathcal{Q}$, and $r > 0$. To be more precise, in a metric space $(\mathcal{Q},d)$, we require $\log N(\mathrm{B}_\delta(m),d,r) \leq \left(\frac{C\delta}{r}\right)^D$ for some constants $C, D > 0$ and all $0 < r < \delta$, which is the same assumption as in [ALP20]. We note, that this requirement could be weakened by using the optimal bound on Rademacher (or Bernoulli) processes [BL14] at the cost of a more complicated and less comprehensible condition.

In the classical Fréchet mean case, where $(\mathcal{Q},d)$ is a metric space and the cost function is $\mathfrak{c} = d^2$, the empirical process that has to be bounded consists of functions of the form $y \mapsto d(y,q)^2$ for $q \in \mathcal{Q}$. To apply some classical empirical process results, one requires a Lipschitz condition on these functions. In [PM19a] and [ALP20] this *Lipschitz condition* is fulfilled by

$$d(y,q)^2 - d(y,p)^2 \leq 2\,\mathsf{diam}(\mathcal{Q})d(q,p) \tag{4.1}$$

for all $y,q,p \in \mathcal{Q}$. Thus, a finite diameter is required. We show, that one can instead require that

$$d(y,q)^2 - d(y,p)^2 - d(z,q)^2 + d(z,p)^2 \leq 2d(y,z)d(q,p) \tag{4.2}$$

holds for all $y, z, q, p \in \mathcal{Q}$ and then bound the supremum of the empirical process even if $\mathsf{diam}(\mathcal{Q}) = \infty$. Equation (4.2) is a special instance of what we call *quadruple inequality*.

Roughly speaking, the transition from Lipschitz to quadruple condition removes certain squared terms and the right hand side by adding and subtracting further squared terms on the left hand side. This is related to the idea of defining the Fréchet mean as minimizer of $q \mapsto \mathbb{E}[d(Y,q)^2 - d(Y,o)^2]$ for an arbitrary fixed point $o \in \mathcal{Q}$ instead of $q \mapsto \mathbb{E}[d(Y,q)^2]$. Then, for existence of the Fréchet mean, only a first moment condition on Y is required instead of a second moment condition, see [Stu03, Acknowledgement to Lutz Mattner].

The inequality (4.2) does not hold in every metric space. But it characterizes Hadamard spaces among geodesic metric spaces, see [BN08]. In Hadamard spaces, (4.2) is known as *Reshetnyak's quadruple inequality* [Stu03] or *quadrilateral inequality* [BN08] and can be interpreted as generalization of the Cauchy–Schwarz inequality to metric spaces [BN08]. Note that our results are not restricted to geodesic metric spaces.

In (subsets of) Hadamard spaces $(\mathcal{Q}, d)$, we can not only utilize the quadruple inequality with the squared metric d^2 (4.2). But we show that for d^α with $\alpha \in [1,2]$, we also obtain a version of the quadruple inequality, namely

$$d(y,q)^\alpha - d(y,p)^\alpha - d(z,q)^\alpha + d(z,p)^\alpha \leq 4\alpha 2^{-\alpha} d(y,z)^{\alpha-1} d(q,p)\,, \tag{4.3}$$

for all $y, z, q, p \in \mathcal{Q}$, see Theorem 4.10. We show that the constant $4\alpha 2^{-\alpha}$ is optimal. Similar to equation (4.1), one can easily show – using the mean value theorem – that

$$d(y,q)^\alpha - d(y,p)^\alpha \leq \alpha\, \mathsf{diam}(\mathcal{Q})^{\alpha-1} d(q,p)$$

for $\alpha > 0$, $q, p, y \in \mathcal{Q}$, where $(\mathcal{Q}, d)$ is an arbitrary metric space. The proof of equation (4.3) is much more complicated, see appendix 4.G.

We state our convergence rate results in a general way, where observations live in a space $\mathcal{Y}$ and a cost function $\mathfrak{c}\colon \mathcal{Y} \times \mathcal{Q} \to \mathbb{R}$ is minimized over $\mathcal{Q}$. The quadruple inequality then reads

$$\mathfrak{c}(y,q) - \mathfrak{c}(y,p) - \mathfrak{c}(z,q) + \mathfrak{c}(z,p) \leq \mathfrak{a}(y,z)\mathfrak{b}(q,p)$$

for all $y, z \in \mathcal{Y}$ and $q, p \in \mathcal{Q}$ and an arbitrary function $\mathfrak{a}\colon \mathcal{Y} \times \mathcal{Y} \to [0,\infty)$ and a pseudo-metric $\mathfrak{b}\colon \mathcal{Q} \times \mathcal{Q} \to [0,\infty)$. This general formulation includes, among others, arbitrary bounded metric spaces, Hadamard spaces (including Euclidean and non-Euclidean spaces) with a power metric d^α, $\alpha \in [1,2]$, and regression settings with $\mathcal{Q} \neq \mathcal{Y}$, where observations $(x,y) \in \mathcal{Y}$ are described by regression functions $(x \mapsto q(x)) \in \mathcal{Q}$.

Furthermore, some trivial statements in appendix 4.B show that the quadruple inequality is stable under many operations such as taking subsets, limits, or product spaces.

We prove – via a peeling device – nonasymptotic finite sample bounds in probability, Theorem 4.1. We do not achieve exponential concentration as [ALP20], but our results can be applied in cases where the cost function is not bounded by a finite constant, i.e., in metric spaces with infinite diameter. Furthermore, we show two ways of obtaining bounds in expectation: One – nonasymptotic – under the assumption of a stronger

version quadruple inequality, Theorem 4.5; the other – asymptotic – with a stricter entropy condition but a weak quadruple inequality, Theorem 4.7.

Aside from the application in Hadamard spaces (including the use of the power inequality, Theorem 4.10), we illustrate our results in different toy examples: Euclidean spaces and infinite dimensional Hilbert spaces. In (convex subsets of) Hilbert spaces the Fréchet mean is equal to the expectation. Thus, these examples are interesting as a benchmark, because we can compare results from our general Fréchet mean approach to exact results. In two additional examples, we apply our results to nonconvex subsets of Hilbert spaces and to Hadamard spaces.

4.1.2 Outline

We start by presenting the finite sample bounds of Theorem 4.1 (bounds in probability) and Theorem 4.5 (bounds in expectation) in the abstract setting in section 4.2. The different versions of the quadruple inequality are discussed in section 4.3. This part concludes with the statement of Theorem 4.7 (alternative route to rates in expectation). In section 4.4, we apply the abstract results in different settings: Euclidean spaces, infinite dimensional Hilbert spaces, nonconvex sets, and Hadamard spaces. Finite sample bounds and rates of convergence for power Hadamard metrics and the power inequality, Theorem 4.10, are presented in section 4.5.

4.2 Abstract Results

In this section, we prove finite sample bounds for the Fréchet mean in a very general setting, see section 4.2.1. For bounds in probability Theorem 4.1 is stated in section 4.2.2 and for bounds in expectation Theorem 4.5 is stated in section 4.2.3. The proofs can be found in appendix 4.A. Some remarks on further extensions are given in section 4.2.4.

4.2.1 Setting

Here we define an *Abstract Setting* in which we will state our most general results. This setting of the generalized Fréchet mean is similar to what is used in [Huc11; EH19] and section 1.3.4.

Let $\mathcal{Q}$ be a set, which is called *descriptor space*. Let $(\mathcal{Y}, \Sigma_{\mathcal{Y}})$ be a measurable space, which is called *data space*. Let Y be a $\mathcal{Y}$-valued random variable. Let $\mathfrak{c}\colon \mathcal{Y} \times \mathcal{Q} \to \mathbb{R}$ be a function such that $y \to \mathfrak{c}(y, q)$ is measurable for every $q \in \mathcal{Q}$. We call $\mathfrak{c}$ *cost function*. Define $F\colon \mathcal{Q} \to \mathbb{R}, q \mapsto \mathbb{E}[\mathfrak{c}(Y, q)]$, assuming that $\mathbb{E}[|\mathfrak{c}(Y, q)|] < \infty$ for all $q \in \mathcal{Q}$. The function F is called *objective function*. Let $n \in \mathbb{N}$. Let $Y_1, \dots, Y_n$ be independent copies of Y. Define $F_n\colon \mathcal{Q} \to \mathbb{R}, q \mapsto \frac{1}{n}\sum_{i=1}^n \mathfrak{c}(Y_i, q)$. We call F_n *empirical objective function*. Let $\mathfrak{l}\colon \mathcal{Q} \times \mathcal{Q} \to [0, \infty)$ be a function such that $\mathfrak{l}(m, q)$ measures the *loss* of choosing q given that the true value is m.

We want to bound $\mathfrak{l}(m, m_n)$ for $m \in \arg\min_{q\in\mathcal{Q}} F(q)$ and $m_n \in \arg\min_{q\in\mathcal{Q}} F_n(q)$.

4.2.2 Finite Sample Bounds in Probability

For our result on finite sample bounds in probability, we make some assumptions, which are listed in the following. We denote the "closed" ball with center $o \in \mathcal{Q}$ of radius $r > 0$ in the set $\mathcal{Q}$ with respect to an arbitrary distance function $d\colon \mathcal{Q} \times \mathcal{Q} \to [0, \infty)$ as $\mathrm{B}_r(o, d) := \{q \in \mathcal{Q}\colon d(o, q) \leq r\}$.

Assumptions.

EXISTENCE:
It holds $\mathbb{E}[|\mathfrak{c}(Y, q)|] < \infty$ for all $q \in \mathcal{Q}$. There are $m_n \in \arg\min_{q \in \mathcal{Q}} F_n(q)$ measurable and $m \in \arg\min_{q \in \mathcal{Q}} F(q)$.

GROWTH:
There are constants $\gamma > 0$ and $c_{\mathfrak{g}} > 0$ such that $F(q) - F(m) \geq c_{\mathfrak{g}} \mathfrak{l}(m, q)^\gamma$ for all $q \in \mathcal{Q}$.

WEAK QUADRUPLE:
There are a function $\mathfrak{a}\colon \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ measurable and a pseudo-metric $\mathfrak{b}\colon \mathcal{Q} \times \mathcal{Q} \to [0, \infty)$, such that, for all $p, q \in \mathcal{Q}$, $y, z \in \mathcal{Y}$, it holds

$$\mathfrak{c}\overline{yq} - \mathfrak{c}\overline{zq} - \mathfrak{c}\overline{yp} + \mathfrak{c}\overline{zp} \;\leq\; \mathfrak{a}(y, z)\, \mathfrak{b}(q, p)\,,$$

where we use the notation $\mathfrak{c}\overline{yq} := \mathfrak{c}(y, q)$. We call $\mathfrak{a}$ the *data distance* and $\mathfrak{b}$ the *descriptor metric*.

MOMENT:
Let $\zeta \geq 1$. Set

$$\mathfrak{M}(\zeta) := \begin{cases} \mathbb{E}\big[\mathfrak{a}(Y', Y)^\zeta\big]\,, & \text{if } \zeta \geq 2\,, \\ \mathbb{E}\big[\mathfrak{a}(Y', Y)^2\big]^{\frac{\zeta}{2}}\,, & \text{if } \zeta \leq 2\,, \end{cases}$$

where Y' is an independent copy of Y. It holds $\mathfrak{M}(\zeta) < \infty$.

ENTROPY:
There are $\alpha, \beta > 0$ with $\frac{\alpha}{\beta} < \gamma$ such that

$$\sqrt{\log N(\mathrm{B}_\delta(m, \mathfrak{l}), \mathfrak{b}, r)} \leq c_{\mathfrak{e}} \frac{\delta^\alpha}{r^\beta}$$

for a constant $c_{\mathfrak{e}} > 0$ and all $\delta, r > 0$.

Here

$$N(A, \mathfrak{b}, r) = \min\left\{ k \in \mathbb{N} \;\middle|\; \exists q_1, \ldots, q_k \in \mathcal{Q}\colon A \subseteq \bigcup_{j=1}^{k} \mathrm{B}_r(q_j, \mathfrak{b}) \right\},$$

is the *covering number* of $A \subseteq \mathcal{Q}$ with respect to $\mathfrak{b}$-balls $\mathrm{B}_r(\cdot, \mathfrak{b})$ of radius r. ENTROPY is essentially the same condition as in [ALP20], but written down for the setting of the

generalized Fréchet mean instead of the classical Fréchet mean in metric spaces.

We shortly discuss other assumptions before stating the theorem for finite sample bounds in probability. The measurability assumptions can be weakened by using the *outer expectation*, see [VW96]. In [BDG07], the GROWTH condition is called *margin condition*; in chapter 5 it is referred to as VARIANCE or *variance inequality*. It is called *low noise assumption* in [ALP20]. If GROWTH holds for every distribution of Y and we are in the traditional setting of the (not generalized) Fréchet mean, it implies that the metric space $\mathcal{Q}$ has nonpositive curvature: Assume that $(\mathcal{Q}, d)$ is a complete *geodesic space* [Stu03, Definition 1.1], i.e., every pair of points y_1, y_2 has a *mid-point* m, i.e., $\overline{y_1,m} = \overline{y_2,m} = \frac{1}{2}\overline{y_1,y_2}$, where we use the notation $\overline{q,p} := d(q,p)$. Set $\mathcal{Y} = \mathcal{Q}$, $\mathfrak{c} = d^2$, and $\mathfrak{l} = d$. If $\mathbb{P}(Y = y_1) = \mathbb{P}(Y = y_2) = \frac{1}{2}$ with $y_1, y_2 \in \mathcal{Q}$, the Fréchet mean $m \in \mathcal{Q}$ of Y is the mid-point between y_1 and y_2. If we assume that the growth condition holds for every distribution of Y, in particular, for every uniform 2-point distribution, with $c_{\mathfrak{g}} = 1$ and $\gamma = 2$, then

$$\frac{1}{2}\overline{y_1,q}^2 + \frac{1}{2}\overline{y_2,q}^2 - \frac{1}{2}\overline{y_1,m}^2 - \frac{1}{2}\overline{y_2,m}^2 \geq \overline{m,q}^2 \,.$$

As m is the mid-point between y_1 and y_2, we obtain

$$\overline{m,q}^2 \leq \frac{1}{2}\overline{y_1,q}^2 + \frac{1}{2}\overline{y_2,q}^2 - \frac{1}{4}\overline{y_1,y_2}^2 \,,$$

which implies nonpositive curvature of the space $(\mathcal{Q}, d)$, see [Stu03, Definition 2.1]. Such spaces are called *Hadamard spaces*. Aside from the GROWTH condition they also fulfill the quadruple inequality, which we discuss in section 4.3.2.3. The WEAK QUADRUPLE-condition will be discussed in detail in section 4.3. Among other things, we will show that it holds in a nice way in all Hadamard spaces, which include the Euclidean spaces.

The following theorem states finite sample bounds for the estimator m_n to the true value m measured with respect to the loss function $\mathfrak{l}$.

Theorem 4.1 (Finite samples bounds in probability)**.** In the *Abstract Setting* of section 4.2.1, assume that following conditions hold: EXISTENCE, GROWTH, WEAK QUADRUPLE, MOMENT, ENTROPY. Define

$$\eta_{\beta,n} := \begin{cases} n^{-\frac{1}{2}} & \text{for } \beta < 1 \,, \\ n^{-\frac{1}{2}} \log(n+1) & \text{for } \beta = 1 \,, \\ n^{-\frac{1}{2\beta}} & \text{for } \beta > 1 \,. \end{cases}$$

Then, for all $t > 0$, it holds

$$\mathbb{P}\left(\eta_{\beta,n}^{-\frac{1}{\gamma - \frac{\alpha}{\beta}}} \, \mathfrak{l}(m, m_n) \geq t \right) \leq c \, \mathfrak{M}(\zeta) \, t^{-\zeta(\gamma - \frac{\alpha}{\beta})}$$

where $c > 0$ depends on $\alpha, \beta, \gamma, c_{\mathfrak{e}}, c_{\mathfrak{g}}, \zeta$.

The proof can be found in appendix 4.A.

Without loss of generality, one can choose $\gamma = 1$ by using the loss $\mathfrak{l}' = \mathfrak{l}^\gamma$. This is consistent with the result: If Growth and Entropy are fulfilled with $\mathfrak{l}, \alpha, \beta, \gamma$, then they are also fulfilled with $\mathfrak{l}' = l^\gamma, \alpha' = \frac{\alpha}{\gamma}, \beta' = \beta, \gamma' = 1$, which gives the same result. We keep this redundancy in the parameters of the theorem for convenience.

A common way of stating rates of convergence in probability is the $\mathbf{O}_\mathbb{P}$-notation, as in the following corollary. Note that the $\mathbf{O}_\mathbb{P}$-result is asymptotic and, thus, weaker than the nonasymptotic Theorem 4.1.

Corollary 4.2. In the *Abstract Setting* of section 4.2.1, assume that following conditions hold: Existence, Weak Quadruple, Growth, Moment with $\zeta = 1$, Entropy. Then

$$\mathfrak{l}(m, m_n) = \mathbf{O}_\mathbb{P}\left(\eta_{\beta,n}^{\frac{1}{\gamma - \frac{\alpha}{\beta}}}\right)$$

with $\eta_{\beta,n}$ as in Theorem 4.1.

It is possible to weaken the assumptions in Corollary 4.2. In particular, we can restrict the Growth and Entropy conditions to hold only in a neighborhood of m if we already know that $\mathfrak{l}(m_n, m) \in \mathbf{o}_\mathbb{P}(1)$.

In Theorem 4.1, the probability of large losses decays polynomially. If the exponent $\zeta(\gamma - \frac{\alpha}{\beta})$ is strictly greater than 1, we can integrate the tail probabilities to obtain a bound on the expectation of the loss.

Corollary 4.3. Let $\kappa \geq 1$. In the *Abstract Setting* of section 4.2.1, assume that following conditions hold: Existence, Weak Quadruple, Growth, Moment with $\zeta > \kappa(\gamma - \frac{\alpha}{\beta})^{-1}$, Entropy. Set $\xi := \zeta(\gamma - \frac{\alpha}{\beta})\kappa^{-1}$. Then

$$\eta_{\beta,n}^{-\frac{\kappa}{\gamma - \frac{\alpha}{\beta}}} \mathbb{E}[\mathfrak{l}(m, m_n)^\kappa] \leq c' \frac{\xi}{\xi - 1} \mathfrak{M}(\zeta)^{\frac{1}{\xi}}.$$

The proof can be found in appendix 4.A.

Corollary 4.3 may require unnecessarily high moments as ξ needs to be *strictly* larger than 1. In the next section, we present a more direct approach to finite sample bounds in expectation, that requires weaker moment conditions, at least in some settings.

4.2.3 Finite Sample Bounds in Expectation

For obtaining finite sample bounds in expectation directly, we need slightly modified, stronger assumptions.

Assumptions.

Strong Quadruple:
Define $\dot{\mathcal{Q}} := \mathcal{Q} \setminus \mathrm{B}_0(m, \mathfrak{l}) = \{q \in \mathcal{Q} \colon \mathfrak{l}(m, q) > 0\}$. There are functions $\mathfrak{b}_m \colon \dot{\mathcal{Q}} \times$

$\dot{\mathcal{Q}} \to [0,\infty)$ (possibly depending on m) and $\mathfrak{a}\colon \mathcal{Y}\times\mathcal{Y}\to[0,\infty)$ with $\mathfrak{a}$ measurable and $\xi \in (0,\gamma)$, such that, for all $p,q \in \dot{\mathcal{Q}}$, $y,z\in\mathcal{Y}$, it holds

$$\frac{\mathfrak{c}\overline{yq} - \mathfrak{c}\overline{ym} - \mathfrak{c}\overline{zq} + \mathfrak{c}\overline{zm}}{\mathfrak{l}(m,q)^{\xi}} - \frac{\mathfrak{c}\overline{yp} - \mathfrak{c}\overline{ym} - \mathfrak{c}\overline{zp} + \mathfrak{c}\overline{zm}}{\mathfrak{l}(m,p)^{\xi}} \le \mathfrak{a}(y,z)\,\mathfrak{b}_m(q,p)\,,$$

Assume that $\mathfrak{b}_m$ is a pseudo-metric on $\dot{\mathcal{Q}}$. We call $\mathfrak{a}$ the *data distance* and $\mathfrak{b}_m$ the *strong quadruple metric* at m.

Strong Moment:
For $\zeta > 0$, set

$$\mathfrak{M}(\zeta) := \begin{cases} \mathbb{E}\big[\mathfrak{a}(Y',Y)^{\zeta}\big]\,, & \text{if } \zeta \ge 2\,,\\ \mathbb{E}\big[\mathfrak{a}(Y',Y)^{2}\big]^{\frac{\zeta}{2}}\,, & \text{if } \zeta \le 2\,, \end{cases}$$

where Y' is an independent copy of Y. Let $\kappa \ge \gamma - \xi$ and assume $\mathfrak{M}\Big(\frac{\kappa}{\gamma-\xi}\Big) < \infty$.

Strong Entropy:
It holds $D := \mathsf{diam}(\dot{\mathcal{Q}}, \mathfrak{b}_m) < \infty$ and there is $\beta > 0$ such that

$$\sqrt{\log N(\dot{\mathcal{Q}}, \mathfrak{b}_m, r)} \le c_{\mathsf{e}} \left(\frac{D}{r}\right)^{\beta}$$

for all $r \in (0, D)$.

For later use in the application to Hilbert spaces, section 4.4.2, and for Theorem 4.5, we state the entropy part of Theorem 4.5 in a more general way than in Theorem 4.1. To this end, we need to introduce different measures of entropy.

Definition 4.4 (Measures of Entropy)**.**

(i) Given a set $\mathcal{Q}$ an *admissible sequence* is an increasing sequence $(\mathcal{A}_k)_{k\in\mathbb{N}_0}$ of partitions of $\mathcal{Q}$ such that $\mathcal{A}_0 = \mathcal{Q}$ and $\mathsf{card}(\mathcal{A}_k) \le 2^{2^k}$ for $k \ge 1$.

By an increasing sequence of partitions we mean that every set of $\mathcal{A}_{k+1}$ is contained in a set of $\mathcal{A}_k$. We denote by $A_k(q)$ the unique element of $\mathcal{A}_k$ which contains $q \in \mathcal{Q}$.

(ii) Let $(\mathcal{Q},\mathfrak{b})$ be a pseudo-metric space. Define

$$\gamma_2(\mathcal{Q},\mathfrak{b}) := \inf \sup_{q\in\mathcal{Q}} \sum_{k=0}^{\infty} 2^{\frac{k}{2}}\, \mathsf{diam}(A_k(q),\mathfrak{b})\,,$$

where the infimum is taken over all admissible sequences in $\mathcal{Q}$ and

$$\mathsf{diam}(A,\mathfrak{b}) := \sup_{q,p\in A} \mathfrak{b}(q,p)$$

for $A \subseteq \mathcal{Q}$.

(iii) Let $(Q, \mathfrak{b})$ be a pseudo-metric space and $n \in \mathbb{N}$. Define

$$\mathsf{entr}_n(Q, \mathfrak{b}) := \inf_{\epsilon > 0} \left(\epsilon \sqrt{n} + \int_\epsilon^\infty \sqrt{\log N(Q, \mathfrak{b}, r)} \mathrm{d}r \right) .$$

Items (i) and (ii) are basic definitions from [Tal14]. Item (iii) is just a convenient notation.

Theorem 4.5 (Finite sample bounds in expectation)**.** In the *Abstract Setting* of section 4.2.1, assume that following conditions hold: Existence, Growth, Strong Quadruple, Strong Moment. Then, it holds

$$\mathbb{E}[\mathfrak{l}(m, m_n)^\kappa] \;\leq\; c n^{-\frac{\kappa}{2(\gamma-\xi)}} \mathfrak{M}\left(\frac{\kappa}{\gamma - \xi}\right) \min(\mathsf{entr}_n(\mathcal{Q}, \mathfrak{b}_m), \gamma_2(\mathcal{Q}, \mathfrak{b}_m))^{\frac{\kappa}{\gamma-\xi}} \;,$$

where $c > 0$ depends only on $\kappa, \gamma, \xi, c_{\mathfrak{g}}$.

If additionally Strong Entropy holds, then

$$\mathbb{E}[\mathfrak{l}(m, m_n)^\kappa] \;\leq\; C \, \mathfrak{M}\left(\frac{\kappa}{\gamma - \xi}\right) D^{\frac{\kappa}{\gamma-\xi}} \, \eta_{\beta,n}^{\frac{\kappa}{\gamma-\xi}} \;,$$

where

$$\eta_{\beta,n} := \begin{cases} n^{-\frac{1}{2}} & \text{for } \beta < 1 \,, \\ n^{-\frac{1}{2}} \log(n+1) & \text{for } \beta = 1 \,, \\ n^{-\frac{1}{2\beta}} & \text{for } \beta > 1 \,, \end{cases}$$

and $C > 0$ depends only on $\kappa, \beta, \gamma, \xi, c_{\mathfrak{g}}$.

The proof can be found in appendix 4.A.

As in Theorem 4.1 the statement contains some redundancy. E.g., by using the loss $\tilde{\mathfrak{l}} = \mathfrak{l}^\xi$ we set $\xi = 1$ without loss of generality. Then the growth exponent and the resulting rate of convergence will scale accordingly.

4.2.4 Further Extensions

In general $M := \arg\min_{q \in \mathcal{Q}} \mathbb{E}[\mathfrak{c}(Y, q)]$ is some subset of $\mathcal{Q}$. One can also extend the main theorems of this paper to deal with a the whole set of Fréchet means and Fréchet mean estimators. To do that, the Growth condition has to be stated as growth of the minimal distance to M. Furthermore, some of the statements and assumptions made in the theorems and proofs have to be modified so that the hold uniformly over all $m \in M$. Additionally, one has to think about the right notion of convergence for sets. We found that those results hard to read without significantly increasing insight into the problem, which is why we chose to stick with unique Fréchet means and only remark that an extension to Fréchet mean sets is possible.

One can also consider ε- $\arg\min$-sets, i.e., the sets of elements which minimize a function up to an $\varepsilon > 0$. If one chooses $m_n \in \varepsilon_n\text{-}\arg\min_{q\in\mathcal{Q}} F_n(q)$ with $\varepsilon_n \to 0$ fast enough, the convergence rate is of the same as for the absolute minimizer.

4.3 Quadruple Inequalities

Recall the definition of the weak and strong quadruple inequalities. Let $(\mathcal{Q}, \mathfrak{b})$ be a pseudo-metric space (*descriptor space* with *descriptor metric*), $(\mathcal{Y}, \Sigma_\mathcal{Y})$ a measurable space (*data space*), $\mathfrak{c}\colon \mathcal{Y} \times \mathcal{Q} \to \mathbb{R}$ such that $y \mapsto \mathfrak{c}(y, q)$ is measurable for every $q \in \mathcal{Q}$ (*cost function*), $\mathfrak{a}\colon \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ measurable (*data distance*), $m \in \mathcal{Q}$ (reference point, usually the Fréchet mean), $\mathfrak{l}\colon \mathcal{Q} \times \mathcal{Q} \to [0, \infty)$ (*loss*), $\xi > 0$ (*rate parameter*), $\dot{\mathcal{Q}} = \{q \in \mathcal{Q}\colon \mathfrak{l}(m, q) > 0\}$ $\mathfrak{b}_m\colon \dot{\mathcal{Q}} \times \dot{\mathcal{Q}} \to [0, \infty)$ a pseudo-metric on $\dot{\mathcal{Q}}$ (*strong quadruple metric* at m). We write ${}^\mathfrak{c}\overline{yq} := \mathfrak{c}(y, q)$.

(a) The tuple $(\mathcal{Q}, \mathcal{Y}, \mathfrak{c}, \mathfrak{a}, \mathfrak{b})$ fulfills the *(weak) quadruple inequality* if and only if for all $p, q \in \mathcal{Q}$, $y, z \in \mathcal{Y}$ it holds

$$ {}^\mathfrak{c}\overline{yq} - {}^\mathfrak{c}\overline{zq} - {}^\mathfrak{c}\overline{yp} + {}^\mathfrak{c}\overline{zp} \leq \mathfrak{a}(y, z)\mathfrak{b}(q, p)\,. $$

(b) The tuple $(\mathcal{Q}, \mathcal{Y}, \mathfrak{c}, \mathfrak{l}^\xi, \mathfrak{a}, \mathfrak{b}_m)$ fulfills the *strong quadruple inequality* at $m \in \mathcal{Q}$ if and only if for all $p, q \in \dot{\mathcal{Q}}$, $y, z \in \mathcal{Y}$ it holds

$$ \frac{{}^\mathfrak{c}\overline{yq} - {}^\mathfrak{c}\overline{ym} - {}^\mathfrak{c}\overline{zq} + {}^\mathfrak{c}\overline{zm}}{\mathfrak{l}(m, q)^\xi} - \frac{{}^\mathfrak{c}\overline{yp} - {}^\mathfrak{c}\overline{ym} - {}^\mathfrak{c}\overline{zp} + {}^\mathfrak{c}\overline{zm}}{\mathfrak{l}(m, p)^\xi} \leq \mathfrak{a}(y, z)\mathfrak{b}_m(q, p)\,. $$

There are a couple of trivial stability results for quadruple inequalities, see appendix 4.B.

In section 4.3.1 we compare the quadruple inequality with a more common Lipschitz property. The simplest advantageous applications of the quadruple inequality are in inner product spaces and quasi-inner product spaces, as is discussed in section 4.3.2. We conclude with Theorem 4.7 in section 4.3.3, which yields rates of convergence in expectation under the assumption of only a weak quadruple inequality instead of a strong one as in Theorem 4.5.

4.3.1 Bounded Spaces and Smooth Cost Function

Let $(\mathcal{Q}, d)$ be a metric space and use the notation $\overline{q,p} = d(q, p)$. For obtaining convergence rates in probability for the Fréchet mean estimator, [PM19a] use

$$ \overline{y,q}^2 - \overline{y,p}^2 = (\overline{y,q} - \overline{y,p})\,(\overline{y,q} + \overline{y,p}) \leq 2\overline{q,p}\,\mathsf{diam}(\mathcal{Q}) $$

for all $q, p, y \in \mathcal{Q}$. In the proof of Theorem 4.1, we have replaced this bound by the weak quadruple inequality, i.e.,

$$ {}^\mathfrak{c}\overline{yq} - {}^\mathfrak{c}\overline{yp} - {}^\mathfrak{c}\overline{zq} + {}^\mathfrak{c}\overline{zp} \leq \mathfrak{a}(y, z)\mathfrak{b}(q, p)\,. $$

This generalizes the results by [PM19a] as for bounded metric spaces $(\mathcal{Q}, d)$ and cost function $\mathfrak{c} = d^2$, the weak quadruple inequality holds with $\mathfrak{a}(y, z) = 4\,\mathsf{diam}(\mathcal{Q})$ and $\mathfrak{b} = d$:

$$\overline{y,q}^2 - \overline{y,p}^2 - \overline{z,q}^2 + \overline{z,p}^2 \leq \left|\overline{y,q}^2 - \overline{y,p}^2\right| + \left|\overline{z,q}^2 - \overline{z,p}^2\right| \leq 4\overline{q,p}\,\mathsf{diam}(\mathcal{Q})\,.$$

More generally, if we can show Lipschitz continuity in the second argument of the cost function, i.e., ${}^{\mathfrak{c}}\overline{yq} - {}^{\mathfrak{c}}\overline{yp} \leq \mathfrak{a}(y)\mathfrak{b}(q, p)$, then the quadruple inequality holds with data distance $\mathfrak{a}(y) + \mathfrak{a}(z)$ and descriptor metric $\mathfrak{b}$. But this might lead to an unnecessarily large bound. We will see in section 4.3.2.3 that at least for certain metric spaces, we can find a bound via the quadruple inequality that does not involve the diameter of the space and, thus, allows for meaningful results in unbounded spaces.

4.3.2 Relation to Inner Product and Cauchy–Schwarz Inequality

4.3.2.1 Inner Product Space

Let $(\mathcal{Q}, d)$ be a metric space such that d comes from an inner product $\langle \cdot\,,\, \cdot \rangle$ on $\mathcal{Q}$, i.e., $\mathcal{Q}$ is a subset of am inner product space and $d(y, q)^2 = \langle y - q\,,\, y - q \rangle$. Use $\mathcal{Y} = \mathcal{Q}$ and the squared metric as cost function, $\mathfrak{c} = d^2$. Then

$$\begin{aligned} {}^{\mathfrak{c}}\overline{yq} - {}^{\mathfrak{c}}\overline{zq} - {}^{\mathfrak{c}}\overline{yp} + {}^{\mathfrak{c}}\overline{zp} &= -2\,\langle y - z\,,\, q - p \rangle \\ &\leq 2\|q - p\|\|y - z\|\,. \end{aligned}$$

Here the Cauchy–Schwarz inequality gives rise to an instance of the weak quadruple inequality. The very general framework that we impose also allows for a more flexible bound: If $\mathcal{Q} \subseteq \mathbb{H}$ is the subset of an infinite dimensional, separable Hilbert space $\mathbb{H}$, we can use a weighted Cauchy–Schwarz inequality: Let $s = (s_k)_{k\in\mathbb{N}} \subseteq (0, \infty)$. Then

$${}^{\mathfrak{c}}\overline{yq} - {}^{\mathfrak{c}}\overline{zq} - {}^{\mathfrak{c}}\overline{yp} + {}^{\mathfrak{c}}\overline{zp} \leq 2\|y - z\|_{s^{-1}}\|q - p\|_s\,,$$

where $\|x\|_s^2 = \sum_{k=1}^{\infty} s_k^2 x_k^2$ with generalized Fourier coefficients $(x_k)_{k\in\mathbb{N}}$ with respect to a fixed orthonormal basis of $\mathbb{H}$.

For the strong quadruple inequality, we set $\xi = 1$, $\mathfrak{l}(q, p) = \|q - p\|$ and obtain

$$\begin{aligned} &\frac{{}^{\mathfrak{c}}\overline{yq} - {}^{\mathfrak{c}}\overline{ym} - {}^{\mathfrak{c}}\overline{zq} + {}^{\mathfrak{c}}\overline{zm}}{\mathfrak{l}(m, q)} - \frac{{}^{\mathfrak{c}}\overline{yp} - {}^{\mathfrak{c}}\overline{ym} - {}^{\mathfrak{c}}\overline{zp} + {}^{\mathfrak{c}}\overline{zm}}{\mathfrak{l}(m, p)} \\ &= -2\left\langle y - z, \frac{q - m}{\|q - m\|} - \frac{p - m}{\|p - m\|} \right\rangle \\ &\leq 2\|y - z\| \left\| \frac{q - m}{\|q - m\|} - \frac{p - m}{\|p - m\|} \right\|. \end{aligned}$$

Thus, the strong quadruple inequality hold with $\mathfrak{a}(y, z) = 2\|y - z\|$ and $\mathfrak{b}_m(q, p) = \left\| \frac{q-m}{\|q-m\|} - \frac{p-m}{\|p-m\|} \right\|$. The pseudo-metric $\mathfrak{b}_m$ first projects the points q and p onto the surface of unit ball around m and then measures their Euclidean distance.

The analogous result for the weighted Cauchy–Schwarz inequality is

$$\begin{aligned}&\frac{\mathfrak{c}\overline{yq}-\mathfrak{c}\overline{ym}-\mathfrak{c}\overline{zq}+\mathfrak{c}\overline{zm}}{\mathfrak{l}(m,q)}-\frac{\mathfrak{c}\overline{yp}-\mathfrak{c}\overline{ym}-\mathfrak{c}\overline{zp}+\mathfrak{c}\overline{zm}}{\mathfrak{l}(m,p)}\\&\leq 2\|y-z\|_{s^{-1}}\left\|\frac{q-m}{\|q-m\|}-\frac{p-m}{\|p-m\|}\right\|_s.\end{aligned}$$

4.3.2.2 Bregman Divergence

Let $\mathcal{Q}\subseteq\mathbb{R}^r$ be a closed convex set. Let $\psi\colon\mathcal{Q}\to\mathbb{R}$ be a continuously differentiable and strictly convex function. The Bregman divergence $D_\psi\colon\mathcal{Q}\times\mathcal{Q}\to[0,\infty)$ associated with ψ for points $y,q\in\mathcal{Q}$ is defined as $D_\psi(y,q)=\psi(y)-\psi(q)-\langle\nabla\psi(q)\,,\,y-q\rangle$. It is the difference between the value of ψ at point y and the value of the first-order Taylor expansion of ψ around point q evaluated at point y. It is well-known, that the minimizer m of $q\mapsto\mathbb{E}[D_\psi(Y,q)]$ for a random variable Y with $\mathbb{E}[D_\psi(Y,q)]<\infty$ for all $q\in\mathcal{Q}$ is the expectation $m=\mathbb{E}[Y]$, see [BGW05, Theorem 1]. The Bregman divergence $\mathfrak{c}=D_\psi$ fulfills the weak quadruple inequality:

$$\begin{aligned}D_\psi(y,q)-D_\psi(z,q)-D_\psi(y,p)+D_\psi(z,p)&=\langle\nabla\psi(q)-\nabla\psi(p)\,,\,y-z\rangle\\&\leq\|\nabla\psi(q)-\nabla\psi(p)\|\|y-z\|.\end{aligned}$$

Similarly, we obtain a version of the strong quadruple inequality with $\xi=1$, $\mathfrak{l}(q,p)=\|q-p\|$,

$$\begin{aligned}&\frac{\mathfrak{c}\overline{yq}-\mathfrak{c}\overline{ym}-\mathfrak{c}\overline{zq}+\mathfrak{c}\overline{zm}}{\mathfrak{l}(m,q)}-\frac{\mathfrak{c}\overline{yp}-\mathfrak{c}\overline{ym}-\mathfrak{c}\overline{zp}+\mathfrak{c}\overline{zm}}{\mathfrak{l}(m,p)}\\&=\left\langle y-z,\frac{\nabla\psi(q)-\nabla\psi(m)}{\|q-m\|}-\frac{\nabla\psi(p)-\nabla\psi(m)}{\|p-m\|}\right\rangle\\&\leq\|y-z\|\left\|\frac{\nabla\psi(q)-\nabla\psi(m)}{\|q-m\|}-\frac{\nabla\psi(p)-\nabla\psi(m)}{\|p-m\|}\right\|.\end{aligned}$$

4.3.2.3 Hadamard Spaces and Quasi-Inner Product

Let $(\mathcal{Q},d)$ be a metric space. Use the notation $\overline{q,p}:=d(q,p)$. We use the squared metric as the cost function $\mathfrak{c}(y,q)=d(y,q)^2=\overline{y,q}^2$. One particularly nice version of the weak quadruple inequality with this cost function is

$$\overline{y,q}^2-\overline{y,p}^2-\overline{z,q}^2+\overline{z,p}^2\leq 2\,\overline{y,z}\,\overline{q,p}\,.$$

Let us call this inequality the *nice quadruple inequality*. As seen before, this holds for subsets of inner product spaces. It also plays an important role for geodesic metric spaces. In this section, we paraphrase some results of [BN08]. In particular, we state that the nice quadruple inequality characterizes CAT(0)-spaces.

Let $(\mathcal{Q},d)$ be a metric space. A *curve* is a continuous mapping $\gamma\colon[a,b]\to\mathcal{Q}$, where $[a,b]$ is a closed interval. The *length* of the curve $\gamma\colon[a,b]\to\mathcal{Q}$ is

$$L(\gamma):=\sup\left\{\sum_{i=1}^{I}d(\gamma(t_{i-1}),\gamma(t_i))\,\middle|\,a=t_0<t_1<\cdots<t_I=b,I\in\mathbb{N}\right\}.$$

A curve $\gamma\colon [a,b] \to \mathcal{Q}$ is called a *geodesic* if $L(\gamma) = d(\gamma(a), \gamma(b))$. A metric space is called *geodesic*, if any two points $q, p \in \mathcal{Q}$ can be joined by a geodesic $\gamma\colon [a,b] \to \mathcal{Q}$ with $\gamma(a) = q, \gamma(b) = p$. A *midpoint* of two points $q, p \in \mathcal{Q}$ is a point $m \in \mathcal{Q}$ such that $\overline{q,m} = \overline{p,m} = \frac{1}{2}\overline{q,p}$. A complete metric space is a geodesic space if and only if all pairs of points have a midpoint, see [Stu03, Proposition 1.2]. Now, let $(\mathcal{Q}, d)$ be a geodesic metric space. For any triple of points $a, b, c \in \mathcal{Q}$ one can construct a *comparison triangle* in the Euclidean plane with corners $a', b', c' \in \mathbb{R}^2$, such that $\overline{a,b} = \|b' - a'\|$, $\overline{a,c} = \|c' - a'\|$, and $\overline{b,c} = \|c' - b'\|$. A geodesic metric space $(\mathcal{Q}, d)$ is called *CAT(0)* if and only if for every triple of points $a, b, c \in \mathcal{Q}$ with comparison triangle (a', b', c') following condition holds: For every point d on a geodesic connecting a and b, it holds $\overline{d,c} \leq \|c' - d'\|$, where $d' \in \mathbb{R}^2$ is the point on the edge of the comparison triangle between a' and b' such that $\|d' - a'\| = \overline{a,d}$. A complete CAT(0)-space is called *Hadamard space* or *global NPC space* (**n**on**p**ositive **c**urvature).

A metric space $(\mathcal{Q}, d)$ is said to fulfill the *NPC-inequality* if and only if for all $y_1, y_2 \in \mathcal{Q}$ there exists a point $m \in \mathcal{Q}$ such that for all $q \in \mathcal{Q}$ it holds $\overline{m,q}^2 \leq \frac{1}{2}\overline{y_1,q}^2 + \frac{1}{2}\overline{y_2,q}^2 - \frac{1}{4}\overline{y_1,y_2}^2$. Then m is the midpoint of y_1 and y_2.

A characterization of CAT(0)-spaces can be found in [Stu03, Section 2]: A metric space is CAT(0) if and only if it fulfills the NPC-inequality.

Another characterization of CAT(0)-spaces by the nice quadruple inequality is given in [BN08, Corollary 3]: A geodesic space is CAT(0) if and only if it fulfills the nice quadruple inequality.

In [BN08], the authors define the *quadrilateral cosine* for $q, p, y, z \in \mathcal{Q}$ as

$$\operatorname{cosq}(\vec{yz}, \vec{qp}) := \frac{\overline{y,q}^2 - \overline{y,p}^2 - \overline{z,q}^2 + \overline{z,p}^2}{-2\,\overline{y,z}\,\overline{q,p}}\,.$$

Obviously, the statement $\operatorname{cosq}(\vec{yz}, \vec{qp}) \leq 1$ for all $q, p, y, z \in \mathcal{Q}$ is equivalent to the nice quadruple inequality. To further motivate this notation and compare it with inner product spaces, they introduce a *quasilinearization* of the metric space and a *quasi-inner product*: Define $\langle \vec{yz}\,,\, \vec{qp} \rangle_d = \operatorname{cosq}(\vec{yz}, \vec{qp})\, \|\vec{yz}\|_d \|\vec{qp}\|_d$, where $\|\vec{yz}\|_d := \overline{y,z}$. Thus, the nice quadruple inequality can be viewed as the Cauchy–Schwarz inequality of the quasi-inner product.

4.3.3 Weak Implies Strong

The weak quadruple inequality is well justified as a condition: Aside from allowing to establish rates in probability (Theorem 4.1), it can be interpreted as a form of Cauchy–Schwarz inequality (section 4.3.2.3), it is fulfilled in a large class of metric spaces (bounded metric spaces, Hadamard spaces, appendix 4.B), and the power inequality (Theorem 4.10) implies even more applications with a nice interpretation in statistics (section 4.5.2).

The case for the strong quadruple inequality, which we use in Theorem 4.5 to establish rates in expectation, seems much weaker. Although it can be established in Hilbert spaces, see section 4.3.2.1, it is not directly clear whether we can have a suitable version for Hadamard spaces or a power inequality.

The next section examines the strong quadruple inequality in Hadamard spaces and concludes with a negative result. Thereafter, we discuss an approach to infer convergence rates in expectation when only assuming the weak quadruple inequality by showing that a weak quadruple inequality imply certain strong quadruple inequalities. This approach is executed to obtain Theorem 4.7 for convergence rates in expectation, where the result holds only asymptotically, in contrast to Theorem 4.5.

4.3.3.1 Projection Metric

In Euclidean spaces, we can take $\mathfrak{b}_m(q,p) = \left\| \frac{q-m}{\|q-m\|} - \frac{p-m}{\|p-m\|} \right\|$ as the strong quadruple metric. This pseudo-metric can be written down only depending on the metric (not the norm or vector space operations) as

$$d_m^{\mathsf{proj}}(q,p) := \sqrt{\frac{\overline{q,p}^2 - (\overline{q,m} - \overline{p,m})^2}{\overline{q,m}\,\overline{p,m}}}\,, \quad d_m^{\mathsf{proj}}(q,p) = \mathfrak{b}_m(q,p)\,.$$

The metric $d_m^{\mathsf{proj}}(q,p)$ can be defined in any metric space. Unfortunately, it does not yield a strong quadruple inequality in non-Euclidean Hadamard spaces in the same way as in Euclidean spaces. See appendix 4.D for details.

4.3.3.2 Power Metric

To establish rates of convergence in expectation for the $\mathfrak{c}$-Fréchet mean, given that a weak quadruple inequality holds, we first show that some version of the strong quadruple inequality is implied by the weak one, Lemma 4.6. Unfortunately, we obtain a strong quadruple distance $\mathfrak{b}_m$ such that the measure of entropy $\mathsf{entr}(\mathcal{Q}, \mathfrak{b}_m)$ might be infinite. To solve this problem, we define an increasing sequence of sets $\mathcal{Q}_n$ such that $\mathcal{Q}_n \subseteq \mathcal{Q}_{n+1}$ and $\bigcup_{n\in\mathbb{N}} \mathcal{Q}_n = \mathcal{Q}$ with distances $\mathfrak{b}_{m,n}$ such that the strong quadruple inequality is fulfilled on $\mathcal{Q}_n$ with strong quadruple distance $\mathfrak{b}_{m,n}$, and $\mathsf{entr}(\mathcal{Q}_n, \mathfrak{b}_{m,n})$ is finite and can be suitably controlled in n. This allows us to prove an asymptotic result for the rate of convergence in expectation, Theorem 4.7.

Lemma 4.6. Assume $(\mathcal{Q}, \mathcal{Y}, \mathfrak{a}, \mathfrak{b}, \mathfrak{c})$ fulfills the weak quadruple inequality. Let $\xi \in [0,1]$. Then

$$\frac{{}^{\mathfrak{c}}\overline{yq} - {}^{\mathfrak{c}}\overline{ym} - {}^{\mathfrak{c}}\overline{zq} + {}^{\mathfrak{c}}\overline{zm}}{\mathfrak{b}(q,m)^{\xi}} - \frac{{}^{\mathfrak{c}}\overline{yp} - {}^{\mathfrak{c}}\overline{ym} - {}^{\mathfrak{c}}\overline{zp} + {}^{\mathfrak{c}}\overline{zm}}{\mathfrak{b}(p,m)^{\xi}} \leq 2^{\xi}\,\mathfrak{a}(y,z)\,\mathfrak{b}(q,p)^{1-\xi} \tag{4.4}$$

for all $y, z, q, p, m \in \mathcal{Q}$ with $\mathfrak{b}(q,m), \mathfrak{b}(p,m) > 0$.

See appendix 4.C for a proof. We would like to have ξ large, i.e., close to 1, to obtain the same rate of convergence in expectation as in probability. We achieve that by defining sequences $\xi_n \nearrow 1$ and $\mathcal{Q}_n \nearrow \mathcal{Q}$, and control the entropy of $\mathcal{Q}_n$ with respect to $\mathfrak{b}^{1-\xi_n}$.

To state the result, we have to modify the Entropy and the Existence condition. Recall the definition of the objective function $F(q) = \mathbb{E}[{}^{\mathfrak{c}}\overline{Yq}]$ and the empirical objective function $F_n(q) = \frac{1}{n}\sum_{i=1}^n {}^{\mathfrak{c}}\overline{Y_i q}$.

Assumptions.

EXISTENCE':
It holds $\mathbb{E}[|\mathfrak{c}(Y,q)|] < \infty$ for all $q \in \mathcal{Q}$. Let $o \in \mathcal{Q}$. Define $R_n := n$ and $\mathcal{Q}_n := \mathrm{B}_{R_n}(o, \mathfrak{b})$. There are $m_n^{\mathcal{Q}_n} \in \arg\min_{q \in \mathcal{Q}_n} F_n(q)$ measurable and $m \in \arg\min_{q \in \mathcal{Q}} F(q)$.

SMALL ENTROPY:
There are $\beta, c_{\mathsf{e}} > 0$ such that for $\delta > 0$ large enough

$$\sqrt{\log N(\mathrm{B}_\delta(o, \mathfrak{b}), \mathfrak{b}, r)} \leq c_{\mathsf{e}} \log\left(\frac{\delta}{r}\right)^\beta$$

for all $r > 0$.

Note that the SMALL ENTROPY condition is much stronger than ENTROPY, which we assumed in Theorem 4.1. In Euclidean subspaces $\mathcal{Q} \subseteq \mathbb{R}^b$, it holds

$$N(r, \mathrm{B}_\delta(0, d), d) \leq \left(\frac{3\delta}{r}\right)^b$$

for all $R > r > 0$ [Pol90, section 4]. Thus, SMALL ENTROPY is fulfilled in Euclidean spaces.

Theorem 4.7 (Convergence rate in expectation)**.** In the *Abstract Setting* of section 4.2.1 with loss $\mathfrak{l} = \mathfrak{b}$, where $\mathfrak{b}$ is a pseudo-metric, and rate parameter $\xi = 1$, assume that following conditions hold: EXISTENCE', GROWTH with $\gamma > 1$, WEAK QUADRUPLE, STRONG MOMENT with $\kappa > \gamma - 1$, SMALL ENTROPY. Then

$$\mathbb{E}\left[\mathfrak{b}(m, m_n^{\mathcal{Q}_n})^\kappa\right] = \mathbf{O}\left(\left(n^{-\frac{1}{2}} \log(n)^\beta\right)^{\frac{\kappa}{\gamma-1}}\right).$$

See appendix 4.A for the proof.

4.4 Application of the Abstract Results

We apply the abstract results of the previous theorems in this section. We first consider two toy examples – Euclidean spaces, section 4.4.1 and infinite dimensional Hilbert spaces, section 4.4.2 – to better understand the result and compare them to optimal bounds. Then we discuss two more involved settings: The Fréchet mean for nonconvex subsets of Euclidean spaces, section 4.4.3, and for Hadamard spaces, section 4.4.4.

4.4.1 Euclidean Spaces

Let $\mathcal{Q} \subseteq \mathbb{R}^b$ be convex with the Euclidean metric $d(p,q) = \|p - q\|$. Choose $\mathcal{Y} = \mathcal{Q}$, $\mathfrak{c} = d^2$, $\mathfrak{l} = d$, $\xi = 1$. Let Y be a $\mathcal{Q}$-valued random variable with $\mathbb{E}[\|Y\|^2] < \infty$. Then

the Fréchet mean equals the expectation $m = \mathbb{E}[Y] \in \mathcal{Q}$. We can easily calculate

$$\mathbb{E}\Big[\|Y-q\|^2 - \|Y-m\|^2\Big] = \|q-m\|^2 .$$

Thus, the Growth condition is fulfilled with $\gamma = 2$. The space has the strong quadruple inequality at every point with data distance $\mathfrak{a}(y,z) = 2\|y-z\|$ and strong quadruple distance $\mathfrak{b}_m(p,q) = \left\| \frac{q-m}{\|q-m\|} - \frac{p-m}{\|p-m\|} \right\|$, see section 4.3.2.1. Thus, Theorem 4.5 implies

$$\begin{aligned}
\mathbb{E}\left[\left\| \mathbb{E}[Y] - \frac{1}{n}\sum_{i=1}^n Y_i \right\|^2 \right] &= \mathbb{E}\Big[\mathfrak{l}(m,m_n)^2\Big] \\
&\le C n^{-1}\, \mathsf{entr}_n(\mathcal{Q},\mathfrak{b}_m)^2 \mathbb{E}\Big[\mathfrak{a}(Y',Y)^2\Big] \\
&\le C' b \frac{1}{n} \mathbb{E}\Big[\|Y-\mathbb{E}[Y]\|^2\Big] ,
\end{aligned}$$

where we used $N(r, \mathrm{B}_R(0,d), d) \le \left(\frac{3R}{r}\right)^b$ for all $R > r > 0$ [Pol90, section 4] to fulfill Strong Entropy. The constants $C, C' > 0$ are universal. Compare this with the result that one obtains by direct calculations, i.e.,

$$\mathbb{E}\left[\left\| \mathbb{E}[Y] - \frac{1}{n}\sum_{i=1}^n Y_i \right\|^2 \right] = \frac{1}{n} \mathbb{E}\Big[\|Y-\mathbb{E}[Y]\|^2\Big] .$$

We pay an extra dimension factor b when using the Fréchet mean approach instead of direct calculations. This comes from the use of the Cauchy–Schwarz inequality, which powers the strong quadruple inequality in Euclidean spaces.

4.4.2 Hilbert Spaces

Let $\mathbb{H}$ be an infinite dimensional Hilbert space and $\mathcal{Q} \subseteq \mathbb{H}$ convex. Let $d(p,q)^2 = \|p-q\|^2 = \langle p-q\,,\, p-q\rangle$. Choose $\mathcal{Y} = \mathcal{Q}$, $\mathfrak{c} = d^2$, $\mathfrak{l} = d$, $\xi = 1$. Let Y be a $\mathcal{Q}$-valued random variable with $\mathbb{E}[\|Y\|^2] < \infty$. As in the Euclidean case, the Fréchet mean m equals the expectation $\mathbb{E}[Y]$, the Growth condition holds with $\gamma = 2$, and the strong quadruple inequality is fulfilled with $\mathfrak{a}(y,z) = 2\|y-z\|$ and pseudometric $\mathfrak{b}_m(p,q) = \left\| \frac{q-m}{\|q-m\|} - \frac{p-m}{\|p-m\|} \right\|$.

Unfortunately, Strong Entropy is not fulfilled on $\mathbb{H}$ if $\dim(\mathbb{H}) = \infty$. By introducing a weight sequence, we can make $\mathfrak{b}_m$ smaller by making $\mathfrak{a}$ larger: Assume that the Hilbert space $\mathbb{H}$ is separable and thus admits a countable basis. Let $s = (s_k)_{k\in\mathbb{N}} \subseteq (0,\infty)$. In section 4.3.2.1, we derived that the strong quadruple condition holds with $\mathfrak{a}(y,z) = 2\|y-z\|_{s^{-1}}$ and $\mathfrak{b}^s_m(p,q) = \left\| \frac{q-m}{\|q-m\|} - \frac{p-m}{\|p-m\|} \right\|_s$. Then $\mathsf{entr}_n(\mathbb{H},\mathfrak{b}^s_m) \le \gamma_2(\mathbb{H},\mathfrak{b}^s_m) \le \gamma_2(\mathcal{E}_s,d)$, where

$$\mathcal{E}_s = \left\{ h \in \mathbb{H} \colon \sum_{k=1}^\infty \frac{h_k^2}{s_k^2} \le 1 \right\} .$$

There is a universal constant $c > 0$ such that $\gamma_2(\mathcal{E}_s, d)^2 \leq c \sum_{k=1}^{\infty} s_k^2$, see [Tal14, Proposition 2.5.1]. As a condition on the variance term, we need

$$\mathbb{E}\Big[\|Y - \mathbb{E}[Y]\|_{s^{-1}}^2\Big] = \|\sigma\|_{s^{-1}}^2 = \sum_{k=1}^{\infty} \sigma_k^2 s_k^{-2} < \infty\,,$$

where $\sigma_k^2 := \mathbb{V}[Y_k]$ and $\sigma = (\sigma_k)_{k\in\mathbb{N}}$. Similar to the Euclidean case, Theorem 4.5 implies

$$\mathbb{E}\left[\left\|\mathbb{E}[Y] - \frac{1}{n}\sum_{i=1}^{n} Y_i\right\|^2\right] = \mathbb{E}\Big[\mathfrak{l}(m, m_n)^2\Big] \leq C\frac{1}{n}\|s\|_{\ell_2}^2 \|\sigma\|_{s^{-1}}^2\,,$$

where $\|s\|_{\ell_2}^2 = \sum_{k=1}^{\infty} s_k^2$.

Direct calculations yield a better result:

$$\mathbb{E}\left[\left\|\mathbb{E}[Y] - \frac{1}{n}\sum_{i=1}^{n} Y_i\right\|^2\right] = \frac{1}{n}\|\sigma\|_{\ell_2}^2\,.$$

As in the Euclidean case, we pay a factor related to the dimension for using the more generally applicable Fréchet mean approach instead of using the inner product for direct calculations.

4.4.3 Nonconvex Subsets

Assume we are in the setting of section 4.4.2 and the mentioned conditions for convergence are fulfilled. But now we want to take $\mathcal{Q} \subseteq \mathbb{H}$ not necessarily convex and $\mathcal{Y} = \mathbb{H}$. Assume that Existence of the Fréchet mean $m \in \mathcal{Q}$ is fulfilled. The expectation $\mu := \mathbb{E}[Y] \in \mathbb{H}$ might not be an element of Q. Then the Fréchet mean m is the closest projection of μ to $\mathcal{Q}$, in the sense that

$$\operatorname*{arg\,min}_{q\in\mathcal{Q}} \mathbb{E}[\|Y - q\|^2] = \operatorname*{arg\,min}_{q\in\mathcal{Q}} \|\mu - q\|\,.$$

To get the same rate as in section 4.4.2, we mainly need to be concerned with the Growth condition, as the quadruple condition holds in all subsets. For $q \in \mathbb{H}$, simple calculations show

$$\mathbb{E}\Big[\|Y - q\|^2 - \|Y - m\|^2\Big] = \|\mu - q\|^2 - \|\mu - m\|^2\,.$$

We want to find a lower bound of this term in the form of $c_{\mathsf{g}}\|q - m\|^{\gamma}$ for constants $\gamma, c_{\mathsf{g}} > 0$. For $a > 1$, it holds

$$\begin{aligned}
&a\,\|\mu - q\|^2 - a\,\|\mu - m\|^2 - \|q - m\|^2 \\
&= (a-1)\left\|q - \left(\mu + \frac{\mu - m}{a-1}\right)\right\|^2 - \frac{a^2}{a-1}\,\|m - \mu\|^2\,.
\end{aligned}$$

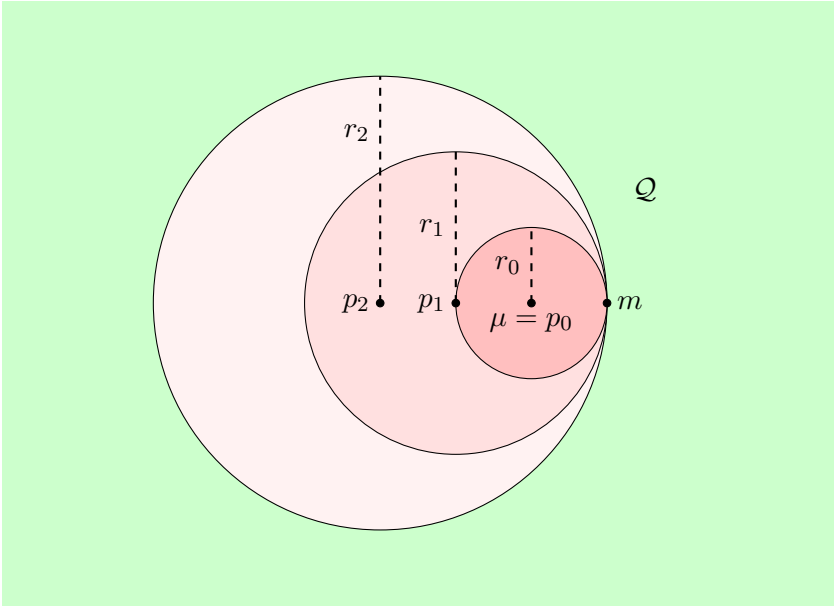

Figure 4.1: If $\mathcal{Q} \cap \mathrm{B}_{r_0}(p_0) \neq \emptyset$, the point m cannot be the Fréchet mean of a distribution with expectation μ. To fulfill the `Growth` condition, we need $\mathcal{Q} \cap \mathrm{B}_{r_1}(p_1) = \emptyset$ for a ball with larger radius $r_1 > r_0$ and adjusted center p_1. Increasing the radius further, $r_2 > r_1$, only improves the constant c_{g} of the `Growth` condition, but not the exponent γ.

Thus, $\|\mu - q\|^2 - \|\mu - m\|^2 \geq \frac{1}{a} \|q - m\|^2$ if and only if

$$\left\| q - \left(\mu + \frac{\mu - m}{a - 1} \right) \right\| \geq \frac{a}{a - 1} \|\mu - m\| \ .$$

Equivalently, the Growth condition holds with $\gamma = 2$ and $c_{\mathrm{g}} \in (0, 1)$ if and only if

$$\left\| q - \left(\mu + \frac{c_{\mathrm{g}}}{1 - c_{\mathrm{g}}} (\mu - m) \right) \right\| \geq \frac{1}{1 - c_{\mathrm{g}}} \|\mu - m\|$$

for all $q \in \mathcal{Q}$, i.e., if and only if $\mathcal{Q} \cap \mathrm{B}_r(p) = \emptyset$, where $r = \frac{1}{1 - c_{\mathrm{g}}} \|\mu - m\|$ and $p = \mu + \frac{1 - c_{\mathrm{g}}}{c_{\mathrm{g}}} (\mu - m)$. Note that $\|p - m\| = r$. This is illustrated in Figure 4.1. We have answered the question of how $\mathcal{Q}$ may look like, given the location of μ and m. Possibly more interesting is the question of, given $\mathcal{Q}$, where may μ be located so that m can be estimated with the same rate as for convex sets. We will answer this question only informally via a description similar to a *medial axis transform* [CCM97]:

For simplicity assume $\mathcal{Q} = \mathbb{R}^2 \setminus A$, where A is a nonempty, open, and simply connected set with border ∂A that is parameterized by the continuous function $\gamma \colon [0, 1] \to \partial A$. Roll a ball along the border on the inside of A. Make the ball as large as possible at any point so that it is fully contained in A and touches the border at point $\gamma(t)$.

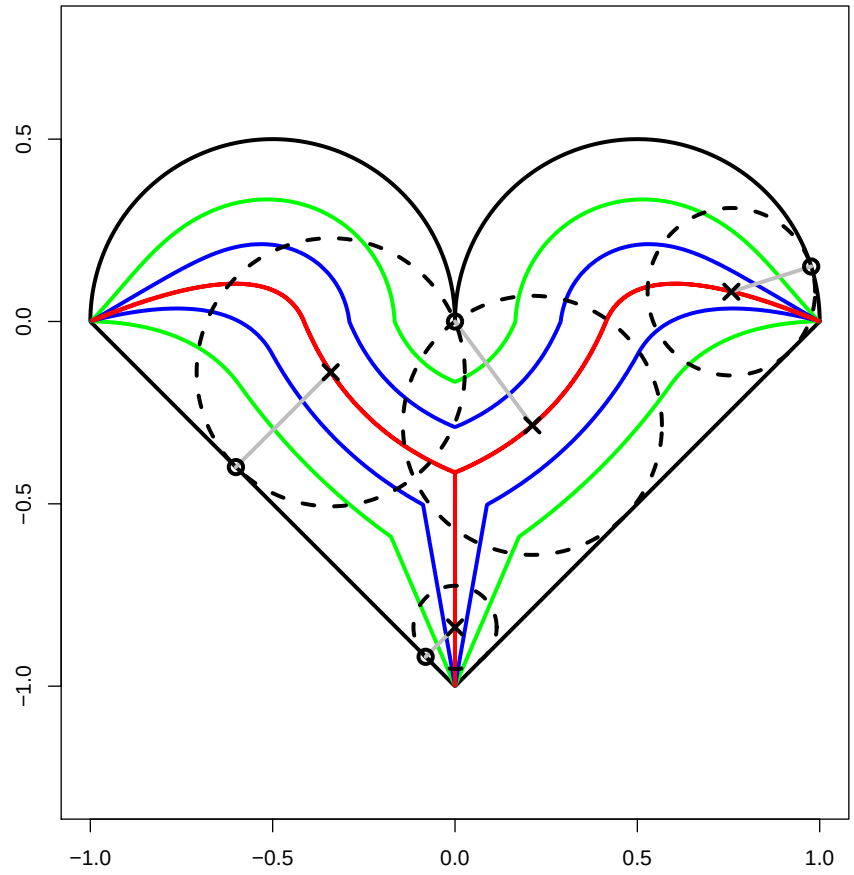

Figure 4.2: Let $A \subseteq \mathbb{R}^2$ be the set enclosed by the heart (solid black lines). Let $\mathcal{Y} = \mathbb{R}^2$ and $\mathcal{Q} = \mathbb{R}^2 \setminus A$. We consider a distribution on $\mathbb{R}^2$ with mean $\mu \in \mathcal{Y}$ and Fréchet mean $m \in \mathcal{Q}$ with respect to the Euclidean metric and the descriptor space $\mathcal{Q}$. The green, blue, and red lines show $p_\epsilon(t)$ for $\epsilon = 0.6, 0.3, 0$.

Denote the center of the ball as $c\colon [0,1] \to A$ and the radius as $r\colon [0,1] \to [0,\infty)$. Take $\epsilon \in (0,1)$ and trace the point $p_\epsilon\colon [0,1] \to A$ on the radius connecting the center of the ball $c(t)$ and the border $\gamma(t)$ such that it divides the radius into two pieces of length $\overline{p_\epsilon(t),c(t)} = \epsilon r(t)$ and $\overline{p_\epsilon(t),\gamma(t)} = (1-\epsilon)r(t)$. If μ lies on the outside of the set prescribed by $p\colon [0,1] \to A$, it can be estimated with the same rate as for convex sets. This is illustrated in Figure 4.2. The set of all centers $\mathcal{C} := \{c(t) \,|\, t \in [0,1]\}$, also called the *medial axis* ot *cut locus*, is critical: The closer μ is to $\mathcal{C}$, the larger the guaranteed error bound for the estimator. In particular, we cannot guarantee consistency of the estimator if $\mu \in \mathcal{C}$. A very similar phenomenon is described in [BP03, section 3]. The authors consider a Riemannian manifold $\mathcal{Q}$ that is embedded in an Euclidean space $\mathcal{Y}$. The *extrinsic mean* of a distribution on $\mathcal{Q}$ is the projection of the mean μ in $\mathcal{Y}$ to $\mathcal{Q}$. The points $\mathcal{C}$ are called *focal points*. It is shown [BP03, Theorem 3.3] that in many cases

the *intrinsic mean*, i.e, the Fréchet mean in $\mathcal{Q}$ with respect to the Riemannian metric on $\mathcal{Q}$, is equal to the extrinsic mean, i.e, the Fréchet mean in $\mathcal{Q}$ with respect to the Euclidean metric on $\mathcal{Y}$.

The conditions described above are connected to the term reach of a set [Fed59]. The reach of $\mathcal{Q} \subseteq \mathbb{R}^b$ is the largest $\epsilon > 0$ (possibly ∞) such that $\inf_{q \in \mathcal{Q}} d(x, q) < \epsilon$ implies that $x \in \mathbb{R}^b$ has a unique projection to $\mathcal{Q}$, i.e., a unique point $x_{\mathcal{Q}}$ with $d(x, x_{\mathcal{Q}}) = \inf_{q \in \mathcal{Q}} d(x, q)$. If the distance of the mean μ to $\mathcal{Q}$ is less than the reach of $\mathcal{Q}$, then the GROWTH condition holds with $\gamma = 2$. Thus, the rate of convergence is upper bounded by $cn^{-\frac{1}{2}}$ for some $c > 0$. Note that convex sets have infinite reach and exhibit this upper bound for any distribution with finite second moment.

By considering the growth condition $\|\mu - q\|^2 - \|\mu - m\|^2 \geq c_{\mathfrak{g}} \|q - m\|^\gamma$, one can also find examples of subspaces where the growth exponent for specific distributions is different from 2.

4.4.4 Hadamard Spaces

Let $(\mathcal{Q}, d)$ be a Hadamard space. A definition of Hadamard spaces is given in section 4.3.2.3. Use the notation $\overline{y,q} = d(y, q)$. For our purposes the most notable property of Hadamard spaces is that they fulfill the nice quadruple property, i.e., $\overline{y,q}^2 - \overline{y,p}^2 - \overline{z,q}^2 + \overline{z,p}^2 \leq 2\,\overline{y,z}\,\overline{q,p}$. In the following subsections, we will see how this translates to convergence rates for the Fréchet mean estimator and use the power inequality to obtain results for a generalized Fréchet mean with cost function d^{2a} for $a \in [\frac{1}{2}, 1]$.

For an introduction to Hadamard spaces see [Bač14a]. A survey of recent developments can be found in [Bac18]. In [BN08] the authors characterize Hadamard spaces by the nice quadruple inequality and discuss a quasilinearzation of these spaces by observing that the left hand side of the nice quadruple inequality behaves like an inner product to some extend. [Stu03] shows how some important theorems of probability theory in Euclidean spaces, like the law of large numbers and Jensen's inequality, translate to non-Euclidean Hadamard spaces. In [Stu02] martingale theory on Hadamard spaces is discussed.

Turning to more applied topics, [Bač14b] shows algorithms for calculating the Fréchet mean in Hadamard spaces with cost function d^{2a} for $a = \frac{1}{2}$ and $a = 1$. An important application of Hadamard spaces in Bioinformatics are phylogenetic trees [BHV01]. See also [Bac18, section 6.3] for a quick overview. Another application of Hadamard spaces is taking means in the manifold of positive definite matrices, e.g., in diffusion tensor imaging. But note that, as the underlying space is a differentiable manifold, one an use gradient-based approaches, see [PFA06].

Further examples of Hadamard spaces include Hilbert spaces, the Poincaré disc, complete metric trees, complete simply-connected Riemannian manifolds of nonpositive sectional curvature. See also [Stu03, section 3]. Let $(\mathcal{Q}, d)$ be a Hadamard space. We use $\mathcal{Q}$ as data space as well as descriptor space, i.e., $\mathcal{Q} = \mathcal{Y}$. The cost function is $\mathfrak{c} = d^2$, the loss $\mathfrak{l} = d$. As described in section 4.3.2.3 the weak quadruple inequality holds with $\mathfrak{a} = 2d$ and $\mathfrak{b} = d$, i.e., $(\mathcal{Q}, d)$ fulfills the nice quadruple inequality. Let Y be a random variable with values in $\mathcal{Q}$. Let $Y_1, \ldots, Y_n$ be iid copies of Y.

If $\mathbb{E}[d(Y,q)^2] < \infty$ for one $q \in \mathcal{Q}$, then it is also finite for every $q \in \mathcal{Q}$ and the Fréchet mean $m \in \arg\min_{q\in\mathcal{Q}} \mathbb{E}[d(Y,q)^2]$ exists and is unique, see [Stu03, Proposition 4.3]. The same holds true for the estimator $m_n \in \arg\min_{q\in\mathcal{Q}} \sum_{i=1}^n d(Y_i,q)^2$. Thus, EXISTENCE is fulfilled.

Here, we chose a second moment condition, because we will need it for estimation anyway. But note that choosing the cost function as $\mathfrak{c}(y,q) = d(y,q)^2 - d(y,o)^2$ for a fixed, arbitrary point $o \in \mathcal{Q}$ allows us to require only a finite first moment for EXISTENCE and the resulting Fréchet mean coincides with the d^2-Fréchet mean if the second moment is finite. This is described in more detail and utilized in [Stu03].

Furthermore, the GROWTH-condition holds in Hadamard spaces with $\gamma = 2$ and $c_{\mathsf{g}} = 1$, see [Stu03, Proposition 4.4]. Thus, we obtain following corollary of Theorem 4.1.

Corollary 4.8 (Convergence rate in probability)**.** Assume MOMENT with $\zeta = 2$ and $\mathfrak{a} = 2d$ and ENTROPY with $\mathfrak{b} = d$ and $\alpha = \beta$. Define

$$\eta_{\beta,n} := \begin{cases} n^{-\frac{1}{2}} & \text{for } \beta < 1\,, \\ n^{-\frac{1}{2}} \log(n+1) & \text{for } \beta = 1\,, \\ n^{-\frac{1}{2\beta}} & \text{for } \beta > 1\,. \end{cases}$$

Then, for all $s > 0$, it holds

$$\mathbb{P}\Big(\eta_{\beta,n}^{-1} d(m,m_n) \geq s\Big) \leq c\,\mathbb{E}[d(Y,Y')^2]\, s^{-2}\,,$$

with a constant $c > 0$ depending only on β and c_{e}. In particular,

$$d(m,m_n) = \mathbf{O}_{\mathbb{P}}(\eta_{\beta,n})\ .$$

As described in section 4.3.2.3, it may be difficult to find a version of the strong quadruple inequality such that the same rate can be derived for convergence in expectation. Thus, instead of trying to apply Theorem 4.5, we utilize (i) Corollary 4.3 and (ii) Theorem 4.7, respectively.

Corollary 4.9.

(i) Let $\epsilon > 0$. Assume $\mathbb{E}\big[d(Y,Y')^{2+\epsilon}\big] < \infty$. Assume ENTROPY with $\mathfrak{b} = d$ and $\alpha = \beta < 1$. Then it holds

$$\mathbb{E}\Big[d(m,m_n)^2\Big] \leq c\,\mathbb{E}\Big[d(Y,Y')^{2+\epsilon}\Big]^{\frac{2}{2+\epsilon}} \frac{1}{\epsilon n}$$

for a constant $c > 0$ depending only on β.

(ii) Assume $\mathbb{E}\big[d(Y,Y')^2\big] < \infty$. Let $o \in \mathcal{Q}$. Assume SMALL ENTROPY with $\mathfrak{b} = d$.

Let $\tilde{m}_n \in \arg\min_{q \in \mathrm{B}_n(o)} \sum_{i=1}^n d(Y_i, q)^2$. Then it holds

$$\mathbb{E}\Big[d(m, \tilde{m}_n)^2\Big] = \mathbf{O}\Big(\frac{1}{n}\log(n)^{2\beta}\Big) \,.$$

4.5 Power Fréchet Means in Hadamard Spaces

In this section, we demonstrate the great utility of the theory developed in the previous sections by providing finite sample bounds and rates of convergence for power Fréchet means with power $\alpha \in [1, 2]$ in Hadamard spaces. To the best knowledge of the author, this result, which first appeared in [Sch19b], was not know before, not even in the Euclidean spaces. It relies on an asymmetric weak quadruple inequality for power metrics that is shown to hold in Hadamard spaces. This power inequality seems to be a deep result; its the theorem with the longest proof in this thesis.

Recall from section 2.5, that for a strong law of large numbers to hold for α-Fréchet means with $\alpha \geq 1$, we require $\mathbb{E}[\overline{Y,o}^{\alpha-1}] < \infty$. We will show that for a parametric rate of the convergence we require $\mathbb{E}[\overline{Y,o}^{2(\alpha-1)}] < \infty$ for $\alpha \in [1, 2]$ in Hadamard spaces.

First, we provide a suitable quadruple inequality in section 4.5.1. Then in section 4.5.2, we use it with the theory of rates for generalized Fréchet means to derive the result.

4.5.1 Power Inequality

If the metric space $(\mathcal{Q}, d)$ fulfills the nice quadruple inequality, i.e, $\overline{y,q}^2 - \overline{y,p}^2 - \overline{z,q}^2 + \overline{z,p}^2 \leq 2\,\overline{y,z}\,\overline{q,p}$, where $\overline{y,q} = d(y, q)$, then $(\mathcal{Q}, d^a)$, $a \in [\frac{1}{2}, 1]$, also fulfills a weak quadruple inequality with a suitably adapted bound. According to [DD16], the metric d^a is called *power transform metric* or *snowflake transform metric*.

Theorem 4.10 (Power Inequality)**.** Let $(\mathcal{Q}, d)$ be a metric space. Use the short notation $\overline{q,p} := d(q, p)$. Let $q, p, y, z \in \mathcal{Q}$, $a \in [\frac{1}{2}, 1]$. Assume

$$\overline{yq}^2 - \overline{yp}^2 - \overline{zq}^2 + \overline{zp}^2 \leq 2\,\overline{y,z}\,\overline{q,p}\,. \tag{4.5}$$

Then

$$\overline{y,q}^{2a} - \overline{y,p}^{2a} - \overline{z,q}^{2a} + \overline{z,p}^{2a} \leq 8a2^{-2a}\,\overline{y,z}^{2a-1}\,\overline{q,p}\,. \tag{4.6}$$

In particular, if the metric space $(\mathcal{Q}, d)$ fulfills the nice quadruple inequality and $a \in [\frac{1}{2}, 1]$, then the weak quadruple inequality for $\mathfrak{c} = d^{2a}$ is fulfilled with $\mathfrak{a} = 8a2^{-2a}d^{2a-1}$ and $\mathfrak{b} = d$.

Following the intermediate step Lemma 4.27 (appendix 4.G) in the proof of Theorem 4.10, one can easily show a similar result if the constant on the right hand side of equation (4.5) is larger than 2. Only the constant $8a2^{-2a}$ on the right hand side of equation (4.6) changes.

The theorem applies to subsets of Hadamard spaces. But note that it is not required that $\mathcal{Q}$ is geodesic, but can consist of only the points q, p, y, z. As a statement purely about metric spaces, it might be of interest outside the realm of statistics.

In Corollary 4.11 (section 4.3.2.3) it is used to show rates of convergence for the Fréchet mean estimator of the power transform metric d^a. There the asymmetry of the exponents of the factors on the right hand side of (4.6) is essential for proving the result under weak assumption.

Unfortunately, the only proof of this statement that the author was able to derive (see appendix 4.G) is very long and does not give much insight into the problem as it mostly consists of distinguishing many cases and then using simple calculus. The author is convinced that a more appealing proof is possible.

The concave function $[\frac{1}{2}, 1] \to (0, \infty), a \mapsto 8a2^{-2a}$ is maximal at $a_0 = (2\ln(2))^{-1} \approx 0.721$ with $8a_0 2^{-2a_0} = \frac{4}{\mathrm{e}\ln(2)} \leq 2.123$. Thus, the constant factor in the bound is very close to 2, but 2 is not sufficient.

In appendix 4.E, we show that $8a2^{-2a}$ is the optimal constant, and that we cannot extend Theorem 4.10 to $a > 1$ or $a < \frac{1}{2}$. Of course, for $a \in (0, \frac{1}{2}]$, we have $\overline{y,q}^{2a} - \overline{y,p}^{2a} - \overline{z,q}^{2a} + \overline{z,p}^{2a} \leq 2\,\overline{q,p}^{2a}$ as d^{2a} is a metric, which obeys the triangle inequality.

It is not known to the author whether the nice quadruple inequality in $(\mathcal{Q}, d)$ does or does not imply the nice quadruple inequality in $(\mathcal{Q}, d^a)$ for $a \in (\frac{1}{2}, 1)$, i.e.,

$$\overline{y,q}^{2a} - \overline{y,p}^{2a} - \overline{z,q}^{2a} + \overline{z,p}^{2a} \leq 2\,\overline{y,z}^{a}\,\overline{q,p}^{a}\,.$$

4.5.2 Rates of Convergence

Let $(\mathcal{Q}, d)$ is a Hadamard space and $a \in [\frac{1}{2}, 1)$. Then $(\mathcal{Q}, d^a)$ is not Hadamard, but fulfills a weak quadruple inequality: Fix an arbitrary point $o \in \mathcal{Q}$. We use the cost function $\mathfrak{c}(y, q) = d^{2a}(y, q) - d^{2a}(y, o)$ and the loss $\mathfrak{l} = d$. Then the weak quadruple inequality holds with $\mathfrak{a}(y, z) = 8a2^{-2a} d(y, z)^{2a-1}$ and $\mathfrak{b} = d$.

We need to choose the cost function $d^{2a}(y, q) - d^{2a}(y, o)$ instead of $d^{2a}(y, q)$ to obtain a result with minimal moment requirement. To fulfill Moment we need that $\mathbb{E}[d(Y, Y')^{2(2a-1)}]$ is finite and for Existence, we need $\mathbb{E}[|\mathfrak{c}(Y, q)|] < \infty$. We fulfill both by assuming that $\mathbb{E}[d(Y, o)^{2(2a-1)}] < \infty$. Then the both conditions are satisfied: On one hand, it holds $\mathbb{E}[d(Y, Y')^{2(2a-1)}] \leq 2\mathbb{E}[d(Y, o)^{2(2a-1)}]$. On the other hand, using the tight power bound of Lemma 4.31 (appendix section 4.G),

$$\overline{y,q}^{2a} - \overline{y,o}^{2a} \leq 2a\,\overline{q,o}\left(\frac{\overline{y,q} + \overline{y,o}}{2}\right)^{2a-1}$$

and thus

$$|\mathfrak{c}(Y, q)| \leq 2a\,\overline{q,o}\left(\frac{\overline{q,o}}{2} + \overline{Y,o}\right)^{2a-1},$$

which implies $\mathbb{E}[|\mathfrak{c}(Y, q)|] < \infty$. But $\mathbb{E}[d(Y, q)^{2a}]$ might be infinite as $2a > 2(2a - 1)$.

Theorem 4.1 with $\zeta = 2$ implies following corollary.

Corollary 4.11 (Bounds in probability for power mean)**.** Assume:

(i) EXISTENCE: Let $a \in [\frac{1}{2}, 1]$. Let $o \in \mathcal{Q}$ be an arbitrary fixed point. Assume there are $m_n \in \arg\min_{q\in\mathcal{Q}} \frac{1}{n}\sum_{i=1}^n \left(\overline{Y_i,q}^{2a} - \overline{Y_i,o}^{2a}\right)$ measurable and $m \in \arg\min_{q\in\mathcal{Q}} \mathbb{E}[\overline{Y,q}^{2a} - \overline{Y,o}^{2a}]$.

(ii) GROWTH: There are constants $c_{\mathsf{g}} > 0, \gamma \in (1,\infty)$ such that $\mathbb{E}[\overline{Y,q}^{2a}] - \mathbb{E}[\overline{Y,m}^{2a}] \geq c_{\mathsf{g}} d(m,q)^\gamma$ for all $q \in \mathcal{Q}$.

(iii) MOMENT: $\mathbb{E}[\overline{Y,q}^{2(2a-1)}] < \infty$ for one (and thus for all) $q \in \mathcal{Q}$.

(iv) ENTROPY: There is $\beta > 0$ such that

$$\sqrt{\log N(\mathrm{B}_\delta(m,d),d,r)} \leq c_{\mathsf{e}} \left(\frac{\delta}{r}\right)^\beta$$

for all $\delta, r > 0$.

Then, for all $s > 0$, it holds

$$\mathbb{P}\left(\eta_{\beta,n}^{-\frac{1}{\gamma-1}}\, \overline{m,m_n} \geq s\right) \leq c\, \mathbb{E}[\overline{Y,o}^{2(2a-1)}]\, s^{-2(\gamma-1)}\,,$$

where $c > 0$ depends only on $\beta, \gamma, c_{\mathsf{e}}$. In particular,

$$d(m,m_n) = \mathbf{O}_{\mathbb{P}}\left(\eta_{\beta,n}^{-\frac{1}{\gamma-1}}\right)\,.$$

For $\beta < 1$ (true in many spaces, e.g., in Euclidean spaces) and $\gamma = 2$, we obtain the parametric rate of convergence, $d(m,m_n) = \mathbf{O}_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$.

Note that the moment condition becomes weaker as a gets smaller and vanishes for $a = \frac{1}{2}$, where, in the Euclidean case, the Fréchet mean is the median.

EXISTENCE of m_n and m is a purely technical condition, as one will usually only be able to minimize the objective functions up to an $\epsilon > 0$ and the set of ϵ-minimizers is always nonempty.

The GROWTH condition is more interesting. It seems possible to choose $\gamma = 2$ for all $a \in [\frac{1}{2}, 1]$ in many circumstances – at least under some conditions on the distribution of Y. But precise statements of this sort are unknown to the author. If γ really can be chosen independently of a, then the rate is the same for all $a \in [\frac{1}{2}, 1]$. In the Euclidean case, this is manifested in the fact that we can estimate median ($a = \frac{1}{2}$) and mean ($a = 1$) and all statistics "in between" ($a \in (\frac{1}{2}, 1)$) with the same rate (under some conditions), but with less restrictive moment assumptions for smaller powers a.

Similarly to the corollary above, we can apply Corollary 4.3 or Theorem 4.7 to obtain rates in expectation.

Appendix of Chapter 4

4.A Proofs of Theorem 1, 2, and 4

4.A.1 Proof of Theorem 4.1 and Corollary 4.3

Define

$$\Delta_n(\delta) := \sup_{q \in \mathcal{Q}\colon \mathfrak{l}(m,q) \le \delta} F(q) - F(m) - F_n(q) + F_n(m)\,.$$

Results similar to following Lemma are well known in the M-estimation literature. The proof relies on the *peeling device*, see [Gee00].

Lemma 4.12 (Weak argmin transform)**.** Assume Growth. Let $\zeta \ge 1$. Assume that there are constants $\xi \in (0, \gamma)$, $h_n \ge 0$ such that $\mathbb{E}[\Delta_n(\delta)^\zeta] \le \left(h_n \delta^\xi\right)^\zeta$ for all $\delta > 0$. Then

$$\mathbb{P}(\mathfrak{l}(m, m_n) \ge s) \le c \left(h_n s^{-(\gamma-\xi)}\right)^\zeta\,,$$

where $c > 0$ depends only on $c_{\mathsf{g}}, \gamma, \xi, \zeta$.

Proof. Let $0 < a < b$. If $\mathfrak{l}(m, m_n) \in [a, b]$, we have

$$c_{\mathsf{g}} a^\gamma \le c_{\mathsf{g}} \mathfrak{l}(m, m_n)^\gamma \le F(m_n) - F(m) \le F(m_n) - F(m) - F_n(m_n) + F_n(m) \le \Delta_n(b)\,.$$

Let $s > 0$. For $k \in \mathbb{N}_0$, set $a_k := s2^k$. It holds

$$\begin{aligned}
&\mathbb{P}(\mathfrak{l}(m, m_n) \ge s) \\
&\le \sum_{k=0}^{\infty} \mathbb{P}(\mathfrak{l}(m, m_n) \in [a_k, a_{k+1}]) \\
&\le \sum_{k=0}^{\infty} \mathbb{P}\big(c_{\mathsf{g}} a_k^\gamma \le \Delta_n(a_{k+1})\big)\,.
\end{aligned}$$

We use Markov's inequality and the bound on $\mathbb{E}[\Delta_n(\delta)^\zeta]$ to obtain

$$\begin{aligned}\mathbb{P}\big(c_{\mathfrak{g}} a_k^\gamma \le \Delta_n(a_{k+1})\big) &\le \frac{\mathbb{E}\Big[\Delta_n(a_{k+1})^\zeta\Big]}{\big(c_{\mathfrak{g}} a_k^\gamma\big)^\zeta} \\ &\le \left(\frac{h_n a_{k+1}^\xi}{c_{\mathfrak{g}} a_k^\gamma}\right)^\zeta \\ &= \Big(2 c_{\mathfrak{g}}^{-1} h_n s^{-(\gamma-\xi)} 2^{-k(\gamma-\xi)}\Big)^\zeta .\end{aligned}$$

As $\gamma - \xi > 0$, we get

$$\begin{aligned}\mathbb{P}(\mathfrak{l}(m, m_n) \ge s) &\le \Big(2 c_{\mathfrak{g}}^{-1} h_n s^{-(\gamma-\xi)}\Big)^\zeta \sum_{k=0}^{\infty} 2^{-k\zeta(\gamma-\xi)} \\ &= \Big(2 c_{\mathfrak{g}}^{-1} h_n s^{-(\gamma-\xi)}\Big)^\zeta \frac{1}{1 - 2^{-\zeta(\gamma-\xi)}} .\end{aligned}$$

□

Lemma 4.13. Let $\zeta \ge 1$. Assume Moment, Weak Quadruple, and Entropy. Then

$$\mathbb{E}[\Delta_n(\delta)^\zeta] \le c\, \mathfrak{M}(\zeta) \, \Big(\delta^{\frac{\alpha}{\beta}} \eta_{\beta,n}\Big)^\zeta$$

where Y' is an independent copy of Y, $c > 0$ is a constant depending only on $\beta, c_{\mathsf{e}}, \zeta$, and

$$\eta_{\beta,n} := \begin{cases} n^{-\frac{1}{2}} & \text{for } \beta < 1 , \\ n^{-\frac{1}{2}} \log(n+1) & \text{for } \beta = 1 , \\ n^{-\frac{1}{2\beta}} & \text{for } \beta > 1 . \end{cases}$$

Proof. Recall the notation $\mathfrak{c}\overline{yq} := \mathfrak{c}(y, q)$, $F(q) = \mathbb{E}[\mathfrak{c}\overline{Yq}]$, $F_n(q) = \frac{1}{n}\sum_{i=1}^n \mathfrak{c}\overline{Y_i q}$. Define

$$Z_i(q) := \frac{1}{n}\Big(\mathbb{E}\Big[\mathfrak{c}\overline{Yq} - \mathfrak{c}\overline{Ym}\Big] - \mathfrak{c}\overline{Y_i q} + \mathfrak{c}\overline{Y_i m}\Big) .$$

Thus, $\Delta_n(\delta) = \sup_{q \in \mathrm{B}_\delta(m,\mathfrak{l})} \sum_{i=1}^n Z_i(q)$. The Moment condition together with the Weak Quadruple condition imply that Z_i are integrable. Let $(Z_1', \dots, Z_n')$ be an independent copy of $(Z_1, \dots, Z_n)$, where $(Y_1', \dots, Y_n')$ is an independent copy of $(Y_1, \dots, Y_n)$. By Weak Quadruple it holds

$$\begin{aligned}&n^2 \left(Z_i(q) - Z_i(p) - Z_i'(q) + Z_i'(p)\right)^2 \\ &= \Big(\mathfrak{c}\overline{Y_i q} - \mathfrak{c}\overline{Y_i p} - \mathfrak{c}\overline{Y_i' q} + \mathfrak{c}\overline{Y_i' p}\Big)^2 \\ &\le \mathfrak{b}(q,p)^2 \mathfrak{a}(Y_i, Y_i')^2 .\end{aligned}$$

Furthermore,

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\mathfrak{a}(Y_i,Y_i')^2\right)^{\frac{\zeta}{2}}\right] \leq \mathfrak{M}(\zeta)\,.$$

Thus, Theorem 4.24 (appendix 4.F) implies

$$\mathbb{E}[\Delta_n(\delta)^\zeta] \leq c_1\mathfrak{M}(\zeta)\left(\frac{1}{\sqrt{n}}\,\mathsf{entr}_n(\mathrm{B}_\delta(m,\mathfrak{l}),\mathfrak{b})\right)^\zeta,$$

where entr_n is defined in Definition 4.4.

To bound $\mathsf{entr}_n(\mathrm{B}_\delta(m,\mathfrak{l}),\mathfrak{b})$ by applying Lemma 4.25 (appendix 4.F), we need to find an upper bound on $\mathsf{diam}(\mathrm{B}_\delta(m,\mathfrak{l}),\mathfrak{b})$. Set $r_0 := (2c_\mathsf{e}\delta^\alpha)^{\frac{1}{\beta}}$. It fulfills $c_\mathsf{e}\frac{\delta^\alpha}{r_0^\beta} < \sqrt{\log(2)}$. Thus, ENTROPY implies $N(\mathrm{B}_\delta(m,\mathfrak{l}),\mathfrak{b},r_0) < 2$. As the covering number is an integer, $N(\mathrm{B}_\delta(m,\mathfrak{l}),\mathfrak{b},r_0) = 1$, which implies, $\mathsf{diam}(\mathrm{B}_\delta(m,\mathfrak{l}),\mathfrak{b}) \leq 2r_0 =: D_\delta$. Rewriting the `Entropy`-condition in terms of D_δ yields

$$\sqrt{\log N(\mathrm{B}_\delta(m,\mathfrak{l}),\mathfrak{b},r)} \leq c_\beta\left(\frac{D_\delta}{r}\right)^\beta$$

for a constant $c_\beta > 0$ depending only on β and c_e.

Together with Lemma 4.25 (appendix 4.F) we get

$$\mathbb{E}[\Delta_n(\delta)^\zeta] \leq c\mathfrak{M}(\zeta)\left(\delta^{\frac{\alpha}{\beta}}\eta_{\beta,n}\right)^\zeta$$

for a constant $c > 0$ depending only on $\beta, c_\mathsf{e}, \zeta$. □

Proof of Theorem 4.1. Combine Lemma 4.12 and Lemma 4.13. □

Proof of Corollary 4.3. Theorem 4.1 yields

$$\begin{aligned}\eta_{\beta,n}^{-\frac{\kappa}{\gamma-\frac{\alpha}{\beta}}}\,\mathbb{E}[\mathfrak{l}(m,m_n)^\kappa] &= \int_0^\infty \mathbb{P}\left(\eta_{\beta,n}^{-\frac{1}{\gamma-\frac{\alpha}{\beta}}}\,\mathfrak{l}(m,m_n) \geq t^{\frac{1}{\kappa}}\right)\mathrm{d}t \\ &\leq \int_0^\infty \min\left(1, c\,\mathfrak{M}(\zeta)\,t^{-\xi}\right)\mathrm{d}t\,.\end{aligned}$$

In general for $a > 1$, $b > 0$, we have

$$\int \min(1, bt^{-a})\mathrm{d}t = \frac{a}{a-1}b^{\frac{1}{a}}\,.$$

The proof is concluded by applying this statement and noting that $\xi > 1$. □

4.A.2 Proof of Theorem 4.5

To state the next Lemma, which will be used to prove Theorem 4.5, we introduce an intermediate condition, which we call CLOSENESS.

Assumptions.

CLOSENESS:

There is $\xi \in (0, \gamma)$ and a random variable $H_n \geq 0$, such that

$$F(q) - F(m) - F_n(q) + F_n(m) \leq H_n \mathfrak{l}(m, q)^{\xi} \tag{4.7}$$

for all $q \in \mathcal{Q}$ almost surely.

Lemma 4.14. Assume CLOSENESS and GROWTH, and let $\kappa > 0$. Then,

$$\mathbb{E}[\mathfrak{l}(m, m_n)^{\kappa}] \leq c\mathbb{E}\left[H_n^{\frac{\kappa}{\gamma - \xi}}\right],$$

where $c > 0$ depends only on $c_{\mathsf{g}}, \gamma, \xi, \kappa$.

Proof. We use GROWTH and the fact that m_n minimizes F_n to obtain

$$\begin{aligned} c_{\mathsf{g}} \mathfrak{l}(m, m_n)^{\gamma} &\leq F(m_n) - F(m) \\ &\leq F(m_n) - F(m) - F_n(m_n) + F_n(m) \\ &\leq H_n \mathfrak{l}(m, m_n)^{\xi}, \end{aligned}$$

where we applied the CLOSENESS condition in the last step. Thus,

$$c_{\mathsf{g}} \mathfrak{l}(m, m_n)^{\gamma - \xi} \leq H_n,$$

which implies the claimed inequality. □

Define

$$X(q) := \frac{F_n(q) - F_n(m) - F(q) + F(m)}{\mathfrak{l}(m, q)^{\xi}}.$$

Lemma 4.15. Let $\zeta \geq 1$. Assume STRONG MOMENT and STRONG QUADRUPLE. Then

$$\mathbb{E}\left[\sup_{q \in \mathcal{Q}} |X(q)|^{\zeta}\right] \leq c n^{-\frac{\zeta}{2}} \mathfrak{M}(\zeta) \min(\mathsf{entr}_n(\mathcal{Q}, \mathfrak{b}_m), \gamma_2(\mathcal{Q}, \mathfrak{b}_m))^{\zeta},$$

where $c > 0$ is a constant depending only on ζ. Additionally, assume STRONG ENTROPY. Then

$$\mathbb{E}\left[\sup_{q \in \mathcal{Q}} |X(q)|^{\zeta}\right] \leq C \mathfrak{M}(\zeta) D^{\zeta} \eta_{n,\beta}^{\zeta},$$

where $C > 0$ is a constant depending only on $\zeta, \beta, c_{\mathsf{e}}$, and

$$\eta_{\beta,n} := \begin{cases} n^{-\frac{1}{2}} & \text{for } \beta < 1\,, \\ n^{-\frac{1}{2}} \log(n+1) & \text{for } \beta = 1\,, \\ n^{-\frac{1}{2\beta}} & \text{for } \beta > 1\,. \end{cases}$$

Proof. Define

$$Z_i(q) := \frac{1}{n} \frac{\mathfrak{c}\overline{Y_i q} - \mathfrak{c}\overline{Y_i m} - \mathbb{E}\left[\mathfrak{c}\overline{Y q} - \mathfrak{c}\overline{Y m}\right]}{\mathfrak{l}(m,q)^{\xi}}\,.$$

Thus, $X(q) = \sum_{i=1}^n Z_i(q)$. The Strong Moment condition together with the Strong Quadruple condition imply that Z_i integrable. Let $(Z_1', \ldots, Z_n')$ be an independent copy of $(Z_1, \ldots, Z_n)$, where $(Y_1', \ldots, Y_n')$ is an independent copy of $(Y_1, \ldots, Y_n)$. By Strong Quadruple it holds

$$\begin{aligned} & n^2 \left(Z_i(q) - Z_i(p) - Z_i'(q) + Z_i'(p)\right)^2 \\ &= \left(\frac{\mathfrak{c}\overline{Y_i q} - \mathfrak{c}\overline{Y_i m} - \mathfrak{c}\overline{Y_i' q} + \mathfrak{c}\overline{Y_i' m}}{\mathfrak{l}(m,q)^{\xi}} - \frac{\mathfrak{c}\overline{Y_i p} - \mathfrak{c}\overline{Y_i m} - \mathfrak{c}\overline{Y_i' p} + \mathfrak{c}\overline{Y_i' m}}{\mathfrak{l}(m,p)^{\xi}} \right)^2 \\ &\leq \mathfrak{b}_m(q,p)^2 \mathfrak{a}(Y_i, Y_i')^2\,. \end{aligned}$$

Furthermore,

$$\mathbb{E}\left[\left(\frac{1}{n} \sum_{i=1}^n \mathfrak{a}(Y_i, Y_i')^2 \right)^{\frac{\zeta}{2}} \right] \leq \mathfrak{M}(\zeta)$$

with $\mathfrak{M}(\zeta) < \infty$ due to the assumption Strong Moment. Thus, Theorem 4.24 (appendix 4.F) implies

$$\mathbb{E}\left[\sup_{q \in \mathcal{Q}} |X(q)|^{\zeta} \right] \leq c n^{-\frac{\zeta}{2}} \mathfrak{M}(\zeta) \min(\mathsf{entr}_n(\mathcal{Q}, \mathfrak{b}_m), \gamma_2(\mathcal{Q}, \mathfrak{b}_m))^{\zeta}\,.$$

Strong Entropy together with Lemma 4.25 (appendix 4.F) yield

$$\mathbb{E}\left[\sup_{q \in \mathcal{Q}} |X(q)|^{\zeta} \right] \leq C \mathfrak{M}(\zeta) \left(D \eta_{n,\beta}\right)^{\zeta}$$

for a constant $C > 0$ depending only on $\beta, \zeta, c_{\mathsf{e}}$. □

Proof of Theorem 4.5. Using $H_n := \sup_{q \in \mathcal{Q}} |X(q)|$ in Lemma 4.14 fulfills the Closeness condition by definition of X. Next, apply Lemma 4.15 with $\zeta := \frac{\kappa}{\gamma - \xi}$ to conclude the proof. □

4.A.3 Proof of Theorem 4.7

Lemma 4.16. The condition Small Entropy implies

$$\mathsf{entr}_n(\mathrm{B}_R(o,\mathfrak{b}),\mathfrak{b}^{1-\xi}) \le cR^{1-\xi}(1-\xi)^{-\beta}$$

for $\xi \in (0,1)$, where $c>0$ depends only on β, c_{e}.

Proof. Obviously, it holds

$$\mathsf{entr}_n(Q,\mathfrak{b}^{1-\xi}) \le \int_0^\infty \sqrt{\log N(Q,\mathfrak{b}^{1-\xi},r)}\mathrm{d}r$$

for any set $Q \subseteq \mathcal{Q}$. Furthermore,

$$N(Q,\mathfrak{b}^{1-\xi},r) = N(Q,\mathfrak{b},r^{\frac{1}{1-\xi}})\,,$$

which yields

$$\int_0^\infty \sqrt{\log N(Q,\mathfrak{b}^{1-\xi},r)}\mathrm{d}r = (1-\xi)\int_0^\infty s^{-\xi}\sqrt{\log N(Q,\mathfrak{b},s)}\mathrm{d}s$$

Thus, for $Q := \mathrm{B}_R(o,\mathfrak{b})$, we obtain, using the Small Entropy condition,

$$\mathsf{entr}_n(Q,\mathfrak{b}^{1-\xi}) \le c_{\mathsf{e}}(1-\xi)\int_0^R r^{-\xi}\log\Big(\frac{R}{r}\Big)^\beta \mathrm{d}r\,.$$

To calculate the integral, we substitute $s := \frac{r}{R}$ and get

$$\int_0^R r^{-\xi}\log\Big(\frac{R}{r}\Big)^\beta \mathrm{d}r = R^{1-\xi}\int_0^1 s^{-\xi}\log\Big(\frac{1}{s}\Big)^\beta \mathrm{d}s\,.$$

For general $a\in(0,1), b>0$ it holds

$$\int_0^1 x^{-a}\log\Big(\frac{1}{x}\Big)^b \mathrm{d}x = (1-a)^{-b-1}\Gamma(b+1)\,,$$

where $\Gamma(\cdot)$ is the Gamma function. Thus,

$$\int_0^1 s^{-\xi}\log\Big(\frac{1}{s}\Big)^\beta \mathrm{d}s \le c_\beta(1-\xi)^{-\beta-1}$$

for a constant $c_\beta>0$ depending only on β. Putting everything together, we obtain

$$\mathsf{entr}_n(Q,\mathfrak{b}^{1-\xi}) \le cR^{1-\xi}(1-\xi)^{-\beta}\,. \qquad \square$$

Lemma 4.17. Set $\xi_n := 1 - \log(n)^{-1}$. Then

$$\left(n^{-\frac{1}{2}} (1-\xi_n)^{-\beta}\right)^{\frac{1}{\gamma-\xi_n}} \leq c_\gamma n^{-\frac{1}{2(\gamma-1)}} \log(n)^{\frac{\beta}{\gamma-1}}$$

where $c_\gamma > 0$ is a constant depending only on γ.

Proof. We have

$$\begin{aligned}\left(n^{-\frac{1}{2}} (1-\xi_n)^{-\beta}\right)^{\frac{1}{\gamma-\xi_n}} &= \left(n^{-\frac{1}{2(\gamma-\xi_n)}}\right) \log(n)^{\frac{\beta}{\gamma-\xi_n}} \\ &= \left(n^{-\frac{1}{2\left(z+\frac{1}{\log(n)}\right)}}\right) \log(n)^{\frac{\beta}{z+\frac{1}{\log(n)}}},\end{aligned}$$

where $z = \gamma - 1$. We use

$$\log(n)^{\frac{\beta}{z+\frac{1}{\log(n)}}} \leq \log(n)^{\frac{\beta}{z}},$$

$$n^{-\frac{1}{2(z+\frac{1}{\log(n)})}} = n^{-\frac{\log(n)}{2z\log(n)+2}} = \exp\left(-\frac{\log(n)^2}{2z\log(n)+2} + \frac{\log(n)}{2z}\right) n^{-\frac{1}{2z}},$$

and

$$\begin{aligned}-\frac{\log(n)^2}{2z\log(n)+2} + \frac{\log(n)}{2z} &= \frac{\log(n)}{2z(z\log(n)+1)} \\ &\leq \frac{1}{2z^2},\end{aligned}$$

to obtain

$$\left(n^{-\frac{1}{2}} (1-\xi_n)^{-\beta}\right)^{\frac{1}{\gamma-\xi_n}} \leq \exp\left(\frac{1}{2z^2}\right) n^{-\frac{1}{2z}} \log(n)^{\frac{\beta}{z}}. \qquad \square$$

Proof of Theorem 4.7. For $n \in \mathbb{N}, n \geq 3$, set $\xi_n := 1 - \log(n)^{-1}$, $Q_n := \mathrm{B}_{R_n}(o, \mathfrak{b})$, and $R_n := n$. For n large enough, the Existence' condition implies the existence of $m_n^{Q_n} \in \arg\min_{q \in Q_n} F_n(q)$ and $m^{Q_n} \in \arg\min_{q \in Q_n} F(q)$.

Theorem 4.5 implies

$$\mathbb{E}\Big[\mathfrak{b}(m^{Q_n}, m_n^{Q_n})^\kappa\Big] \leq C n^{-\frac{\kappa}{2(\gamma-\xi_n)}} \mathsf{entr}_n(Q_n, \mathfrak{b}^{1-\xi_n})^{\frac{\kappa}{\gamma-\xi_n}} \mathfrak{M}\left(\frac{\kappa}{\gamma-\xi_n}\right),$$

for n large enough. Note, that $C > 0$ can be chosen independently of n (even for ξ_n depending on n).

In Strong Moment we require $\kappa \geq \gamma - 1$, because then $x \mapsto x^{\frac{\kappa}{\gamma-1}}$ is convex, which is needed for the symmetrization argument in the proof of Theorem 4.5. But, if $\kappa = \gamma - 1$, then $\frac{\kappa}{\gamma-\xi_n} < 1$, and Theorem 4.5 cannot be applied directly. For this technical reason, we assumed $\kappa > \gamma - 1$, so that $\kappa \geq \gamma - \xi_n$ for n large enough.

By Small Entropy and Lemma 4.16 there is $c_\beta > 0$ such that for $n \in \mathbb{N}$ large enough, it holds

$$\mathsf{entr}_n(\mathrm{B}_{R_n}(o, \mathfrak{b}), \mathfrak{b}^{1-\xi_n}) \leq c_\beta R_n^{1-\xi_n} (1-\xi_n)^{-\beta} .$$

Using $R_n^{1-\xi_n} = n^{\frac{1}{\log(n)}} = \exp(1)$ together with Lemma 4.17, we obtain

$$\mathbb{E}\Big[\mathfrak{b}(m^{Q_n}, m_n^{Q_n})^\kappa\Big] \leq C' \left(n^{-\frac{1}{2}} \log(n)^\beta\right)^{-\frac{\kappa}{\gamma-1}} \mathfrak{M}\Big(\frac{\kappa}{\gamma-\xi_n}\Big) .$$

As $\lim_{n\to\infty} \mathfrak{M}\Big(\frac{\kappa}{\gamma-\xi_n}\Big) = \mathfrak{M}\Big(\frac{\kappa}{\gamma-1}\Big)$, we have

$$\mathbb{E}\Big[\mathfrak{b}(m^{Q_n}, m_n^{Q_n})^\kappa\Big] \leq C'' n^{-\frac{\kappa}{2(\gamma-\xi_n)}} \left(R_n^{1-\xi_n} (1-\xi_n)^{-\beta}\right)^{\frac{\kappa}{\gamma-\xi_n}} \mathfrak{M}\Big(\frac{\kappa}{\gamma-1}\Big) .$$

Finally, there is a $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$ it holds $m \in Q_n$, which implies $m = m^{Q_n}$. Thus,

$$\mathbb{E}\Big[\mathfrak{b}(m, m_n^{Q_n})^\kappa\Big] = \mathbf{O}\left(\left(n^{-\frac{1}{2}} \log(n)^\beta\right)^{-\frac{\kappa}{\gamma-1}}\right) . \qquad \square$$

4.B Stability of Quadruple Inequalities

We present some trivial stability results for quadruple inequalities. The notation we use here is introduced in the beginning of section 4.3.

Subsets:
If $(\mathcal{Q}, \mathcal{Y}, \mathfrak{c}, \mathfrak{a}, \mathfrak{b})$ fulfills the weak quadruple inequality, then so does $(\mathcal{Q}', \mathcal{Y}', \mathfrak{c}, \mathfrak{a}, \mathfrak{b})$ with $\mathcal{Q}' \subseteq \mathcal{Q}$, $\mathcal{Y}' \subseteq \mathcal{Y}$.

Images:
Assume $(\mathcal{Q}, \mathcal{Y}, \mathfrak{c}, \mathfrak{a}, \mathfrak{b})$ fulfills the weak quadruple inequality and $f\colon \mathcal{Y}' \to \mathcal{Y}$, $g\colon \mathcal{Q}' \to \mathcal{Q}$. Then $(\mathcal{Q}', \mathcal{Y}', \mathfrak{c}', \mathfrak{a}', \mathfrak{b}')$ fulfills the weak quadruple inequality with $\mathfrak{c}'(y,q) = \mathfrak{c}(f(y), g(q))$, $\mathfrak{a}'(y,z) = \mathfrak{a}(f(y), f(z))$, $\mathfrak{b}'(q,p) = \mathfrak{b}(g(q), g(p))$.

Limits:
Let $(\mathcal{Q}, \mathcal{Y}, \mathfrak{c}_i, \mathfrak{a}_i, \mathfrak{b}_i)$ fulfill the weak quadruple inequality for $i \in \mathbb{N}$ and assume for all $q, p \in \mathcal{Q}$ and $y, z \in \mathcal{Y}$ the point-wise limits

$$\begin{aligned}
\mathfrak{a}(y,z) &:= \lim_{i\to\infty} \mathfrak{a}_i(y,z) \\
\mathfrak{b}(q,p) &:= \lim_{i\to\infty} \mathfrak{b}_i(q,p) \\
\mathfrak{c}(y,q) &:= \lim_{i\to\infty} \mathfrak{c}_i(y,q)
\end{aligned}$$

exist. Then $(\mathcal{Q}, \mathcal{Y}, \mathfrak{c}, \mathfrak{a}, \mathfrak{b})$ also fulfills the weak quadruple inequality.

Similar results hold for the strong quadruple inequality. For the following results it may not be so easy to obtain an analog for the strong quadruple inequality.

Product Spaces:

If $(\mathcal{Q}_i, \mathcal{Y}_i, \mathfrak{c}_i, \mathfrak{a}_i, \mathfrak{b}_i)$ fulfill the weak quadruple inequality for all $i \in \mathbb{N}$, then so does $(\mathcal{Q}, \mathcal{Y}, \mathfrak{c}, \mathfrak{a}, \mathfrak{b})$ where $\mathcal{Q} = \bigtimes_{i\in\mathbb{N}} \mathcal{Q}_i$, $\mathcal{Y} = \bigtimes_{i\in\mathbb{N}} \mathcal{Y}_i$, $\mathfrak{c} = \sum_{i=1}^{\infty} \mathfrak{c}_i$, $\mathfrak{a} = \|(\mathfrak{a}_i)_{i\in\mathbb{N}}\|_{\ell^2}$, $\mathfrak{b} = \|(\mathfrak{b}_i)_{i\in\mathbb{N}}\|_{\ell^2}$.

Proof. It holds

$$\begin{aligned}
{}^{\mathfrak{c}}\overline{yq} - {}^{\mathfrak{c}}\overline{zq} - {}^{\mathfrak{c}}\overline{yp} + {}^{\mathfrak{c}}\overline{zp} &= \sum_{i=1}^{\infty} ({}^{\mathfrak{c}_i}\overline{y_i,q_i} - {}^{\mathfrak{c}_i}\overline{z_i,q_i} - {}^{\mathfrak{c}_i}\overline{y_i,p_i} + {}^{\mathfrak{c}_i}\overline{z_i,p_i}) \\
&\leq \sum_{i=1}^{\infty} \mathfrak{a}_i(y_i, z_i)\mathfrak{b}_i(q_i, p_i) \\
&\leq \mathfrak{a}(y, z)\mathfrak{b}(q, p)\,,
\end{aligned}$$

using the Cauchy–Schwarz inequality. □

Measure Spaces:

Let $(\Omega, \mathcal{A}, \mu)$ be a measure space. Assume $(\mathcal{Q}, \mathcal{Y}, \mathfrak{c}(\omega), \mathfrak{a}(\omega), \mathfrak{b}(\omega))$ fulfills the weak quadruple inequality for every $\omega \in \Omega$. Let $s, t > 0$ with $\frac{1}{s} + \frac{1}{t} = 1$. Let $L(\Omega, \mathcal{Q})$ be the set of measurable functions from Ω to $\mathcal{Q}$, define $L(\Omega, \mathcal{Y})$ analogously. For $q, p \in L(\Omega, \mathcal{Q})$, $y, z \in L(\Omega, \mathcal{Y})$, let

$$\begin{aligned}
\mathfrak{C}(y, q) &:= \int \mathfrak{c}(\omega, y(\omega), q(\omega))\,\mathrm{d}\mu(\omega)\,, \\
\mathfrak{A}(y, z) &:= \left(\int \mathfrak{a}(\omega; y(\omega), z(\omega))^t\,\mathrm{d}\mu(\omega)\right)^{\frac{1}{t}}\,, \\
\mathfrak{B}(q, p) &:= \left(\int \mathfrak{b}(\omega; q(\omega), p(\omega))^s\,\mathrm{d}\mu(\omega)\right)^{\frac{1}{s}}\,,
\end{aligned}$$

where we implicitly assume that the necessary measurablity and integrability conditions are fulfilled. Then

$$(L(\Omega, \mathcal{Q}), L(\Omega, \mathcal{Y}), \mathfrak{C}, \mathfrak{A}, \mathfrak{B})$$

also fulfills the quadruple inequality.

Proof. It holds

$$\begin{aligned}
&\mathfrak{C}(y, q) - \mathfrak{C}(z, q) - \mathfrak{C}(y, p) + \mathfrak{C}(z, p) \\
&= \int \mathfrak{c}(\omega; y(\omega), q(\omega)) - \mathfrak{c}(\omega; y(\omega), p(\omega)) \\
&\qquad - \mathfrak{c}(\omega; z(\omega), q(\omega)) + \mathfrak{c}(\omega; z(\omega), p(\omega))\mathrm{d}\mu(\omega) \\
&\leq \int \mathfrak{a}(\omega; y(\omega), z(\omega))\mathfrak{b}(\omega; q(\omega), p(\omega))\mathrm{d}\mu(\omega) \\
&\leq \mathfrak{A}(y, z)\mathfrak{B}(q, p)\,,
\end{aligned}$$

by Hölder's inequality. □

Minima:

Let $(\mathcal{Q}, \mathcal{Y}, \mathfrak{c}, \mathfrak{a}, \mathfrak{b})$ fulfill the weak quadruple inequality. Let $\tilde{\mathcal{Y}} \subseteq 2^{\mathcal{Y}}$. Define the cost function $\mathfrak{C}\colon \tilde{\mathcal{Y}} \times \mathcal{Q} \to \mathbb{R}$ by $\mathfrak{C}(\mathbf{y}, q) = \inf_{y \in \mathbf{y}} \mathfrak{c}(y, q)$ and $\mathfrak{A}(\mathbf{y}, \mathbf{z}) = \sup_{y \in \mathbf{y}, z \in \mathbf{z}} \mathfrak{a}(y, z)$ assuming the infinma and suprema are finite. Then $(\mathcal{Q}, \tilde{\mathcal{Y}}, \mathfrak{C}, \mathfrak{A}, \mathfrak{b})$ fulfills the weak quadruple inequality.

Proof. Let $\mathbf{y}, \mathbf{z} \in \tilde{\mathcal{Y}}$ and $q, p \in \mathcal{Q}$. Assume there are $y_q, y_p \in \mathbf{y}, z_q, z_p \in \mathbf{z}$ such that $\mathfrak{C}(\mathbf{y}, q) = {}^{\mathfrak{c}}\overline{y_q q}$, $\mathfrak{C}(\mathbf{y}, p) = {}^{\mathfrak{c}}\overline{y_p p}$, $\mathfrak{C}(\mathbf{z}, q) = {}^{\mathfrak{c}}\overline{z_q q}$, and $\mathfrak{C}(\mathbf{z}, p) = {}^{\mathfrak{c}}\overline{z_p p}$. Then

$$\begin{aligned}
\mathfrak{C}(\mathbf{y}, q) - \mathfrak{C}(\mathbf{y}, p) - \mathfrak{C}(\mathbf{z}, q) + \mathfrak{C}(\mathbf{z}, p) &= {}^{\mathfrak{c}}\overline{y_q q} - {}^{\mathfrak{c}}\overline{y_p p} - {}^{\mathfrak{c}}\overline{z_q q} + {}^{\mathfrak{c}}\overline{z_p p} \\
&\leq {}^{\mathfrak{c}}\overline{y_p q} - {}^{\mathfrak{c}}\overline{y_p p} - {}^{\mathfrak{c}}\overline{z_q q} + {}^{\mathfrak{c}}\overline{z_q p} \\
&\leq \mathfrak{a}(y_p, z_q)\mathfrak{b}(q, p) \\
&\leq \mathfrak{A}(\mathbf{y}, \mathbf{z})\mathfrak{b}(q, p)\,.
\end{aligned}$$

If the infima are not attained, one can follow the same proof with minimizing sequences. □

In many interesting problems the setting is opposite to what was described before, i.e., $\mathfrak{C}\colon \mathcal{Y} \times \tilde{\mathcal{Q}} \to \mathbb{R}$, $(y, \mathbf{q}) \mapsto \inf_{q \in \mathbf{q}} \mathfrak{c}(y, q)$, where $\tilde{\mathcal{Q}} \subseteq 2^{\mathcal{Q}}$: the elements of the descriptor space are subsets and the elements of data space are points. Examples are k-means, where $\tilde{\mathcal{Q}}$ consists of k-tuples of points in $\mathcal{Q}$, or fitting hyperplanes. A quadruple inequality with $\sup_{q \in \mathbf{q}, p \in \mathbf{p}} \mathfrak{b}(q, p)$ as the descriptor distance can be established. Unfortunately, this is usually not useful, as the entropy condition cannot be fulfilled with distances of this type. The framework described in this chapter can still be applied using inequalities as for bounded spaces, see section 4.3.1. But we cannot directly use the advantage of quadruple inequalities over Lipschitz-continuity.

4.C Proof of Lemma 4.6

We first state and prove two simple lemmas for some simple arithmetic expressions and then use those for the proof of Lemma 4.6.

Lemma 4.18. Let $A, B \in \mathbb{R}$, $a, b, c, r \geq 0$, $s, t > 0$. Assume $t \geq s \;\Leftrightarrow\; b \geq a$. Assume $|A| \leq ra$, $|B| \leq rb$, $|A - B| \leq rc$. Then

$$\left|\frac{A}{s} - \frac{B}{t}\right| \leq r\frac{\min(s,t)c + |s-t|\min(a,b)}{st}\,.$$

Proof. For $t \geq s$, using the bound on A and on $A - B$ implies $\frac{A}{s} - \frac{B}{t} \leq r\frac{(t-s)a+sc}{st}$. Similarly, for $s \geq t$, we get $\frac{A}{s} - \frac{B}{t} \leq r\frac{tc+(s-t)b}{st}$ by using the bound on B and $A - B$.

Together, we obtain

$$\frac{A}{s} - \frac{B}{t} \leq r \frac{\min(s,t)c + |s-t| \min(a,b)}{st} .$$

We finish the proof by pointing out the symmetry between (A, a, s) and (B, b, t). □

Lemma 4.19. Let $a, b, c > 0$, $\beta \in [0,1]$. Assume $a \leq b$, $b \leq a + c$, $c \leq a + b$. Then

$$\frac{ca^\beta + \left(b^\beta - a^\beta\right) a}{a^\beta b^\beta} \leq 2^\beta c^{1-\beta} .$$

Proof. The statement is trivial for $\beta \in \{0, 1\}$. So let $\beta \in (0, 1)$.
Case I, $c \leq a$: Define $f(x) = 1 - x - (1+x)^\beta (1 - x^{1-\beta})$. Then $f(0) = f(1) = 0$ and $f''(x) = -(1-\beta)\beta x^{-\beta-1}(x+1)^{\beta-2}(1 - x^{\beta+1}) \leq 0$ for $x \in (0,1)$. Thus, $f(x) \geq 0$ for $x \in [0,1]$. In particular $f(\frac{c}{a}) \geq 0$, which implies $a - c \geq (a+c)^\beta (a^{1-\beta} - c^{1-\beta}) \geq b^\beta (a^{1-\beta} - c^{1-\beta})$. Thus,

$$\frac{ca^\beta + \left(b^\beta - a^\beta\right) a}{a^\beta b^\beta} \leq c^{1-\beta} .$$

Case II, $c \geq a$: As $1 - \beta \leq 1$ and $c - a \geq 0$, we have $(c-a)^{1-\beta} + a^{1-\beta} \leq 2^\beta c^{1-\beta}$. Multiplying by $(c-a)^\beta$ and using $c - a \leq b$, we get $c - a \leq b^\beta \left(2^\beta c^{1-\beta} - a^{1-\beta}\right)$. Thus,

$$\frac{ca^\beta + \left(b^\beta - a^\beta\right) a}{a^\beta b^\beta} \leq 2^\beta c^{1-\beta} .$$ □

Proof of Lemma 4.6. Applying Lemma 4.18 to the left hand side of equation (4.4), yields

$$\frac{\mathfrak{c}\overline{yq} - \mathfrak{c}\overline{ym} - \mathfrak{c}\overline{zq} + \mathfrak{c}\overline{zm}}{\mathfrak{b}(q,m)^\xi} - \frac{\mathfrak{c}\overline{yp} - \mathfrak{c}\overline{ym} - \mathfrak{c}\overline{zp} + \mathfrak{c}\overline{zm}}{\mathfrak{b}(p,m)^\xi} \leq \mathfrak{a}(y,z)\, \tilde{\mathfrak{b}}_{m,\xi}(q,p)$$

where

$$\tilde{\mathfrak{b}}_{m,\xi}(q,p) = \frac{\min(\mathfrak{b}\overline{qm}, \mathfrak{b}\overline{pm})^\xi\, \mathfrak{b}\overline{qp} + \left|\mathfrak{b}\overline{qm}^\xi - \mathfrak{b}\overline{pm}^\xi\right| \min(\mathfrak{b}\overline{qm}, \mathfrak{b}\overline{pm})}{\mathfrak{b}\overline{qm}^\xi\, \mathfrak{b}\overline{pm}^\xi}$$

with the short notation $\mathfrak{b}\overline{qp} := \mathfrak{b}(q,p)$, for all $y, z, q, p, m \in \mathcal{Q}$. Applying Lemma 4.19 yields $\tilde{\mathfrak{b}}_{m,\xi}(q,p) \leq 2^\xi \mathfrak{b}(q,p)^{1-\xi}$. □

4.D Projection Metric Counter Example

We take a tripod $(\mathcal{Q}, d)$ as a simple example of a non-Euclidean Hadamard space, see [Stu03, Example 3.2], and show that it does not fulfill the strong quadruple inequality

	y	z	q	p	m
pod number	1	1	2	1	3
distance from center	ε	0	r	ε	r

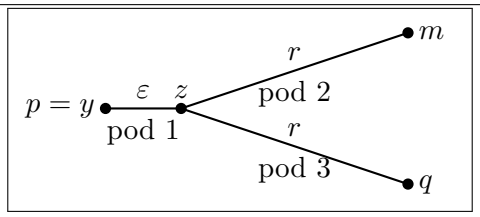

Figure 4.3: Tripod counter-example for the strong quadruple inequality

with the projection metric

$$d_m^{\mathsf{proj}}(q,p) := \sqrt{\frac{\overline{q,p}^2 - (\overline{q,m} - \overline{p,m})^2}{\overline{q,m}\,\overline{p,m}}}\,.$$

Let $r > \varepsilon > 0$ and define y, z, q, p, o on a tripod as in Figure 4.3. We take $\mathfrak{c} = d^2$, $\xi = 1$, $\mathfrak{l} = d$, $\mathfrak{a} = Kd$, and $\mathfrak{b}_m = d_m^{\mathsf{proj}}$. Then

$$\frac{\mathfrak{c}\overline{yq} - \mathfrak{c}\overline{ym} - \mathfrak{c}\overline{zq} + \mathfrak{c}\overline{zm}}{\mathfrak{l}(q,m)} - \frac{\mathfrak{c}\overline{yp} - \mathfrak{c}\overline{ym} - \mathfrak{c}\overline{zp} + \mathfrak{c}\overline{zm}}{\mathfrak{l}(p,m)} = 2\varepsilon\,,$$

$$\mathfrak{a}(y,z)\mathfrak{b}_m(q,p) = K\varepsilon\sqrt{2\frac{\varepsilon}{r+\varepsilon}}\,.$$

If the strong quadruple inequality holds, then

$$K \geq \sqrt{2\frac{r+\varepsilon}{\varepsilon}} \xrightarrow{\varepsilon \searrow 0} \infty\,.$$

Thus, d_m^{proj} is not a suitable candidate for the strong quadruple distance in general Hadamard spaces.

4.E Optimality of Power Inequality

We show that $8\alpha 2^{-2\alpha}$ is the optimal constant, and that we cannot extend Theorem 4.10 to $\alpha > 1$ or $\alpha < \frac{1}{2}$. Let $\epsilon \in (0,1)$ and $(\mathcal{Q}, d)$ be a metric space with $q, p, y, z \in \mathcal{Q}$ such that for each case below the distances have the values written down in Table 4.1. One can easily show that in all three cases the necessary triangle inequalities and the nice quadruple inequality hold.

Case	$\overline{y,q}$	$\overline{y,p}$	$\overline{z,q}$	$\overline{z,p}$	$\overline{y,z}$	$\overline{q,p}$
(a)	$1-\epsilon$	$1-3\epsilon$	$1-2\epsilon$	1	$2-3\epsilon$	2ϵ
(b)	1	ϵ	$1-\epsilon$	2ϵ	2ϵ	1
(c)	2ϵ	ϵ	1	1	1	ϵ

Table 4.1: Distances of four points $y, z, q, p \in \mathcal{Q}$ for showing lower bounds of the constant in Theorem 4.10.

(a) For $\alpha \in [\frac{1}{2}, 1]$ it holds

$$\begin{aligned}
&\lim_{\epsilon \searrow 0} \frac{\overline{y,q}^{2\alpha} - \overline{y,p}^{2\alpha} - \overline{z,q}^{2\alpha} + \overline{z,p}^{2\alpha}}{\overline{y,z}^{2\alpha-1}\,\overline{q,p}} \\
&= \lim_{\epsilon \searrow 0} \frac{(1-\epsilon)^{2\alpha} - (1-3\epsilon)^{2\alpha} - (1-2\epsilon)^{2\alpha} + 1}{(2\epsilon)(2-3\epsilon)^{2\alpha-1}} \\
&= \lim_{\epsilon \searrow 0} 2\alpha \frac{-(1-\epsilon)^{2\alpha-1} + 3(1-2\epsilon)^{2\alpha-1} + 2(1-3\epsilon)^{2\alpha-1}}{2(2-3\epsilon)^{2\alpha-1} + 2\epsilon(2\alpha-1)(2-3\epsilon)^{2\alpha-2}} \\
&= 8\alpha 2^{-2\alpha}\,.
\end{aligned}$$

Thus, the constant $8\alpha 2^{-2\alpha}$ in Theorem 4.10 is optimal.

(b) For $\alpha > 1$ it holds

$$\begin{aligned}
&\lim_{\epsilon \searrow 0} \frac{\overline{y,q}^{2\alpha} - \overline{y,p}^{2\alpha} - \overline{z,q}^{2\alpha} + \overline{z,p}^{2\alpha}}{\overline{y,z}^{2\alpha-1}\,\overline{q,p}} \\
&= \lim_{\epsilon \searrow 0} \frac{-2\epsilon^{2\alpha-1} + 2(1-\epsilon)^{2\alpha-1} + 2(2\epsilon)^{2\alpha-1}}{2^{2\alpha-1}\epsilon^{2\alpha-2}} \\
&= \infty\,.
\end{aligned}$$

Thus, there is no power inequality in the form of Theorem 4.10 for $\alpha > 1$.

(c) For $\alpha \in (0, \frac{1}{2})$ it holds

$$\begin{aligned}
\lim_{\epsilon \searrow 0} \frac{\overline{y,q}^{2\alpha} - \overline{y,p}^{2\alpha} - \overline{z,q}^{2\alpha} + \overline{z,p}^{2\alpha}}{\overline{y,z}^{2\alpha-1}\,\overline{q,p}} &= \lim_{\epsilon \searrow 0} \frac{(2\epsilon)^{2\alpha} - \epsilon^{2\alpha} - 1 + 1}{2\epsilon} \\
&= \lim_{\epsilon \searrow 0} \frac{1}{2}(2^{2\alpha} - 1)\epsilon^{2\alpha-1} \\
&= \infty\,.
\end{aligned}$$

Thus, there is no power inequality in the form of Theorem 4.10 for $\alpha < \frac{1}{2}$.

4.F Chaining

Recall the measures of entropy γ_2 and entr_n defined in Definition 4.4. We add another useful entry to this list.

Definition 4.20 (Bernoulli Bound)**.** For $T \subseteq \mathbb{R}^n$ define

$$b(T) := \inf\left\{ \sup_{t\in T_1} \|t\|_1 + \gamma_2(T_2) : \; T_1, T_2 \subseteq \mathbb{R}^n, T \subseteq T_1 + T_2 \right\},$$

where $\gamma_2(T_2) = \gamma_2(T_2, d_2)$ for the Euclidean metric d_2 on $\mathbb{R}^n$, $\|t\|_1 = \sum_{i=1}^n |t_i|$, and $T_1 + T_2 = \{t_1 + t_2 \colon t_1 \in T_1, t_2 \in T_2\}$.

We write down the Bernoulli bound for powers of the Bernoulli process. [BL14] show that the bound can be reversed (up to an universal constant). Thus, this step can be regarded as optimal.

Theorem 4.21 (Bernoulli bound)**.** Let $\sigma_1, \dots, \sigma_n$ be independent random signs, i.e., $\mathbb{P}(\sigma_i = \pm 1) = \frac{1}{2}$. For $t \in R^n$ set $\tilde{X}_t := \sum_{i=1}^n \sigma_i t_i$. Let $T \subseteq \mathbb{R}^n$. Let $\kappa \geq 1$. Then

$$\mathbb{E}\left[\sup_{t\in T} \left| \tilde{X}_t \right|^\kappa \right] \leq c_\kappa b(T)^\kappa ,$$

where c_κ depends only on κ.

Proof. Let $T_1, T_2 \subseteq \mathbb{R}^n$ such that $T \subseteq T_1 + T_2$. As $(a+b)^\kappa \leq 2^{\kappa-1}(a^\kappa + b^\kappa)$ for all $a, b \geq 0$, we can split the supremum into two parts,

$$\mathbb{E}\left[\sup_{t\in T} \left| \tilde{X}_t \right|^\kappa \right] \leq 2^{\kappa-1} \left(\mathbb{E}\left[\sup_{t\in T_1} \left| \tilde{X}_t \right|^\kappa \right] + \mathbb{E}\left[\sup_{t\in T_2} \left| \tilde{X}_t \right|^\kappa \right] \right) .$$

The first term is bounded using the 1-norm, $\mathbb{E}\left[\sup_{t\in T_1} \left| \tilde{X}_t \right|^\kappa \right] \leq \sup_{t\in T_1} \|t\|_1^\kappa$. For the second we use Talagrand's generic chaining bound for the supremum of the subgaussian process $\mathbb{E}\left[\sup_{t\in T_2} \left| \tilde{X}_t \right|^\kappa \right] \leq c'_\kappa \gamma_2(T_2)^\kappa$, see [Tal14]. We obtain

$$\mathbb{E}\left[\sup_{t\in T} \left| \tilde{X}_t \right|^\kappa \right] \leq c_\kappa \left(\sup_{t\in T_1} \|t\|_1^\kappa + \gamma_2(T_2)^\kappa \right) \leq c_\kappa \left(\sup_{t\in T_1} \|t\|_1 + \gamma_2(T_2) \right)^\kappa . \qquad \square$$

Lemma 4.22 (Lipschitz connection)**.** Let $(\mathcal{Q}, b)$ be a pseudo-metric space. Assume there are function $f_i \colon \mathcal{Q} \to \mathbb{R}$ such that $|f_i(q) - f_i(p)| \leq a_i b(q,p)$. Let $T := \{(f_i(q))_{i=1,\dots,n} \colon q \in \mathcal{Q}\}$. Set $a = (a_1, \dots, a_n)$. Then

$$b(T) \leq C\|a\|_2 \min(\mathsf{entr}_n(\mathcal{Q}, b), \gamma_2(\mathcal{Q}, b)) ,$$

where $C > 0$ is an universal constant.

Proof. For $\epsilon > 0$, choose $\mathcal{Q}_2$ to be an ϵ-covering of $\mathcal{Q}$ with respect to b, i.e., for all $q \in \mathcal{Q}$ there is a $p_q \in \mathcal{Q}_2$ such that $b(q, p_q) \leq \epsilon$. For $q \in \mathcal{Q}$ denote $t_q := (f_i(q))_{i=1,\dots,n} \in \mathbb{R}^n$. Define $T_2 := \{t_p \colon p \in \mathcal{Q}_2\}$ and $T_1 := \{t_q - t_{p_q} \colon q \in Q\}$. Then $T \subseteq T_1 + T_2$. The Lipschitz-condition implies $\|t_q - t_p\|_2 \leq \|a\|_2 b(q,p)$ for all $q, p \in \mathcal{Q}$. Thus,

$$\sup_{t \in T_1} \|t\|_1 \leq \sup_{q \in \mathcal{Q}} \sqrt{n} \|t_q - t_{p_q}\|_2 \leq \epsilon \sqrt{n} \|a\|_2 \,.$$

By the properties of γ_2, see [Tal14], we obtain

$$\gamma_2(T_2) \leq c \|a\|_2 \gamma_2(\mathcal{Q}_2, b) \leq c' \|a\|_2 \int_\epsilon^\infty \sqrt{\log N(Q, b, r)} \mathrm{d}r$$

for universal constants $c, c' > 0$. Applying the two inequalities to the definition of $b(T)$ concludes the proof. □

Lemma 4.23 (Symmetrization)**.** Let $\mathcal{Q}$ be set. Let $Z_1, \dots, Z_n$ be centered, independent, and integrable stochastic processes indexed by $\mathcal{Q}$. Let $\Phi \colon \mathbb{R} \to \mathbb{R}$ be a convex, nondecreasing function. Let $(Z_1', \dots, Z_n')$ be an independent copy of $(Z_1, \dots, Z_n)$. Let $\varepsilon_1, \dots, \varepsilon_n$ be iid with $\mathbb{P}(\varepsilon_1 = \pm 1) = \frac{1}{2}$. Then

$$\mathbb{E}\left[\sup_{q \in \mathcal{Q}} \Phi\left(\sum_{i=1}^n Z_i(q)\right)\right] \leq \mathbb{E}\left[\sup_{q \in \mathcal{Q}} \Phi\left(\sum_{i=1}^n \varepsilon_i \left(Z_i(q) - Z_i'(q)\right)\right)\right],$$

assuming measurability of the involved terms.

The symmetrization lemma is well-known. The statement here is an intermediate step of from the proof of [VW96, 2.3.6 Lemma].

Theorem 4.24 (Empirical process bound)**.** Let $(\mathcal{Q}, b)$ be a separable pseudo-metric space. Let $Z_1, \dots, Z_n$ be centered, independent, and integrable stochastic processes indexed by $\mathcal{Q}$ with a $q_0 \in \mathcal{Q}$ such that $Z_i(q_0) = 0$ for $i = 1, \dots, n$. Let $(Z_1', \dots, Z_n')$ be an independent copy of $(Z_1, \dots, Z_n)$. Assume the following Lipschitz-property: There is a random vector A with values in $\mathbb{R}^n$ such that

$$|Z_i(q) - Z_i(p) - Z_i'(q) + Z_i'(p)| \leq A_i b(q,p)$$

for $i = 1, \dots, n$ and all $q, p \in \mathcal{Q}$. Let $\kappa \geq 1$. Then

$$\mathbb{E}\left[\sup_{q \in \mathcal{Q}} \left|\sum_{i=1}^n Z_i(q)\right|^\kappa\right] \leq C\, \mathbb{E}[\|A\|_2^\kappa]\, \min(\mathsf{entr}_n(\mathcal{Q}, b), \gamma_2(\mathcal{Q}, b))^\kappa \,,$$

where $C > 0$ is an universal constant.

Proof. Use Lemma 4.23. Then apply Theorem 4.21 and Lemma 4.22 conditionally on $Z_1, \ldots, Z_n$. □

Lemma 4.25. Let $(\mathcal{Q}, b)$ be a pseudo-metric space. Let $D > 0$ such that $\mathsf{diam}(\mathcal{Q}, b) \leq D < \infty$. Let $\beta > 0$. Assume

$$\sqrt{\log(N(\mathcal{Q}, b, r))} \leq c_{\mathsf{e}} \left(\frac{D}{r}\right)^{\beta}$$

for all $0 < r < D$.

(i) If $\beta < 1$ then $\mathsf{entr}_n(\mathcal{Q}, b) \leq \frac{c_{\mathsf{e}} D}{1-\beta}$.

(ii) If $\beta = 1$ then $\mathsf{entr}_n(\mathcal{Q}, b) \leq c'_{\mathsf{e}} D \log(n+1)$, where $c' > 0$ depends only on c_{e}.

(iii) If $\beta > 1$ then $\mathsf{entr}_n(\mathcal{Q}, b) \leq c_{\mathsf{e}}^{\frac{1}{\beta}} \frac{\beta}{\beta-1} n^{-\frac{1}{2\beta}+\frac{1}{2}}$.

In particular

$$n^{-\frac{1}{2}}\, \mathsf{entr}_n(\mathcal{Q}, b) \;\leq\; c\, D\, \eta_{\beta,n}\,,$$

where c depends only on c_{e} and β and

$$\eta_{\beta,n} := \begin{cases} n^{-\frac{1}{2}} & \text{for } \beta < 1\,, \\ n^{-\frac{1}{2}} \log(n+1) & \text{for } \beta = 1\,, \\ n^{-\frac{1}{2\beta}} & \text{for } \beta > 1\,. \end{cases}$$

The proof consists of calculating the entropy integral with the given bound on the covering numbers and, for $\beta \geq 1$, choosing the minimizing starting point of the integral $\epsilon > 0$.

4.G Proof of the Power Inequality, Theorem 4.10

Let $(\mathcal{Q}, d)$ be a metric space. Use the short notation $\overline{q,p} := d(q,p)$. Let $q, p, y, z \in \mathcal{Q}$, $\alpha \in [\frac{1}{2}, 1]$. Assume

$$\overline{yq}^2 - \overline{yp}^2 - \overline{zq}^2 + \overline{zp}^2 \leq 2\, \overline{y,z}\, \overline{q,p}\,.$$

The goal of this section is to prove

$$\overline{y,q}^{2\alpha} - \overline{y,p}^{2\alpha} - \overline{z,q}^{2\alpha} + \overline{z,p}^{2\alpha} \leq 8\alpha 2^{-2\alpha}\, \overline{y,z}^{2\alpha-1}\, \overline{q,p}\,.$$

4.G.1 Arithmetic Form

Theorem 4.10 will be proven in the form of Lemma 4.26.

Lemma 4.26. Let $a, b, c \geq 0$, $r, s \in [-1, 1]$, and $\alpha \in [\frac{1}{2}, 1]$. Then

$$\begin{aligned} &a^{2\alpha} - c^{2\alpha} - \left(a^2 - 2rab + b^2\right)^{\alpha} + \left(c^2 - 2scb + b^2\right)^{\alpha} \\ &\leq 8\alpha 2^{-2\alpha} b \max(ra - sc, |a - c|)^{2\alpha - 1} . \end{aligned}$$

The advantage of using Lemma 4.26 to prove Theorem 4.10 is, that we do not need to consider a system of additional conditions for describing that the real values in the inequality are distances, which have to fulfill the triangle inequality. The disadvantage is, that we loose the possibility for a geometric interpretation of the proof.

Lemma 4.27. Lemma 4.26 implies Theorem 4.10.

Proof. Three points from an arbitrary metric space can be embedded in the Euclidean plane so that the distances are preserved. Thus, the cosine formula of Euclidean geometry can be applied to the three points $y, p, q \in \mathcal{Q}$: It holds

$$\overline{y,q}^2 = \overline{y,p}^2 + \overline{q,p}^2 - 2s\,\overline{y,p}\,\overline{q,p}\,,$$

where $s := \cos(\measuredangle ypq)$ with the angle $\measuredangle ypq$ in the Euclidean plane. Similarly

$$\overline{z,q}^2 = \overline{z,p}^2 + \overline{q,p}^2 - 2r\,\overline{z,p}\,\overline{q,p}\,,$$

where $r := \cos(\measuredangle zpq)$. Thus,

$$\begin{aligned} &\overline{y,q}^{2\alpha} - \overline{y,p}^{2\alpha} - \overline{z,q}^{2\alpha} + \overline{z,p}^{2\alpha} \\ &= \left(\overline{y,p}^2 + \overline{q,p}^2 - 2s\,\overline{y,p}\,\overline{q,p}\right)^{\alpha} - \left(\overline{z,p}^2 + \overline{q,p}^2 - 2r\,\overline{z,p}\,\overline{q,p}\right)^{\alpha} - \overline{y,p}^{2\alpha} + \overline{z,p}^{2\alpha} \\ &= \left(c^2 + b^2 - 2scb\right)^{\alpha} - \left(a^2 + b^2 - 2rab\right)^{\alpha} - c^{2\alpha} + a^{2\alpha}\,, \end{aligned}$$

where $a := \overline{z,p}$, $c := \overline{y,p}$, $b := \overline{q,p}$. Hence, Lemma 4.26 yields

$$\overline{y,q}^{2\alpha} - \overline{y,p}^{2\alpha} - \overline{z,q}^{2\alpha} + \overline{z,p}^{2\alpha} \leq 8\alpha 2^{-2\alpha} b \max(ra - sc, |a - c|)^{2\alpha - 1} . \tag{4.8}$$

The assumption of Theorem 4.10 states $\overline{yq}^2 - \overline{yp}^2 - \overline{zq}^2 + \overline{zp}^2 \leq 2\,\overline{y,z}\,\overline{q,p}$. This implies

$$2b\,(ra - sc) = \left(c^2 + b^2 - 2scb\right) - \left(a^2 + b^2 - 2rab\right) - c^2 + a^2 \leq 2b\,\overline{y,z}\,.$$

Therefore, $ra - sc \leq \overline{y,z}$ (or $b = 0$, but then $q = p$ and Theorem 4.10 becomes trivial). Furthermore, the triangle inequality implies $|a - c| = |\overline{z,p} - \overline{y,p}| \leq \overline{y,z}$. Thus, we obtain

$$\max(ra - sc, |a - c|) \leq \overline{y,z}\,. \tag{4.9}$$

Finally, (4.8) and (4.9) together yield

$$\overline{y,q}^{2\alpha} - \overline{y,p}^{2\alpha} - \overline{z,q}^{2\alpha} + \overline{z,p}^{2\alpha} \leq 8\alpha 2^{-2\alpha}\, \overline{q,p}\, \overline{y,z}^{2\alpha-1}\,. \qquad \square$$

The remaining part of this section is dedicated to proving Lemma 4.26.

The proof of Lemma 4.26 can be described as *brute force*. We will distinguish many different cases, i.e., certain bounds on a, b, c, r, s, e.g., $a \leq c$ and $a > c$. In each case, we try to simplify the inequality step by step until we can solve it easily. Mostly, the simplification consists of taking some derivative and showing that the derivative is always negative (or always positive). Then we only need to show the inequality at one extremal point. This process may have to be iterated. It is often not clear immediately which derivative to take in order to simplify the inequality. Even after finishing the proof there seems to be no deeper reason for distinguishing the cases that are considered. Thus, unfortunately, the proof does not create a deeper understanding of the result.

4.G.2 First Proof Steps and Outline of the Remaining Proof

We want to show Lemma 4.26 to prove Theorem 4.10. We refer to the left hand side of the inequality, $a^{2\alpha} - c^{2\alpha} - \left(a^2 - 2rab + b^2\right)^{\alpha} + \left(c^2 - 2scb + b^2\right)^{\alpha}$, as LHS. By RHS we, of course, mean the right hand side, $8\alpha 2^{-2\alpha} b \max(ra - sc, |a - c|)^{2\alpha-1}$.

For $\max(ra - sc, |a - c|) = 0$ we have $a = c$ and $r \leq s$. Thus, LHS ≤ 0. If $\max(ra - sc, |a - c|) > 0$, LHS and RHS are continuous in all parameters. Thus, it is enough to show the inequality on a dense set. In particular, we can and will ignore certain special cases in the following which might introduce technical problems, e.g., "0^0".

We have to distinguish the cases $|a-c| = \max(ra-sc, |a-c|)$ and $ra-sc = \max(ra-sc, |a-c|)$. We further distinguish $a \geq c$ and $c \geq a$. Some trivial implications for each case are recorded in following lemma.

Lemma 4.28 ($ra - sc$ vs $|a - c|$)**.** Let $a, b, c \geq 0$, $r, s \in [-1, 1]$, and $\alpha \in [\frac{1}{2}, 1]$. Then

$$\begin{aligned} ra - sc \geq a - c \quad &\Leftrightarrow \quad s \leq (r-1)\frac{a}{c} + 1 \quad \Leftrightarrow \quad r \geq (s-1)\frac{c}{a} + 1\,, \\ ra - sc \geq c - a \quad &\Leftrightarrow \quad s \leq (r+1)\frac{a}{c} - 1 \quad \Leftrightarrow \quad r \geq (s+1)\frac{c}{a} - 1\,. \end{aligned}$$

In the upcoming two subsections we consider less trivial consequences for the two cases $|a - c| \leq ra - sc$ and $|a - c| \geq ra - sc$.

4.G.2.1 The Case $|a - c| \leq ra - sc$

Consider the case $ra - sc \geq |a - c|$. The next lemma shows convexity in r of the function "LHS minus RHS". This means, we only have to check values of r on the border of its domain.

Lemma 4.29 (Convexity in r)**.** Let $a, b, c \geq 0,\ s, r \in [-1, 1],\ \alpha \in [\frac{1}{2}, 1]$. Assume $ra - sc \geq 0$. Define

$$F(r,s) := a^{2\alpha} - c^{2\alpha} - \left(a^2 - 2rab + b^2\right)^{\alpha} + \left(c^2 - 2scb + b^2\right)^{\alpha} - 8\alpha 2^{-2\alpha} b(ra - sc)^{2\alpha - 1}\,.$$

Then

$$\partial_r^2 F(r,s) \geq 0\,.$$

We will want to show $F(r,s) \leq 0$ for certain restrictions on r and s. Lemma 4.29 implies that F is convex in r. Thus, only extreme values of r need to be checked. We will make use of this fact in Remark 4.30 below. Note, neither $\partial_s^2 F(r,s) \geq 0$ nor $\partial_s^2 F(r,s) \leq 0$ for all a, b, c, s, r.

Proof. Define

$$\ell(r,s) := a^{2\alpha} - c^{2\alpha} - \left(a^2 - 2rab + b^2\right)^{\alpha} + \left(c^2 - 2scb + b^2\right)^{\alpha}\,.$$

It holds

$$\partial_r \ell(r,s) = 2ab\alpha \left(a^2 - 2rab + b^2\right)^{\alpha - 1}\,.$$

Define $h(r,s) := 8\alpha 2^{-2\alpha} b\,(ra - sc)^{2\alpha - 1}$. It holds

$$\partial_r h(r,s) = 8\alpha(2\alpha - 1)2^{-2\alpha} ba\,(ra - sc)^{2\alpha - 2}\,.$$

It holds $F(r,s) = \ell(r,s) - h(r,s)$. It holds

$$f(r) := \frac{\partial_r \ell(r,s) - \partial_r h(r,s)}{2ab\alpha} = \left(a^2 - 2rab + b^2\right)^{\alpha - 1} - (2\alpha - 1)\left(\frac{ra - sc}{2}\right)^{2\alpha - 2}$$

and

$$\partial_r f(r) = -2ab(\alpha - 1)\left(a^2 - 2rab + b^2\right)^{\alpha - 2} - \frac{1}{2}a(2\alpha - 1)(2\alpha - 2)\left(\frac{ra - sc}{2}\right)^{2\alpha - 3}\,.$$

It holds $2ab\left(a^2 - 2rab + b^2\right)^{\alpha - 2} \geq 0$ and $(\alpha - 1) \leq 0$. Thus,

$$-2ab(\alpha - 1)\left(a^2 - 2rab + b^2\right)^{\alpha - 2} \geq 0\,.$$

It holds $\frac{1}{2}a(2\alpha - 1)\left(\frac{ra - sc}{2}\right)^{2\alpha - 3} \geq 0$ and $(2\alpha - 2) \leq 0$. Thus, $-\frac{1}{2}a(2\alpha - 1)(2\alpha - 2)\left(\frac{ra - sc}{2}\right)^{2\alpha - 3} \geq 0$. Hence, $\partial_r f(r) \geq 0$. Hence, $\partial_r^2 F(r,s) \geq 0$. □

4.G.2.2 The Case $|a-c| \geq ra - sc$

In the case $|a-c| \geq ra - sc$, the RHS does not depend on s or r. Thus, we maximize the LHS with respect to r and s and only need to show the inequality for this maximized term. Define

$$\ell(r,s) := a^{2\alpha} - c^{2\alpha} - \left(a^2 - 2rab + b^2\right)^\alpha + \left(c^2 - 2scb + b^2\right)^\alpha .$$

It holds

$$\max_{s \geq s_0, r \leq r_0} \ell(r,s) = \ell(r_0, s_0) .$$

Distinguish the two cases $a \geq c$ and $a \leq c$.
Case 1: $a \geq c$. For fixed $r \in [-1,1]$, set $s = s_{\mathsf{min}}(r) = (r-1)\frac{a}{c} + 1$, cf Lemma 4.28. Define

$$\begin{aligned} f(r) &:= \ell(r, s_{\mathsf{min}}(r)) \\ &= a^{2\alpha} - c^{2\alpha} - \left(a^2 - 2rab + b^2\right)^\alpha + \left(c^2 - 2rab + 2ab - 2cb + b^2\right)^\alpha . \end{aligned}$$

Then

$$\frac{f'(r)}{2ab\alpha} = \left(a^2 - 2rab + b^2\right)^{\alpha-1} - \left(c^2 - 2rab + 2ab - 2cb + b^2\right)^{\alpha-1} .$$

Case 1.1: $a^2 \leq c^2 + 2ab - 2cb$. Then

$$\begin{aligned} a^2 - 2rab + b^2 &\leq c^2 - 2rab + 2ab - 2cb + b^2 , \\ \left(a^2 - 2rab + b^2\right)^{\alpha-1} &\geq \left(c^2 - 2rab + 2ab - 2cb + b^2\right)^{\alpha-1} . \end{aligned}$$

Thus, $f'(r) \geq 0$. In this case, we need to show

$$a^{2\alpha} - c^{2\alpha} - |a-b|^{2\alpha} + |c-b|^{2\alpha} = f(1) \leq 8\alpha 2^{-2\alpha} b (a-c)^{2\alpha-1} .$$

Case 1.2: $a^2 \geq c^2 + 2ab - 2cb$. Then

$$\begin{aligned} a^2 - 2rab + b^2 &\geq c^2 - 2rab + 2ab - 2cb + b^2 , \\ \left(a^2 - 2rab + b^2\right)^{\alpha-1} &\leq \left(c^2 - 2rab + 2ab - 2cb + b^2\right)^{\alpha-1} . \end{aligned}$$

Thus, $f'(r) \leq 0$. The relevant values are $r = r_{\mathsf{min}} = 1 - 2\frac{c}{a}$, with $s = s_{\mathsf{min}}(r) = -1$. In this case, we need to show

$$a^{2\alpha} - c^{2\alpha} - \left((a-b)^2 + 4cb\right)^\alpha + (c+b)^{2\alpha} = f(r_{\mathsf{min}}) \leq 8\alpha 2^{-2\alpha} b (a-c)^{2\alpha-1} .$$

Case 2: $a \leq c$. For fixed $r \in [-1,1]$, set $s = s_{\mathsf{min}}(r) = (r+1)\frac{a}{c} - 1$. Define

$$\begin{aligned} f(r) &:= \ell(r, s_{\mathsf{min}}(r)) \\ &= a^{2\alpha} - c^{2\alpha} - \left(a^2 - 2rab + b^2\right)^\alpha + \left(c^2 - 2rab - 2ab + 2cb + b^2\right)^\alpha . \end{aligned}$$

Then

$$\frac{f'(r)}{2ab\alpha} = \left(a^2 - 2rab + b^2\right)^{\alpha-1} - \left(c^2 - 2rab - 2ab + 2cb + b^2\right)^{\alpha-1} .$$

Case 2.1: $a^2 \leq c^2 - 2ab + 2cb$. Then

$$a^2 - 2rab + b^2 \leq c^2 - 2rab - 2ab + 2cb + b^2 ,$$
$$\left(a^2 - 2rab + b^2\right)^{\alpha-1} \geq \left(c^2 - 2rab - 2ab + 2cb + b^2\right)^{\alpha-1} .$$

Thus, $f'(r) \geq 0$. The critical value is $r = r_{\mathsf{max}} = 1$, with $s = s_{\mathsf{min}}(r) = 2\frac{a}{c} - 1$. In this case, we need to show

$$a^{2\alpha} - c^{2\alpha} - |a-b|^{2\alpha} + \left((c+b)^2 - 4ab\right)^{\alpha} = f(1) \leq 8\alpha 2^{-2\alpha} b(c-a)^{2\alpha-1} .$$

Case 2.2: $a^2 \geq c^2 - 2ab + 2cb$. This cannot happen for $a \leq c$.

4.G.2.3 Outline

The previous considerations reduce Lemma 4.26 to certain special cases. Here we summarize what is left to show and outline how this is achieved in the upcoming sections.

Remark 4.30 (What we need to show)**.** Define

$$F(r,s) := a^{2\alpha} - c^{2\alpha} - \left(a^2 - 2rab + b^2\right)^{\alpha} + \left(c^2 - 2scb + b^2\right)^{\alpha} - 8\alpha 2^{-2\alpha} b(ra - sc)^{2\alpha-1} .$$

First consider $a \geq c$. By Lemma 4.29, if $|a-c| \leq ra - sc$, we need to show $F(r,s) \leq 0$ for extreme values of $r \in [-1,1]$. For maximal r, we want to show

(i) $F(1,s) \leq 0$ for all $s \in [-1,1]$ and $a \geq c$.

This also covers case 1.1 of section 4.G.2.2 when $s = 1$. The value r is minimal if $r = (s-1)\frac{c}{a} + 1$. Then $ra - sc = a - c$. As reasoned in case 1.2 of section 4.G.2.2, which is also covered by this, we then need to show

(ii) $F(1 - 2\frac{c}{a}, -1) \leq 0$ for $a \geq c$.

In the case $a \leq c$, $|a-c| \leq ra - sc$, the maximal value of r is 1. Then, $s \leq 2\frac{a}{c} - 1$ by Lemma 4.29. Thus, we need

(iii) $F(1,s) \leq 0$ for all $s \in [-1, 2\frac{a}{c} - 1]$ and $a \leq c$.

This includes case 2.1 of section 4.G.2.2). The minimal value of r in this case is $(s+1)\frac{c}{a} - 1$ (Lemma 4.29). Then $ra - sc = c - a$ and we can argue as in case 2 of section 4.G.2.2, which leaves us with case 2.1, which is already covered. As cases 1.1, 1.2, and 2.1 of section 4.G.2.2 are covered, we also have $|a-c| \geq ra - sc$ covered.

Thus, the only relevant instances of F are

$$F\left(1-2\frac{c}{a},-1\right) = a^{2\alpha} - c^{2\alpha} - \left((a-b)^2+4bc\right)^{\alpha} + (b+c)^{2\alpha} - 8\alpha 2^{-2\alpha} b(a-c)^{2\alpha-1}\,,$$
$$F(1,s) = a^{2\alpha} - c^{2\alpha} - (a-b)^{2\alpha} + \left(c^2-2scb+b^2\right)^{\alpha} - 8\alpha 2^{-2\alpha} b(a-sc)^{2\alpha-1}\,.$$

Item (ii) is discussed in section 4.G.7. Items (i) and (iii) are shown in the following way:

(a) $b \geq 2sc$: Lemma 4.34 (Merging Lemma) + Lemma 4.31 (Tight Power Bound)

(b) $b \leq 2sc$ and $sc \leq a-b$: Lemma 4.35 (Merging Lemma) + Lemma 4.31 (Tight Power Bound)

(c) $b \leq 2sc$ and $sc \geq a-b$ and $a \geq c$: Lemma 4.39

(d) $b \leq 2sc$ and $sc \geq a-b$ and $a \leq c$, $sc \leq 2a-c$ and $b \leq 2a-2sc$: Lemma 4.40

(e) $b \leq 2sc$ and $sc \geq a-b$ and $a \leq c$, $sc \leq 2a-c$ and $b \geq 2a-2sc$ and $a \leq b$: Lemma 4.43

(f) $b \leq 2sc$ and $sc \geq a-b$ and $a \leq c$, $sc \leq 2a-c$ and $b \geq 2a-2sc$ and $a \geq b$: Lemma 4.45

All six cases together cover $s \in [-1,1]$ with $a \geq c$ and $s \in [-1, 2\frac{a}{c}-1]$ with $a \leq c$.

The proofs consist of distinguishing many different cases and applying simple analysis methods in each case. Nonetheless, finding the poofs is often quite hard, as the inequalities are usually very tight and the right steps necessary for the proof are hard to guess.

As intermediate steps we can, in some cases, use two lemmas: the Tight Power Bound, see section 4.G.3, and the Merging Lemma, see 4.G.4. The remaining cases that cannot be solved via Tight Power Bound and Merging Lemma will be discussed in sections 4.G.6 and 4.G.7.

4.G.3 Tight Power Bound

Following lemma presents a very useful inequality in three different forms. It gives a hint to why the power $\ldots^{2\alpha-1}$ comes up in the RHS of Lemma 4.26.

Lemma 4.31 (Tight Power Bound)**.** Let $x, y \geq 0$.

(i) If $a \in [1,2]$, $x \geq y$, then

$$2^a x^{a-1} y \quad \leq \quad (x+y)^a - (x-y)^a \quad \leq \quad 2a x^{a-1} y\,.$$

(ii) If $a \in [1,2]$, then

$$(x+y)^a - |x-y|^a \quad \leq \quad 2a \min(xy^{a-1}, x^{a-1}y)\,.$$

(iii) If $a \in [1,2]$, $x \geq y$, then

$$(x+y)^{a-1}(x-y) \quad \leq \quad x^a - y^a \quad \leq \quad a(x-y)\left(\frac{x+y}{2}\right)^{a-1}\,.$$

This result is slightly stronger than the application of the mean value theorem to the function $x \mapsto x^a$, which yields $x^a - y^a \leq a(x-y)z^{a-1}$ for all $x \geq y \geq 0$ and $a > 0$, where $z \in [y,x]$.

Proof. Assume $x \geq y$. Set $z = \frac{y}{x} \in [0,1]$. Define

$$f(z) = \frac{(1+z)^a - (1-z)^a}{z}\,.$$

If we can show $f(z) \leq 2a$, then

$$\begin{aligned} & (1+z)^a - (1-z)^a \leq 2az \\ \Rightarrow \quad & (x+zx)^a - (x-zx)^a \leq 2ax^a z \\ \Rightarrow \quad & (x+y)^a - (x-y)^a \leq 2ax^{a-1}y\,. \end{aligned}$$

It holds

$$f'(z) = \frac{g(z)}{z^2}\,,$$

where

$$g(z) = az\left((1+z)^{a-1} + (1-z)^{a-1}\right) - \left((1+z)^a - (1-z)^a\right)\,.$$

It holds

$$g'(z) = az(a-1)\left((1+z)^{a-2} - (1-z)^{a-2}\right) \leq 0\,.$$

Thus, $g(z) \leq g(0) = 0$. Thus, $f'(z) \leq 0$. Thus, for all $z_0 \in [0,1]$,

$$f(z_0) \leq \lim_{z \searrow 0} f(z) \overset{\mathsf{L'H}}{=} \lim_{z \searrow 0} \frac{a(1+z)^{a-1} + a(1-z)^{a-1}}{1} = 2a\,,$$

where $\mathsf{L'H}$ indicates the use of L'Hospital's rule. Furthermore, $f(z_0) \geq f(1) = 2^a$, which implies the lower bound. This finishes the proof for (i). The other parts follow immediately. □

4.G.4 Merging Lemma

In many cases (i.e., with additional assumption on a, b, c, r or s), we prove the inequality of Lemma 4.26 by applying first a merging lemma to the LHS to reduce the four

summands to two summands of a specific form. Then we apply the Tight Power Bound. The Merging Lemma is discussed in this section.

4.G.4.1 Simple Merging Lemma

Lemma 4.32 (Simple Merging Lemma)**.** Let $\alpha \in [\frac{1}{2}, 1]$, $b \geq 0$, $a, c \in \mathbb{R}$. Then

$$|a|^{2\alpha} - |c|^{2\alpha} - |a-b|^{2\alpha} + |c-b|^{2\alpha} \leq 2^{1-2\alpha}\Big((a-c+b)^{2\alpha} - |a-c-b|^{2\alpha}\Big)\mathbb{1}_{a-c>0}\,.$$

Proof. For $\tilde{\alpha} \geq 1$, the function $\mathbb{R} \to \mathbb{R}$, $x \mapsto |x|^{\tilde{\alpha}} - |x-1|^{\tilde{\alpha}}$ is increasing. It holds $2\alpha \geq 1$. Thus, if $a \leq c$, then

$$|a|^{2\alpha} - |a-b|^{2\alpha} \leq |c|^{2\alpha} - |c-b|^{2\alpha}\,.$$

This shows the inequality for the case $a \leq c$.

Set $q := a - b$ and define

$$g(b) := |q+b|^{2\alpha} - |c|^{2\alpha} - |q|^{2\alpha} + |c-b|^{2\alpha} - 2\left(\left(\frac{q-c}{2}+b\right)^{2\alpha} - \left(\frac{q-c}{2}\right)^{2\alpha}\right).$$

It holds $g(0) = 0$ and

$$\frac{g'(b)}{2\alpha} = \mathsf{sgn}(q+b)|q+b|^{2\alpha-1} - \mathsf{sgn}(c-b)|c-b|^{2\alpha-1} - 2\left(\frac{q-c}{2}+b\right)^{2\alpha-1}.$$

Case 1: $\mathsf{sgn}(q+b) = +1$, $\mathsf{sgn}(c-b) = +1$:

$$\frac{g'(b)}{2\alpha} = (q+b)^{2\alpha-1} - (c-b)^{2\alpha-1} - 2\left(\frac{q-c}{2}+b\right)^{2\alpha-1},$$

$$(q+b)^{2\alpha-1} - (c-b)^{2\alpha-1} \leq (q+b-(c-b))^{2\alpha-1} \leq 2\left(\frac{q-c}{2}+b\right)^{2\alpha-1}.$$

Case 2: $\mathsf{sgn}(q+b) = -1$, $\mathsf{sgn}(c-b) = -1$:

$$\frac{g'(b)}{2\alpha} = (b-c)^{2\alpha-1} - (-q-b)^{2\alpha-1} - 2\left(\frac{q-c}{2}+b\right)^{2\alpha-1},$$

$$(b-c)^{2\alpha-1} - (-q-b)^{2\alpha-1} \leq (b-c-(-q-b))^{2\alpha-1} \leq 2\left(\frac{q-c}{2}+b\right)^{2\alpha-1}.$$

Case 3: $\mathsf{sgn}(q+b) = +1$, $\mathsf{sgn}(c-b) = -1$:

$$\frac{g'(b)}{2\alpha} = (q+b)^{2\alpha-1} + (b-c)^{2\alpha-1} - 2\left(\frac{q-c}{2}+b\right)^{2\alpha-1},$$

$$(q+b)^{2\alpha-1} + (b-c)^{2\alpha-1} \le 2\left(\frac{q-c}{2}+b\right)^{2\alpha-1}.$$

Case 4: $\mathsf{sgn}(q+b) = -1$, $\mathsf{sgn}(c-b) = +1$:

$$\frac{g'(b)}{2\alpha} = -(-q-b)^{2\alpha-1} - (c-b)^{2\alpha-1} - 2\left(\frac{q-c}{2}+b\right)^{2\alpha-1},$$

$$-(-q-b)^{2\alpha-1} - (c-b)^{2\alpha-1} \le 0\,.$$

Together: In every case, we have $g'(b) \le 0$ and $g(0) = 0$. Thus,

$$g(b) \le 0\,. \qquad \square$$

4.G.4.2 $ra - sc$–Merging Lemma

Lemma 4.33. Let $\alpha \in [0,1]$.

(i) Let $b, c \ge 0$, $s \in [-1,1]$. Assume $2sc \le b$. Then

$$-c^{2\alpha} + (c^2 - 2scb + b^2)^\alpha \le -|sc|^{2\alpha} + |sc-b|^{2\alpha}\,.$$

(ii) Let $a, b \ge 0$, $r \in [-1,1]$. Assume $2ra \ge b$. Then

$$a^{2\alpha} - (a^2 - 2rab + b^2)^\alpha \le |ra|^{2\alpha} - |ra-b|^{2\alpha}\,.$$

Proof. The function $t \mapsto (t+1)^\alpha - t^\alpha$, $t \ge 0$ is nonincreasing for all $\alpha \in [0,1]$. It holds $0 \le s^2c^2 \le c^2$ and

$$x := -2scb + b^2 \ge 0\,.$$

Thus,

$$\left(c^2 + x\right)^\alpha - \left(c^2\right)^\alpha \le \left((sc)^2 + x\right)^\alpha - \left((sc)^2\right)^\alpha.$$

Thus,

$$(c^2 - 2scb + b^2)^\alpha - c^{2\alpha} \le |-sc+b|^{2\alpha} - |sc|^{2\alpha}\,.$$

For the second part, set $x := 2rab - b^2$, $y := a^2 - x \ge 0$, $\tilde{y} := |ra|^2 - x \ge 0$. The condition $2ra \ge b$ implies $x \ge 0$. Thus, as before,

$$\begin{aligned} a^{2\alpha} - (a^2 - 2rab + b^2)^\alpha &= (y+x)^\alpha - y^\alpha \\ &\le (\tilde{y}+x)^\alpha - \tilde{y}^\alpha \\ &= |ra|^{2\alpha} - |ra-b|^{2\alpha}\,. \end{aligned} \qquad \square$$

Lemma 4.34 ($ra-sc$–Merging Lemma)**.** Let $\alpha \in [\frac{1}{2}, 1]$. Let $a, b, c \geq 0$, $r, s \in [-1, 1]$.

(i) Assume $2ra \geq b$, $s \in \{-1, 1\}$. Then

$$\begin{aligned} &a^{2\alpha} - c^{2\alpha} - (a^2 - 2rab + b^2)^\alpha + (c^2 - 2scb + b^2)^\alpha \\ &\leq 2^{1-2\alpha}\Big((ra - sc + b)^{2\alpha} - |ra - sc - b|^{2\alpha}\Big)\mathbb{1}_{ra-sc>0}\,. \end{aligned}$$

(ii) Assume $b \geq 2sc$, $r \in \{-1, 1\}$. Then

$$\begin{aligned} &a^{2\alpha} - c^{2\alpha} - (a^2 - 2rab + b^2)^\alpha + (c^2 - 2scb + b^2)^\alpha \\ &\leq 2^{1-2\alpha}\Big((ra - sc + b)^{2\alpha} - |ra - sc - b|^{2\alpha}\Big)\mathbb{1}_{ra-sc>0}\,. \end{aligned}$$

(iii) Assume $2ra \geq b \geq 2sc$. Then

$$\begin{aligned} &a^{2\alpha} - c^{2\alpha} - (a^2 - 2rab + b^2)^\alpha + (c^2 - 2scb + b^2)^\alpha \\ &\leq 2^{1-2\alpha}\left((ra - sc + b)^{2\alpha} - |ra - sc - b|^{2\alpha}\right)\mathbb{1}_{ra-sc>0}\,. \end{aligned}$$

Proof. The lemma above and the simple merging lemma imply

$$\begin{aligned} &a^{2\alpha} - c^{2\alpha} - (a^2 - 2rab + b^2)^\alpha + (c^2 - 2scb + b^2)^\alpha \\ &\leq (ra)^{2\alpha} - (sc)^{2\alpha} - (ra - b)^{2\alpha} + (sc - b)^{2\alpha} \\ &\leq 2^{1-2\alpha}\left((ra - sc + b)^{2\alpha} - |ra - sc - b|^{2\alpha}\right)\mathbb{1}_{ra-sc>0}\,. \end{aligned}$$

□

4.G.4.3 $a - sc$–Merging Lemma

Lemma 4.34 covers the case $\frac{1}{2}b \geq sc$. The following lemma covers $\frac{1}{2}b \leq sc$ under the additional restriction $sc \leq a - b$.

Lemma 4.35 ($a - sc$–Merging Lemma)**.** Let $\alpha \in [\frac{1}{2}, 1]$. Let $a, b, c \geq 0$, $s \in [-1, 1]$. Assume $\frac{1}{2}b \leq sc \leq a - b$. Then

$$a^{2\alpha} - c^{2\alpha} - (a - b)^{2\alpha} + (c^2 - 2scb + b^2)^\alpha \leq 2^{1-2\alpha}\Big((a - sc + b)^{2\alpha} - (a - sc - b)^{2\alpha}\Big)\,.$$

Proof. Set $\delta := a - b$. Define

$$f(\delta) = (\delta + b)^{2\alpha} - c^{2\alpha} - \delta^{2\alpha} + (c^2 - 2scb + b^2)^{\alpha} - 2\left(\left(\frac{\delta - sc}{2} + b\right)^{2\alpha} - \left(\frac{\delta - sc}{2}\right)^{2\alpha}\right).$$

Then

$$\frac{f'(\delta)}{2\alpha} = (\delta + b)^{2\alpha-1} - \delta^{2\alpha-1} - \left(\frac{\delta - sc}{2} + b\right)^{2\alpha-1} + \left(\frac{\delta - sc}{2}\right)^{2\alpha-1}.$$

It holds

$$\begin{aligned} \delta + b &\geq \delta\,, \\ \frac{\delta - sc}{2} &\leq \frac{\delta - sc}{2} + b\,, \\ (\delta + b) + \frac{\delta - sc}{2} &= \delta + \left(\frac{\delta - sc}{2} + b\right). \end{aligned}$$

Thus,

$$(\delta + b)^{2\alpha-1} + \left(\frac{\delta - sc}{2}\right)^{2\alpha-1} \leq \delta^{2\alpha-1} + \left(\frac{\delta - sc}{2} + b\right)^{2\alpha-1}.$$

Thus, $f'(\delta) \leq 0$.
The next lemma shows $f(sc) \leq 0$. Thus, $f(\delta) \leq 0$ for all $\delta \geq sc$. □

Lemma 4.36. Let $x, a, b, c \geq 0$. Assume $b \leq 2x$, $x + b \geq c$, $x \leq c$. Then

$$(x + b)^{2\alpha} + (c^2 - 2xb + b^2)^{\alpha} \leq c^{2\alpha} + x^{2\alpha} + 2b^{2\alpha}.$$

Proof. Define

$$\begin{aligned} g(x) &:= (x + b)^{2\alpha} + (c^2 - 2xb + b^2)^{\alpha} - c^{2\alpha} - x^{2\alpha} - 2b^{2\alpha}\,, \\ h(x) &:= \frac{g'(x)}{2\alpha} = (x + b)^{2\alpha-1} - x^{2\alpha-1} - b(c^2 - 2xb + b^2)^{\alpha-1}\,. \end{aligned}$$

It holds

$$h'(x) = (2\alpha - 1)(x + b)^{2\alpha-2} - (2\alpha - 1)x^{2\alpha-2} + 2(\alpha - 1)b^2(c^2 - 2xb + b^2)^{\alpha-2}\,.$$

As $2\alpha - 2 \leq 0$ and $2\alpha - 1 \geq 0, (2\alpha - 1)(x + b)^{2\alpha-2} - (2\alpha - 1)x^{2\alpha-2} \leq 0$. As $\alpha - 1 \leq 0$, $2(\alpha - 1)b^2(c^2 - 2xb + b^2)^{\alpha-2} \leq 0$. Thus, $h'(x) \leq 0$.
It holds $x \geq x_{\mathsf{min}} := \max(\frac{b}{2}, c - b)$. For checking $h(x_{\mathsf{min}}) \leq 0$ and $g(x_{\mathsf{min}}) \leq 0$, we distinguish $x_{\mathsf{min}} = c - b$ and $x_{\mathsf{min}} = \frac{b}{2}$.
Case 1, $c - b \leq \frac{b}{2}$:

If $c - b \leq \frac{b}{2}$, then $c \leq \frac{3}{2}b \leq (1+\sqrt{3})b$, $c^2 - 2cb - 2b^2 \leq 0$, and

$$\begin{aligned} h(c-b) &= c^{2\alpha-1} - (c-b)^{2\alpha-1} - b(c^2 - 2(c-b)b + b^2)^{\alpha-1} \\ &= c^{2\alpha-1} - (c-b)^{2\alpha-1} - b(c^2 - 2cb - b^2)^{\alpha-1} \\ &\leq b^{2\alpha-1} - b(c^2 - 2cb - b^2)^{\alpha-1} \\ &\leq b\left(\left(b^2\right)^{\alpha-1} - \left(c^2 - 2cb - b^2\right)^{\alpha-1}\right) \end{aligned}$$

And, thus, $h(c-b) \leq 0$ as $c^2 - 2cb - 2b^2 \leq 0$. Furthermore,

$$\begin{aligned} g(c-b) &= -(c-b)^{2\alpha} + (c^2 - 2cb - b^2)^{\alpha} - 2b^{2\alpha} \\ &= -(c-b)^{2\alpha} + (c^2 - 2cb - 2b^2 + b^2)^{\alpha} - 2b^{2\alpha} \\ &\leq -(c-b)^{2\alpha} + b^{2\alpha} - 2b^{2\alpha} \\ &= -(c-b)^{2\alpha} - b^{2\alpha} \\ &\leq 0\,. \end{aligned}$$

Thus, $g(x) \leq 0$ for all valid x.
Case 2, $c - b \geq \frac{b}{2}$:
If $c - b \geq \frac{b}{2}$, then $c \geq b$ and

$$\begin{aligned} h\left(\frac{b}{2}\right) &= \left(\frac{3}{2}b\right)^{2\alpha-1} - \left(\frac{1}{2}b\right)^{2\alpha-1} - b(c^2)^{\alpha-1} \\ &\leq \left(\left(\frac{3}{2}\right)^{2\alpha-1} - \left(\frac{1}{2}\right)^{2\alpha-1}\right) b^{2\alpha-1} - b(c^2)^{\alpha-1} \\ &\leq b^{2\alpha-1} - b(c^2)^{\alpha-1} \\ &\leq b^{2\alpha-1} - b(b^2)^{\alpha-1} \\ &\leq 0\,, \end{aligned}$$

$$\begin{aligned} g\left(\frac{b}{2}\right) &= \left(\frac{3}{2}b\right)^{2\alpha} - c^{2\alpha} - \left(\frac{1}{2}b\right)^{2\alpha} + c^{2\alpha} - 2b^{2\alpha} \\ &\leq \left(\left(\frac{9}{4}\right)^{\alpha} - \left(\frac{1}{2}\right)^{\alpha} - 2\right) b^{2\alpha} \\ &\leq 0\,. \end{aligned}$$

Thus, $g(x) \leq 0$ for all valid x. □

4.G.5 Application of Tight Power Bound and Merging Lemma

Whenever a Merging Lemma holds, we apply it as a first step and then use the Tight Power Bound, Lemma 4.31, to obtain

$$\begin{aligned}
&a^{2\alpha} - c^{2\alpha} - (a^2 - 2rab + b^2)^\alpha + (c^2 - 2scb + b^2)^\alpha \\
&\leq 2^{1-2\alpha} \left((ra - sc + b)^{2\alpha} - |ra - sc - b|^{2\alpha} \right) \\
&\leq 4\alpha 2^{1-2\alpha} (ra - sc)^{2\alpha - 1} b \,.
\end{aligned}$$

In particular, we have finished the proof of Lemma 4.26 in following cases:

- $ra \geq sc$ and $s, r \in \{-1, 1\}$: Lemma 4.32,
- $2ra \geq b$ and $s \in \{-1, 1\}$; or $b \geq 2sc$ and $r \in \{-1, 1\}$; or $2ra \geq b \geq 2sc$: Lemma 4.34,
- $\frac{1}{2}b \leq sc \leq a - b$ and $r = 1$: Lemma 4.35.

4.G.6 The Case $ra - sc \geq |a - c|$

4.G.6.1 The Case $a \geq c$

First we prove two simple lemmas, before we solve this case.

Lemma 4.37. Let $a \geq b \geq 0$, $d \geq c \geq 0$, and $\alpha \in [0, 1]$. Then

$$a^\alpha - b^\alpha - c^\alpha + d^\alpha \leq 2^{1-\alpha} (a - b - c + d)^\alpha \,.$$

Proof. As $a \geq b$, $d \geq c$, $\alpha \leq 1$,

$$a^\alpha - b^\alpha + d^\alpha - c^\alpha \leq (a - b)^\alpha + (d - c)^\alpha \,.$$

Furthermore, by concavity of $x \mapsto x^\alpha$,

$$(a - b)^\alpha + (d - c)^\alpha \leq 2^{1-\alpha} (a - b + d - c)^\alpha \,. \qquad \square$$

Lemma 4.38. Let $a \geq b \geq c \geq d \geq 0$, $a + d \geq b + c$, and $\alpha \in [0, 1]$. Then

$$a^\alpha - b^\alpha - c^\alpha + d^\alpha \leq (a - b - c + d)^\alpha \,.$$

Proof. Define $f(x, y) = x^\alpha + y^\alpha - (x + y)^\alpha$ for $x, y \geq 0$. Then $\partial_x f(x, y) = \alpha(x^{\alpha-1} - (x + y)^{\alpha-1}) \geq 0$ and similarly $\partial_y f(x, y) \geq 0$. Set $\delta := a - b$ and $\epsilon := c - d$. The assumptions ensure $\delta \geq \epsilon \geq 0$. Then,

$$f(b, \delta) \geq f(b, \epsilon) \geq f(d, \epsilon) \,.$$

Thus,

$$\begin{aligned}0 &\geq f(d,\epsilon) - f(b,\delta)\\ &= d^\alpha + \epsilon^\alpha - (d+\epsilon)^\alpha - b^\alpha - \delta^\alpha + (b+\delta)^\alpha\\ &= d^\alpha + \epsilon^\alpha - c^\alpha - b^\alpha - \delta^\alpha + a^\alpha\,.\end{aligned}$$

With this we get

$$\begin{aligned}d^\alpha - c^\alpha - b^\alpha + a^\alpha &\leq \delta^\alpha - \epsilon^\alpha\\ &\leq (\delta-\epsilon)^\alpha\\ &= (a-b-c+d)^\alpha\,.\end{aligned}$$

□

For $a \geq c$, the remaining case is solved by following lemma.

Lemma 4.39. Let $\alpha \in [0,1]$. Let $a,b,c \geq 0$, $s \in [-1,1]$. Assume $\frac{1}{2}b \leq sc$, $sc \geq a-b$, and $a \geq c$. Then

$$a^{2\alpha} - c^{2\alpha} - (a-b)^{2\alpha} + (c^2 - 2scb + b^2)^\alpha \leq 2b^\alpha(a-sc)^\alpha \leq 2b(a-sc)^{2\alpha-1}\,.$$

Proof. Because $a \geq c$ and $\frac{1}{2}b \leq sc$, we have $a - b \geq a - 2sc \geq a - 2c \geq -c$. Hence, $a^2 \geq \max(c^2, (a-b)^2)$. Thus, applying either Lemma 4.37 (if $c^2 - 2scb + b^2$ is larger then either c^2 or $(a-b)^2$) or Lemma 4.38 yields

$$\begin{aligned}&a^{2\alpha} - c^{2\alpha} - (a-b)^{2\alpha} + (c^2 - 2scb + b^2)^\alpha\\ &\leq 2^{1-\alpha}\left(a^2 - c^2 - (a-b)^2 + (c^2 - 2scb + b^2)\right)^\alpha\\ &= 2b^\alpha(a-sc)^\alpha\,.\end{aligned}$$

The condition $0 \leq a - sc \leq b$ implies

$$2b^\alpha(a-sc)^\alpha \leq 2b(a-sc)^{2\alpha-1}\,.$$

□

4.G.6.2 The Case $a \leq c$

For the case $c \geq a$, we only need $ra - sc \geq c - a$ (for $r = 1$), i.e., $sc \leq 2a - c$. Assume $c \geq a$, $sc \geq a - b$, $\frac{1}{2}b \leq sc$, and $sc \leq 2a - c$. Then

$$\begin{aligned}c^2 &\geq c^2 - 2scb + b^2 \geq (a-b)^2\\ c^2 &\geq a^2 \geq (a-b)^2\end{aligned}$$

We distinguish $\frac{1}{2}b \leq a - sc$ and $\frac{1}{2}b \geq a - sc$.

Lemma 4.40 ($\frac{1}{2}b \le a - sc$)**.** Let $\alpha \in [0,1]$. Let $a, b, c \ge 0$, $s \in [-1,1]$. Assume $\frac{1}{2}b \le sc$, $sc \ge a - b$, $c \ge a$, $sc \le 2a - c$, and $\frac{1}{2}b \le a - sc$. Then

$$a^{2\alpha} - c^{2\alpha} - (a-b)^{2\alpha} + (c^2 - 2scb + b^2)^\alpha \le 2b^\alpha (a - sc)^\alpha \,.$$

Proof. The conditions imply

$$\max\Big(\frac{1}{2}b,\, a - b\Big) \le sc \le \min\Big(a - \frac{1}{2}b,\, 2a - c\Big) \,.$$

In particular, $\frac{1}{2}b \le a - \frac{1}{2}b$, and $a - b \le 2a - c$. Thus, $a + b \ge c \ge a \ge b$.

Fix $a, b, c \ge 0$. Assume $c \ge a$. Let $x \in [0, a)$. Then $a - x > 0$ and $c^2 - 2xb + b^2 > 0$. Define

$$f(x) := a^{2\alpha} - c^{2\alpha} - (a-b)^{2\alpha} + (c^2 - 2xb + b^2)^\alpha - 2b^\alpha (a-x)^\alpha \,.$$

It holds

$$\frac{f'(x)}{2\alpha b} = -(c^2 - 2xb + b^2)^{\alpha-1} + b^{\alpha-1}(a-x)^{\alpha-1} \,.$$

Furthermore,

$$\begin{aligned}
& a - c \le 0 \le (c-b)^2 \\
\Rightarrow\quad & 2ab - 2cb \le 0 \le c^2 + b^2 - 2cb \\
\Rightarrow\quad & xb \le 2ab - cb \le c^2 + b^2 - cb \le c^2 + b^2 - ab \\
\Rightarrow\quad & ab - xb \le c^2 - 2xb + b^2 \\
\Rightarrow\quad & b^{\alpha-1}(a-x)^{\alpha-1} \ge (c^2 - 2xb + b^2)^{\alpha-1} \\
\Rightarrow\quad & f'(x) \ge 0 \,.
\end{aligned}$$

We define

$$g(c) := f(a - \frac{1}{2}b) = a^{2\alpha} - c^{2\alpha} - (a-b)^{2\alpha} + (c^2 - 2ab + 2b^2)^\alpha - 2^{1-\alpha} b^{2\alpha} \,.$$

Thus,

$$\frac{g'(c)}{2\alpha c} = -c^{2\alpha-2} + \Big(c^2 - 2ab + 2b^2\Big)^{\alpha-1} \ge 0 \,.$$

Define

$$\begin{aligned}
h(a,b) &:= g(a+b) \\
&= a^{2\alpha} - (a+b)^{2\alpha} - (a-b)^{2\alpha} + (a^2 + 3b^2)^\alpha - 2^{1-\alpha} b^{2\alpha} \\
&= a^{2\alpha} - (a+b)^{2\alpha} - (a-b)^{2\alpha} + (a^2 + 3b^2)^\alpha - 2\left(\frac{b^2}{2}\right)^\alpha \,.
\end{aligned}$$

The next lemma shows $h(a,b) \le 0$ for $a \ge b$. Thus, $g(c) \le 0$. Thus, $f(x) \le 0$. □

Lemma 4.41. Let $\alpha \in [0,1]$, $a, b \geq 0$. Assume $a \geq b$. Then

$$a^{2\alpha} + (a^2 + 3b^2)^\alpha \leq (a+b)^{2\alpha} + (a-b)^{2\alpha} + 2\left(\frac{b^2}{2}\right)^\alpha .$$

Proof. Set $x = \frac{b}{a} \in [0,1]$. Define

$$f(\alpha, x) = 1 + \left(1 + 3x^2\right)^\alpha - (1+x)^{2\alpha} - (1-x)^{2\alpha} - 2\left(\frac{x^2}{2}\right)^\alpha .$$

It holds

$$\begin{aligned}
\frac{\partial_\alpha f(\alpha, x)}{\alpha} &= \left(1 + 3x^2\right)^\alpha \log\left(1 + 3x^2\right) - (1+x)^{2\alpha} \log\left((1+x)^2\right) \\
&\quad - (1-x)^{2\alpha} \log\left((1-x)^2\right) - 2\left(\frac{x^2}{2}\right)^\alpha \log\left(\frac{x^2}{2}\right) \\
&=: g(x, \alpha) ,
\end{aligned}$$

$$\begin{aligned}
\frac{\partial_\alpha g(\alpha, x)}{\alpha} &= \left(1 + 3x^2\right)^\alpha \log\left(1 + 3x^2\right)^2 - (1+x)^{2\alpha} \log\left((1+x)^2\right)^2 \\
&\quad - (1-x)^{2\alpha} \log\left((1-x)^2\right)^2 - 2\left(\frac{x^2}{2}\right)^\alpha \log\left(\frac{x^2}{2}\right)^2 \\
&\leq \left(1 + 3x^2\right)^\alpha \log\left(1 + 3x^2\right)^2 - (1+x)^{2\alpha} \log\left((1+x)^2\right)^2 \\
&=: h(x, \alpha) .
\end{aligned}$$

For $x \in [0,1]$, it holds $1 + 3x^2 \leq (1+x)^2$. Thus,

$$\left(\frac{1+3x^2}{(1+x)^2}\right)^\alpha \leq 1 \leq \left(\frac{\log((1+x)^2)}{\log(1+3x^2)}\right)^2 .$$

Thus,

$$\left(1 + 3x^2\right)^\alpha \log\left(1 + 3x^2\right)^2 \leq (1+x)^{2\alpha} \log\left((1+x)^2\right)^2$$

Thus, $h(x, \alpha) \leq 0$ and $\partial_\alpha g(\alpha, x) \leq 0$. Thus, $g(\alpha, x) \geq g(1, x)$ and

$$\begin{aligned}
g(1, x) &= \left(1 + 3x^2\right) \log\left(1 + 3x^2\right) - (1+x)^2 \log\left((1+x)^2\right) \\
&\quad - (1-x)^2 \log\left((1-x)^2\right) - 2\left(\frac{x^2}{2}\right) \log\left(\frac{x^2}{2}\right) \\
&=: \ell(x) .
\end{aligned}$$

The next lemma shows $\ell(x) \geq 0$. Thus, $g(\alpha, x) \geq g(1, x) \geq 0$. Thus $\partial_\alpha f(\alpha, x) \geq 0$. Thus, $f(\alpha, x) \leq f(1, x)$ and

$$f(1,x) = 1 + \left(1+3x^2\right) - (1+x)^2 - (1-x)^2 - 2\left(\frac{x^2}{2}\right) = 0\,.$$

Thus, $f(\alpha, x) \leq 0$. □

Lemma 4.42. Let $x \in [0,1]$. Define

$$\begin{aligned} f(x) := \left(1+3x^2\right)\log\left(1+3x^2\right) - (1+x)^2\log\left((1+x)^2\right) \\ - (1-x)^2\log\left((1-x)^2\right) - x^2\log\left(\frac{x^2}{2}\right)\,. \end{aligned}$$

Then

$$f(x) \geq 0\,.$$

Proof. Let us first calculate some derivatives:

$$\begin{aligned} f(x) &= x^2\log\left(\frac{2\left(1+3x^2\right)^3}{x^2\left(1-x^2\right)^2}\right) - 4x\log\left(\frac{1+x}{1-x}\right) + \log\left(\frac{1+3x^2}{\left(1-x^2\right)^2}\right)\,, \\ f'(x) &= 2x\log\left(\frac{2\left(1+3x^2\right)^3}{x^2\left(1-x^2\right)^2}\right) - 4\log\left(\frac{1+x}{1-x}\right)\,, \\ f''(x) &= 2\log\left(\frac{2\left(1+3x^2\right)^3}{x^2\left(1-x^2\right)^2}\right) - \frac{12}{1+3x^2}\,, \\ f'''(x) &= \frac{4\left(9x^4+24x^2-1\right)}{x(1-x)(1+x)\left(3x^2+1\right)^2}\,, \\ f^{(4)}(x) &= \frac{4\left(81x^8+324x^6-186x^4+36x^2+1\right)}{x^2\left(1-x^2\right)^2\left(3x^2+1\right)^3}\,. \end{aligned}$$

We consider the cases $x \in [0, \frac{1}{10}]$ and $x \in [\frac{1}{10}, 1]$ separately, and start with the latter. For $x_0 \in (0,1)$, define

$$g_{x_0}(x) := f(x_0) + f'(x_0)\,(x-x_0) + \frac{1}{2}f''(x_0)\,(x-x_0)^2 + \frac{1}{6}f'''(x_0)\,(x-x_0)^3\,.$$

Then the Taylor-Expansion for $x \in (0,1)$ is

$$f(x) = g_{x_0}(x) + \frac{1}{24}f^{(4)}(\xi(x,x_0))\,(x-x_0)^4\,,$$

with suitable $\xi(x, x_0)$. One can show that $81x^4 + 324x^3 - 186x^2 + 36x^1 + 1 > 0$ for all

$x \geq 0$. In particular, $f^{(4)}(x) \geq 0$ for $x \in [0,1]$. Thus,

$$f(x) \geq g_{x_0}(x)\,.$$

We use $x_0 = \frac{1}{3}$:

$$\begin{aligned} f\left(\frac{1}{3}\right) &= \frac{10}{3}\log(3) - \frac{47}{9}\log(2)\,, \\ f'\left(\frac{1}{3}\right) &= 2\log(3) - \frac{10}{3}\log(2)\,, \\ f''\left(\frac{1}{3}\right) &= -9 + 2\log(2) + 6\log(3)\,, \\ f'''\left(\frac{1}{3}\right) &= \frac{27}{2}\,. \end{aligned}$$

One can show that $g_{\frac{1}{3}}(x) \geq 0$ for $x \geq \frac{1}{10}$. Thus, $f(x) \geq 0$ for $x \in [\frac{1}{10}, 1]$.

The case $x \in [0, \frac{1}{10}]$ is left. One can show

$$4\log\left(\frac{1+x}{1-x}\right) \leq 10x \leq 2x\log\left(\frac{2\left(1+3x^2\right)^3}{x^2\left(1-x^2\right)^2}\right)$$

for $x \in [0, \frac{1}{10}]$. This implies $f'(x) \geq 0$. Together with $f(0) = 0$, this yields $f(x) \geq 0$ for $x \in [0, \frac{1}{10}]$. □

Lemma 4.43 ($\frac{1}{2}b \geq a - sc$ and $a \leq b$)**.** Let $\alpha \in [\frac{1}{2}, 1]$. Let $a, b, c \geq 0$, $s \in [-1, 1]$. Assume

$$0 \leq c - a \leq a - sc \leq \frac{b}{2} \leq sc \leq 2a - c \leq a \leq c$$

and $a \leq b$. Then

$$a^{2\alpha} - c^{2\alpha} - |a-b|^{2\alpha} + (c^2 - 2scb + b^2)^{\alpha} \leq 2b(a - sc)^{2\alpha - 1}\,.$$

Proof. Define

$$f(a,b,y,w) := a^{2\alpha} - (y+a)^{2\alpha} - (b-a)^{2\alpha} + ((y+a)^2 - 2wb + b^2)^{\alpha} - 2b(a-w)^{2\alpha-1}\,.$$

It holds

$$\partial_y f(a,b,y,w) = -2\alpha(y+a)^{2\alpha-1} + \alpha\left(2(y+a)\right)\left((y+a)^2 - 2wb + b^2\right)^{\alpha-1}\,.$$

Because of $\frac{b}{2} \leq w$, we have

$$(y+a)^2 \geq (y+a)^2 - 2wb + b^2\,.$$

Thus $\partial_y f(a,b,y,w) \geq 0$. Thus, for $y \in [0, a-w]$, it holds $f(a,b,y,w) \leq f(a,b,a-w,w)$. It holds

$$\begin{aligned}
&f(a,b,a-w,w)\\
&= a^{2\alpha} - (2a-w)^{2\alpha} - (b-a)^{2\alpha} + ((2a-w)^2 - 2wb + b^2)^{\alpha} - 2b(a-w)^{2\alpha-1}\\
&=: g(a,b,w)\,,
\end{aligned}$$

$$\begin{aligned}
&\partial_b g(a,b,w)\\
&= -2\alpha(b-a)^{2\alpha-1} + 2\alpha(b-w)((2a-w)^2 - 2wb + b^2)^{\alpha-1} - 2(a-w)^{2\alpha-1}\\
&\leq 2\alpha h(a,b,w)\,,
\end{aligned}$$

$$h(a,b,w) = -(b-a)^{2\alpha-1} + (b-w)((2a-w)^2 - 2wb + b^2)^{\alpha-1} - (a-w)^{2\alpha-1}\,.$$

The conditions $0 \leq a-w \leq \frac{b}{2} \leq w$ and $a \leq b$ imply $w \leq a \leq b$. It holds,

$$\begin{aligned}
\partial_a^2 h(a,b,w) = &-(2\alpha-1)(2\alpha-2)(b-a)^{2\alpha-3}\\
&+ 4(\alpha-1)(\alpha-2)(2a-w)^2(b-w)((2a-w)^2 - 2wb + b^2)^{\alpha-3}\\
&- (2\alpha-1)(2\alpha-2)(a-w)^{2\alpha-3}\\
&\geq 0\,,
\end{aligned}$$

$$\begin{aligned}
h(w,b,w) &= -(b-w)^{2\alpha-1} + (b-w)((2w-w)^2 - 2wb + b^2)^{\alpha-1} - (w-w)^{2\alpha-1}\\
&= -(b-w)^{2\alpha-1} + (b-w)^{2\alpha-1} = 0\,,
\end{aligned}$$

$$h(b,b,w) = -(b-b)^{2\alpha-1} + (b-b)((2b-w)^2 - 2wb + b^2)^{\alpha-1} - (b-b)^{2\alpha-1} \leq 0\,.$$

Thus, $h(a,b,w) \leq 0$ for all $a \in [w,b]$. Thus, $\partial_b g(a,b,w) \leq 0$. The conditions for g are $0 \leq a-w \leq \frac{b}{2} \leq w \leq a \leq b$. As $a \leq b$ and $\frac{b}{2} \leq w$, we have $a \leq 2w$ and thus $a \geq 2a-2w$.

$$\begin{aligned}
g(a,a,w) = &\, a^{2\alpha} - (2a-w)^{2\alpha} - (a-a)^{2\alpha} + ((2a-w)^2 - 2wa + a^2)^{\alpha} -\\
&2a(a-w)^{2\alpha-1}\\
= &\, a^{2\alpha} - (2a-w)^{2\alpha} + (5a^2 - 6aw + w^2)^{\alpha} - 2a(a-w)^{2\alpha-1}\,.
\end{aligned}$$

Set $w = a-y$, $y \in [0,a]$. It holds

$$\begin{aligned}
\ell(a,y) &:= a^{2\alpha} + (5a^2 - 6a(a-y) + (a-y)^2)^{\alpha} - (a+y)^{2\alpha} - 2ay^{2\alpha-1}\\
&= a^{2\alpha} + (4ay + y^2)^{\alpha} - (a+y)^{2\alpha} - 2ay^{2\alpha-1}\,.
\end{aligned}$$

It holds $g(a,a,w) = \ell(a,y)$. Under the condition $0 \leq y \leq a$, it holds $\ell(a,y) \leq 0$, cf next lemma. □

Lemma 4.44. Let $\alpha \in [\frac{1}{2}, 1]$, $a, y \geq 0$. Assume $y \leq a$. Then

$$a^{2\alpha} + (4ay + y^2)^\alpha \leq (a+y)^{2\alpha} + 2ay^{2\alpha-1}\,.$$

Proof. For $y \leq a$ define

$$g(a,y) := 2a^{2\alpha-1} + 4^\alpha y^\alpha a^{\alpha-1} - 2(a+y)^{2\alpha-1} - 2y^{2\alpha-1}\,.$$

It holds

$$\partial_a g(a,y) = 2(2\alpha-1)a^{2\alpha-2} + 4^\alpha(\alpha-1)y^\alpha a^{\alpha-2} - 2(2\alpha-1)(a+y)^{2\alpha-2}\,,$$

$$\partial_y \partial_a g(a,y) = 4^\alpha(\alpha-1)\alpha y^{\alpha-1}a^{\alpha-2} - 2(2\alpha-1)(2\alpha-2)(a+y)^{2\alpha-3}\,.$$

Set $z := \frac{y}{a} \in [0,1]$. Then

$$\frac{\partial_y \partial_a g(a,y)}{a^{2\alpha-3}} = (\alpha-1)\left(4^\alpha \alpha z^{\alpha-1} - 4(2\alpha-1)(1+z)^{2\alpha-3}\right)\,.$$

One can show

$$\frac{4^\alpha \alpha}{4(2\alpha-1)} \geq 1 \geq \frac{(1+z)^{2\alpha-3}}{z^{\alpha-1}}\,,$$

$$4^\alpha \alpha z^{\alpha-1} \geq 4(2\alpha-1)\,(1+z)^{2\alpha-3}\,.$$

Thus, $\partial_y \partial_a g(a,y) \leq 0$. Thus, $\partial_a g(a,y) \leq \partial_a g(a,0)$

$$\partial_a g(a,0) = 2(2\alpha-1)a^{2\alpha-2} - 2(2\alpha-1)(a)^{2\alpha-2} = 0$$

Thus, $\partial_a g(a,y) \leq 0$. Thus, $g(a,y) \leq g(y,y)$, and

$$g(y,y) = \left(2 + 4^\alpha - 2^{2\alpha} - 2\right) y^{2\alpha-1} = 0\,.$$

Thus, $g(a,y) \leq 0$. It holds

$$\begin{aligned} &a^{2\alpha} + (4ay+y^2)^\alpha - (a+y)^{2\alpha} - 2ay^{2\alpha-1} \\ &\leq a^{2\alpha} + (4a)^\alpha y^\alpha + y^{2\alpha} - (a+y)^{2\alpha} - 2ay^{2\alpha-1} \\ &=: f(a,y)\,, \end{aligned}$$

$$\partial_a f(a,y) = 2\alpha a^{2\alpha-1} + 4^\alpha \alpha y^\alpha a^{\alpha-1} - 2\alpha(a+y)^{2\alpha-1} - 2y^{2\alpha-1} \leq \alpha g(a,y) \leq 0\,.$$

Thus,

$$f(a,y) \leq f(y,y) = (1 + 5^\alpha - 4^\alpha - 2)\, y^{2\alpha} \leq 0\,. \qquad \square$$

Lemma 4.45 ($\frac{1}{2}b \geq a - sc$ and $a \geq b$)**.** Let $\alpha \in [\frac{1}{2}, 1]$. Let $a, b, c \geq 0$, $s \in [-1, 1]$. Assume

$$0 \leq c - a \leq a - sc \leq \frac{b}{2} \leq sc \leq 2a - c \leq a \leq c$$

and $a \geq b$. Then

$$a^{2\alpha} - c^{2\alpha} - (a-b)^{2\alpha} + (c^2 - 2scb + b^2)^{\alpha} \leq 2b(a - sc)^{2\alpha - 1}\,.$$

Proof. Define

$$f(a, b, c, w) := a^{2\alpha} - c^{2\alpha} - (a-b)^{2\alpha} + (c^2 - 2wb + b^2)^{\alpha} - 2b(a-w)^{2\alpha-1}\,.$$

It holds

$$\partial_c f(a, b, c, w) = 2\alpha \left(-c^{2\alpha-1} + c(c^2 - 2wb + b^2)^{\alpha-1}\right)\,.$$

Because of $2w \geq b$, it holds $\partial_c f(a, b, c, w) \leq 0$.

$$f(a, b, a, w) = -(a-b)^{2\alpha} + (a^2 - 2wb + b^2)^{\alpha} - 2b(a-w)^{2\alpha-1}\,.$$

Set $x := a - w$. The conditions for x are $0 \leq x \leq \frac{b}{2} \leq w$ and $b \leq x + w$. Define

$$g(x, b, w) := -(x + w - b)^{2\alpha} + ((x+w)^2 - 2wb + b^2)^{\alpha} - 2bx^{2\alpha-1}\,.$$

It holds

$$\partial_w g(x, b, w) = -2\alpha(x + w - b)^{2\alpha-1} + 2\alpha(x + w - b)((x+w)^2 - 2wb + b^2)^{\alpha-1}\,.$$

It holds

$$(x + w - b)^2 - \left((x+w)^2 - 2wb + b^2\right) = -2bx \leq 0\,.$$

Thus,

$$-2\alpha(x + w - b)^{2\alpha-2} + 2\alpha \left((x+w)^2 - 2wb + b^2\right)^{\alpha-1} \leq 0\,.$$

As $a \geq b$ and thus $x + w - b \geq 0$, $\partial_w g(x, b, w) \leq 0$. It holds $w \geq b - x \geq \frac{b}{2}$ and

$$g(x, b, b - x) = (2xb)^{\alpha} - 2bx^{2\alpha-1} \leq 0$$

because $x \leq b$. □

4.G.7 The Case $|a - c| \geq ra - sc$

Here, we show item (ii) of Remark 4.30, which is case 1.2 of section 4.G.2.2. In case 1.2 we assume $a \geq c$ and $a^2 \geq c^2 + 2ab - 2cb$. The latter is equivalent to $a + c \geq 2b$.

Lemma 4.46 (Case 1.2)**.** Let $\alpha \in [\frac{1}{2}, 1]$. Let $a, b, c \geq 0$. Assume $a \geq c$ and $a + c \geq 2b$. Then

$$a^{2\alpha} - c^{2\alpha} - \left((a-b)^2 + 4cb\right)^{\alpha} + (c+b)^{2\alpha} \leq 8\alpha 2^{-2\alpha} b(a-c)^{2\alpha-1}\,.$$

This lemma follows from the upcoming Lemma 4.47 ($a \geq c$, $a + c \geq 2b$, and $a \geq b + c$) and Lemma 4.48 ($a \geq c$, $a + c \geq 2b$, and $a \leq b + c$).

Lemma 4.47 (Case 1.2, Merging)**.** Let $\alpha \in [\frac{1}{2}, 1]$. Let $a, b, c \geq 0$. Assume $a \geq b + c$. Then

$$a^{2\alpha} - c^{2\alpha} - \left((a-b)^2 + 4cb\right)^{\alpha} + (c+b)^{2\alpha} \leq 2^{1-2\alpha}\left((a-c+b)^{2\alpha} - (a-c-b)^{2\alpha}\right)$$

Proof. Set $\delta := a - b \geq c \geq 0$ and define

$$f(b) := (\delta+b)^{2\alpha} - c^{2\alpha} - \left(\delta^2 + 4cb\right)^{\alpha} + (c+b)^{2\alpha} - 2^{1-2\alpha}\left((\delta - c + 2b)^{2\alpha} - (\delta - c)^{2\alpha}\right).$$

Then

$$\frac{f'(b)}{2\alpha} = (\delta+b)^{2\alpha-1} - 2c\left(\delta^2 + 4cb\right)^{\alpha-1} + (c+b)^{2\alpha-1} - 2^{2-2\alpha}(\delta - c + 2b)^{2\alpha-1}\,,$$

$$\frac{f'(b)}{2\alpha} \leq (\delta + 2b + c)^{2\alpha-1} - (\delta + 2b - c)^{2\alpha-1} - 2c\left(\delta^2 + 4cb\right)^{\alpha-1} =: g(c)\,,$$

$$\begin{aligned} g'(c) = (2\alpha-1)(\delta+2b+c)^{2\alpha-2} + (2\alpha-1)(\delta+2b-c)^{2\alpha-2} - \\ 2\left(\delta^2+4cb\right)^{\alpha-1} - 8cb(\alpha-1)\left(\delta^2+4cb\right)^{\alpha-2}\,, \end{aligned}$$

$$g''(c) = (2\alpha-1)(2\alpha-2)(\delta+2b+c)^{2\alpha-3} - (2\alpha-1)(2\alpha-2)(\delta+2b-c)^{2\alpha-3} - A\,,$$

where

$$\begin{aligned} A :=& 2(4b)(\alpha-1)\left(\delta^2+4cb\right)^{\alpha-2} 8b(\alpha-1)\left(\delta^2+4cb\right)^{\alpha-2} + \\ & 8cb(4b)(\alpha-1)(\alpha-2)\left(\delta^2+4cb\right)^{\alpha-3} \\ =& 16b(\alpha-1)\left(\delta^2+4cb\right)^{\alpha-3}\left(\delta^2+4cb+2cb(\alpha-2)\right). \end{aligned}$$

Thus, $g''(c) \geq 0$. It holds $0 \leq c \leq \delta$ and

$$g(0) = (\delta+2b)^{2\alpha-1} - (\delta+2b)^{2\alpha-1} - 0 = 0\,,$$

$$g(\delta) = (2\delta + 2b)^{2\alpha-1} - (2b)^{2\alpha-1} - 2\delta \left(\delta^2 + 4\delta b\right)^{\alpha-1} =: h(\delta)\,,$$

$$h'(\delta) = 2(2\alpha-1)(2\delta+2b)^{2\alpha-2} - 2\left(\delta^2+4\delta b\right)^{\alpha-1} - 2\delta(\alpha-1)(2\delta+4b)\left(\delta^2+4\delta b\right)^{\alpha-2}\,.$$

For $\alpha \in [\frac{1}{2}, 1]$, we have

$$\begin{aligned}
&\left(\delta^2+4\delta b\right)^{\alpha-1} + \delta(\alpha-1)(2\delta+4b)\left(\delta^2+4\delta b\right)^{\alpha-2} \\
&= \left(\delta^2+4\delta b\right)^{\alpha-2}\left(\delta^2+4\delta b + \delta(\alpha-1)(2\delta+4b)\right) \\
&\geq \left(\delta^2+4\delta b\right)^{\alpha-2}\left(\delta^2+4\delta b + -\delta(1\delta+2b)\right) \geq 0\,.
\end{aligned}$$

Thus,

$$h'(\delta) \leq 0\,.$$

It holds

$$h(0) = (2b)^{2\alpha-1} - (2b)^{2\alpha-1} - 0 = 0\,.$$

Thus, $h(\delta) \leq 0$, thus, $g(b=\delta) \leq 0$, thus $g(b) \leq 0$, thus $f'(b) \leq 0$

$$f(0) = (\delta)^{2\alpha} - c^{2\alpha} - (\delta)^{2\alpha} + (c)^{2\alpha} - 2^{1-2\alpha}\left((\delta-c)^{2\alpha} - (\delta-c)^{2\alpha}\right) = 0\,.$$

Thus, $f(b) \leq 0$. □

We can write the three conditions $a \geq c$, $a + c \geq 2b$, and $a \leq b + c$ as $b \geq a - c \geq 2\max(0, b - c)$.

Lemma 4.48 (Case 1.2, $a \leq b + c$)**.** Let $\alpha \in [\frac{1}{2}, 1]$. Let $a, b, c \geq 0$. Assume $b \geq a - c \geq 2\max(0, b - c)$. Then

$$a^{2\alpha} - c^{2\alpha} - \left((a-b)^2 + 4cb\right)^{\alpha} + (c+b)^{2\alpha} \leq 2b^{2\alpha-1}(a-c) \leq 2b(a-c)^{2\alpha-1}\,.$$

Proof. As $b \geq a - c$ and $2\alpha - 1 \in [0, 1]$, it holds $b^{2\alpha-1}(a-c) \leq b(a-c)^{2\alpha-1}$. Define

$$f(a) := a^{2\alpha} - c^{2\alpha} - \left((a-b)^2 + 4cb\right)^{\alpha} + (c+b)^{2\alpha} - 2b^{2\alpha-1}(a-c)\,.$$

It holds

$$\begin{aligned}
f'(a) &= 2\alpha a^{2\alpha-1} - 2\alpha(a-b)\left((a-b)^2+4cb\right)^{\alpha-1} - 2b^{2\alpha-1} \\
&\overset{a\geq b,c \text{ and } 2\alpha-2\leq 0}{\leq} 2\alpha a^{2\alpha-1} - 2\alpha(a-b)\,(a+b)^{2\alpha-2} - 2b^{2\alpha-1} \\
&= 2\left(\alpha a^{2\alpha-1} - \alpha\frac{a-b}{a+b}\,(a+b)^{2\alpha-1} - b^{2\alpha-1}\right)\,.
\end{aligned}$$

Set $x = \left(\frac{a-b}{a+b}\right)^{\frac{1}{2\alpha-1}} \leq 1$, $y = \alpha^{\frac{1}{2\alpha-1}} \leq 1$. Then

$$\begin{aligned}
\alpha a^{2\alpha-1} - \alpha\frac{a-b}{a+b}(a+b)^{2\alpha-1} - b^{2\alpha-1} &\leq (ya)^{2\alpha-1} - (xya+xyb)^{2\alpha-1} - b^{2\alpha-1} \\
&\leq (ya - xya - xyb)^{2\alpha-1} - b^{2\alpha-1} \\
&\leq (ya - xya - xyb - b)^{2\alpha-1} \\
&= ((y-xy)a - (xy+1)b)^{2\alpha-1} \\
&\leq (a-b)^{2\alpha-1} \\
&\leq 0\,.
\end{aligned}$$

Thus, $f'(a) \leq 0$. Thus, only need to show $f(b) \leq 0$. Assume $b \geq c$. Then

$$\begin{aligned}
f(b) &= b^{2\alpha} - c^{2\alpha} - (4cb)^{\alpha} + (c+b)^{2\alpha} - 2b^{2\alpha-1}(b-c) \\
&= -c^{2\alpha} - (4cb)^{\alpha} + (c+b)^{2\alpha} - b^{2\alpha} + 2b^{2\alpha-1}c \\
&\leq (c+b)^{2\alpha} - b^{2\alpha} - c^{2\alpha} - (4^{\alpha}-2)\,(cb)^{\alpha} \\
&= (c+b)^{2\alpha} - (b^{\alpha}-c^{\alpha})^2 - 4^{\alpha}\,(cb)^{\alpha}\ .
\end{aligned}$$

Thus, the next lemma implies $f(b) \leq 0$. □

Lemma 4.49. Let $\alpha \in [\frac{1}{2}, 1]$, $x, y \geq 0$. Then

$$(x+y)^{2\alpha} - (x^{\alpha} - y^{\alpha})^2 \leq (4xy)^{\alpha}\,.$$

We need two further lemmas before we prove this inequality.

Lemma 4.50. For $s \in [0, \frac{1}{2}]$ it holds

$$\frac{1-s}{s} \leq \frac{\log(s)}{\log(1-s)}\,.$$

Proof. Define

$$f(s) := s\log(s) - (1-s)\log(1-s)\,.$$

It hold

$$f''(s) = \frac{1}{s} - \frac{1}{1-s}\,.$$

Thus, $f''(s) \geq 0$ for $s \leq \frac{1}{2}$. It holds $f(0) = f(\frac{1}{2}) = 0$. Thus, $f(s) \leq 0$. Thus,

$$s\log(s) \leq (1-s)\log(1-s)\,.$$

Because of $\log(1-s) \le 0$, thus implies

$$\frac{1-s}{s} \le \frac{\log(s)}{\log(1-s)}\,.$$

□

Lemma 4.51. Let $a, b \ge 0$, $x \in [1,2]$. Define

$$f(x) := \frac{a^x - b^x}{(a+b)^x}\,.$$

Assume $a \ge b$. Then $f''(x) \le 0$. In particular,

$$\inf_{x\in[1,2]} f(x) = f(1) = f(2) = \frac{a^2-b^2}{(a+b)^2} = \frac{a-b}{a+b}\,.$$

Proof. It holds

$$f''(x) = (a+b)^{-x}\left(a^x \log\left(\frac{a}{a+b}\right)^2 - b^x \log\left(\frac{b}{a+b}\right)^2\right)\,.$$

Set $s = \frac{b}{a+b}$. Then $1-s = \frac{a}{a+b}$. Then Lemma 4.50 implies

$$\frac{a}{b} \le \frac{\log\left(\frac{b}{a+b}\right)}{\log\left(\frac{a}{a+b}\right)}\,.$$

Thus,

$$\left(\frac{a}{b}\right)^x \le \left(\frac{a}{b}\right)^2 \le \left(\frac{\log\left(\frac{b}{a+b}\right)}{\log\left(\frac{a}{a+b}\right)}\right)^2\,.$$

Thus,

$$a^x \log\left(\frac{a}{a+b}\right)^2 \le b^x \log\left(\frac{b}{a+b}\right)^2\,.$$

Thus, $f''(x) \le 0$. □

Proof of Lemma 4.49. For $z \ge 1$ define

$$f(z) := \left(z + 2 + z^{-1}\right)^\alpha - z^\alpha - z^{-\alpha}\,.$$

We will show that $f(z) \le 4^\alpha - 2$. This implies

$$\frac{(z+1)^{2\alpha} - z^{2\alpha} - 1}{z^\alpha} \le 4^\alpha - 2\,.$$

Thus,

$$(z+1)^{2\alpha} \leq (4^\alpha - 2)\, z^\alpha + z^{2\alpha} + 1\,.$$

By setting $z = \frac{x}{y}$ for $x \geq y$, we obtain

$$(x+y)^{2\alpha} - (x^\alpha - y^\alpha)^2 \leq (4xy)^\alpha\,.$$

The condition $x \geq y$ can be dropped because of symmetry. It remains to show that $f(z) \leq 4^\alpha - 2$ is indeed true. It holds $f(1) = 4^\alpha - 2$. To finish the proof, we will show $f'(z) \leq 0$. Define

$$g(z) := (z^2-1)(z+2)^{2\alpha} - (z+1)^2(z^{2\alpha}-1)\,.$$

Then

$$f'(z)\frac{z^{\alpha+2}\left(z+2+z^{-1}\right)}{\alpha} = g(z)\,.$$

We show $g(z) \leq 0$, and therefore $f'(z) \leq 0$, by applying Lemma 4.51 with $a = z$, $b = 1$, and $x = 2\alpha$:

$$\frac{z^x - 1^x}{(z+1)^x} \geq \frac{z^2 - 1^2}{(z+1)^2}\,,$$

which implies

$$\left(z^{2\alpha}-1\right)(z+1)^2 \geq \left(z^2-1\right)(z+1)^{2\alpha}\,.$$

□

According to Remark 4.30, we have now finally finished to proof of Lemma 4.26 and therefore of Theorem 4.10.

5 Regression in Non-Euclidean Spaces

Contents

5.1 Introduction

We have established rates convergence for FMs in general settings in chapter 4. Now assume the FM depends on a covariate x and we want to know how the FM changes with x. Essentially, the object of interest is a conditional FM, which is a function from the space of covariates to the metric space in which the FM lives. Here, we consider the simple case of x being an element of the unit interval and observing data at fixed points. In this setting we develop different regression techniques and – similarly to chapter 4 – try to find rates of convergence in general settings.

To be more precise, our goal is to estimate an unknown function $[0,1] \to \mathcal{Q}, t \mapsto m_t$, where $(\mathcal{Q}, d)$ is a metric space. To this end, we have access to independent data $(x_i, y_i)_{i=1,\dots,n}$. We assume that the covariates are fixed, e.g., $x_i = \frac{i}{n}$, and y_i is a random variable with values in $\mathcal{Q}$ such that its Fréchet mean is equal to m_{x_i}, i.e., $m_{x_i} = \arg\min_{q \in \mathcal{Q}} \mathbb{E}[d(y_i, q)^2]$.

Nonparametric regression with metric target values is developed, e.g., in [Hei09; Dav+10; PM19a]. [LM19] present a regression technique with regularization by total variation. [SHS10] discuss nonparametric regression techniques between Riemannian manifolds. [LMP20] develop an additive regression model with responses in spaces of of symmetric positive-definite matrices with a generalization to Riemannian manifolds. Based on the notion of geodesics, [Fle13] introduces an analog of linear regression in symmetric Riemannian manifolds. These results are generalized and extended in [Cor+17].

5.1.1 Settings

We will present our results in three levels of abstraction: the hypersphere $\mathbb{S}^k$, certain classes of metric spaces $\mathcal{Q}$ like Hadamard spaces and metric spaces of finite diameter, and an even more general setting which is governed by what kinds of meaningful statements can be proven for abstract mathematical objects.

Hypersphere. Let $k \in \mathbb{N}$. Let $\mathbb{S}^k := \{x \in \mathbb{R}^{k+1} \colon |x| = 1\}$ be the hypersphere with radius 1 as a subset of $\mathbb{R}^{k+1}$. We equip $\mathbb{S}^k$ with its intrinsic metric $d(q,p) := \arccos(q^\top p)$. Let $\mathsf{T}\mathbb{S}^k := \bigcup_{q \in \mathbb{S}^k} (\{q\} \times \mathsf{T}_q \mathbb{S}^k)$ be the tangent bundle, where $\mathsf{T}_q \mathbb{S}^k := \{v \in \mathbb{R}^{k+1} \mid q^\top v = 0\}$ is the tangent space at $q \in \mathbb{S}^k$. The exponential map is $\mathsf{Exp} \colon \mathsf{T}\mathbb{S}^k \to \mathbb{S}^k$, $(q, v) \mapsto \cos(|v|)q + \sin(|v|)\frac{v}{|v|}$, where $|\cdot|$ denotes the Euclidean norm. Geodesics can be represented by a tuple $(p, v) \in \mathsf{T}\mathbb{S}^k$ as $x \mapsto \mathsf{Exp}(p, xv)$.

For $t \in [0,1]$, let Y_t be a $\mathbb{S}^k$-valued random variable. Let the regression function $m \colon [0,1] \to \mathbb{S}^k$ be a minimizer $m_t \in \arg\min_{q \in \mathbb{S}^k} \mathbb{E}[d(Y_t, q)^2]$. Let $x_i = \frac{i}{n}$ and let

$(y_i)_{i=1,\dots,n}$ be independent random variables with values in $\mathbb{S}^k$ such that y_i has the same distribution as Y_{x_i}.

Metric. Let $(\mathcal{Q}, d)$ be a metric space. For $t \in [0,1]$, let Y_t be a $\mathcal{Q}$-valued random variable with finite second moment, i.e., $\mathbb{E}[d(Y_t, q)^2] < \infty$ for all $t \in [0,1]$ and $q \in \mathcal{Q}$. Let the regression function $m\colon [0,1] \to \mathcal{Q}$ be a minimizer $m_t \in \arg\min_{q \in \mathcal{Q}} \mathbb{E}[d(Y_t, q)^2]$. Let $x_i = \frac{i}{n}$ and let $(y_i)_{i=1,\dots,n}$ be independent random variables with values in $\mathcal{Q}$ such that y_i has the same distribution as Y_{x_i}.

General. Let $\mathcal{X}$ be the space of covariates, $\mathcal{Y}$ a set called data space, $\mathcal{Q}$ a set called descriptor space. Let $\mathfrak{c}\colon \mathcal{Y} \times \mathcal{Q} \to \mathbb{R}$ be a cost function, $\mathfrak{l}\colon \mathcal{Q} \times \mathcal{Q} \to \mathbb{R}$ be a loss function. For $t \in \mathcal{X}$, let Y_t be a $\mathcal{Y}$-valued random variable with finite expected cost, i.e., $\mathbb{E}[|\mathfrak{c}(Y_t, q)|] < \infty$ for all $t \in \mathcal{X}$ and $q \in \mathcal{Q}$. Let the regression function $m\colon \mathcal{X} \to \mathcal{Q}$ be a minimizer $m_t \in \arg\min_{q \in \mathcal{Q}} \mathbb{E}[\mathfrak{c}(Y_t, q)]$. Let $x_i \in \mathcal{X}$ be deterministic and let $(y_i)_{i=1,\dots,n}$ be independent random variables with values in $\mathcal{Y}$ such that y_i has the same distribution as Y_{x_i}.

5.1.2 Two Approaches

To construct an estimator for m in these settings, one may try to adapt a known Euclidean estimator to the new scenario. Two prominent approaches to this task are Fréchet regression [PM19a] and geodesic regression [Fle13].

The Fréchet approach. The regression function m_t is the Fréchet mean of Y_t, i.e., the minimizer of $\mathbb{E}[d(Y_t, q)^2]$ over $q \in \mathcal{Q}$. In Fréchet regression, we estimate the function $t \mapsto \mathbb{E}[d(Y_t, q)^2]$ for every fixed $q \in \mathcal{Q}$ by an Euclidean estimator $t \mapsto \hat{F}_t(q)$ using the data $(x_i, z_{q,i})_{i=1,\dots,n} \subseteq [0,1] \times \mathbb{R}$ with $z_{q,i} = d(y_i, q)^2$. In this step we may use one of the standard parametric or nonparametric regression estimators for certain classes of functions $[0,1] \to \mathbb{R}$. Then $\hat{F}_t(q)$ is minimized over $q \in \mathcal{Q}$ for a fixed t to obtain the estimator $\hat{m}_t$.

The Geodesic approach. Assume our metric space $\mathcal{Q}$ is equipped with an exponential map $\mathsf{Exp}\colon \Theta \to \mathcal{Q}$, where $\Theta \subseteq \mathsf{T}\mathcal{Q} \subseteq \mathcal{Q} \times \mathbb{R}^k$ is a subset of the tangent bundle of $\mathcal{Q}$. A geodesic starting in point $p \in \mathcal{Q}$ and continuing in the direction $v \in \mathsf{T}_p\mathcal{Q}$ of the tangent space of $\mathcal{Q}$ at p can be described as a function $\mathbb{R} \to \mathcal{Q}$, $x \mapsto \mathsf{Exp}(p, xv)$ with $(p, v) \in \Theta$. In geodesic regression with covariates $x_i \in \mathbb{R}$, we minimize the empirical squared error

$$\sum_{i=1}^{n} d(y_i, \mathsf{Exp}(p, x_i v))^2$$

over $(p, v) \in \Theta$ to find the best fitting geodesic. All forms of geodesic regression built on this criterion or a modification of it. For example, we can extend it to multivariate regression

$$\sum_{i=1}^{n} d\left(y_i, \mathsf{Exp}\left(p, \sum_{j=1}^{k} x_{i,j} v_j\right)\right)^2,$$

where $x_i \in \mathbb{R}^k$ and $v_1, \dots, v_k \in \mathsf{T}_p\mathcal{Q}$ or more general feature regression

$$\sum_{i=1}^{n} d\left(y_i, \mathsf{Exp}\left(p, \sum_{j=1}^{k} \psi_j(x_i) v_j\right)\right)^2 ,$$

where $x_i \in \mathcal{X}$ for an arbitrary space of covariates $\mathcal{X}$ and features $\psi_j \colon \mathcal{X} \to \mathbb{R}$. Furthermore, we may introduce weights $w_{i,t}$, e.g., $w_{i,t} = K((x_i - t)/h)$ for a kernel K and a bandwidth $h > 0$ to localize the procedure, and obtain (for one-dimensional covariates)

$$\hat{m}_t \in \operatorname*{arg\,min}_{(p,v)\in\Theta} \sum_{i=1}^{n} w_i d(y_i, \mathsf{Exp}(p, x_i v))^2 .$$

In this paper, we do not require the existence of an exponential map in the sense of Riemannian geometry. Instead, we introduce a link function $g \colon \Theta \times \mathcal{X} \to \mathcal{Q}$ for a set of covariates $\mathcal{X}$ and a set Θ, which we can think of as parameterizing geodesics. We then minimize $\sum_{i=1}^{n} d(y_i, g(\theta, x_i))^2$. For $\mathcal{X} \subseteq \mathbb{R}$, this generalizes the setting used above via $\Theta \subseteq \mathsf{T}\mathcal{Q}$ and $g((p, v), x) = \mathsf{Exp}(p, xv)$.

5.1.3 Contributions

We compare the two approaches of geodesic (`Geo`) and Frechet (`Fre`) regression on three regression estimators, namely linear regression (`Lin`), local linear regression (`Loc`), and the trigonometric orthogonal series projection estimator (`Tri`). This makes six estimation procedures, which we refer to as `LinGeo`, `LinFre`, `LocGeo`, `LocFre`, `TriGeo`, and `TriFre`. For the resulting estimators, which we denote as $\hat{m}_t$, our goal is to show explicit finite sample bounds of the form $\mathbb{E}[d(m_t, \hat{m}_t)^2] \leq Cn^{-\alpha}$ (in the metric setting), where $C > 0$ is a constant. We are not interested in optimal universal constants, e.g., whether $C = 2$ or $C = 2000$, but the dependence on further parameters, like a moment bound, is to be explicit.

- `LinGeo` (section 5.2): For standard geodesic regression in symmetric Riemannian manifolds, [Fle13] shows existence and uniqueness of the estimator as well as equivalence of the least squares estimator and the maximum likelihood estimator with Gaussian errors. [Cor+17] prove asymptotic normality results in this setting. We show $\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} d(m_{x_i}, \hat{m}_{x_i})^2] \leq Cn^{-1}$ for $n \in \mathbb{N}$ and a constant $C > 0$ in Hadamard spaces Corollary 5.4 and general metric spaces of finite diameter Corollary 5.5. These results are derived from an even more general statement, Theorem 5.2.

- `LinFre` (section 5.3): Among other Fréchet regression methods, linear (or global) Fréchet regression is developed in [PM19a]. We want to apply this estimator in a model where the regression function is a geodesic in a non-Euclidean space. For this setup, we show a negative result. The estimator of the objective function $t \mapsto \mathbb{E}[d(Y_t, q)^2]$ is only consistent in standard spaces, Theorem 5.8. Our simulations show inconsistency of `LinFre` on the sphere in our model. As an alternative, we suggest a modified estimator, `LinCos`, which maximizes the cosine of

the distance instead of minimizing the squared distance. The simulations suggest consistency of `LinCos`. We also give some theoretical justification. But we do not investigate rates of convergence, as `LinCos` is a method specific to the sphere $\mathbb{S}^2$ (with possible extensions to the hypersphere $\mathbb{S}^k$ and hyperbolic spaces) and not a regression technique for more general nonstandard spaces, which is the topic of this chapter.

- `LocGeo` (section 5.4): We apply the approach of geodesic regression to the well-known local linear estimator and arrive at a new estimator, `LocGeo`. We show $\mathbb{E}[d(m_t, \hat{m}_t)^2] \leq Cn^{-\frac{2\beta}{2\beta+1}}$ for all $t \in [0,1]$, $n \in \mathbb{N}$, a smoothness parameter $\beta \in (1,2]$, and a constant $C > 0$, Theorem 5.12, Corollary 5.14, Corollary 5.15. For this result, we assume a smoothness condition, which generalizes the Hölder condition that is common for local linear estimators. It demands that the true function $t \mapsto m_t$ can be locally approximated at t by a geodesic up to an error of order $|x-t|^\beta$ for x close to t.

- `LocFre` (section 5.5): [PM19a] introduce local constant (Nadaray–Watson) and local linear Fréchet regression for general metric spaces. For the local linear estimator, they show $d(\hat{m}_t, m_t) \in \mathbf{O}_{\mathbb{P}}(n^{-\frac{2}{5}})$ and a more general version of this result, see Corollary 1 in their article. This result is refined in [CM20] to a bound that is uniform in t. We show, for a general local polynomial Fréchet estimator of order $\ell \in \mathbb{N}_0$, that $\mathbb{E}[d(m_t, \hat{m}_t)^2] \leq Cn^{-\frac{2\beta}{2\beta+1}}$ for a constant $C > 0$ and a smoothness parameter $\beta > \ell$, Theorem 5.17, Corollary 5.19, Corollary 5.20. Our results are slightly more general with conditions slightly less demanding. Furthermore, bounds in expectation for finite n are stronger than in $\mathbf{O}_{\mathbb{P}}$. Similar to [PM19a], we demand a smoothness condition not directly on $t \mapsto m_t$, but on the change of the probability density of Y_t in t.

- `TriGeo` (section 5.6): The application of the geodesic approach to the trigonometric projection estimator yields a new estimator, `TriGeo`. We are not able to derive results on rates of convergence. We argue, that this estimator may be sub-optimal as the properties that make it appealing in Euclidean spaces are lost in nonstandard spaces. Nonetheless, we include the estimator in our simulation study.

- `TriFre` (section 5.7): We apply the approach of Fréchet regression to the trigonometric projection estimator and arrive at a new estimator, `TriFre`. We show $\mathbb{E}[\int_0^1 d(m_t, \hat{m}_t)^2 \mathrm{d}t] \leq Cn^{-\frac{2\beta}{2\beta+1}}$ for a smoothness parameter $\beta \geq 1$ and a constant $C > 0$, Theorem 5.23, Corollary 5.24, Corollary 5.25. As for `LocFre` the smoothness condition is a requirement on the change of the density of Y_t in t.

We briefly summarize these results for the two approaches.

- **Results on the Fréchet approach.** For the nonparametric estimators, we can make assumptions that ensure smoothness of the objective function. Thus, it can be approximated well by a finite cut-off of the expansion in the trigonometric series (`TriFre`) or locally by a linear function (`LocFre`). Together with a so-called

variance inequality, this yields the nonparametric rate of convergence of the minimizers, i.e., the nonparametric Fréchet regression estimators. As the objective function may not be linear in a geodesic model on non-Euclidean spaces, applying the Fréchet approach to linear regression (`LinFre`) may yield an inconsistent estimator in these cases.

- **Results on the geodesic approach.** The geodesic approach applied to linear regression (`LinGeo`) is a least squares estimator for geodesics. In a model where the regression function is a geodesic with a corresponding least squares property with respect to the noise, the resulting estimator is well-suited. Localizing the procedure (`LocGeo`) allows to estimate functions that can locally be approximated well by geodesics. On one hand, the rationale for these two procedures is almost straight forward. On the other hand, one can hardly justify the trigonometric basis of $\mathbb{L}_2[0,1]$ as meaningful features (`TriGeo`), because the basis functions loose the orthogonality property after their output is mapped to a non-Euclidean target space.

The comparison of theses estimation procedures underlines the importance to have a versatile tool belt when tackling new challenges: There is not one approach alone that solves the problem of nonstandard regression in every scenario. For a simple geodesic model, only the geodesic approach leads to a consistent estimator. For trigonometric regression the results are basically reversed: We can prove rates of convergence only for the Fréchet approach. For local linear estimation both approaches seem equally well suited. This comparison of geodesic and Fréchet approach on three different Euclidean estimators leads to three different outcomes. Thus, focusing on one setting alone would not reveal the complexities in the general comparison of the two approaches.

Our goal is to make all theorems as general as reasonably possible. This manifests in quite abstract statements. To get a gist of the meaning of the abstract objects, we start most sections by a corollary of a general theorem on the sphere: Corollary 5.1, Corollary 5.11, Corollary 5.16, and Corollary 5.22. These corollaries illustrate our results and show that they are indeed applicable to explicit interesting nonstandard spaces. Furthermore, abstract assumptions of the general theorems are justified by showing that they are fulfilled on the sphere.

The sphere is also the metric space used in our simulation study, section 5.8. To fulfill a variance inequality, which is an assumption for all our results, we introduce a new family of distributions on the sphere, the contracted uniform distributions. All estimators are implemented using the statistical programming language R [R D08]. The resulting package is freely available at `https://github.com/ChristofSch/spheregr`. Our experiments confirm and illustrate the theoretical findings.

The proofs (appendix 5.A) partially built upon techniques developed in [Sch19b]. Therein a so-called *weak quadruple inequality* is assumed to prove rates of convergence for the (generalized) Fréchet mean without requiring that the descriptor space $\mathcal{Q}$ is bounded. We fulfill this condition by definition of our moment conditions. Generally, the major tools to prove results in this setting are empirical process theory with chaining, e.g. [VW96] or [Tal14] and appendix 5.B, and a technique called *slicing* or *peeling*, e.g.,

[Gee00]. The proofs for local regression techniques follow the Euclidean version in [Tsy08, section 1.6] as far as possible, for trigonometric regression we make use of [Tsy08, section 1.7].

5.1.4 Notation and Conventions

We use a lower case c for universal constants $c > 0$. If the value depends on a variable, we indicate this by an index, e.g., c_κ is a constant that depends only on κ. We do not introduce or define every such constant. They are silently understood to take an appropriate value. Furthermore, the value may vary between two occurrences of such a constant. Alternatively we may use c', c'', ... for the same purpose.

A capital C indicates a constant that has further meaning, which is usually described by a three letter index, e.g., we may require a moment condition $\mathbb{E}[d(Y_t, m_t)^2] \le C_{\mathsf{Mom}}$ for all t to be fulfilled. For simplicity, we assume these constants to be ≥ 1, so that $C_{\mathsf{Abc}}^2 + C_{\mathsf{Abc}} C_{\mathsf{Xyz}} \le c C_{\mathsf{Abc}}^2 C_{\mathsf{Xyz}}$.

Assumptions are named in small caps, e.g., MOMENT. Different assumptions in different sections may have the same name. Hence, assumptions always refer to the assumptions defined in the same section. Nonetheless, the names are consistent across the sections insofar as assumptions with the same name are – if not identical – expressions of the same underlying requirement.

There is a silently underlying probability space $(\Omega, \Sigma_\Omega, \mathbb{P})$. If a random variable, say Y, has values in a set, say $\mathcal{Y}$, that set is silently understood to be a measurable space $(\mathcal{Y}, \Sigma_\mathcal{Y})$ and the random variable is a measurable map $Y\colon (\Omega, \Sigma_\Omega) \to (\mathcal{Y}, \Sigma_\mathcal{Y})$.

In each section the estimator of the regression function at t is denoted as $\hat{m}_t$. It depends on n and potentially on further parameters like a bandwidth h, which will not be indicated in the notation but should be clear in the context.

Let $(\mathcal{Q}, d)$ be a metric space. To shorten the notation, we sometimes write $\overline{q,p}$ instead of $d(q,p)$ for $q, p \in \mathcal{Q}$. Define the ball $\mathrm{B}(o, d, \delta) := \{q \in \mathcal{Q}\colon d(q,o) < \delta\}$ and the diameter $\mathsf{diam}(\mathcal{Q}, d) := \sup_{q,p\in\mathcal{Q}} d(q,p)$.

For a vector $v \in \mathbb{R}^k$, we denote its Euclidean norm by $|v|$. For a matrix $A \in \mathbb{R}^{k\times\ell}$, we denote its operator norm by $\|A\|_{\mathsf{op}} := \sup_{v\in\mathbb{R}^\ell, |v|=1} |Av|$.

5.2 Linear Geodesic Regression

We apply the geodesic approach to linear regression, which yields the standard geodesic regression, `LinGeo`, introduced in [Fle13].

5.2.1 Hypersphere

Before we present the general and abstract results, we illustrate them in the hypersphere setting, see section 5.1.1. Let $\Lambda \in [1, \infty)$. Let $\Theta := \{(p,v) \in \mathsf{T}\mathbb{S}^k \,|\, |v| \le \Lambda\}$. The regression function $m\colon [-1,1] \to \mathbb{S}^k$ is assumed to be a geodesic $m_t = \mathsf{Exp}(p^*, tv^*)$, $(p^*, v^*) \in \Theta$. We observe $(x_i, y_i)_{i=1,\dots,n}$ on a regular grid $(x_i)_{i=1,\dots,n}$ of $[-1,1]$ (instead of $[0,1]$).

We estimate the starting point p^* and velocity vector v^* by the least squares method in $\mathbb{S}^k$, i.e.,

$$(\hat{p}, \hat{v}) = \operatorname*{arg\,min}_{(p,v)\in\Theta} \frac{1}{n}\sum_{i=1}^{n} d(y_i, \mathsf{Exp}(p, x_i v))^2 \,.$$

The estimated curve then is $t \mapsto \hat{m}_t = \mathsf{Exp}(\hat{p}, t\hat{v})$.

Using our general theory in the next section, we obtain following corollary.

Corollary 5.1 (`LinGeo` Hypersphere)**.** Assume there is $C_{\mathsf{Vlo}} \geq 1$ with $C_{\mathsf{Vlo}}^{-1} d(m_t, q)^2 \leq \mathbb{E}[d(Y_t, q)^2 - d(Y_t, m_t)^2]$ for all $t \in [-1, 1]$ and $q \in \mathbb{S}^k$. Then

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} d(m_{x_i}, \hat{m}_{x_i})^2\right] \leq C\frac{1}{n}\,,$$

where $C := ckC_{\mathsf{Vlo}}\Lambda^{-2}$.

5.2.2 General

We now present a result in the general setting of section 5.1.1. Let Θ be a space of parameters. Let $g\colon \mathcal{X} \times \Theta \to \mathcal{Q}$ be the link function. Our model assumption is $g(t, \theta^*) = m_t$ for the true parameter $\theta^* \in \Theta$. The canonical M-estimator of θ^* is

$$\hat{\theta} \in \operatorname*{arg\,min}_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \mathfrak{c}(y_i, g(x_i, \theta))\,.$$

The resulting plug-in estimator for the regression function m_t is $\hat{m}_t = g(t, \hat{\theta})$.

We introduce some further notation. For $\theta, \tilde{\theta}, \theta_0 \in \Theta$, $x_1, \ldots, x_n, t \in \mathcal{X}$, $y, \tilde{y} \in \mathcal{Y}$, define

$$\begin{aligned}
\mathfrak{c}_t(y, \theta) &:= \mathfrak{c}(y, g(t, \theta)) \\
\Diamond_t(y, \tilde{y}, \theta, \tilde{\theta}) &:= \mathfrak{c}_t(y, \theta) - \mathfrak{c}_t(\tilde{y}, \theta) - \mathfrak{c}_t(y, \tilde{\theta}) + \mathfrak{c}_t(\tilde{y}, \tilde{\theta}) \\
\mathbf{x} &:= (x_1, \ldots, x_n) \\
\mathrm{B}_{\mathbf{x}}(\theta_0, \mathfrak{l}, \delta) &:= \left\{\theta \in \Theta\colon \frac{1}{n}\sum_{i=1}^{n} \mathfrak{l}(g(x_i, \theta), g(x_i, \theta_0)) \leq \delta\right\} \\
\mathfrak{a}_t(y, \tilde{y}) &:= \sup_{\theta_1, \theta_2 \in \Theta, \theta_1 \neq \theta_2} \frac{\Diamond_t(y, \tilde{y}, \theta_1, \theta_2)}{\mathfrak{b}(\theta_1, \theta_2)}\,.
\end{aligned}$$

Assumptions.

- VARIANCE: There is $C_{\mathsf{Vlo}} \in [1, \infty)$ such that $C_{\mathsf{Vlo}}^{-1}\mathfrak{l}(m_t, q) \leq \mathbb{E}[\mathfrak{c}(Y_t, q) - \mathfrak{c}(Y_t, m_t)]$ for all $t \in \mathcal{X}$ and $q \in \mathcal{Q}$.

- ENTROPY: There are $T_n \geq 0$, $C_{\mathsf{Ent}} \in [1, \infty)$, and $\xi \in (0, 1)$ such that $\gamma_2(\mathrm{B}_{\mathbf{x}}(\theta_0, \mathfrak{l}, \delta), \mathfrak{b}) \leq B\delta^{\xi}$ for all $\delta \geq T_n$ and $\theta_0 \in \Theta$, where γ_2 is a measure of entropy defined in Definition 5.55 (appendix 5.B).

- MOMENT: There are $\kappa \geq 2$ and $C_{\mathsf{Mom}} \in [1, \infty)$ such that $\mathbb{E}[\mathfrak{a}_t(Y_t, m_t)^{\kappa}]^{\frac{1}{\kappa}} \leq C_{\mathsf{Mom}}$ for all $t \in \mathcal{X}$.

Theorem 5.2 (`LinGeo` General)**.** Assume VARIANCE, ENTROPY, MOMENT. Assume $\kappa(1 - \xi) > 1$. Then

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \mathfrak{l}(m_{x_i}, \hat{m}_{x_i})\right] \leq Cn^{-\frac{1}{2(1-\xi)}} + C_{\mathsf{Vlo}} T_n \,,$$

where $C := c_{\kappa,\xi} C_{\mathsf{Vlo}} \left(C_{\mathsf{Ent}} C_{\mathsf{Mom}}\right)^{\frac{1}{1-\xi}}$.

In many settings like in Euclidean linear regression with $\mathfrak{c}(y, q) = |y - q|^2$, $\mathfrak{l}(m, q) = |m - q|^2$, we have $\xi = \frac{1}{2}$ and retrieve the parametric rate of convergence.

Remark 5.3.

- VARIANCE:
  This condition is also called *variance inequality* (or GROWTH in chapter 4) and is well-known in the context of Fréchet means in Alexandrov spaces, [Stu03; Oht12; Gou+19]. VARIANCE is a condition on the noise distribution and the geometry of involved spaces. It can be viewed as a quantitative version of the condition of unique Fréchet means m_x of Y_x. The variance inequality not only ensures uniqueness of m_x, it also requires the objective function $\mathbb{E}[d(Y, q)^2]$ (in the metric space setting) or $\mathbb{E}[\mathfrak{c}(Y, q)]$ (in the general setting) to grow quadratically in the distance of a test point q to the minimizer m_x (metric) or linearly in $\mathfrak{l}$ (general). Intuitively, this is fulfilled when the noise distribution is not too similar to a distribution that has nonunique Fréchet means.

  In the metric setting, VARIANCE is true in Hadamard spaces [Stu03, Proposition 4.4], which are geodesic metric spaces with nonpositive curvature and include the Euclidean spaces. For a variance inequality in spaces of nonnegative curvature, see [ALP20, Theorem 3.3].

- ENTROPY:
  ENTROPY restricts the size of the sets $\mathrm{B}_{\mathbf{x}}(\theta^*, \mathfrak{l}, \delta)$. It can be viewed as a quantitative version of the requirement that these sets are totally bounded.

  We use Talagrand's γ_2 [Tal14] (Definition 5.55 in section 5.B) to formulate the

entropy condition. Let $(\mathcal{Q}, d)$ be a metric space and $\mathcal{B} \subseteq \mathcal{Q}$. It holds

$$\gamma_2(\mathcal{B}, \mathfrak{b}) \leq \int_0^\infty \sqrt{\log(N(\mathcal{B}, \mathfrak{b}, r))} \mathrm{d}r \,,$$

where the integral is called *entropy integral* and

$$N(\mathcal{B}, \mathfrak{b}, r) = \min \left\{ k \in \mathbb{N} \,\middle|\, \exists q_1, \dots, q_k \in \mathcal{Q} \colon \mathcal{B} \subseteq \bigcup_{j=1}^{k} \mathrm{B}(q_j, \mathfrak{b}, r) \right\}$$

is the *covering number*. Thus, we can use bounds on the entropy integral to fulfill Entropy, which is more common in the statistics literature. In some circumstances γ_2 is strictly lower than the entropy integral [Tal14, Exercise 4.3.11]. One can further weaken the entropy condition as done in [ALP20] and chapter 4 at the cost of worse rates of convergence, but it is not clear whether these results are optimal.

- Moment:
  We partially follow the approach of chapter 4 to prove the theorem. Therein a so-called *weak quadruple inequality* is assumed to prove rates of convergence for the (generalized) Fréchet mean. We fulfill this condition by the definition of $\mathfrak{a}_x$, i.e., it holds $\Diamond_x(y, z, \theta, \tilde{\theta}) \leq \mathfrak{b}(\theta, \tilde{\theta}) \mathfrak{a}_x(y, z)$. This inequality can be understood as a generalization of Cauchy-Schwarz inequality: If $\Theta = \mathcal{Y} = \mathbb{R}^k$ and $\mathfrak{c}_x(y, \theta) = |y - \theta|^2$, then

  $$\Diamond_x(y, z, \theta, \tilde{\theta}) = 2\langle y - z, \tilde{\theta} - \theta \rangle \leq \mathfrak{a}_x(y, z) \mathfrak{b}(\theta, \tilde{\theta})$$

  if $\mathfrak{b}(\theta, \tilde{\theta}) = |\tilde{\theta} - \theta|$ and $\mathfrak{a}_x(y, z) = 2\,|y - z|$. Moment then is nothing but the condition that the κ-th moment of the noise distribution is finite.

5.2.3 Corollaries

Next, we apply Theorem 5.2 to the metric setting of section 5.1.1. We will replace Entropy by two conditions that compare the distances of the metric space to the Euclidean distance.

Assumptions.

- MetricUp: There is $C_{\mathsf{Mup}} \in [1, \infty)$ such that $d(g(x, \theta), g(x, \tilde{\theta})) \leq C_{\mathsf{Mup}} |\theta - \tilde{\theta}|$ for all $x \in \mathcal{X}$, $\theta, \tilde{\theta} \in \Theta$.
- MetricLo: There are $C_{\mathsf{Res}}, C_{\mathsf{Mlo}} \in [1, \infty)$ such that, for all $\theta, \tilde{\theta} \in \Theta$,

  $$\frac{1}{n} \sum_{i=1}^{n} d(g(x_i, \theta), g(x_i, \tilde{\theta}))^2 \geq C_{\mathsf{Mlo}}^{-1} |\theta - \tilde{\theta}|^2 - C_{\mathsf{Res}} n^{-1} \,.$$

If we assume that $\mathcal{Q}$ is a Hadamard space, VARIANCE holds and we can set $\mathfrak{a}_x = d$ in MOMENT.

Corollary 5.4 (`LinGeo` Hadamard)**.** Let $(\mathcal{Q}, d)$ be a Hadamard space. Assume METRICUP, METRICLO, and MOMENT with $\mathfrak{a}_x = d$. Then

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} d(m_{x_i}, \hat{m}_{x_i})^2\right] \leq Cn^{-1},$$

where $C := c_\kappa \mathsf{d}_\Theta C_{\mathsf{Mlo}} C_{\mathsf{Mup}}^2 C_{\mathsf{Mom}}^2 C_{\mathsf{Res}}$.

In bounded metric spaces, i.e., metric spaces $(\mathcal{Q}, d)$ with $\mathsf{diam}(\mathcal{Q}) < \infty$, MOMENT is trivial, but VARIANCE needs to be assumed.

Corollary 5.5 (`LinGeo` Bounded)**.** Let $(\mathcal{Q}, d)$ be a bounded metric space. Assume METRICUP, METRICLO, and VARIANCE. Then

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} d(m_{x_i}, \hat{m}_{x_i})^2\right] \leq Cn^{-1}$$

where $C := c\mathsf{d}_\Theta C_{\mathsf{Vlo}} C_{\mathsf{Mlo}} C_{\mathsf{Mup}}^2 C_{\mathsf{Res}}\, \mathsf{diam}(\mathcal{Q})^2$

5.3 Linear Fréchet Regression

First, we directly apply the Fréchet approach to linear regression, which leads to an estimator, `LinFre`, introduced in [PM19a], that may be inconsistent in some intuitively sensible models on nonstandard spaces. Then, with a more relaxed interpretation of the Fréchet approach applied on the sphere, we introduce cosine regression, `LinCos`.

5.3.1 Model and Consistency

We use the metric model of section 5.1.1 with covariates in $[-1, 1]$ (instead of $[0, 1]$). The approach of Fréchet regression, as proposed in [PM19a], is to estimate $F_t(q) := \mathbb{E}[d(Y_t, q)^2]$ and then minimize that estimator over q to obtain an estimation $\hat{m}_t$. When applying this idea to the concept of linear regression, we estimate the function $t \mapsto F_t(q)$ for each $q \in \mathcal{Q}$ using a linear regression estimator on the real-valued quantities $d(y_i, q)^2$.

As a first step to validate this approach, we apply it to the case $\mathcal{Q} = \mathbb{R}^{\mathsf{d}_\mathcal{Y}}$ with the Euclidean metric. To make this work, we estimate not $F_t(q)$ but $F_t(q, o) := F_t(q) - F_t(o)$ for all $q \in \mathcal{Q}$ and a fixed element $o \in \mathcal{Q}$. This is necessary to have a linear objective in the case of a linear model as following calculations show: Assume $Y_t = \beta_0 + t\beta_1 + \epsilon$, where $\beta_0, \beta_1 \in \mathbb{R}^{\mathsf{d}_\mathcal{Y}}$, and ϵ is a centered random variable in $\mathbb{R}^{\mathsf{d}_\mathcal{Y}}$ with finite second moment.

Denote $\beta = (\beta_0, \beta_1) \in \mathbb{R}^{\mathtt{d}_\mathcal{Y} \times 2}$ and $t^\cdot = (1, t) \in \mathbb{R}^2$. Then, for $o, q \in \mathbb{R}^{\mathtt{d}_\mathcal{Y}}$,

$$F_t(q) = \mathbb{E}[|Y_t - q|^2] = |\beta t^\cdot|^2 + \mathbb{E}[|\epsilon|^2] + |q|^2 - 2(\beta t^\cdot)^\top q\,,$$
$$F_t(q, o) = |q|^2 - |o|^2 - 2(\beta t^\cdot)^\top (q - o)\,.$$

For fixed $q, o \in \mathcal{Q}$ the function $t \mapsto F_t(q, o)$ is linear in t, whereas $t \mapsto F_t(q)$ is quadratic. Note that $\arg\min_q F_t(q) = \arg\min_q F_t(q, o)$.

With these considerations in mind, we define the linear Fréchet regression estimator $\hat{m}_t$ in the metric setting as follows:

$$\mathbf{X}_n := \begin{pmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{pmatrix} \in \mathbb{R}^{2 \times n}\,, \qquad \hat{F}_t(q, o) := \sum_{i=1}^n w(t, x_i) \left(d(y_i, q)^2 - d(y_i, o)^2 \right)\,,$$
$$B_n := \mathbf{X}_n \mathbf{X}_n^\top\,, \qquad \hat{m}_t \in \operatorname*{arg\,min}_{q \in \mathcal{Q}} \hat{F}_t(q, o)\,,$$
$$w(t, x) := t^{\cdot\top} B_n^{-1} x^\cdot\,,$$

where $o \in \mathcal{Q}$ is an arbitrary fixed element. The empirical objective $\hat{F}_t(q, o)$ is the linear regression estimator of $F_t(q, o)$: Define $\hat{\theta}(q, o) := \arg\min_{\theta \in \mathbb{R}^{\mathtt{d}_\mathcal{Y} \times 2}} \frac{1}{n} \sum_{i=1}^n (z_i - \theta_0 - \theta_1 x_i)^2$ with $z_i := d(y_i, q)^2 - d(y_i, o)^2$. Then $\hat{F}_t(q, o) = \hat{\theta}(q, o)^\top t^\cdot$. [PM19a] showed (in a slightly different model) that, under some conditions, $\hat{m}_t$ converges to $\bar{m}_t$ for $n \to \infty$, where

$$B := \int_{-1}^1 x^\cdot x^{\cdot\top} \mathrm{d}x = \begin{pmatrix} 2 & 0 \\ 0 & \frac{2}{3} \end{pmatrix}\,, \qquad \bar{F}_t(q, o) := \int_{-1}^1 \mathbb{E}\Big[w(t, x) \left(d(Y_x, q)^2 - d(Y_x, o)^2 \right) \Big] \mathrm{d}x\,,$$
$$w(t, x) := t^{\cdot\top} B^{-1} x^\cdot = \frac{1}{2} + \frac{3}{2} x t\,, \quad \bar{m}_t := \operatorname*{arg\,min}_{q \in \mathcal{Q}} \bar{F}_t(q, o)\,.$$

The objective $t \mapsto \bar{F}_t(q, o)$ is the linear approximation of $t \mapsto F_t(q, o)$ in the least squares sense. In particular, if $t \mapsto F_t(q, o)$ is linear for every $q \in \mathcal{Q}$, then $\bar{F} = F$ and $\bar{m}_t = m_t$.

These considerations yield two ways of extending the idea of a linear model to arbitrary metric spaces.

Definition 5.6.

(i) The distributions of Y_x for $x \in [-1, 1]$ follow the *strict linear Fréchet regression model*, if $\bar{F} = F$.

(ii) The distributions of Y_x for $x \in [-1, 1]$ follow the *relaxed linear Fréchet regression model*, if $\bar{m} = m$.

As the idea of Fréchet regression is to estimate $\bar{F}$, this function being the true objective would give rise to a meaningful model, in which the linear Fréchet regression estimator is consistent due to the aforementioned results in [PM19a]. But in the end, we are only interested in the minimizers. Thus, it is sufficient (and necessary) to have $\bar{m} = m$ for a model, so that the linear Fréchet regression estimator is consistent.

To further affirm that these model assumptions are reasonable, the following proposition shows that they generalize the Euclidean linear model assumptions. As calculated above $F_t(q, o)$ is linear in t for Euclidean spaces. As $\bar{F}$ is the linear approximation of F, we have $F = \bar{F}$.

Proposition 5.7. Any Euclidean linear model, $Y_x = \beta_0 + \beta_1 x + \epsilon$, $x \in \mathbb{R}$, $\beta_0, \beta_1 \in \mathbb{R}^{\mathsf{d}_\mathcal{Y}}$, $\mathbb{E}[\epsilon] = 0 \in \mathbb{R}^{\mathsf{d}_\mathcal{Y}}$, follows the *strict linear Fréchet regression model* for the space $(\mathcal{Q}, d) = (\mathbb{R}^{\mathsf{d}_\mathcal{Y}}, |\cdot|)$.

It is not clear how to write a generalization of a model equation like $Y_x = \beta_0 + \beta_1 x + \epsilon$ in arbitrary metric spaces (except on Riemannian manifolds with an exponential map at hand, where $Y_x = \mathsf{Exp}(m_x, \epsilon), m_x = \mathsf{Exp}(\beta_0, x\beta_1)$ seems to be meaningful). But there are some elements of the model that should reasonably be included: If the metric space is a geodesic space, we should demand that a meaningful regression estimator is consistent at least in the no-noise settings $Y_t = \gamma_t$ for global geodesics $\gamma\colon [-1, 1] \to \mathcal{Q}, t \mapsto \gamma_t$, as this is the simplest distribution with a nonconstant regression function. Furthermore, geodesics in Euclidean spaces are linear functions, which can be estimated by linear Fréchet regression, as linear Fréchet regression is equivalent to linear regression in Euclidean spaces.

Unfortunately, Hilbert spaces are essentially the only spaces where this no-noise setting fulfills the *strict linear Fréchet regression model*, as the following theorem shows.

To state the theorem, we first need to further extend our knowledge of metric geometry from section 1.2.5. A **minimizing geodesic** between two points $q, p \in \mathcal{Q}$ is a geodesic $\gamma\colon [a, b] \to \mathcal{Q}$ with $L(\gamma) = d(\gamma(a), \gamma(b))$ and $\gamma(a) = q$, $\gamma(b) = p$. A geodesic $\gamma\colon [a, b] \to \mathcal{Q}$ is **extendible** (through both ends) if there is $\epsilon > 0$ and a geodesic $\tilde{\gamma}\colon [a - \epsilon, b + \epsilon] \to \mathcal{Q}$ such that $\tilde{\gamma}_{|[a,b]} = \gamma$. A geodesic space $(\mathcal{Q}, d)$ is **geodesically complete**, if it is complete and all geodesics are extendible.

Theorem 5.8 (`LinFre` inconsistency)**.** Let $(\mathcal{Q}, d)$ be a nonempty geodesic space. It is also a Hilbert space if and only if it is geodesically complete, and for each minimizing geodesic $(\gamma_t)_{t\in[-1,1]}$, the *strict linear Fréchet regression model* holds for the no-noise setting $Y_t = \gamma_t$.

Theorem 5.8 shows that it does not make sense to assume the strict linear Fréchet regression model in non-Euclidean spaces. It is not clear to the author, whether a statement similar to Theorem 5.8 holds for the relaxed model. But simulations in appendix 5.8 indicate inconsistency of linear Fréchet regression on the sphere.

5.3.2 Parametric Cosine Regression

One important aspect of linear regression is that the regression function has a simple parametric form. The idea behind Fréchet regression is to apply regression not directly to the regression function $t \mapsto m_t$, but to the objective function $t \mapsto F_t(q)$. Thus, for a

generalization of parametric regression in the sense of Fréchet regression, we would want to target function $t \mapsto F_t(q)$ to have a simple parametric form.

On the sphere $\mathbb{S}^2$, instead of minimizing the squared distance, we will maximize the cosine of the distance (or minimize the hyperbolic cosine in the hyperbolic plane $\mathbb{H}^2$), which will yield a simple parametric form of the objective function. It holds

$$\cos(x) = 1 - \frac{1}{2}x^2 + \mathbf{O}(x^4)\,, \qquad \cosh(x) = 1 + \frac{1}{2}x^2 + \mathbf{O}(x^4)\,.$$

Thus, minimizing $\mathbb{E}[\overline{Y,q}^2]$ is closely related to maximizing $\mathbb{E}[\cos(\overline{Y,q})]$ or minimizing $\mathbb{E}[\cosh(\overline{Y,q})]$. Furthermore, using the cosine on $\mathbb{S}^2$ (or hyperbolic cosine in $\mathbb{H}^2$) seems appealing as laws of cosines hold in $\mathbb{S}^2$ and $\mathbb{H}^2$ analogously to the Euclidean space: In a triangle with side lengths a, b, c and angle α opposing side c, it holds, in the respective space with intrinsic metric,

$$\begin{aligned} c^2 &= a^2 + b^2 - 2ab\cos(\alpha) && \text{Euclidean}\,,\\ \cos(c) &= \cos(a)\cos(b) + \sin(a)\sin(b)\cos(\alpha) && \text{sphereical}\,,\\ \cosh(c) &= \cosh(a)\cosh(b) - \sinh(a)\sinh(b)\cos(\alpha) && \text{hyperbolic}\,. \end{aligned}$$

We will only discuss $\mathcal{Q} = \mathbb{S}^2$ with intrinsic metric d. Similar considerations are valid for the hyperbolic space. In our new model, we replace the Fréchet mean, $\arg\min_q \mathbb{E}[\overline{Y,q}^2]$, of a random variable Y in $\mathbb{S}^2$ by the cosine mean, $\arg\max_q \mathbb{E}[\cos(\overline{Y,q})]$. For distributions with enough symmetry, the cosine mean can be characterized easily.

Proposition 5.9. Let Y be a random variable with values in $\mathbb{S}^2 = \{(\vartheta, \varphi) \in [-\frac{\pi}{2}, \frac{\pi}{2}] \times [0, 2\pi)\}$ such that its distribution is symmetric with respect to rotation around one axis, without loss of generality the axis connecting $(-\frac{\pi}{2}, 0)$ and $(\frac{\pi}{2}, 0)$. The distribution of Y is given by $\mathbb{P}(Y \in B) = \frac{1}{2\pi}\int\int_0^{2\pi} \mathbb{1}_B(\vartheta, \varphi)\mathrm{d}\varphi\mathrm{d}\nu(\vartheta)$ for all measurable $B \subseteq \mathbb{S}^2$, where ν is a probability measure on $[-\frac{\pi}{2}, \frac{\pi}{2}]$. Let $A := \int \sin(\vartheta)\mathrm{d}\nu(\vartheta)$. Let $M := \arg\max_{q\in\mathbb{S}^2} \mathbb{E}[\cos(\overline{Y,q})]$ the set of cosine means. Then it holds,

- $A < 0$ if and only if $M = \{(-\frac{\pi}{2}, 0)\}$,
- $A = 0$ if and only if $M = \mathbb{S}^2$,
- $A > 0$ if and only if $M = \{(\frac{\pi}{2}, 0)\}$.

As expected from a mean-value, the cosine mean of a symmetric distribution is its center. If the regression function is equal to a geodesic, the objective function has a simple parametric form.

Proposition 5.10. Let $\gamma\colon \mathbb{R} \to \mathbb{S}^2, s \mapsto \gamma_s$ be a unit-speed geodesic. Assume that $m_t := \arg\max_{q\in\mathbb{S}^2} \mathbb{E}[\cos(\overline{Y_t,q})]$ is unique for all $t \in [0, 1]$. Assume $m_t = \gamma_{t_0+\lambda t}$ with $t_0 \in [0, 2\pi)$ and $\lambda \in [0, \infty)$. For $q \in \mathbb{S}^2$, let $s_q \in [0, 2\pi)$ be such that $\min_s \overline{q,\gamma_s} = \overline{q,\gamma_{s_q}}$. Then

$$\mathbb{E}[\cos(\overline{Y_t,q})] = A_q\cos(B_q + \lambda t) = a_q\cos(\lambda t) + b_q\sin(\lambda t)\,,$$

where

$$A_q := \mathbb{E}[\cos(\overline{Y_t,m_t})]\cos(\overline{\gamma_{s_q},q})\,, \qquad a_q := A_q\cos(B_q)\,,$$
$$B_q := t_0 - s_q\,, \qquad b_q := -A_q\sin(B_q)\,.$$

Proposition 5.10 shows that following model is appropriate for estimating geodesics on the sphere. For $t \in [0,1]$, let Y_t be a $\mathbb{S}^2$-valued random variable. Let the regression function $m\colon [0,1] \to \mathbb{S}^2$ be a maximizer $m_t \in \arg\max_{q\in\mathbb{S}^2} \mathbb{E}[\cos(\overline{Y_t,q})]$. Let $\Lambda \in (0,\infty)$ and assume that $t \mapsto m_t$ is a geodesic with speed bounded by Λ. Assume that $\mathbb{E}[\cos(\overline{Y_t,m_t})]$ does not depend on t. Let $x_i := \frac{i}{n}$ and let $(y_i)_{i=1,\dots,n}$ be independent random variables with values in $\mathbb{S}^2$ such that y_i has the same distribution as Y_{x_i}.

Set $z_{q,i} := \cos(\overline{y_i,q})$ and define the least squares estimators

$$(\hat{a}_{q,\lambda},\hat{b}_{q,\lambda}) \in \operatorname*{arg\,min}_{a,b\in[-1,1]} \frac{1}{n}\sum_{i=1}^{n} \left(z_{q,i} - a\cos(\lambda x_i) - b\sin(\lambda x_i)\right)^2\,,$$
$$\hat{\lambda} \in \operatorname*{arg\,min}_{\lambda\in[0,\Lambda]} \int_{\mathbb{S}^2} \frac{1}{n}\sum_{i=1}^{n} \left(z_{q,i} - \hat{a}_{q,\lambda}\cos(\lambda x_i) - \hat{b}_{q,\lambda}\sin(\lambda x_i)\right)^2 \mathrm{d}\mu(q)\,,$$

where μ is the Hausdorff measure on $\mathbb{S}^2$ (for an implementation it is enough to use a three points measure $\mu = \frac{1}{3}(\delta_{q_1} + \delta_{q_2} + \delta_{q_3})$ with $q_1, q_2, q_3 \in \mathbb{S}^2$ not on the same geodesic). Now set $\hat{a}_q = \hat{a}_{q,\hat{\lambda}}$, $\hat{b}_q = \hat{b}_{q,\hat{\lambda}}$ and $\hat{F}_t(q) = \hat{a}_q\cos(\hat{\lambda}t) + \hat{b}_q\sin(\hat{\lambda}t)$. The `LinCos`-estimator for m_t is $\hat{m}_t \in \arg\max_{q\in\mathcal{Q}} \hat{F}_t(q)$.

We do not investigate `LinCos` deeply, as it mainly serves to illustrate the comparison of `LinFre` to `LinGeo`. Moreover, it does not fit into the scheme of this chapter, as we want to compare general regression methods which are not limited to one specific metric space and `LinCos` is only applicable in $\mathbb{S}^2$ (with possible extensions to hyper-spheres and hyperbolic spaces). But note that, for fixed q, the estimation of a_q, b_q, and λ is well-studied in the literature on *sinusoidal regression*, see e.g., [NK13].

5.4 Local Geodesic Regression

We apply the geodesic approach to the classical local linear estimator and arrive at a new procedure, local geodesic regression or `LocGeo`.

5.4.1 Hypersphere

In the hypersphere setting of section 5.1.1, let $\Theta \subseteq \mathsf{T}\mathbb{S}^k$ be the subset of the tangent bundle with $|v| < \pi$ for all $(q,v) \in \Theta$. This set parameterizes a set of geodesics $x \mapsto \mathsf{Exp}(q, xv)$.

We investigate an estimator that locally fits geodesics. Let $h \geq \frac{2}{n}$. Let $K\colon \mathbb{R} \to \mathbb{R}$ be a function, the kernel, such that $C_{\mathsf{Ker}}^{-1}\mathbb{1}_{[-\frac{1}{2},\frac{1}{2}]}(x) \leq K(x) \leq C_{\mathsf{Ker}}\mathbb{1}_{[-1,1]}(x)$ for a constant $C_{\mathsf{Ker}} \geq 1$ (Kernel condition). For $t \in [0,1]$, define $w_h(t,x) := \frac{1}{h}K(\frac{x-t}{h})$ and $w_i := w_h(t,x_i)(\sum_{j=1}^n w_h(t,x_j))^{-1}$. Let $(\hat{m}_t, \hat{v}_t) \in \arg\min_{(p,v)\in\Theta} \sum_{i=1}^n w_i d(y_i, \mathsf{Exp}(p, \frac{x_i-t}{h}v))^2$.

To be able to estimate m, it must fulfill a Hölder-type SMOOTHNESS condition: Assume there are $\beta > 0$ and $C_{\mathsf{Smo}} \in [1, \infty)$ such that $d(m_x, \mathsf{Exp}(m_t, (x-t)\dot{m}_t)) \leq C_{\mathsf{Smo}} |x-t|^\beta$ for all $x \in [t-h, t+h] \cap [0,1]$, $t \in [0,1]$, where $\dot{m}_t \in \mathsf{T}_{m_t}\mathbb{S}^k$ is the derivative of m_t.

Furthermore, we again assume a VARIANCE condition: There is $C_{\mathsf{Vlo}} \geq 1$ such that $C_{\mathsf{Vlo}}^{-1} d(m_t, q)^2 \leq \mathbb{E}[d(Y_t, q)^2 - d(Y_t, m_t)^2]$ for all $t \in [-1, 1]$ and $q \in \mathbb{S}^k$.

Corollary 5.11 (`LocGeo` Hypersphere)**.** Assume VARIANCE, KERNEL and SMOOTHNESS. Choose $h := n^{-\frac{1}{2\beta+1}}$. Then

$$\mathbb{E}[d(m_t, \hat{m}_t)^2] \leq C n^{-\frac{2\beta}{2\beta+1}}$$

for all $t \in [0,1]$, $n \geq 2$, where $C := ckC_{\mathsf{Ker}}^3 C_{\mathsf{Vlo}}^2 C_{\mathsf{Smo}}^2$.

5.4.2 General

We prove the main theorem of this section in the metric setting of section 5.1.1 instead of the general setting.

We investigate an estimator that locally fits (generalized) geodesics of the form $x \mapsto g(x, \theta)$, where $\theta \in \Theta$ parameterizes geodesics: Let $h \geq \frac{2}{n}$. Let $K \colon \mathbb{R} \to \mathbb{R}$ be a function, the kernel. For $t \in [0,1]$, define $w_h(t,x) := \frac{1}{h} K(\frac{x-t}{h})$, $w_i := w_h(t, x_i)(\sum_{j=1}^n w_h(t, x_j))^{-1}$. Note that w_i depends on n, t, h, K, which is not indicated in the notation. Let Θ be a set. Let $g \colon \mathbb{R} \times \Theta \to \mathcal{Q}$ be a function, the link function. Define $g_i(\theta) := g(\frac{x_i - t}{h}, \theta)$. Let $\hat{\theta}_{t,h} \in \arg\min_{\theta \in \Theta} \sum_{i=1}^n w_i \, d(y_i, g_i(\theta))^2$ and $\hat{m}_t := g(0, \hat{\theta}_{t,h})$.

We show that this estimator attains the classical nonparametric rate of convergence. To formulate the assumptions for this theorem, we first need to define two (semi-)metrics on $\mathcal{Q}$: For $\theta, \tilde{\theta} \in \Theta$, define $\mathfrak{b}(\theta, \tilde{\theta}) := \sup_{x \in [-1,1]} d(g(x, \theta), g(x, \tilde{\theta}))$. For $y, z \in \mathcal{Q}$, define $\mathfrak{a}(y, z) := \sup_{q,p \in \mathcal{Q}, q \neq p} (\overline{y,q}^2 - \overline{y,p}^2 - \overline{z,q}^2 + \overline{z,p}^2)/\overline{q,p}$, where we use the short notation $\overline{q,p} := d(q,p)$.

Assumptions.

- LIPSCHITZ: There is $C_{\mathsf{Lip}} \in [1, \infty)$ such that the function $[-1,1] \to \mathbb{R}$, $x \mapsto d(g(x,\theta), g(x,\tilde{\theta}))^2$ is Lipschitz continuous with constant C_{Lip} for all $\theta, \tilde{\theta} \in \Theta$.
- KERNEL: There are $C_{\mathsf{Kmi}}, C_{\mathsf{Kma}} \in [1, \infty)$ such that

  $$C_{\mathsf{Kmi}}^{-1} \mathbb{1}_{[-\frac{1}{2}, \frac{1}{2}]}(x) \leq K(x) \leq C_{\mathsf{Kma}} \mathbb{1}_{[-1,1]}(x)$$

  for all $x \in \mathbb{R}$.
- SMOOTHNESS: Let $\beta > 0$. There is $C_{\mathsf{Smo}} \in [1, \infty)$ such that $t \mapsto m_t$ belongs to the generalized Hölder class with parameters β and C_{Smo}, i.e., there is $\theta_{t,h} \in \Theta$

such that $d(m_x, g(\frac{x-t}{h}, \theta_{t,h})) \leq C_{\mathsf{Smo}} |x-t|^\beta$ for all $x \in [t-h, t+h] \cap [0,1]$, $t \in [0,1]$.

- VARIANCE: There is $C_{\mathsf{Vlo}} \in [1, \infty)$ such that $C_{\mathsf{Vlo}}^{-1} d(q, m_t)^2 \leq \mathbb{E}[d(Y_t, q)^2 - d(Y_t, m_t)^2]$ for all $q \in \mathcal{Q}$ and $t \in [0,1]$.
- R–VARIANCE: There is $C_{\mathsf{Vup}} \in [1, \infty)$ such that $\mathbb{E}[d(Y_t, q)^2 - d(Y_t, m_t)^2] \leq C_{\mathsf{Vup}} d(q, m_t)^2$ for all $q \in \mathcal{Q}$ and $t \in [0,1]$.
- MOMENT: Let $\kappa > 2$. There is $C_{\mathsf{Mom}} \in [1, \infty)$ such that $\mathbb{E}[\mathfrak{a}(Y_t, m_t)^\kappa]^{\frac{1}{\kappa}} \leq C_{\mathsf{Mom}}$ for all $t \in [0,1]$.
- ENTROPY: For $\theta_0 \in \Theta$ and $\delta > 0$, let

$$\mathcal{B}_\delta(\theta_0) = \left\{ \theta \in \Theta \colon \int_{-\frac{1}{2}}^{\frac{1}{2}} d(g(x,\theta), g(x,\theta_0))^2 \, \mathrm{d}x \leq \delta \right\} .$$

  There is $C_{\mathsf{Ent}} \in [1, \infty)$ such that $\gamma_2(\mathcal{B}_\delta(\theta_0), \mathfrak{b}) \leq C_{\mathsf{Ent}} \delta^{\frac{1}{2}}$ for all $\delta > 0$ and all $\theta_0 \in \Theta$.

Theorem 5.12 (LocGeo General)**.** Assume VARIANCE, SMOOTHNESS, R–VARIANCE, MOMENT, KERNEL, ENTROPY, and LIPSCHITZ. Then, for all $n \in \mathbb{N}$, $h \geq \frac{2}{n}$, and $t \in [0,1]$, it holds

$$\mathbb{E}\left[\int_{-\frac{1}{2}}^{\frac{1}{2}} d\Big(g(x, \hat{\theta}_{t,h}), g(x, \theta_{t,h})\Big)^2 \, \mathrm{d}x \right] \leq C_1 (nh)^{-1} + C_2 h^{2\beta} ,$$

where

$$C_1 := c_\kappa C_{\mathsf{Ent}}^2 C_{\mathsf{Kmi}}^3 C_{\mathsf{Kma}}^3 C_{\mathsf{Lip}} C_{\mathsf{Mom}}^2 C_{\mathsf{Vlo}}^2 ,$$
$$C_2 := c'_\kappa C_{\mathsf{Kma}}^2 C_{\mathsf{Kmi}}^2 C_{\mathsf{Lip}}^2 C_{\mathsf{Smo}}^2 C_{\mathsf{Vlo}} C_{\mathsf{Vup}} .$$

We essentially obtain the classical bound of a squared bias term $h^{2\beta}$ and a variance term $(nh)^{-1}$, which yield the usual nonparametric rate of convergence for an appropriate choice of h.

Remark 5.13.

- LIPSCHITZ:
  In Euclidean spaces LIPSCHITZ bounds the slope of linear functions for the local fit. This is not a restrictive requirement as for increasing number of data points, the fit is done on an increasingly stretched version of the function, which has a lower and lower absolute slope. Thus, for every finite slope, we eventually meet this requirement.

- KERNEL:
  This is a typical condition on kernels for local kernel regression, see also [Tsy08, Lemma 1.5]. It is fulfilled e.g., by the rectangular kernel $\mathbb{1}_{[-\frac{1}{2},\frac{1}{2}]}(x)$ or the Epanechnikov kernel $\frac{3}{4}(1-x^2)\mathbb{1}_{[-1,1]}(x)$. KERNEL likely could be weakened to allow for a greater variety of kernels.

- SMOOTHNESS:
  SMOOTHNESS can be understood as a Hölder-smoothness condition. It bound the residual of the first order approximation of m at t, i.e., the approximation of $x \mapsto m_x$ by a generalized geodesic $x \mapsto g((x-t)/h, \theta)$ for x close to t.

- R-VARIANCE:
  Together with VARIANCE, reverse variance inequality R-VARIANCE requires $\mathbb{E}[d(Y_t, q)^2 - d(Y_t, m_t)^2]$ to behave like $d(q, m_t)^2$ up to constants. [Gou+19, Theorem 8] introduce a variance equality, from which both inequalities may be deduced. R-VARIANCE always holds in proper Alexandrov spaces of nonnegative curvature with $C_{\mathsf{Vup}} = 1$ [Oht12, Theorem 5.2], where a metric space is called *proper* if every closed ball is compact.

For a discussion of VARIANCE, MOMENT, ENTROPY see Remark 5.3 in section 5.2.

5.4.3 Corollaries

Next, we apply Theorem 5.12 to the metric setting of section 5.1.1. We, first need to make further assumptions to be able to relate the bound on the integral of the parameters $\theta_{t,h}$ and $\hat{\theta}_{t,h}$ to the distance of the true and estimated regression function m_t and $\hat{m}_t$.

Assumptions.

- CONNECTION: There is $C_{\mathsf{Con}} \in [1, \infty)$ such that

$$d\Big(g(0,\theta), g(0,\tilde{\theta})\Big)^2 \leq C_{\mathsf{Con}} \int_{-\frac{1}{2}}^{\frac{1}{2}} d\Big(g(x,\theta), g(x,\tilde{\theta})\Big)^2 \,\mathrm{d}x$$

  for all $\theta, \tilde{\theta} \in \Theta$.

- CONVEXITY: The function $x \mapsto d\Big(g(x,\theta), g(x,\tilde{\theta})\Big)^2$ is convex for all $\theta, \tilde{\theta} \in \Theta$.

The theorem bounds $\mathbb{E}[\int_{-\frac{1}{2}}^{\frac{1}{2}} d(g(x,\hat{\theta}_{t,h}), g(x,\theta_{t,h}))^2]$. Note that $g(0,\theta_{t,h}) = m_t$ due to SMOOTHNESS. To obtain a bound on $\mathbb{E}[d(\hat{m}_t, m_t)^2]$, we may require CONNECTION. CONNECTION with $C_{\mathsf{Con}} = 1$ is implied by CONVEXITY due to Jensen's inequality. CONVEXITY is true in Hadamard spaces (including the Euclidean spaces).

[Oht12, Theorem 5.2] implies that in proper Alexandrov spaces of nonnegative curvature, R–VARIANCE holds with $C_{\mathsf{Vup}} = 1$. Of course, MOMENT is trivial in bounded spaces.

Corollary 5.14 (LocGeo Bounded)**.** Let $(\mathcal{Q}, d)$ be a bounded proper Alexandrov space of nonnegative curvature. Assume Variance, Smoothness, Kernel, Entropy, Lipschitz, Connection. Choose $h := n^{-\frac{1}{2\beta+1}}$. Then

$$\mathbb{E}[d(m_t, \hat{m}_t)^2] \leq C n^{-\frac{2\beta}{2\beta+1}}$$

for all $t \in [0,1]$, $n \geq 2^{\frac{2\beta}{2\beta+1}}$, with $C := c_\kappa C_{\mathsf{Con}} C_{\mathsf{Ent}}^2 C_{\mathsf{Kmi}}^3 C_{\mathsf{Kma}}^3 C_{\mathsf{Lip}}^2 C_{\mathsf{Vlo}}^2 C_{\mathsf{Smo}}^2 \operatorname{diam}(\mathcal{Q}, d)^2$.

In Hadamard spaces Variance and Convexity always hold, see [Stu03, Proposition 4.4, Corollary 2.5].

Corollary 5.15 (LocGeo Hadamard)**.** Let $(\mathcal{Q}, d)$ be a Hadamard space and g such that $x \mapsto g(x, \theta)$ is a geodesic. Assume Smoothness, R–Variance, Moment, Kernel, Entropy, Lipschitz. Choose $h := n^{-\frac{1}{2\beta+1}}$. Then

$$\mathbb{E}[d(m_t, \hat{m}_t)^2] \leq C n^{-\frac{2\beta}{2\beta+1}}$$

for all $t \in [0,1]$, $n \geq 2^{\frac{2\beta}{2\beta+1}}$, where $C := c_\kappa C_{\mathsf{Ent}}^2 C_{\mathsf{Kmi}}^3 C_{\mathsf{Kma}}^3 C_{\mathsf{Lip}}^2 C_{\mathsf{Mom}}^2 C_{\mathsf{Smo}}^2 C_{\mathsf{Vup}}$.

5.5 Local Fréchet Regression

We use the principles of Fréchet regression on local polynomial regression. In particular, this yields local linear Fréchet regression, `LocFre`, introduced in [PM19a].

5.5.1 Hypersphere

We use the hypersphere setting of section 5.1.1. Let $K \colon \mathbb{R} \to \mathbb{R}$ be a function, the kernel, such that $C_{\mathsf{Ker}}^{-1} \mathbb{1}_{[-\frac{1}{2}, \frac{1}{2}]}(x) \leq K(x) \leq C_{\mathsf{Ker}} \mathbb{1}_{[-1,1]}(x)$ for a constant $C_{\mathsf{Ker}} \geq 1$ (Kernel condition). For $h > 0$ define $K_h(x) := \frac{1}{h} K(x/h)$, $a_{h,k}(t) := \sum_{j=1}^n (x_j - t)^k K_h(x_j - t)$ and

$$w_{h,i}(t) := \frac{a_{h,2}(t) - (x_i - t) a_{h,1}(t)}{a_{h,0}(t) a_{h,2}(t) - a_{h,1}(t)^2} K_h(x_i - t)\,,$$

whenever the denominator is not 0; in the other case, set $w_i := 0$. The local linear Fréchet regression estimator is $\hat{m}_t \in \arg\min_{q \in \mathbb{S}^k} \sum_{i=1}^n w_{h,i}(t)\, \overline{y_i, q}^2$, where $\overline{q,p} := d(q,p)$.

We need a smoothness assumption to be able to estimate m: For $a > 0$, define $\lfloor a \rfloor$ as the largest integer strictly smaller than a. The Hölder class $\Sigma(\beta, L)$ for $\beta, L > 0$ is defined as the set of $\lfloor \beta \rfloor$-times continuously differentiable functions $f \colon [0,1] \to \mathbb{R}$ with $\left| f^{(\lfloor \beta \rfloor)}(t) - f^{(\lfloor \beta \rfloor)}(x) \right| \leq L\, |x - t|^\beta$ for all $x, t \in [0,1]$. Let μ be a the measure of the uniform distribution on $\mathbb{S}^k$. Assume that for all $t \in [0,1]$, the random variable Y_t has a density $y \mapsto \rho(y|t)$ with respect to μ. Let $\beta \in (1, 2]$. Assume, there is $C_{\mathsf{SmD}} \geq$

1, such that for μ-almost all $y \in \mathbb{S}^k$, $t \mapsto \rho(y|t) \in \Sigma(\beta, C_{\mathsf{SmD}})$) (SmoothDensity). Furthermore, we assume Variance: There is $C_{\mathsf{Vlo}} \in [1,\infty)$ such that $C_{\mathsf{Vlo}}^{-1}\,\overline{q,m_t}^2 \le \mathbb{E}[\overline{Y_t,q}^2 - \overline{Y_t,m_t}^2]$ for all $q \in \mathbb{S}^k$ and $t \in [0,1]$.

Corollary 5.16 (`LocFre` Hypersphere)**.** Assume Variance, SmoothDensity, and Kernel. Let $n \ge n_0$ for a universal constant n_0 and set $h := n^{-\frac{1}{2\beta+1}}$. Then

$$\mathbb{E}[\overline{m_t,\hat{m}_t}^2] \le C n^{-\frac{2\beta}{2\beta+1}}$$

for all $t \in [0,1]$, where $C := ck\,(C_{\mathsf{Vlo}} C_{\mathsf{Ker}} C_{\mathsf{SmD}})^2$.

We obtain the usual nonparametric rate of convergence.

5.5.2 General

The general theorem of this section uses the general setting of 5.1.1, but with a specific loss function: Let d be a metric on $\mathcal{Q}$. Let $\alpha > 1$. We use d^α as loss function. Define $\Diamond(y,z,q,p) := \mathfrak{c}(y,q) - \mathfrak{c}(y,p) - \mathfrak{c}(z,q) + \mathfrak{c}(z,p)$ and $\mathfrak{a}(y,z) := \sup_{q,p\in\mathcal{Q}, q\ne p} \Diamond(y,z,q,p)\overline{q,p}^{-1}$. Let $K\colon \mathbb{R} \to \mathbb{R}$ be a function. For $\ell \in \mathbb{N}_0$, $h = h_n > 0$, and $x, t \in [0,1]$ define

$$\begin{aligned}
\Psi(x) &:= \left(\frac{x^k}{k!}\right)_{k=0,\dots,\ell}, \\
B_{n,t} &:= \frac{1}{nh}\sum_{i=1}^n \Psi\left(\frac{x_i - t}{h}\right)\Psi\left(\frac{x_i - t}{h}\right)^\top K\left(\frac{x_i - t}{h}\right), \\
w_i &:= \Psi(0)^\top B_{n,t}^{-1} \Psi\left(\frac{x_i - t}{h}\right) K\left(\frac{x_i - t}{h}\right),
\end{aligned}$$

whenever $B_{n,t}$ is invertible. Note that w_i depends on n, t, h, K. A local polynomial Fréchet estimator of order ℓ is any element

$$\hat{m}_t \in \operatorname*{arg\,min}_{q\in\mathcal{Q}} \sum_{i=1}^n w_i \mathfrak{c}(y_i, q)\,.$$

Assumptions.

- Moment: There are $\kappa \ge 2$ and $C_{\mathsf{Mom}} \in [1,\infty)$ such that $\mathbb{E}[\mathfrak{a}(Y_t, m_t)^\kappa]^{\frac{1}{\kappa}} \le C_{\mathsf{Mom}}$ for all $x \in [0,1]$.
- Smoothness: Let $\beta > 0$. For all $q, p \in \mathcal{Q}$ there is $L(q,p) > 0$ such that $t \mapsto \mathbb{E}[\mathfrak{c}(Y_t, q) - \mathfrak{c}(Y_t, p)] \in \Sigma(\beta, \overline{q,p}\,L(q,p))$. There is $C_{\mathsf{Smo}} \in [1,\infty)$ such that $\mathbb{E}[L(m_t, \hat{m}_t)^\kappa]^{\frac{1}{\kappa}} \le C_{\mathsf{Smo}}$ for all $t \in [0,1]$.
- Kernel: There is $C_{\mathsf{Kma}} \in [1,\infty)$ such that $K(x) \le C_{\mathsf{Kma}} \mathbb{1}_{[-1,1]}(x)$ for all $x \in \mathbb{R}$. There are $n_0 \in \mathbb{N}, \lambda_0 \in (0,\infty)$ such that $\lambda_{\mathsf{min}}(B_{n,t}) > \lambda_0$ for all

$x \in [-1,1]$, $n \geq n_0$ and the given choice of $h = h_n$, where $\lambda_{\mathsf{min}}(B_{n,t})$ is the smallest eigenvalue of $B_{n,t}$. The constants n_0 and λ_0 give rise to a constant $C_{\mathsf{Ker}} \in [1,\infty)$, see Lemma 5.40.

- VARIANCE: There is $C_{\mathsf{Vlo}} \in [1,\infty)$ such that $C_{\mathsf{Vlo}}^{-1}\,\overline{q,m_t}^{\alpha} \leq \mathbb{E}[\mathfrak{c}(Y_t,q) - \mathfrak{c}(Y_t,m_t)]$ for all $q \in \mathcal{Q}$ and $t \in [0,1]$.
- ENTROPY: There is $C_{\mathsf{Ent}} \in [1,\infty)$ such that $\gamma_2(\mathcal{B},d) \leq C_{\mathsf{Ent}}\,\mathsf{diam}(\mathcal{B})$ for all $\mathcal{B} \subseteq \mathcal{Q}$.

Theorem 5.17 (`LocFre` General)**.** Assume SMOOTHNESS, KERNEL, VARIANCE, ENTROPY, MOMENT and $\kappa > \frac{\alpha}{\alpha-1}$. Let $\ell := \lfloor \beta \rfloor$. Then, for $t \in [0,1]$ and $n \geq n_0$, the local polynomial Fréchet estimator $\hat{m}_t$ of order ℓ fulfills,

$$\mathbb{E}\Big[\overline{m_t,\hat{m}_t}^{\alpha}\Big] \leq \Big(C_1 h^{\beta} + C_2 (nh)^{-\frac{1}{2}}\Big)^{\frac{\alpha}{\alpha-1}} ,$$

where $C_1 := c_{\kappa,\alpha} C_{\mathsf{Vlo}} C_{\mathsf{Ker}} C_{\mathsf{Smo}}$ and $C_2 := c_{\kappa,\alpha} C_{\mathsf{Vlo}} C_{\mathsf{Mom}} C_{\mathsf{Ent}} C_{\mathsf{Ker}}$.

For $\alpha = 2$, we obtain the classical error bound for local polynomial estimators with a bias term h^{β} and a variance term $(nh)^{-\frac{1}{2}}$. The theorem does not necessarily give bounds for different powers α of the distance between estimator and true value, but possibly only for one specific α, which is determined by VARIANCE.

SMOOTHNESS and KERNEL are classical conditions for local polynomial estimators [Tsy08, Proposition 1.13]. VARIANCE, ENTROPY, MOMENT are conditions needed to ensure the rate of convergence for a generalized Fréchet mean, see [Sch19b, Theorem 1]. For a discussion see Remark 5.3 in section 5.2.

Remark 5.18.

- SMOOTHNESS:
  In this theorem, we have to insert a loose bound $\mathbb{E}[L(m_t,\hat{m}_t)^{\kappa}]^{\frac{1}{\kappa}} \leq C_{\mathsf{Smo}} < \infty$ independent of n and h to obtain a bound on $\mathbb{E}\Big[\overline{m_t,\hat{m}_t}^{\alpha}\Big]$ that vanishes for $n \to \infty$ and $h = h_n$ chosen appropriately. In the corollaries below, we see that this is not too difficult to fulfill.

  In Euclidean spaces with $\mathfrak{c} = d^2$, where d is the Euclidean metric, we have $\mathbb{E}[\mathfrak{c}(Y_t,q) - \mathfrak{c}(Y_t,p)] = -2\langle m(t), q-p\rangle + \|q\|^2 - \|p\|^2$ and SMOOTHNESS is equivalent to $m \in \Sigma(L,\beta)$ with $L(q,p) = 2L$.
- KERNEL:
  KERNEL is fulfilled for $C_{\mathsf{Kmi}}^{-1}\mathbb{1}_{[-\frac{1}{2},\frac{1}{2}]}(x) \leq K(x) \leq C_{\mathsf{Kma}}\mathbb{1}_{[-1,1]}(x)$, appropriately chosen h_n, and n large enough, see [Tsy08, Lemma 1.5].

5.5.3 Corollaries

Next, we apply Theorem 5.17 to the metric setting of section 5.1.1. SMOOTHNESS can be replaced by SMOOTHDENSITY, see Lemma 5.41 in the appendix, and BIASMOMENT. To fulfill BIASMOMENT we can assume BOMBOUND.

Assumptions.

- SMOOTHDENSITY: Let μ be a probability measure on $\mathcal{Q}$ with $\int \overline{y,o}^2 \mu(\mathrm{d}y) < \infty$ for an arbitrary $o \in \mathcal{Q}$. Let $\beta > 0$ with $\ell = \lfloor \beta \rfloor$. For μ-almost all $y \in \mathcal{Q}$, there is $L(y) \geq 0$ such that $t \mapsto \rho(y|t) \in \Sigma(\beta, L(y))$. Furthermore there is a constant $C_{\mathsf{SmD}} > 0$, $\int L(y)^2 \mathrm{d}\mu(y) \leq C_{\mathsf{SmD}}^2$.
- BIASMOMENT: Define $H(q,p) = \left(\int (\overline{y,q} + \overline{y,p})^2 \, \mu(\mathrm{d}y) \right)^{\frac{1}{2}}$. There is $C_{\mathsf{Bom}} \in [1, \infty)$ such that $\mathbb{E}[H(\hat{m}_t, m_t)^\kappa]^{\frac{1}{\kappa}} \leq C_{\mathsf{Bom}}$ for all $t \in [0,1]$.
- BOMBOUND: There are $C_{\mathsf{Int}}, C_{\mathsf{Len}} \in [1, \infty)$ such that $\int \overline{y,m_t}^2 \mu(\mathrm{d}y) \leq C_{\mathsf{Int}}^2$ and $\mathfrak{a}(m_t, m_s) \leq C_{\mathsf{Len}}$ for all $s, t \in [0,1]$.

MOMENT is trivial in bounded spaces.

Corollary 5.19 (`LocFre` Bounded)**.** Let $(\mathcal{Q}, d)$ be a bounded metric space. Let $\beta > 0$ with $\ell := \lfloor \beta \rfloor$. Let $\hat{m}_t$ be the local polynomial estimator of order ℓ. Assume VARIANCE, ENTROPY, SMOOTHDENSITY, KERNEL. Set $h := n^{-\frac{1}{2\beta+1}}$. Then

$$\mathbb{E}[\overline{m_t, \hat{m}_t}^2] \leq C n^{-\frac{2\beta}{2\beta+1}}$$

for all $t \in [0,1]$, where $C := c \left(\mathsf{diam}(\mathcal{Q}, d) C_{\mathsf{Vlo}} C_{\mathsf{Ker}} C_{\mathsf{SmD}} C_{\mathsf{Ent}} \right)^2$.

VARIANCE is always true in Hadamard spaces.

Corollary 5.20 (`LocFre` Hadamard)**.** Let $(\mathcal{Q}, d)$ be a Hadamard space. Let $\beta > 0$ with $\ell := \lfloor \beta \rfloor$. Let $\hat{m}_t$ be the local polynomial estimator of order ℓ. Let $\kappa > 2$. Assume MOMENT, ENTROPY, SMOOTHDENSITY, BOMBOUND, KERNEL. Set $h := n^{-\frac{1}{2\beta+1}}$. Then

$$\mathbb{E}[\overline{m_t, \hat{m}_t}^2] \leq C n^{-\frac{2\beta}{2\beta+1}}$$

for all $t \in [0,1]$, where $C := c_\kappa \left(C_{\mathsf{Ker}}^2 C_{\mathsf{Len}} C_{\mathsf{Mom}} C_{\mathsf{Int}} C_{\mathsf{SmD}} C_{\mathsf{Ent}} \right)^2$.

Remark 5.21.

- BOMBOUND:
  We require the length of $[0,1] \to \mathcal{Q}$, $t \mapsto m_t$ to be finite, measured with respect the measure μ and with respect to the pseudo metric $\mathfrak{a}$. This is a mild condition

and should be fulfilled for smooth functions m. See Proposition 5.43 (appendix 5.A.4.3) for a result on Lipschitz continuity of the regression function.

- BiasMoment:
  This is a technical condition that we also use as an intermediate step to prove corollaries in metric spaces for `LocFre`. It is fulfilled in bounded metric spaces and can also be replaced by BomBound.

- SmoothDensity:
  If the noise distribution has a μ-density and this density is smooth enough, SmoothDensity can be interpreted as a smoothness condition on $t \mapsto m_t$: In a Euclidean space $\mathcal{Q} = \mathbb{R}^k$ with a location model $\rho(y|t) = \rho((y - m(t))^2)$, we have $\partial_t \rho(y|t) = -2(y-t)\dot{m}(t)\rho'(y|t)$. Informally, the density should be as least as smooth as the regression function, to view this condition as a typical smoothness assumption on the regression function. It is likely an artifact of the proof that we require the error density to be smooth.

5.6 Trigonometric Geodesic Regression

We apply the principles of geodesic regression to transfer the Euclidean trigonometric series estimator to a new method, `TriGeo`, for nonstandard spaces.

Let $(\psi_\ell)_{\ell \in \mathbb{N}}$ be the trigonometric basis of $\mathbb{L}_2[0,1]$, i.e., for $x \in [0,1]$, $k \in \mathbb{N}$,

$$\psi_1(x) := 1\,, \qquad \psi_{2k}(x) := \sqrt{2}\cos(2\pi k x)\,, \qquad \psi_{2k+1}(x) := \sqrt{2}\sin(2\pi k x)\,.$$

The trigonometric basis is orthonormal, i.e.,

$$\int_0^1 \psi_k(x)\psi_\ell(x)\mathrm{d}x = \delta_{l\ell}$$

for all $\ell, k \in \mathbb{N}$, where $\delta_{k\ell}$ is the Kronecker delta.

In the metric space setting of section 5.1.1 with the assumption of the existence of an exponential map $\mathsf{Exp}(p, \cdot)$, the resulting method is `TriGeo`:

$$(\hat{p}, \hat{v}_1, \ldots, \hat{v}_N) \in \operatorname*{arg\,min}_{p \in \mathcal{Q}, v_\ell \in \mathsf{T}_p\mathcal{Q}} d\left(\mathsf{Exp}\left(p, \sum_{\ell=1}^N \psi_\ell(x_i) v_\ell\right), y_i\right)^2\,,$$
$$\hat{m}(t) := \mathsf{Exp}\left(\hat{p}, \sum_{\ell=1}^N \psi_\ell(t)\hat{v}_\ell\right)\,.$$

For trigonometric series estimators, one usually bounds the mean integrated squared error (MISE), as this makes it possible to utilize the orthogonality property of $(\psi_\ell)_{\ell \in \mathbb{N}}$ in $\mathbb{L}_2[0,1]$. To be able to use the same properties in the metric space setting, one could take the integrated mean squared Euclidean error in the tangent space $\mathsf{T}_o\mathcal{Q}$, where, e.g., $o = m(0)$. Then the problem reduces to the standard Euclidean trigonometric estimator

which is discussed, e.g., in [Tsy08, chapter 1.7]. If we assume that an inverse $\mathsf{Log}(o,\cdot)$ of $\mathsf{Exp}(o,\cdot)$ exists, a smoothness condition should be applied to $t \mapsto \mathsf{Log}(o, m(t))$.

The condition of centered / zero-mean noise of the Euclidean model for trigonometric estimation translates to $\mathbb{E}[\mathsf{Log}_o(Y_t)] = \mathsf{Log}_o(m_t)$. Unfortunately, this seems to be far from the condition of centered noise in our metric setting, as it introduces distortions which highly depend on $o = m(0)$. Compare this to our usual assumption, $m_t = \arg\min_{q\in\mathcal{Q}} \mathbb{E}[d(Y_t,q)^2]$, which implies (under mild assumptions) $\mathbb{E}[\mathsf{Log}_{m_t}(Y_t)] = \mathsf{Log}_{m_t}(m_t) = 0$, cf [Kar77, Theorem 1.2].

We were not able to show a theorem similar to Theorem 5.23 or Theorem 5.12 using our usual settings. Of course, this does not mean that the estimator above will necessarily perform badly.

The estimator was implemented for simulations (section 5.8). This revealed another drawback: High-dimensional nonconvex optimization is required so that `TriGeo` is – by far – the slowest of all tested methods. The MISE values seem to be worse than for the other estimators on average. It is not clear, whether this is due to theoretical disadvantages or a worse outcome of the general purpose optimizer used for finding $(\hat{p}, \hat{v}_1, \dots, \hat{v}_N)$.

5.7 Trigonometric Fréchet Regression

Using the Fréchet approach, we create a new trigonometric estimator, `TriFre`.

Confer section 5.6 for the definition of the trigonometric basis of $\mathbb{L}_2[0,1]$. In every setting, we will require a smoothness condition. The appropriate smoothness class connected to the trigonometric basis $(\psi_k)_{k\in\mathbb{N}}$ is the periodic Sobolev class $W^{\mathsf{per}}(\beta, L)$, see [Tsy08, Definition 1.11]. A function $f(x) = \sum_{k=1}^{\infty} \theta_k \psi_k(x)$ belongs to $W^{\mathsf{per}}(\beta, L)$ if and only if the sequence $\theta = (\theta_k)_{k\in\mathbb{N}}$, $\theta_k = \int_0^1 f(x)\psi_k(x)\mathrm{d}x$, of the Fourier coefficients of f belongs to the ellipsoid $\Theta(\beta, L)$, which is defined as

$$\Theta(\beta, L) := \left\{ \theta \in \ell^2 \colon \sum_{k=1}^{\infty} \theta_k^2 w_k^{-2} \le L^2 \right\},$$

where $w_{2k+1} := w_{2k} := (2k)^{-\beta}$, see [Tsy08, Proposition 1.14].

5.7.1 Hypersphere

We use the hypersphere setting of section 5.1.1. For $N \in \mathbb{N}$, define the vector $\Psi_N := (\psi_k)_{k=1,\dots,N} \colon [0,1] \to \mathbb{R}^N$. For $t \in [0,1]$, $q \in \mathbb{S}^k$ set $\hat{F}_t(q) := \Psi_N(t)^\top \frac{1}{n} \sum_{i=1}^n \Psi_N(x_i)\overline{y_i,q}^2$. The trigonometric Fréchet estimator on the hypersphere is $\hat{m}_t \in \arg\min_{q\in\mathbb{S}^k} \hat{F}_t(q)$.

To be able to estimate m, we require a smoothness condition: Let μ be a the measure of the uniform distribution on $\mathbb{S}^k$. Assume that for all $t \in [0,1]$, the random variable Y_t has a density $y \mapsto \rho(y|t)$ with respect to μ. Let $\beta \ge 1$. Assume, there is $C_{\mathsf{SmD}} \ge 1$, such that for μ-almost all $y \in \mathbb{S}^k$, $t \mapsto \rho(y|t) \in W^{\mathsf{per}}(\beta, C_{\mathsf{SmD}})$ (SmoothDensity). Furthermore, we again assume Variance: There is $C_{\mathsf{Vlo}} \in [1,\infty)$ such that $C_{\mathsf{Vlo}}^{-1} d(q, m_t)^2 \le \mathbb{E}[d(Y_t,q)^2 - d(Y_t,m_t)^2]$ for all $q \in \mathbb{S}^k$ and $t \in [0,1]$.

Corollary 5.22 (`TriFre` Hypersphere)**.** Assume VARIANCE and SMOOTHDENSITY. Set $N := n^{\frac{1}{2\beta+1}}$. Then

$$\mathbb{E}\left[\int_0^1 \overline{m_t,\hat{m}_t}^2 \mathrm{d}t\right] \leq C n^{-\frac{2\beta}{2\beta+1}}$$

for all $n \in \mathbb{N}$, where $C := c_\beta C_{\mathsf{Vlo}}^2 C_{\mathsf{SmD}}^2$.

5.7.2 General

We will only show a theorem in the metric setting of 5.1.1. For $N \in \mathbb{N}$ with $\Psi_N = (\psi_k)_{k=1,\dots,N}$, define $\hat{F}_t(q) := \Psi_N(t)^\top \frac{1}{n} \sum_{i=1}^n \Psi_N(x_i) \overline{y_i,q}^2$. Let $\hat{m}_t \in \arg\min_{q\in\mathcal{Q}} \hat{F}_t(q)$. Essentially, we estimate $t \mapsto F_t(q)$ for every fixed $q \in \mathcal{Q}$ by a trigonometric series estimator described in [Tsy08, section 1.7]. Instead of the unknown function $F_t(q)$, we then minimize $\hat{F}_t(q)$ with respect to q. Out goal is to bound the mean integrated squared error $\mathbb{E}[\int_0^1 \overline{m_t,\hat{m}_t}^2 \mathrm{d}t]$.

For $y, z \in \mathcal{Q}$ define

$$\mathfrak{a}(y,z) := \sup_{q,p\in\mathcal{Q}, q\neq p} \frac{\overline{y,q}^2 - \overline{y,p}^2 - \overline{z,q}^2 + \overline{z,p}^2}{\overline{q,p}}\,.$$

Assumptions.

- SMOOTHDENSITY: Let μ be a probability measure on $\mathcal{Q}$ with $\int \overline{y,o}^2 \mu(\mathrm{d}y) < \infty$ for an arbitrary $o \in \mathcal{Q}$. For all $t \in [0,1]$, the random variable Y_t has a density $y \mapsto \rho(y|t)$ with respect to μ. Let $\beta \geq 1$. For μ-almost all $y \in \mathcal{Y}$, there is $L(y) \geq 0$ such that $t \mapsto \rho(y|t) \in W^{\mathsf{per}}(\beta, L(y))$. Furthermore, there is $C_{\mathsf{SmD}} \in [1,\infty)$ such that $\int L(y)^2 \mathrm{d}\mu(y) \leq C_{\mathsf{SmD}}^2$.
- VARIANCE: There is $C_{\mathsf{Vlo}} \in [1,\infty)$ such that $C_{\mathsf{Vlo}}^{-1} \overline{q,m_t}^2 \leq \mathbb{E}[\overline{Y_t,q}^2 - \overline{Y_t,m_t}^2]$ for all $q \in \mathcal{Q}$ and $t \in [0,1]$.
- MOMENT: Let $\kappa > 2$. There is $C_{\mathsf{Mom}} \in [1,\infty)$ such that $\mathbb{E}[\mathfrak{a}(Y_t, m_t)^\kappa]^{\frac{1}{\kappa}} \leq C_{\mathsf{Mom}}$ for all $t \in [0,1]$.
- BIASMOMENT: Define $H(q,p) = \left(\int (\overline{y,q} + \overline{y,p})^2 \mu(\mathrm{d}y)\right)^{\frac{1}{2}}$. There is $C_{\mathsf{Bom}} \in [1,\infty)$ such that $\mathbb{E}[H(\hat{m}_t, m_t)^\kappa]^{\frac{1}{\kappa}} \leq C_{\mathsf{Bom}}$ for all $t \in [0,1]$.
- ENTROPY: There is $C_{\mathsf{Ent}} \in [1,\infty)$ such that $\gamma_2(\mathcal{B}, d) \leq C_{\mathsf{Ent}} \operatorname{diam}(\mathcal{B})$ for all $\mathcal{B} \subseteq \mathcal{Q}$.

Theorem 5.23 (`TriFre` General)**.** Assume Variance, Moment, BiasMoment, Entropy, SmoothDensity. Then

$$\mathbb{E}\left[\int_0^1 \overline{m_t,\hat{m}_t}^2 \mathrm{d}t\right] \leq C_1 \left(N^{-2\beta} + Nn^{1-2\beta}\right) + C_2 \frac{N}{n},$$

where $C_1 := c_{\kappa,\beta} C_{\mathsf{Vlo}}^2 C_{\mathsf{SmD}}^2 C_{\mathsf{Bom}}^2$ and $C_2 := c_{\kappa,\beta} C_{\mathsf{Vlo}}^2 C_{\mathsf{Mom}}^2 C_{\mathsf{Ent}}^2$.

We obtain a bound with the same rates as in the Euclidean setting, which lead to the classical nonparametric rate of convergence, see corollaries below.

All condition have previously been discussed, see Remark 5.3 and Remark 5.21.

5.7.3 Corollaries

In bounded spaces Moment and BiasMoment are trivial.

Corollary 5.24 (`TriFre` Bounded)**.** Let $(\mathcal{Q}, d)$ be a metric space with $\mathsf{diam}\,\mathcal{Q} < \infty$. Assume Variance, Entropy, SmoothDensity. Set $N := n^{\frac{1}{2\beta+1}}$. Then

$$\mathbb{E}\left[\int_0^1 \overline{m_t,\hat{m}_t}^2 \mathrm{d}t\right] \leq C n^{-\frac{2\beta}{2\beta+1}},$$

where $C := c_\beta C_{\mathsf{Vlo}}^2 C_{\mathsf{SmD}}^2 C_{\mathsf{Ent}}^2 \,\mathsf{diam}(\mathcal{Q})^2$.

In Hadamard spaces, $\mathfrak{a}(y,z) \leq 2d(y,z)$ because of the quadruple inequality [Stu03, Theorem 4.9]. Furthermore, Variance is fulfilled as noted before. Lastly, we replace BiasMoment by BomBound, which introduces an additional $\log(n)$-factor.

Assumptions.

- BomBound: There are $C_{\mathsf{Int}}, C_{\mathsf{Len}} \in [1,\infty)$ such that $\int \overline{y,m_t}^2 \mu(\mathrm{d}y) \leq C_{\mathsf{Int}}^2$ and $\mathfrak{a}(m_t, m_s) \leq C_{\mathsf{Len}}$ for all $s,t \in [0,1]$.

Corollary 5.25 (`TriFre` Hadamard)**.** Let $(\mathcal{Q}, d)$ be a Hadamard metric space. Assume Moment, BomBound, Entropy, SmoothDensity. Set $N = n^{\frac{1}{2\beta+1}}$. Then

$$\mathbb{E}\left[\int_0^1 \overline{m_t,\hat{m}_t}^2 \mathrm{d}t\right] \leq C n^{-\frac{2\beta}{2\beta+1}} \log(n)^2,$$

where $C := c_{\kappa,\beta} C_{\mathsf{SmD}}^2 C_{\mathsf{Mom}}^2 C_{\mathsf{Ent}}^2 C_{\mathsf{Len}}^2 C_{\mathsf{Int}}^2$.

5.8 Simulation

There is a total of 7 methods discussed in this chapter: `LinGeo`, `LinFre`, `LinCos`, `LocGeo`, `LocFre`, `TriGeo`, `TriFre`. To illustrate and compare these methods on the sphere, the R-package `spheregr` was developed. All code used for this paper, including all scripts which create the plots and run and evaluate the experiments shown in this section, are freely available at `https://github.com/ChristofSch/spheregr`.

Each method requires numerical optimization. We use R's general purpose optimizers `stats::optim(method = "L-BFGS-B")` and `stats::optimize()`, both without explicit implementation of derivatives, but with several starting points. The implementations could potentially be improved by using the algorithm presented in [EHW19]. For alternative implementation of geodesic regression, see [SO20].

The parametric methods are much faster than the nonparametric ones and Fréchet methods are faster than geodesic methods, as the optimization problem for geodesics is of higher dimension. We use *leave-one-out cross-validation* to estimate the hyperparameters (h for `LocGeo` and `LocFre`, N for `TriFre`). For `TriGeo` it did not seem feasible to do many repetitions of the experiments with cross-validation in each run. Instead we set $N = 3$ for this method, which seems to be a good choice in many runs. For `LocGeo` and `LocFre`, we use the Epanechnikov-kernel.

5.8.1 Model and Contracted Uniform Distribution

Let $\mathbb{S}^2 = \{x \in \mathbb{R}^3 \colon |x| = 1\}$ be the sphere with radius 1 and intrinsic metric $d(q,p) = \arccos(q^\top p)$. For $t \in [0,1]$, let Y_t be a $\mathbb{S}^2$-valued random variable. Let the regression function $m \colon [0,1] \to \mathbb{S}^2$ be a minimizer $m_t \in \arg\min_{q \in \mathbb{S}^2} \mathbb{E}[\overline{Y_t,q}^2]$. Let $x_i := \frac{i-1}{n-1}$ and let $(y_i)_{i=1,\dots,n}$ be independent random variables with values in $\mathbb{S}^2$ such that y_i has the same distribution as Y_{x_i}.

As distribution of Y_t, we choose the contracted uniform distribution $\mathsf{CntrUnif}(m_t, a)$ with $a \in (0,1)$, which we define next. The contracted uniform distribution is obtained from the uniform distribution on the sphere by moving all points towards a center point along the connecting geodesic by a given fraction of the total distance.

Definition 5.26. Let $a \in [0,1]$. Let (Θ, Φ) be random angles with values in $[0,\pi] \times [0, 2\pi)$ that form a uniform distribution on the sphere, i.e., they are independent, Θ has Lebesgue density $\frac{1}{2}\sin(x)\mathbb{1}_{[0,\pi]}(x)$, and Φ is uniformly distributed on $[0, 2\pi)$. Let

$$Z_a := \begin{pmatrix} \sin(a\Theta)\cos(\Phi) \\ \sin(a\Theta)\sin(\Phi) \\ \cos(a\Theta) \end{pmatrix}.$$

Let $m \in \mathbb{S}^2$. Let $R_m \in O(3) \subseteq \mathbb{R}^{3\times 3}$ be any orthogonal matrix that fulfills $m = R_m e_3$, where $e_3^\top := (0\ 0\ 1)$. Then the **contracted uniform distribution** $\mathsf{CntrUnif}(m, a)$ at m with contraction parameter a is defined as the distribution of $R_m Z_a$.

The matrix R_m in the definition above is not unique, but the symmetry of the distribution of Z_a ensures that the contracted uniform distribution is well-defined.

Two important properties are implied by the following proposition: For $a \in [0,1)$, $m \in \mathbb{S}^2$ is the unique Fréchet mean of $\mathsf{CntrUnif}(m,a)$. Furthermore, VARIANCE is fulfilled with $C_{\mathsf{Vlo}} = (1-a)^{-1}$.

Proposition 5.27 ([Oht12, section 5])**.** Let $(\mathcal{Q}, d)$ be a proper Alexandrov space of nonnegative curvature. Let Y_1 be a random variable with values $\mathcal{Q}$ such that $\mathbb{E}[d(Y,q)^2] < \infty$ for all $q \in \mathcal{Q}$. Let $m \in \arg\min_{q\in\mathcal{Q}} \mathbb{E}[\overline{Y_1,q}^2]$ be any Fréchet mean of Y_1. For $a \in [0,1)$, let $Y_a := \gamma_{m\to Y}(a)$, where, for $y \in \mathcal{Q}$, $\gamma_{m\to y}$ is a geodesic with $\gamma_{m\to y}(0) = m$, $\gamma_{m\to y}(1) = y$. Then

$$(1-a)\overline{q,m}^2 \le \mathbb{E}[\overline{Y_a,q}^2 - \overline{Y_a,m}^2]$$

for all $a \in [0,1]$.

Lastly, we calculate the variance of the contracted uniform distribution. Let $m \in \mathbb{S}^2$, $a \in [0,1]$, and $Y \sim \mathsf{CntrUnif}(m,a)$. Let Z_a and Θ as in Definition 5.26. Then $\mathbb{E}[d(Y,m)^2] = \mathbb{E}[d(Z_a,e_3)^2]$ because of symmetry. The distance does only depend on Θ and is equal to $a\Theta$. Thus, $\mathbb{E}[d(Y,m)^2] = \mathbb{E}[(a\Theta)^2] = \frac{1}{2}a^2\int_0^\pi x^2\sin(x)\mathrm{d}x = \frac{1}{2}(\pi^2-4)a^2$.

5.8.2 Parametric Regression

We draw a random geodesic m with fixed speed. Then we create independent samples $y_i \sim \mathsf{CntrUnif}(m_{x_i}, a)$ to obtain our data $(x_i,y_i)_{i=1,\dots,n}$. Then we calculate the three different parametric regression estimators `LinGeo`, `LinFre`, and `LinCos`.

We will describe points $q \in \mathbb{S}^2 = \{x \in \mathbb{R}^3 \,|\, |x| = 1\}$ via two angles $(\vartheta_q, \varphi_q) \in [0,\pi]\times[0,2\pi)$ such that $q = (\sin(\vartheta_q)\cos(\varphi_q), \sin(\vartheta_q)\sin(\varphi_q), \cos(\vartheta_q))$. We first show some illustrating plots Figure 5.1 and Figure 5.2. We want to depict functions of the form $[0,1] \to [0,\pi]\times[0,2\pi)$, $t \mapsto (\vartheta_{m_t}, \varphi_{m_t})$. The graph of such a function is 3-dimensional and hard to understand on 2D-paper. Creating two plots, one for $[0,1]\to[0,\pi]$, $t\mapsto\vartheta_{m_t}$ and another for $[0,1]\to[0,2\pi)$, $t\mapsto\varphi_{m_t}$, is also difficult to interpret, as one has to always take both graphs into account at the same time. Instead we show the image of the functions $\{(\vartheta_{m_t},\varphi_{m_t})\colon t\in[0,1]\} \subseteq [0,\pi]\times[0,2\pi)$ and encode the dependence on t via color.

The rectangle of the two angles $(\vartheta,\varphi) \in [0,\pi]\times[0,2\pi)$ parameterizing the sphere is the *Mercator projection*. This projection (as any projection of the sphere to the Euclidean plane) distorts the surface of the sphere. This is made visible by the thin gray lines in the plots, which are geodesics and replace the usual grid lines. The plots show the image of $t\mapsto m_t$ (line with black border) and the different estimators $t\mapsto\hat{m}_t$ (lines with colored border). The covariate t is represented by the rainbow color inside each line. To visually compare the deviations of $\hat{m}_t$ from m_t, one has to compare the positions on the lines with the same inner color. But note that distances are distorted: Distances close to the equator ($\vartheta = \frac{1}{2}\pi$) are larger than they appear and smaller at the poles ($\vartheta \in \{0,\pi\}$). The

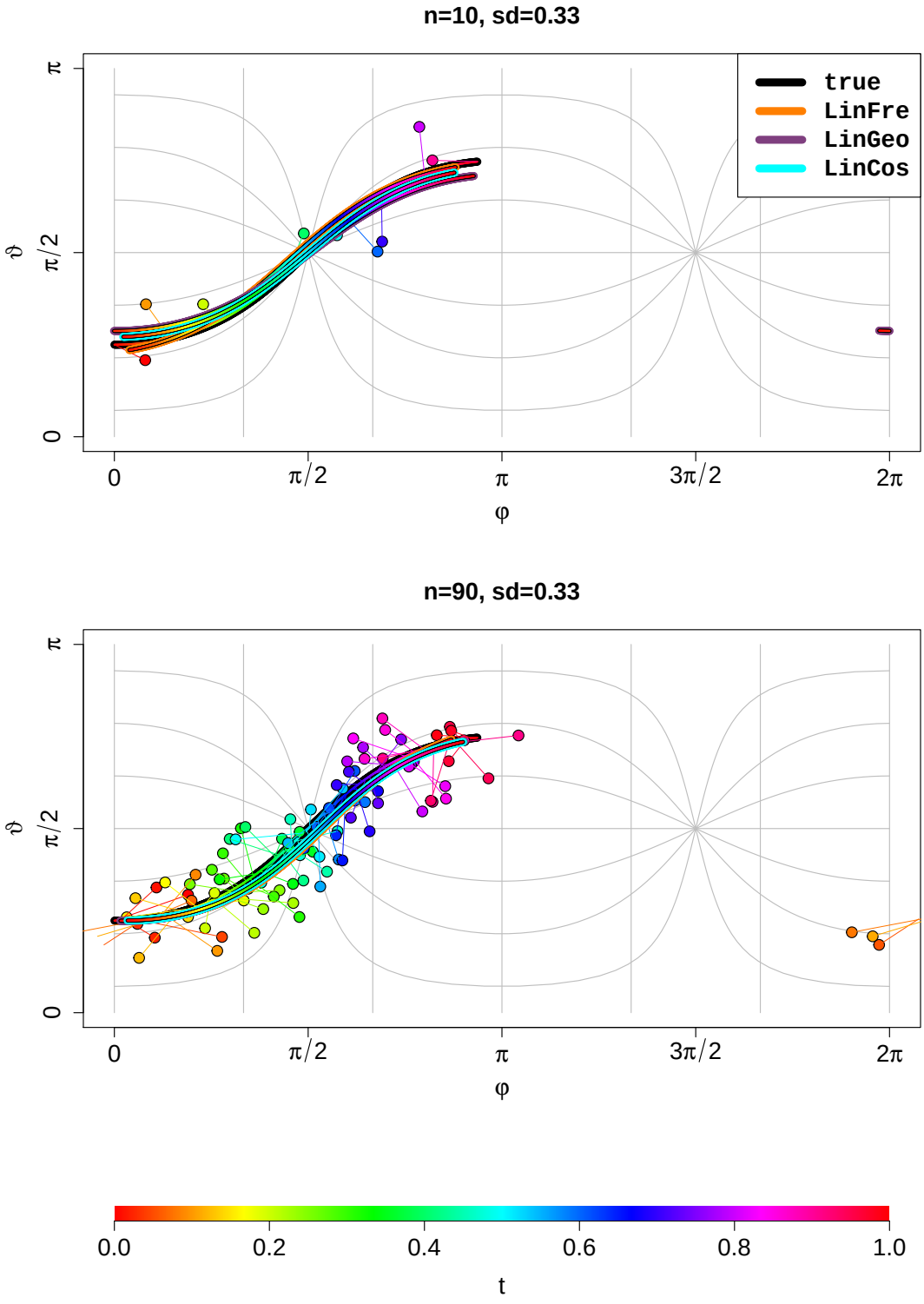

Figure 5.1: For a true geodesic of length 3, we sample $n \in \{10, 90\}$ observations with contracted uniform noise of standard deviation $\texttt{sd} \in \{\frac{1}{3}, 1\}$. Then we apply `LinGeo`, `LinFre`, and `LinCos`. (Part 1.)

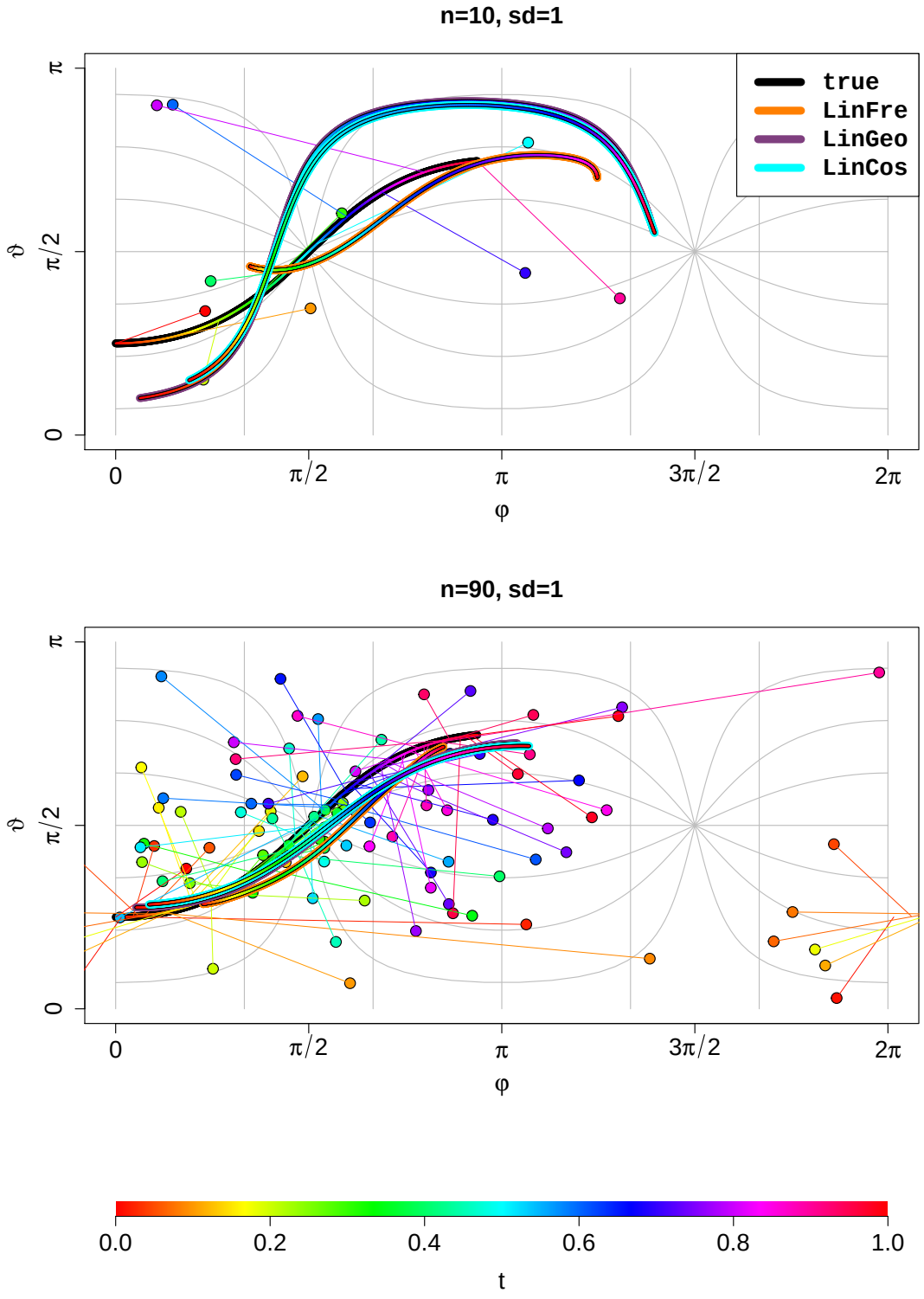

Figure 5.1: (Part 2.)

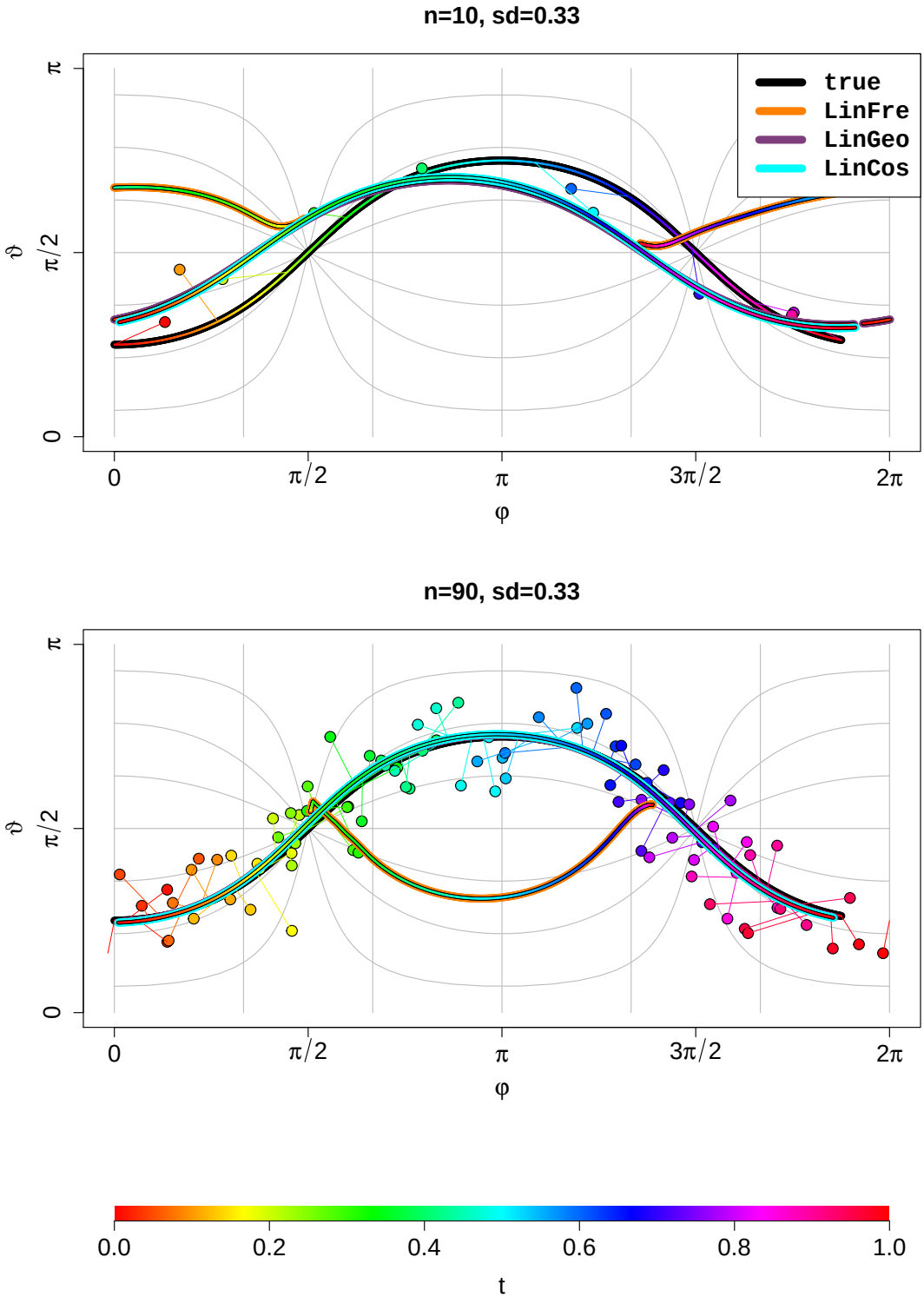

Figure 5.2: For a true geodesic of length 6, we sample $n \in \{10, 90\}$ observations with contracted uniform noise of standard deviation $\texttt{sd} \in \{\frac{1}{3}, 1\}$. Then we apply `LinGeo`, `LinFre`, and `LinCos`. (Part 1.)

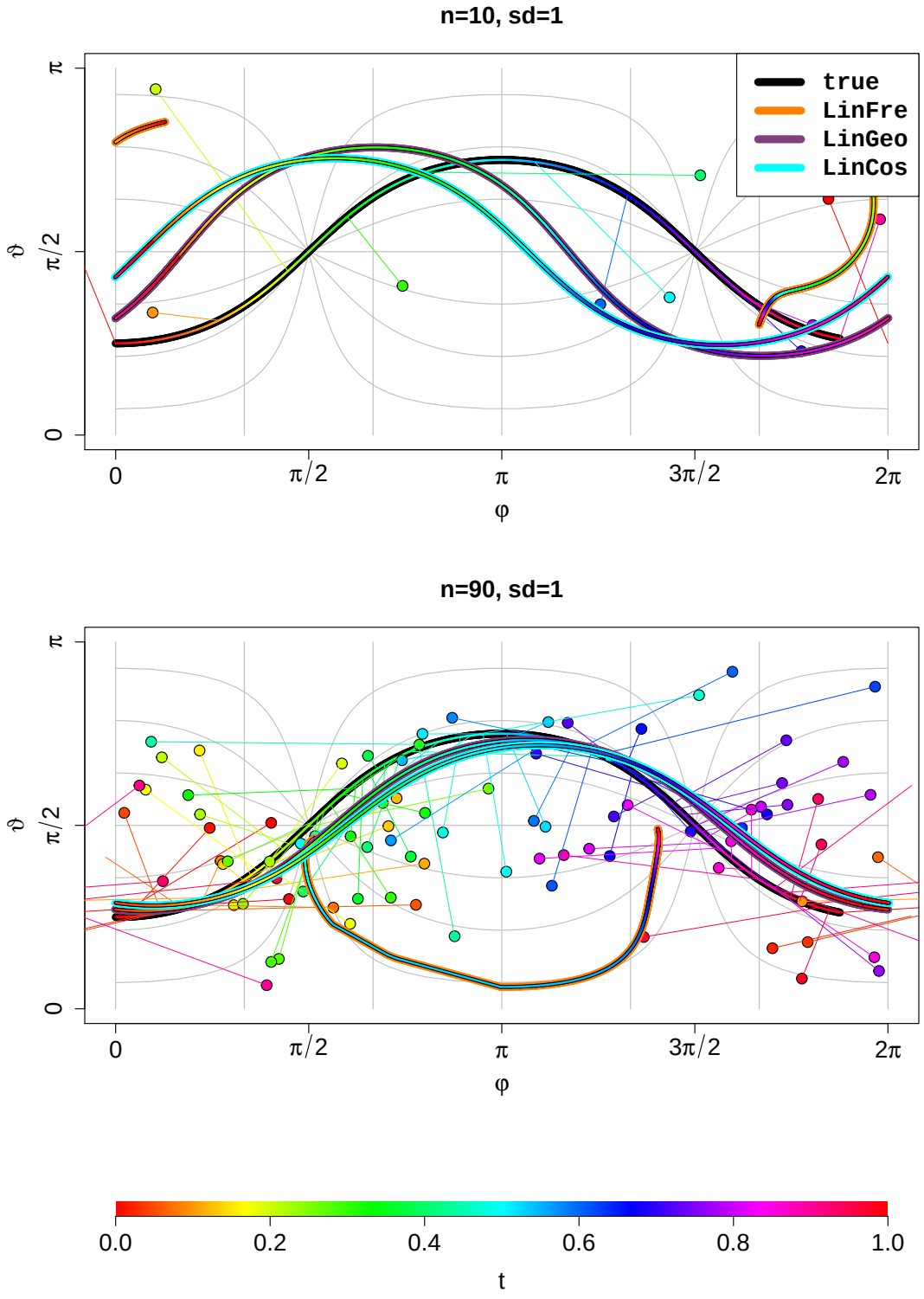

Figure 5.2: (Part 2.)

Setting			MISE		
n	sd	**speed**	linfre	lingeo	lincos
10	0.1	1.00000	0.00185	0.00232	0.00186
100	0.1	1.00000	0.00020	0.00020	0.00020
10	1.0	1.00000	0.26237	0.46893	0.48505
100	1.0	1.00000	0.03029	0.03111	0.03765
10	0.1	3.14159	0.00267	0.00223	0.00211
100	0.1	3.14159	0.00047	0.00019	0.00021
10	1.0	3.14159	0.42652	0.51267	0.48289
100	1.0	3.14159	0.06469	0.03267	0.04360
10	0.1	8.00000	2.21166	0.00231	0.00220
100	0.1	8.00000	2.05709	0.00021	0.00023
10	1.0	8.00000	2.91093	0.48702	0.47977
100	1.0	8.00000	2.35239	0.03090	0.04342

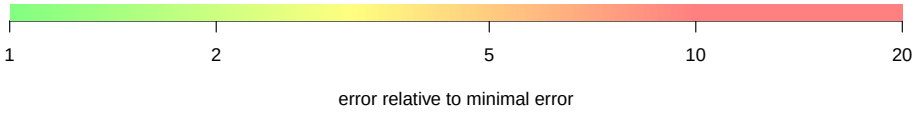

Table 5.1: Approximated MISE values for parametric regression methods. The colors give a visual indication of the MISE value of the given methods divided by the best MISE value in the row.

observations y_i are also color-coded to identify which x_i they belong to. Furthermore, thin colored lines are drawn between y_i and $m(x_i)$.

A geodesic of length 3 (Figure 5.1) is estimated similarly well by all estimators. This is true in different settings. Compare this with the estimation of a length 6 geodesic in Figure 5.2. Only `LinCos` and `LinGeo` perform well but not `LinFre`. This strongly suggests that `LinFre` is not consistent if non-Euclidean properties of the descriptor space play a significant role. Note that the errors in the settings $(n = 10, \texttt{sd} = \frac{1}{3})$ and $(n = 90, \texttt{sd} = 1)$ are similar and $\texttt{sd}^2/n$ is the same in both settings.

Next we repeat this experiment 1000 times for 12 different settings. The setting specifies the number of samples drawn n, the noise standard deviation $\texttt{sd} = \sqrt{\frac{1}{2}(\pi^2 - 4)}\, a$, and the speed of the true geodesic. For each run we calculate the integrated squared error, ISE, $\int_0^1 d(\hat{m}_t, m_t)^2 dt$. Then we take the mean of those 1000 ISE values to approximate the mean integrated squared error, MISE. Table 5.1 shows the results. We can see that for geodesics with small speed, all three methods perform well. For high speed geodesics `LinFre` does not give meaningful results. `LinGeo` is by far the slowest method in our implementation, as it has the most complex optimization problem to solve.

5.8.3 Nonparametric Regression

Next, we investigate the nonparametric methods `LocGeo`, `LocFre`, `TriGeo`, `TriFre`. We test two different regression functions $t \mapsto m_t$. The first one, named *simple* has angles $t \mapsto (\frac{1}{4}\pi, \frac{1}{2} + 2\pi t)$, see Figure 5.3. This seems to be a straight line in the Mercator projection but is a curved function on the sphere and cannot be estimated well by the parametric methods of the previous subsection. This *simple* curve is periodic. The second curve is described by $t \mapsto (\frac{1}{8}\pi + \frac{3}{4}\pi t, \frac{1}{2} + 3\pi t)$. Again this curve is not geodesic. It *spirals* around the sphere, see Figure 5.4, and is not periodic. To estimate nonperiodic functions with `TriGeo` and `TriFre`, which require periodicity, we copy the data and append it in reverse order to estimate the periodic function

$$t \mapsto \begin{cases} m_{2t} & \text{if } t < \frac{1}{2}\,, \\ m_{2-2t} & \text{if } t \geq \frac{1}{2}\,. \end{cases}$$

This may lead to boundary effects.

On a broad scale, all estimators seem to perform similarly, except for a worse outcome for `TriGeo` on the *spiral*. In the setting ($n = 10, \texttt{sd} = 1$) the estimators are not able to come close to the true curve. Compare this to the same setting in the parametric cases, where performance of estimators is still good enough to potentially be useful.

As with the parametric methods, we approximate the MISE values in different settings. The simulations are repeated 500 times. Only the two curves *simple* and *spiral* described above are used. The results are presented in Table 5.2. The more reliable analysis of the approximated MISE-values confirms that all estimators behave similar, except `TriGeo`, which has some bad outcomes. This may have several reasons. We were not able to show an error bound for this method and argued that it may be sub-optimal, i.e., it may be inherently worse than the other methods. We do not use cross-validation for `TriGeo`, as we do for the other methods, but fix $N = 3$. Thus, the comparison might be unfair, because the hyper-parameters are not tuned equally. Lastly, in `TriGeo`, we have to numerically solve an 8-dimensional nonconvex optimization problem (2 dimensions for each of $\hat{p}$, $\hat{v}_1$, $\hat{v}_2$, $\hat{v}_3$). There are 4 dimensions for `LocGeo` and 2 for the Fréchet methods. Our program might return values farther away from the optimum in those methods with higher dimensional optimization problems.

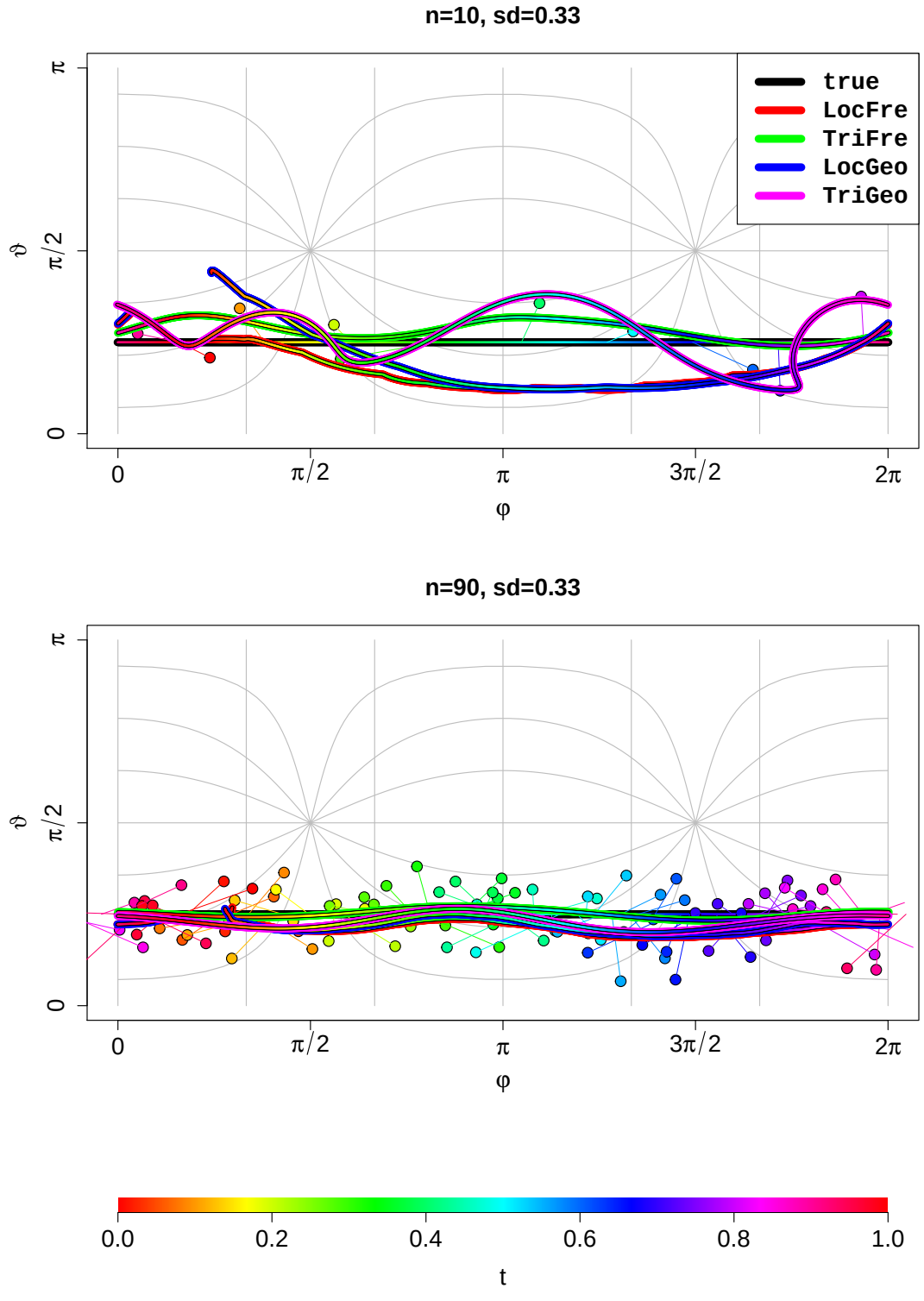

Figure 5.3: For the *simple* curve, we sample $n \in \{10, 90\}$ observations with contracted uniform noise of standard deviation $\texttt{sd} \in \{\frac{1}{3}, 1\}$. Then we apply `LocGeo`, `LocFre`, `TriGeo`, `TriFre`. (Part 1.)

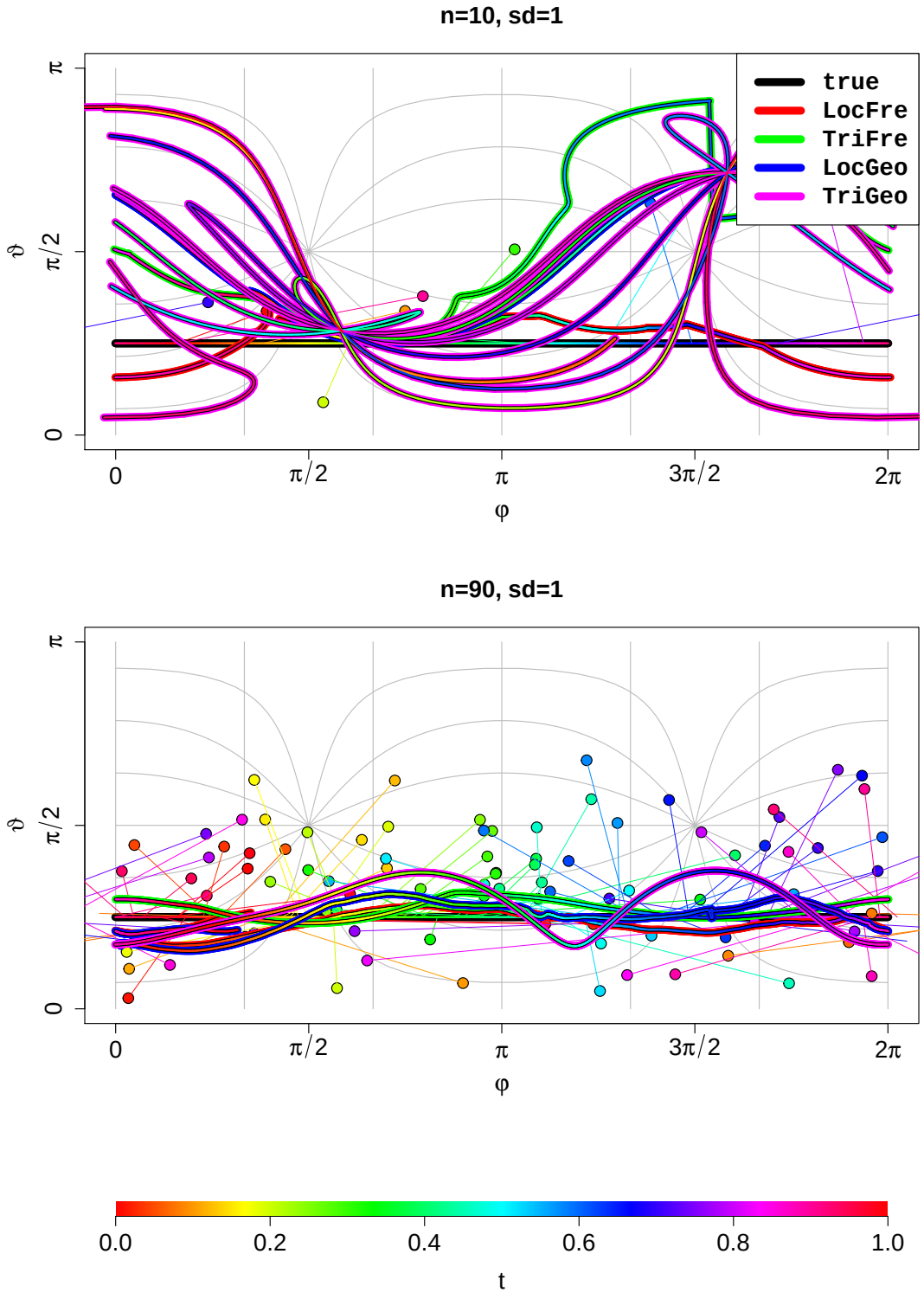

Figure 5.3: (Part 2.)

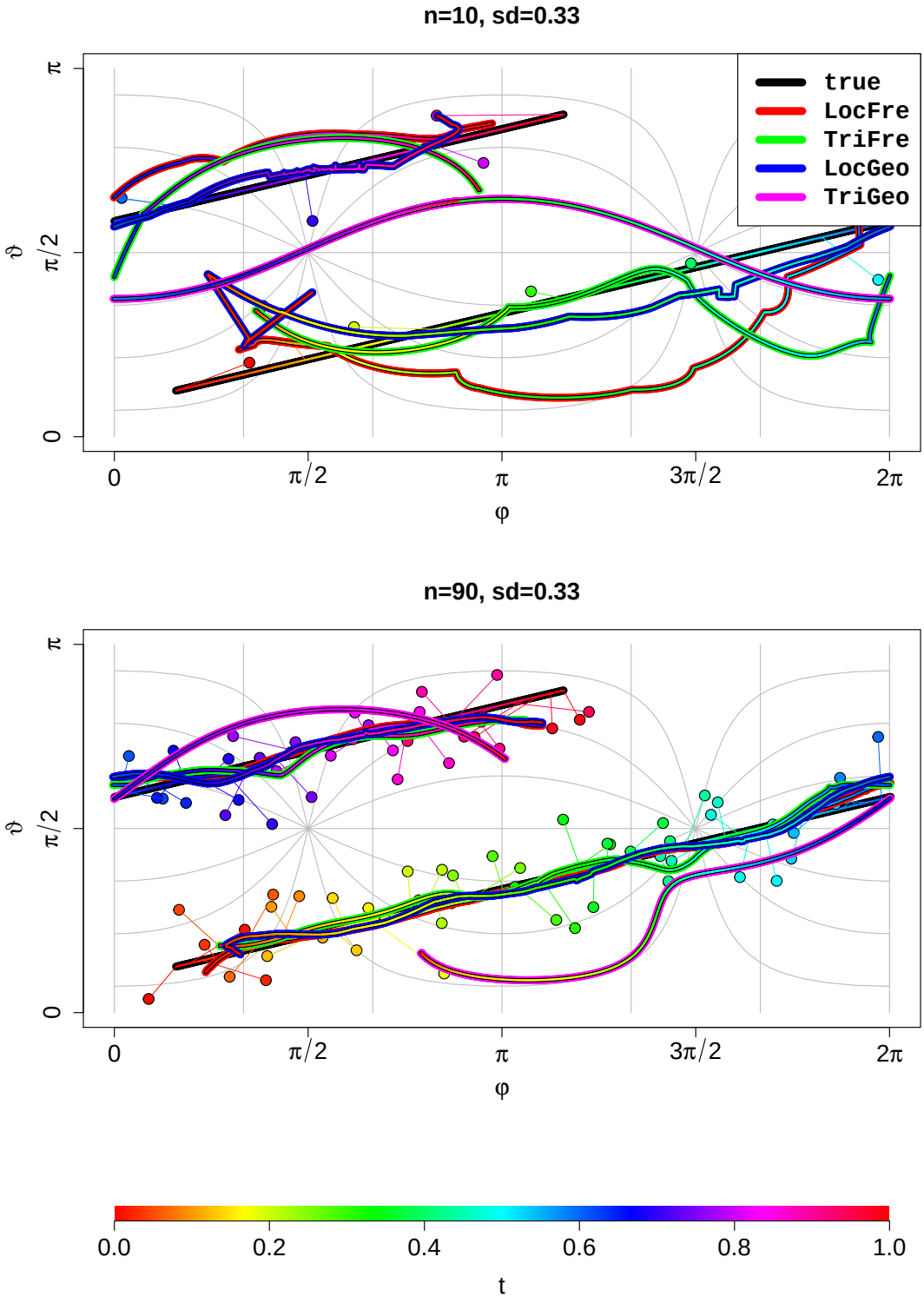

Figure 5.4: For the *spiral*, we sample $n \in \{10, 90\}$ observations with contracted uniform noise of standard deviation `sd` $\in \{\frac{1}{3}, 1\}$. Then we apply `LocGeo`, `LocFre`, `TriGeo`, `TriFre`. (Part 1.)

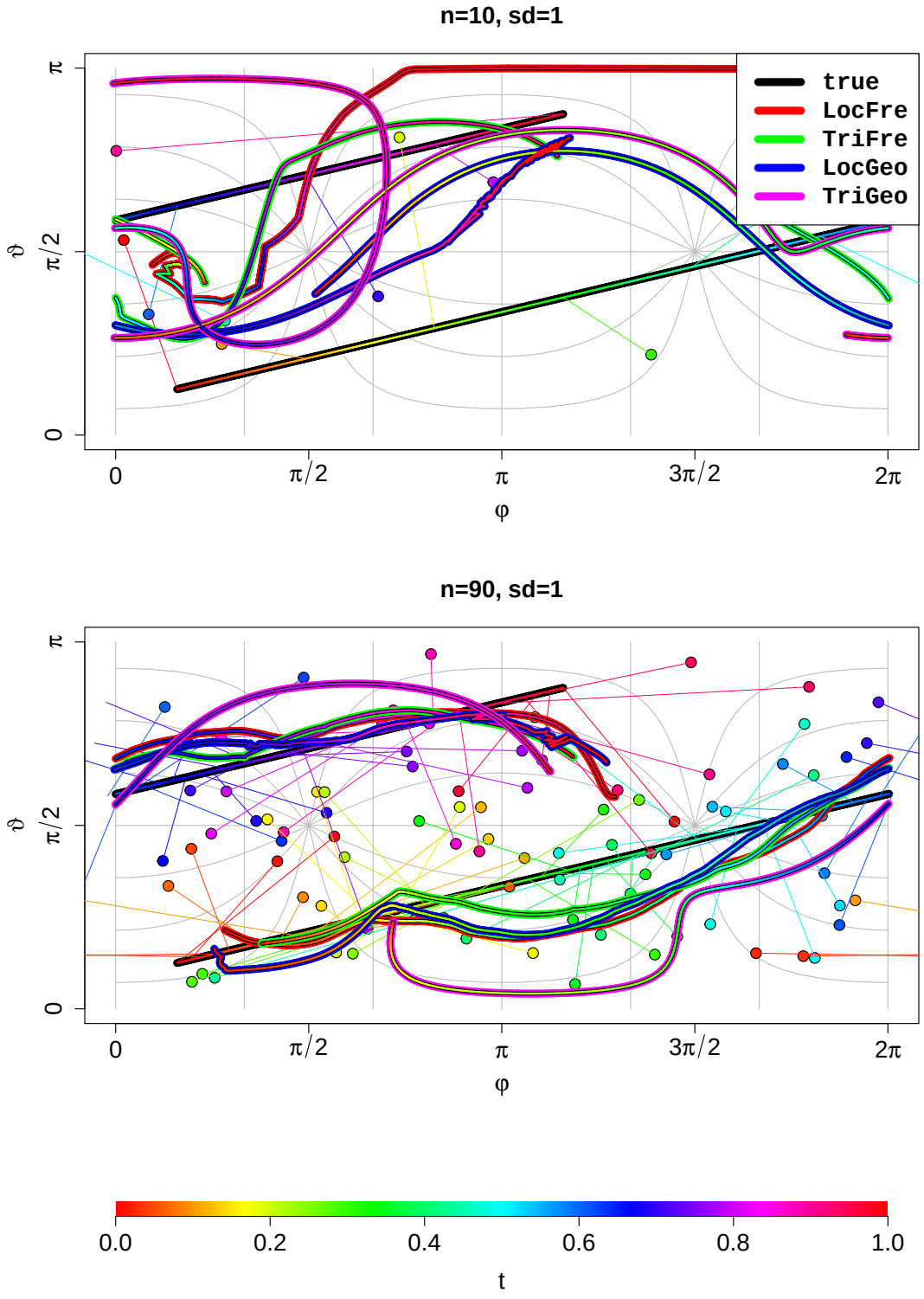

Figure 5.4: (Part 2.)

Setting			MISE			
n	σ	**curve**	`LocFre`	`TriFre`	`LocGeo`	`TriGeo`
20	0.25	simple	0.02070	0.02410	0.02595	0.01397
20	0.25	spiral	0.02899	0.05902	0.03268	0.38623
80	0.25	simple	0.00731	0.00662	0.00851	0.00361
80	0.25	spiral	0.00900	0.01534	0.01008	0.37191
20	1.00	simple	0.34890	0.39052	0.36356	0.86604
20	1.00	spiral	0.56768	0.52354	0.54786	0.91824
80	1.00	simple	0.12056	0.09350	0.11026	0.09228
80	1.00	spiral	0.15185	0.14662	0.14677	0.47189

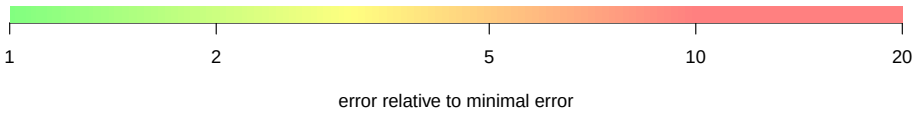

Table 5.2: Approximated MISE values for nonparametric regression methods. The colors give a visual indication of the MISE value of the given methods divided by the best MISE value in the row.

Appendix of Chapter 5

5.A Proofs

5.A.1 Section 5.2: `LinGeo`

5.A.1.1 Theorem

We prove Theorem 5.2. We first apply Variance to relate the difference between the objective functions to the loss between their minimizers. Then chaining is used to bound the objective functions and a peeling device leads to tail bounds on the loss. Lastly, we integrate the tails.

Define the objective functions

$$F_{\mathbf{x}}(\theta) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\mathfrak{c}(Y_{x_i}, g(x_i,\theta))]\,, \qquad F_{\mathbf{x}}(\theta_1,\theta_2) := F_{\mathbf{x}}(\theta_1) - F_{\mathbf{x}}(\theta_2)\,,$$
$$\hat{F}_{\mathbf{x}}(\theta) := \frac{1}{n}\sum_{i=1}^{n} \mathfrak{c}(y_i, g(x_i,\theta))\,, \qquad \hat{F}_{\mathbf{x}}(\theta_1,\theta_2) := \hat{F}_{\mathbf{x}}(\theta_1) - \hat{F}_{\mathbf{x}}(\theta_2)\,.$$

Variance and the minimizing property of $\hat{\theta}$ yield

$$C_{\mathsf{Vlo}}^{-1}\frac{1}{n}\sum_{i=1}^{n} \mathfrak{l}(m_{x_i}, \hat{m}_{x_i}) \le F_{\mathbf{x}}(\hat{\theta},\theta^*) \le F_{\mathbf{x}}(\hat{\theta},\theta^*) - \hat{F}_{\mathbf{x}}(\hat{\theta},\theta^*)\,.$$

Define

$$\Delta_{\mathbf{x}}(\delta) := \sup_{\theta\in \mathrm{B}_{\mathbf{x}}(\theta^*,\mathfrak{l},\delta)} \left(F_{\mathbf{x}}(\hat{\theta},\theta^*) - \hat{F}_{\mathbf{x}}(\hat{\theta},\theta^*)\right)$$

and

$$Z_i(\theta) := \frac{1}{n}\left(\mathbb{E}[\mathfrak{c}_{x_i}(Y_{x_i},\theta) - \mathfrak{c}_{x_i}(Y_{x_i},\theta^*)] - \mathfrak{c}_{x_i}(y_i,\theta) + \mathfrak{c}_{x_i}(y_i,\theta^*)\right)\,.$$

Then $Z_1,\dots,Z_n$ are independent and centered processes with $Z_i(\theta^*) = 0$. They are also integrable due to Moment. By the definition of $\mathfrak{a}_x$, it holds

$$n\left(Z_i(\theta_1) - Z_i(\theta_2) - Z_i'(\theta_1) + Z_i'(\theta_2)\right) \le \mathfrak{b}(\theta_1,\theta_2)\mathfrak{a}_{x_i}(y_i,y_i')\,.$$

Thus, the chaining result of Theorem 5.56 (appendix 5.B) yields

$$\mathbb{E}[\Delta_{\mathbf{x}}(\delta)^\kappa] \le c_\kappa \left(\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\mathfrak{a}_{x_i}(y_i,y_i')^2\right)^{\frac{\kappa}{2}}\right]^{\frac{1}{\kappa}} \gamma_2(\mathrm{B}_{\mathbf{x}}(\theta^*,\mathfrak{l},\delta),\mathfrak{b})n^{-\frac{1}{2}}\right)^\kappa\,.$$

By ENTROPY $\gamma_2(\mathrm{B}_{\mathbf{x}}(\theta^*, \mathfrak{l}, \delta), \mathfrak{b}) \le C_{\mathsf{Ent}}\delta^{\xi}$ for $\delta > T_n$. As $\kappa \ge 2$, by MOMENT,

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\mathfrak{a}_{x_i}(y_i, y_i')^2\right)^{\frac{\kappa}{2}}\right]^{\frac{1}{\kappa}} \le \left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[\mathfrak{a}_{x_i}(y_i, y_i')^{\kappa}\big]\right)^{\frac{1}{\kappa}} \le C_{\mathsf{Mom}}\,.$$

Thus, for $\delta > T_n$,

$$\mathbb{E}[\Delta_{\mathbf{x}}(\delta)^{\kappa}] \le c_{\kappa}\left(n^{-\frac{1}{2}}C_{\mathsf{Ent}}C_{\mathsf{Mom}}\delta^{\xi}\right)^{\kappa}.$$

For $T_n < a < b < \infty$, using Markov's inequality, we obtain

$$\begin{aligned}\mathbb{P}\left(C_{\mathsf{Vlo}}^{-1}\frac{1}{n}\sum_{i=1}^{n}\mathfrak{l}(m_{x_i}, \hat{m}_{x_i}) \in [a, b]\right) &\le \mathbb{P}(a \le \Delta_{\mathbf{x}}(b)) \\ &\le a^{-\kappa}\mathbb{E}[\Delta_{\mathbf{x}}(b)^{\kappa}] \\ &\le c_{\kappa}\left(\frac{n^{-\frac{1}{2}}C_{\mathsf{Ent}}C_{\mathsf{Mom}}b^{\xi}}{a}\right)^{\kappa}.\end{aligned}$$

We use this bound in the peeling device, to obtain a tail bound for $t \ge T_n$:

$$\begin{aligned}\mathbb{P}\left(C_{\mathsf{Vlo}}^{-1}\frac{1}{n}\sum_{i=1}^{n}\mathfrak{l}(m_{x_i}, \hat{m}_{x_i}) > t\right) &\le \sum_{k=0}^{\infty}\mathbb{P}\left(C_{\mathsf{Vlo}}^{-1}\frac{1}{n}\sum_{i=1}^{n}\mathfrak{l}(m_{x_i}, \hat{m}_{x_i}) \in [2^k t, 2^{k+1}t]\right) \\ &\le 2^{\kappa}c_{\kappa}\sum_{k=0}^{\infty}\left(\frac{n^{-\frac{1}{2}}C_{\mathsf{Ent}}C_{\mathsf{Mom}}t^{\xi}2^{k\xi}}{t2^k}\right)^{\kappa} \\ &= 2^{\kappa}c_{\kappa}\left(n^{-\frac{1}{2}}C_{\mathsf{Ent}}C_{\mathsf{Mom}}\right)^{\kappa}t^{\kappa(\xi-1)}\sum_{k=0}^{\infty}\left(2^{\kappa(\xi-1)}\right)^{k} \\ &= c_{\kappa,\xi}\left(n^{-\frac{1}{2}}C_{\mathsf{Ent}}C_{\mathsf{Mom}}\right)^{\kappa}t^{\kappa(\xi-1)}\end{aligned}$$

with $c_{\kappa,\xi} := \frac{2^{\kappa}c_{\kappa}}{1-2^{\kappa(\xi-1)}}$. To obtain the desired bound on the expectation, we integrate the tail probability

$$\begin{aligned}\mathbb{E}\left[C_{\mathsf{Vlo}}^{-1}\frac{1}{n}\sum_{i=1}^{n}\mathfrak{l}(m_{x_i}, \hat{m}_{x_i})\right] &\le T_n + \int_{T_n}^{\infty}\mathbb{P}\left(C_{\mathsf{Vlo}}^{-1}\frac{1}{n}\sum_{i=1}^{n}\mathfrak{l}(m_{x_i}, \hat{m}_{x_i}) > t\right)\mathrm{d}t \\ &\le T_n + \int_{0}^{\infty}\min\Big(1, c_{\kappa,\xi}\left(n^{-\frac{1}{2}}C_{\mathsf{Ent}}C_{\mathsf{Mom}}\right)^{\kappa}t^{\kappa(\xi-1)}\Big)\,\mathrm{d}t\,.\end{aligned}$$

It holds

$$\int_0^{\infty}\min(1, bt^{-a})\mathrm{d}t = \frac{a}{a-1}b^{\frac{1}{a}}$$

for all $a > 1, b > 0$. Now set $a = \kappa(1-\xi)$ and $b = c_{\kappa,\xi}\left(n^{-\frac{1}{2}}C_{\mathsf{Ent}}C_{\mathsf{Mom}}\right)^{\kappa}$. We obtain

$$\begin{aligned}C_{\mathsf{Vlo}}^{-1}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathfrak{l}(m_{x_i}, \hat{m}_{x_i})\right] &\le T_n + \frac{\kappa(1-\xi)}{\kappa(1-\xi)-1}\left(c_{\kappa,\xi}\left(n^{-\frac{1}{2}}C_{\mathsf{Ent}}C_{\mathsf{Mom}}\right)^{\kappa}\right)^{\frac{1}{\kappa(1-\xi)}} \\ &= T_n + c'_{\kappa,\xi}\left(n^{-\frac{1}{2}}C_{\mathsf{Ent}}C_{\mathsf{Mom}}\right)^{\frac{1}{1-\xi}}\end{aligned}$$

with $c'_{\kappa,\xi} = \frac{\kappa(1-\xi)}{\kappa(1-\xi)-1}c_{\kappa,\xi}^{\frac{1}{\kappa(1-\xi)}}$.

5.A.1.2 Corollaries

Proof of Corollary 5.4. In Hadamard spaces, the variance inequality, i.e., $d(q,m)^2 \leq \mathbb{E}[d(Y,q)^2 - d(Y,m)^2]$ for all $q \in \mathcal{Q}$ and $m = \arg\min_{q\in\mathcal{Q}} \mathbb{E}[d(Y,q)^2]$, holds for all distributions of Y with $\mathbb{E}[d(Y,q)^2] < \infty$, [Stu03, Theorem 4.9]. This shows VARIANCE. Furthermore, the quadruple inequality

$$d(y,q)^2 - d(\tilde{y},q)^2 - d(y,\tilde{q})^2 + d(\tilde{y},\tilde{q})^2 \leq 2d(y,\tilde{y})d(q,\tilde{q})$$

holds for all $q,\tilde{q},y,\tilde{y} \in \mathcal{Q}$, [Stu03, Theorem 4.9]. Thus, with METRICUP we get

$$\Diamond_x(y,\tilde{y},\theta,\tilde{\theta}) \leq 2d(y,\tilde{y})d(g(x,\theta),g(x,\tilde{\theta})) \leq 2C_{\mathsf{Mup}}d(y,\tilde{y})\left|\theta-\tilde{\theta}\right| .$$

Hence, we set $\mathfrak{a}_x(y,\tilde{y}) := 2d(y,\tilde{y})C_{\mathsf{Mup}}$ and $\mathfrak{b} := |\cdot - \cdot|$ when applying Theorem 5.2. Next, to check ENTROPY, use METRICLO

$$\frac{1}{n}\sum_{i=1}^{n} \mathfrak{l}(g(x_i,\theta),g(x_i,\theta^*)) = \frac{1}{n}\sum_{i=1}^{n} d(g(x_i,\theta),g(x_i,\theta^*))^2 \geq C_{\mathsf{Mlo}}^{-1}\left|\theta-\tilde{\theta}\right|^2 - T_n ,$$

where $T_n := C_{\mathsf{Res}}n^{-1}$. Thus, for $\delta > T_n$,

$$\mathrm{B}_{\mathbf{x}}(\theta^*,\mathfrak{l},\delta) \subseteq \left\{\theta\in\Theta\colon C_{\mathsf{Mlo}}^{-1}|\theta-\theta^*|^2 \leq 2\delta\right\} = \left\{\theta\in\Theta\colon |\theta-\theta^*| \leq (2C_{\mathsf{Mlo}}\delta)^{\frac{1}{2}}\right\} .$$

From this, together with the bound on γ_2 for Euclidean spaces Lemma 5.57 (appendix 5.B), we obtain

$$\begin{aligned}\gamma_2(\mathrm{B}_{\mathbf{x}}(\theta^*,\mathfrak{l},\delta),\mathfrak{b}) &\leq \gamma_2\Big(\left\{\theta\in\Theta\colon |\theta-\theta^*| \leq (2C_{\mathsf{Mlo}}\delta)^{\frac{1}{2}}\right\},\mathfrak{b}\Big)\\ &\leq c\,(\mathsf{d}_\Theta C_{\mathsf{Mlo}}\delta)^{\frac{1}{2}}\\ &= C_{\mathsf{Ent}}\delta^{\frac{1}{2}}\end{aligned}$$

with $C_{\mathsf{Ent}} := c\,(\mathsf{d}_\Theta C_{\mathsf{Mlo}})^{\frac{1}{2}}$. □

Proof of Corollary 5.5. We want to apply Theorem 5.2. Hence, we have to check its assumptions. VARIANCE is a condition of the corollary. In order show MOMENT, we note

$$d(y,g(x,\theta))^2 - d(y,g(x,\tilde{\theta}))^2 \leq 2\,\mathsf{diam}(\mathcal{Q})d(g(x,\theta),g(x,\tilde{\theta})) \leq 2\,\mathsf{diam}(\mathcal{Q})C_{\mathsf{Mup}}\left|\theta-\tilde{\theta}\right|$$

using the triangle inequality and METRICUP. Thus,

$$\Diamond_x(y,\tilde{y},\theta,\tilde{\theta}) \leq 4\,\mathsf{diam}(\mathcal{Q})C_{\mathsf{Mup}}\left|\theta-\tilde{\theta}\right| .$$

We can set $\mathfrak{a}_x(y,\tilde{y}) := 4\,\mathsf{diam}(\mathcal{Q})C_{\mathsf{Mup}}$ and $\mathfrak{b} := |\cdot - \cdot|$ when applying Theorem 5.2. The

moment condition is trivial, as $\mathfrak{a}_x$ is a finite constant. As before

$$\gamma_2(\mathrm{B}_{\mathbf{x}}(\theta^*, \mathfrak{l}, \delta), \mathfrak{b}) \leq C_{\mathsf{Ent}} \delta^{\frac{1}{2}}$$

with $C_{\mathsf{Ent}} := c\,(\mathsf{d}_\Theta C_{\mathsf{Mlo}})^{\frac{1}{2}}$. □

Next, we want to apply Corollary 5.5 to show Corollary 5.1. To do this, we need to show MetricUp and MetricLo translated to the spherical setting:

- MetricUp:
  There is $C_{\mathsf{Mup}} \in [1, \infty)$ such that $d(\mathsf{Exp}(q, xv), \mathsf{Exp}(p, xu)) \leq C_{\mathsf{Mup}}\,(|p - q| + |u - v|)$ for all $x \in [-1, 1]$, $(q, u), (p, v) \in \mathsf{T}\mathbb{S}^k$.
- MetricLo:
  There are $T_n \geq 0$ and $C_{\mathsf{Mlo}} \in [1, \infty)$ such that $\frac{1}{n} \sum_{i=1}^n d(\mathsf{Exp}(q, x_i v), \mathsf{Exp}(p, x_i u))^2 \geq C_{\mathsf{Mlo}}^{-1} \left(|p - q|^2 + |u - v|^2\right) - C_{\mathsf{Res}} n^{-1}$.

The following lemma shows MetricUp with $C_{\mathsf{Mup}} := 4\pi$. This constant may not be sharp.

Lemma 5.28. Let $(p, u), (q, v) \in \mathsf{T}\mathbb{S}^k$. Then

$$d(\mathsf{Exp}(q, v), \mathsf{Exp}(p, u)) \leq \frac{\pi}{2} |q - p| + 2\pi |v - u| \,.$$

Proof. We can bound the intrinsic metric on the sphere by the extrinsic one,

$$\begin{aligned} d(\mathsf{Exp}(q, v), \mathsf{Exp}(p, u)) &\leq \frac{\pi}{2} |\mathsf{Exp}(q, v) - \mathsf{Exp}(p, u)| \\ &\leq \frac{\pi}{2} \left(|\cos(|v|)q - \cos(|u|)p| + \left| \frac{\sin(|v|)}{|v|} v - \frac{\sin(|u|)}{|u|} u \right| \right) . \end{aligned}$$

For the cos-terms, it holds

$$\begin{aligned} |\cos(|v|)q - \cos(|u|)p| &\leq |\cos(|v|)|\,|q - p| + |p|\,|\cos(|v|) - \cos(|u|)| \\ &\leq |q - p| + ||v| - |u|| \,. \end{aligned}$$

For the sin-terms, let $J(x)$ be the Jacobi matrix of the function $\mathbb{R}^k \to \mathbb{R}^k$, $x \mapsto \frac{\sin(|x|)}{|x|} x$. Then

$$\left| \frac{\sin(|v|)}{|v|} v - \frac{\sin(|u|)}{|u|} u \right| \leq \sup_{x \in \mathbb{R}^k} \|J(x)\|_{\mathsf{op}}\, |u - v| \,.$$

As

$$J(x) = \left(\cos(|x|) - \frac{\sin(|x|)}{|x|} \right) |x|^{-2}\, x x^\top + \frac{\sin(|x|)}{|x|} I_k \,,$$

it holds

$$\|J(x)\|_{\mathsf{op}} \leq \left(|\cos(|x|)| + \left|\frac{\sin(|x|)}{|x|}\right|\right) \| \, |x|^{-2} \, xx^\top \|_{\mathsf{op}} + \left|\frac{\sin(|x|)}{|x|}\right| \|I_k\|_{\mathsf{op}} \leq 3 \, .$$

Thus, $d(\mathsf{Exp}(q,v), \mathsf{Exp}(p,u)) \leq \frac{\pi}{2} \left(|q-p| + ||v| - |u|| + 3\,|u-v|\right)$. □

For MetricLo we prove following lemma.

Lemma 5.29. Let $(p,u),(q,v) \in \mathsf{T}\mathbb{S}^k$ with $|u|\,, |v| \leq \frac{\pi}{2}$. Then

$$\int_{-1}^{1} d_{\mathbb{S}^k}(\mathsf{Exp}(p,xu), \mathsf{Exp}(q,xv))^2 \mathrm{d}x \geq \frac{2}{\pi} \, |p-q|^2 + \frac{8}{\pi^2} \, |v-u|^2 \, .$$

Proof. First we lower bound the intrinsic distance $d_{\mathbb{S}^k}$ by the Euclidean one and use the explicit representation of the Exp-function,

$$d_{\mathbb{S}^k}(\mathsf{Exp}(p,xu), \mathsf{Exp}(q,xv))^2 \geq \left|\cos(x\,|u|)p + \sin(x\,|u|)\frac{u}{|u|} - \cos(x\,|v|)q - \sin(x\,|v|)\frac{v}{|v|}\right|^2 .$$

When integrating after calculating the squared norm, all summands with a $\cos()\sin()$-factor disappear, because of symmetry. Thus, we obtain

$$\begin{aligned}
&\int_{-1}^{1} d_{\mathbb{S}^k}(\mathsf{Exp}(p,xu), \mathsf{Exp}(q,xv))^2 \mathrm{d}x \\
&\geq \int_{-1}^{1} \cos(x\,|u|)^2 p^\top p - 2\cos(x\,|u|)\cos(x\,|v|) p^\top q + \cos(x\,|v|)^2 q^\top q \, \mathrm{d}x \\
&\quad + \int_{-1}^{1} \sin(x\,|u|)^2 \frac{u^\top u}{|u|^2} - 2\sin(x\,|u|)\sin(x\,|v|)\frac{u^\top v}{|u|\,|u|} + \sin(x\,|v|)^2 \frac{v^\top v}{|v|^2} \, \mathrm{d}x \, .
\end{aligned}$$

As $|p| = |q| = 1$, $\cos(x)^2 + \sin(x)^2 = 1$, $2\cos(\alpha)\cos(\beta) = \cos(\alpha-\beta) + \cos(\alpha+\beta)$, and $2\sin(\alpha)\sin(\beta) = \cos(\alpha-\beta) - \cos(\alpha+\beta)$, the right hand side reduces to

$$\int_{-1}^{1} 2 - (\cos(xa) + \cos(xb))\, p^\top q - (\cos(xa) - \cos(xb))\, z \, \mathrm{d}x \, ,$$

where we set $a := |u| - |v|$, $b := |u| + |v|$, and $z := \frac{u^\top v}{|u||v|}$. Integrating yields

$$4 - 2\left(\frac{\sin(a)}{a} + \frac{\sin(b)}{b}\right) q^\top p - 2\left(\frac{\sin(a)}{a} + \frac{\sin(b)}{b}\right) z \, .$$

As $\left(\frac{\sin(a)}{a} + \frac{\sin(b)}{b}\right) \geq \frac{2}{\pi}$ for $|v|, |u| \leq \frac{\pi}{2}$ and as $\frac{1}{2}|q-p|^2 = (1 - q^\top p)$, we have shown

$$\begin{aligned}
&\int_{-1}^{1} d_{\mathbb{S}^k}(\mathsf{Exp}(p, xu), \mathsf{Exp}(q, xv))^2 \mathrm{d}x \\
&\geq \frac{2}{\pi}|p-q|^2 + \left(4 - 2\left(\frac{\sin(a)}{a} + \frac{\sin(b)}{b}\right) - 2\left(\frac{\sin(a)}{a} + \frac{\sin(b)}{b}\right) z\right).
\end{aligned}$$

To complete the proof, we will show $f(a, b, z) \geq 0$ for all $a \in [-\frac{\pi}{2}, \frac{\pi}{2}]$, $b \in [0, \pi]$, and $z \in [-1, 1]$, where

$$f(a,b,z) := 4 - 2\left(\frac{\sin(a)}{a} + \frac{\sin(b)}{b}\right) - 2\left(\frac{\sin(a)}{a} + \frac{\sin(b)}{b}\right) z - c\left(a^2 + b^2 + (a^2 - b^2)z\right)$$

with $c = \frac{4}{\pi^2}$. This suffices as $a^2 + b^2 + (a^2 - b^2)z = 2|v-u|^2$. As f is linear in z, it is minimized either at $z = 1$ or at $z = -1$. It holds

$$f(a,b,1) = 4 - 4\frac{\sin(a)}{a} - ca^2, \qquad f(a,b,-1) = 4 - 4\frac{\sin(b)}{b} - cb^2.$$

Consider the function

$$g(x) := \frac{1 - \frac{\sin(x)}{x}}{x^2} \quad \text{with derivative} \quad g'(x) = \frac{\cos(x) - 2}{x^3}.$$

It is symmetric at 0 and decreasing for positive x. Thus, it attains its minimum on $[-\frac{\pi}{2}, \pi]$ at $x = \pi$. For $c := \frac{4}{\pi^2} = 4g(\pi)$, we hereby have shown $f(x, y, z) \geq 0$ and thus have proven the lemma. □

Proof of Corollary 5.1. We want to apply Corollary 5.5 and have to check its conditions. Variance is the assumption stated in Corollary 5.1. MetricUp is implied by Lemma 5.28. To show MetricLo, let

$$T_n := \left| \frac{1}{2}\int_{-1}^{1} d(\mathsf{Exp}(q, xv), \mathsf{Exp}(p, xu))^2 \mathrm{d}x - \frac{1}{n}\sum_{i=1}^{n} d(\mathsf{Exp}(q, x_i v), \mathsf{Exp}(p, x_i u))^2 \right|.$$

With the use of the Lemma 5.29 above on $(q, v), (p, u) \in \mathsf{T}\mathbb{S}^k$, $|v|, |u| \leq \Lambda$, we obtain

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} d(\mathsf{Exp}(q, x_i v), \mathsf{Exp}(p, x_i u))^2 &\geq \frac{1}{2\Lambda}\int_{-1}^{1} d(\mathsf{Exp}(q, xv/\Lambda), \mathsf{Exp}(p, xu/\Lambda))^2 \mathrm{d}x - T_n \\
&\geq C_{\mathsf{MIo}}\left(|q-p|^2 + |v-u|^2\right) - T_n,
\end{aligned}$$

where $C_{\mathsf{Mlo}} := \frac{1}{\Lambda\pi}$. Finally, for an L-Lipschitz continuous function $f\colon [-1,1] \to \mathbb{R}$,

$$\left| \int_{-1}^{1} f(x)\mathrm{d}x - \frac{2}{n} \sum_{i=1}^{n} f(x_i) \right| \leq \frac{2L}{n} .$$

In the worst case $x \mapsto \mathsf{Exp}(p, xv/\Lambda)$ and $x \mapsto \mathsf{Exp}(q, xu/\Lambda)$ move in opposite directions and the distance changes with a rate of $(|u| + |v|)/\Lambda$. Thus, we obtain

$$T_n \leq \frac{2\left(|u| + |v|\right)^2}{n\Lambda^2} \leq \frac{8\pi^2}{n} = C_{\mathsf{Res}} n^{-1} ,$$

where $C_{\mathsf{Res}} := 8\pi^2$. □

5.A.2 Section 5.3: `LinFre`

Proof of Theorem 5.8. We show that for each pair $y, z \in \mathcal{Q}$ there is a point $m \in \mathcal{Q}$ such that for all $q \in \mathcal{Q}$ it holds

$$d(m,q)^2 = \frac{1}{2} d(y,q)^2 + \frac{1}{2} d(z,q)^2 - \frac{1}{4} d(y,z)^2 .$$

Then [Stu03, Definition 2.1 and Proposition 3.5 (iii)] implies that $\mathcal{Q}$ is a Hilbert space.

Let γ_t be a minimizing geodesic between $y = \gamma_{-1}$ and $z = \gamma_1$. Let $m := \gamma_0$. Let $q \in \mathcal{Q}$ be arbitrary. The *strict linear Fréchet regression model* implies that there are $\theta_0, \theta_1 \in \mathbb{R}$ such that

$$\theta_0 + \theta_1 t = \mathbb{E}[d(Y_t, q)^2 - d(Y_t, m)^2] = d(\gamma_t, q)^2 - d(\gamma_t, m)^2 . \tag{5.1}$$

Adding this equality with $t = +1$ and $t = -1$, we obtain

$$2\theta_0 = d(\gamma_1, q)^2 - d(\gamma_1, m)^2 + d(\gamma_{-1}, q)^2 - d(\gamma_{-1}, m)^2 = d(y,q)^2 + d(z,q)^2 - \frac{1}{2} d(y,z)^2$$

as $d(y,m) = d(z,m) = \frac{1}{2} d(y,z)$. Evaluating (5.1) at $t = 0$ yields $\theta_0 = d(m,q)^2$. Together, we arrive at the result

$$2d(m,q)^2 = d(y,q)^2 + d(z,q)^2 - \frac{1}{2} d(y,z)^2 . \qquad \square$$

Proof of Proposition 5.9. Let $(\alpha, \beta) \in \mathbb{S}^2$. For $\varphi \in [0, 2\pi)$, let $\angle(\varphi, \beta) \in [0, \pi]$ be the

distance of the two angles on the circle. We calculate the objective function,

$$\begin{aligned}\mathbb{E}[\cos(d(Y,(\alpha,\beta)))] &= \frac{1}{2\pi}\int\int_0^{2\pi} \sin(\vartheta)\sin(\alpha) + \sin(\vartheta)\sin(\alpha)\cos(\angle(\varphi,\beta))\mathrm{d}\varphi\mathrm{d}\nu(\vartheta) \\ &= \int \left(\sin(\vartheta)\sin(\alpha) + \frac{1}{\pi}\sin(\vartheta)\sin(\alpha)\int_0^{\pi}\cos(\varphi)\mathrm{d}\varphi\right)\mathrm{d}\nu(\vartheta) \\ &= \int \sin(\vartheta)\sin(\alpha)\mathrm{d}\nu(\vartheta) \\ &= A\sin(\alpha)\,.\end{aligned}$$

Thus, if $A > 0$, $\mathbb{E}[\cos(d(Y,(\alpha,\beta)))]$ is uniquely maximized at $\alpha = \pi/2$, analogously for $A < 0$. If $A = 0$, $\mathbb{E}[\cos(d(Y,(\alpha,\beta)))] = 0$ for all $(\alpha,\beta) \in \mathbb{S}^2$. □

Proof of Proposition 5.10. By the law of cosines

$$\mathbb{E}[\cos(\overline{Y_t,q})] = \cos(\overline{m_t,q})\mathbb{E}[\cos(\overline{Y_t,m_t})] + \sin(\overline{m_t,q})\mathbb{E}[\sin(\overline{Y_t,m_t})\cos(\angle(Y_t,m_t,q))]\,.$$

By Lemma 5.30 below, $\mathbb{E}[\sin(\overline{Y_t,m_t})\cos(\angle(Y_t,m_t,q))] = 0$. By the Pythagorean theorem with $\angle(m_t,\gamma_{s_q},q) = \frac{\pi}{2}$,

$$\cos(\overline{m_t,q}) = \cos(\overline{m_t,\gamma_{s_q}})\cos(\overline{\gamma_{s_q},q})\,.$$

By definition, $\cos(\overline{m_t,\gamma_{s_q}}) = \cos(\overline{\gamma_{t_0+\lambda t},\gamma_{s_q}}) = \cos(B_q + \lambda t)$. It holds

$$\cos(B_q+\lambda t) = \cos(B_q)\cos(\lambda t) - \sin(B_q)\sin(\lambda t)$$

and, thus,

$$\mathbb{E}[\cos(\overline{Y_t,q})] = A_q\cos(B_q+\lambda t) = a_q\cos(\lambda t) + b_q\sin(\lambda t)\,.$$ □

Lemma 5.30. Let $(\mathcal{Q},d)$ be a Alexandrov space of nonpositive or nonnegative curvature [BBI01, section 4]. Let be a geodesic metric space. Let $f\colon [0,\infty) \to \mathbb{R}$ be a continuously differentiable function with derivative f'. Let Y be a random variable with values in $\mathcal{Q}$ such that $\mathbb{E}[|f(d(Y,q))|] < \infty$ and $\mathbb{E}[|f'(d(Y,q))|] < \infty$ for all $q \in \mathcal{Q}$. Let $m \in \arg\max_{q\in\mathcal{Q}} \mathbb{E}[f(d(Y,q))]$. Then $\mathbb{E}[f'(\overline{Y,m})\cos(\angle(Y,m,q))] = 0$, where $\angle(Y,m,q)$ is the angle between Y, m, and q at m.

Proof. Let $(\gamma_t)_{t\in[0,T]}$ be the minimizing unit-speed geodesic between $\gamma_0 = m$ and $\gamma_T = q$.

[BBI01, Theorem 4.5.6] yields $\partial_t d(Y, \gamma_t)_{|t=0} = -\cos(\alpha)$ where $\alpha := \angle(Y, m, q)$. Thus,

$$\begin{aligned} 0 &= \partial_t \mathbb{E}[f(d(Y, \gamma_t))]_{|t=0} \\ &= \mathbb{E}[\partial_t f(d(Y, \gamma_t))]_{|t=0} \\ &= \mathbb{E}[f'(d(Y, \gamma_t))\partial_t d(Y, \gamma_t))]_{|t=0} \\ &= -\mathbb{E}[f'(d(Y, m))\cos(\alpha)]\,. \end{aligned}$$

□

5.A.3 Section 5.4: `LocGeo`

5.A.3.1 Theorem

We prove Theorem 5.12. To this end, we first replace the integral over x by a sum over x_i in Lemma 5.32. Then the comparison of the estimated parameter $\hat{\theta}_{t,h}$ with the best local parameter $\theta_{t,h}$ is replaced by the comparison of $\hat{\theta}_{t,h}$ to the true function m in Lemma 5.34. This is necessary to apply the variance inequality, which makes it possible to translate a bound on the objective functions to a bound on their minimizers, which are elements of the metric space. For the remaining part, we bound a variance term via chaining (Lemma 5.35) and a bias term using the smoothness assumption (Lemma 5.36). These are used in Lemma 5.37, where a peeling device is applied to bound the tails of the error distribution (and via integration also its expectation). This is supplemented by the auxiliary lemmata Lemma 5.38 and Lemma 5.39 . But first we start out with another auxiliary result, Lemma 5.31, which shows that $\mathfrak{a}$ and $\mathfrak{b}$ are semi-metrics.

A map $d\colon \mathcal{Q} \times \mathcal{Q} \to [0, \infty]$ is called *semi-metric* on $\mathcal{Q}$, if d is symmetric with $d(q, q) = 0$ for all $q \in \mathcal{Q}$ and obeys the triangle inequality.

Lemma 5.31. The functions $\mathfrak{a}$ and $\mathfrak{b}$ are semi-metrics on $\mathcal{Q}$ and Θ, respectively.

Proof. Recall $\overline{q,p} = d(q, p)$. All properties for $\mathfrak{a}$ are straight forward. For the triangle inequality, as

$$\frac{\overline{y,q}^2 - \overline{y,p}^2 - \overline{z,q}^2 + \overline{z,p}^2}{\overline{q,p}} = \frac{\overline{y,q}^2 - \overline{y,p}^2 - \overline{v,q}^2 + \overline{v,p}^2}{\overline{q,p}} + \frac{\overline{v,q}^2 - \overline{v,p}^2 - \overline{z,q}^2 + \overline{z,p}^2}{\overline{q,p}}\,,$$

we obtain

$$\begin{aligned} &\sup_{q \neq p} \frac{\overline{y,q}^2 - \overline{y,p}^2 - \overline{z,q}^2 + \overline{z,p}^2}{\overline{q,p}} \\ &\leq \sup_{q \neq p} \frac{\overline{y,q}^2 - \overline{y,p}^2 - \overline{v,q}^2 + \overline{v,p}^2}{\overline{q,p}} + \sup_{q \neq p} \frac{\overline{v,q}^2 - \overline{v,p}^2 - \overline{z,q}^2 + \overline{z,p}^2}{\overline{q,p}}\,. \end{aligned}$$

For $\mathfrak{b}$ the argument is almost identical. □

Using the properties of the kernel, we bound the integrated squared error by a sum.

Lemma 5.32. Assume KERNEL and LIPSCHITZ. Then

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} d\Big(g(x,\theta), g(x,\tilde{\theta})\Big)^2 \,\mathrm{d}x \le 6C_{\mathsf{Kmi}}C_{\mathsf{Kma}}\left(\sum_{i=1}^{n} w_i d\Big(g_i(\theta), g_i(\tilde{\theta})\Big)^2 + \frac{2C_{\mathsf{Lip}}}{nh}\right)$$

for all $\theta, \tilde{\theta} \in \Theta, h \ge \frac{2}{n}$.

Proof. KERNEL implies

$$\frac{C_{\mathsf{Kmi}}^{-1}}{C_{\mathsf{Kma}}\#I_{t,h}} \mathbb{1}_{[-\frac{1}{2},\frac{1}{2}]}\left(\frac{x_i - t}{h}\right) \le w_i \,,$$

where $I_{t,h} := \{i \in \{1,\dots,n\} : t - h \le x_i \le t + h\}$. We bound the difference between the Riemann sum and its corresponding integral using LIPSCHITZ

$$\left|\frac{1}{\#I_{t,\frac{h}{2}}} \sum_{i \in I_{t,h}} d\Big(g\Big(\frac{x_i - t}{h}, \theta\Big), g\Big(\frac{x_i - t}{h}, \tilde{\theta}\Big)\Big)^2 - \int_{-\frac{1}{2}}^{\frac{1}{2}} d\Big(g(x,\theta), g(x,\tilde{\theta})\Big)^2 \,\mathrm{d}x\right| \le \frac{C_{\mathsf{Lip}}}{\#I_{t,\frac{h}{2}}} \,.$$

Thus,

$$\begin{aligned}
\sum_{i=1}^{n} w_i d\Big(g_i(\theta), g_i(\tilde{\theta})\Big)^2 &\ge \frac{C_{\mathsf{Kmi}}^{-1}\#I_{t,\frac{h}{2}}}{C_{\mathsf{Kma}}\#I_{t,h}} \frac{1}{\#I_{t,\frac{h}{2}}} \sum_{i \in I_{t,h}} d\Big(g\Big(\frac{x_i - t}{h}, \theta\Big), g\Big(\frac{x_i - t}{h}, \tilde{\theta}\Big)\Big)^2 \\
&\ge \frac{C_{\mathsf{Kmi}}^{-1}\#I_{t,\frac{h}{2}}}{C_{\mathsf{Kma}}\#I_{t,h}} \left(\int_{-\frac{1}{2}}^{\frac{1}{2}} d\Big(g(x,\theta), g(x,\tilde{\theta})\Big)^2 \,\mathrm{d}x - \frac{C_{\mathsf{Lip}}}{\#I_{t,\frac{h}{2}}}\right) .
\end{aligned}$$

As $h \ge \frac{2}{n}$, we obtain

$$\sum_{i=1}^{n} w_i d\Big(g_i(\theta), g_i(\tilde{\theta})\Big)^2 \ge \frac{C_{\mathsf{Kmi}}^{-1}}{6C_{\mathsf{Kma}}} \left(\int_{-\frac{1}{2}}^{\frac{1}{2}} d\Big(g(x,\theta), g(x,\tilde{\theta})\Big)^2 \,\mathrm{d}x - \frac{2C_{\mathsf{Lip}}}{nh}\right) . \qquad \square$$

The weights w_i have following properties, see [Tsy08, Proposition 1.13].

Lemma 5.33. Assume KERNEL and $h \ge \frac{2}{n}$. Then

$$\begin{aligned}
&w_i \ge 0 \,, \qquad \sum_{i=1}^{n} w_i = 1 \,, \qquad w_i \le \frac{6C_{\mathsf{Kmi}}C_{\mathsf{Kma}}}{nh} \,, \\
&w_i = 0 \text{ if } |x_i - t| > h \,, \qquad \sum_{i=1}^{n} w_i^2 \le \frac{6C_{\mathsf{Kmi}}C_{\mathsf{Kma}}}{nh}
\end{aligned}$$

for all $t \in [0,1]$ and $h \ge \frac{2}{n}$.

Define $U(\theta) := \sum_{i=1}^n w_i d(g_i(\theta), m_{x_i})^2$. We make use of Smoothness to obtain a bound on $\sum_{i=1}^n w_i d(g_i(\hat{\theta}_{t,h}), g_i(\theta_{t,h}))^2$.

Lemma 5.34. Assume Kernel, Lipschitz, Smoothness. Then,

$$\begin{aligned}&\int_{-\frac{1}{2}}^{\frac{1}{2}} d\Big(g(x,\hat{\theta}_{t,h}), g(x,\theta_{t,h})\Big)^2 \,\mathrm{d}x \\ &\leq 6C_{\mathsf{Kmi}}C_{\mathsf{Kma}}\left(U(\hat{\theta}_{t,h}) + 6C_{\mathsf{Kmi}}C_{\mathsf{Kma}}C_{\mathsf{Smo}}^2 h^{2\beta} + \frac{2C_{\mathsf{Lip}}}{nh}\right).\end{aligned}$$

Proof. Lemma 5.32 with Lipschitz states

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} d\Big(g(x,\hat{\theta}_{t,h}), g(x,\theta_{t,h})\Big)^2 \,\mathrm{d}x \leq 6C_{\mathsf{Kmi}}C_{\mathsf{Kma}}\left(\sum_{i=1}^n w_i d\Big(g_i(\hat{\theta}_{t,h}), g_i(\theta_{t,h})\Big)^2 + \frac{2C_{\mathsf{Lip}}}{nh}\right).$$

The remaining sum can be bounded using Smoothness and Kernel by

$$\begin{aligned}\sum_{i=1}^n w_i d\Big(g_i(\hat{\theta}_{t,h}), g_i(\theta_{t,h})\Big)^2 &\leq \sum_{i=1}^n w_i \left(d(g_i(\hat{\theta}_{t,h}), m_{x_i})^2 + d(m_{x_i}, g_i(\theta_{t,h}))^2\right) \\ &\leq \sum_{i=1}^n w_i d(g_i(\hat{\theta}_{t,h}), m_{x_i})^2 + C_{\mathsf{Ker}}C_{\mathsf{Smo}}^2 h^{2\beta}.\end{aligned}$$

Put together, we obtain

$$\begin{aligned}&\int_{-\frac{1}{2}}^{\frac{1}{2}} d\Big(g(x,\hat{\theta}_{t,h}), g(x,\theta_{t,h})\Big)^2 \,\mathrm{d}x \\ &\leq 6C_{\mathsf{Kmi}}C_{\mathsf{Kma}}\left(\sum_{i=1}^n w_i d(g_i(\hat{\theta}_{t,h}), m_{x_i})^2 + 6C_{\mathsf{Kmi}}C_{\mathsf{Kma}}C_{\mathsf{Smo}}^2 h^{2\beta} + \frac{2C_{\mathsf{Lip}}}{nh}\right).\end{aligned}$$ □

Next, we bound a variance term using chaining.

Lemma 5.35. Let $\theta_0 \in \mathcal{B} \subseteq \Theta$. Assume Moment and Kernel. Then,

$$\mathbb{E}\left[\sup_{\theta\in\mathcal{B}}\left|\bar{F}_t(\theta,\theta_0) - \hat{F}_t(\theta,\theta_0)\right|^\kappa\right] \leq c_\kappa\left(2\left(6C_{\mathsf{Kmi}}C_{\mathsf{Kma}}\right)^{\frac{1}{2}} C_{\mathsf{Mom}}C_{\mathsf{Ent}}\gamma_2(\mathcal{B},\mathfrak{b})(nh)^{-\frac{1}{2}}\right)^\kappa.$$

Proof. Define

$$Z_i(\theta) := w_i\left(d(y_i, g_i(\theta))^2 - d(y_i, g_i(\theta_0))^2 - \mathbb{E}\Big[d(y_i, g_i(\theta))^2 - d(y_i, g_i(\theta_0))^2\Big]\right).$$

We set y_i to be independent of Y_{x_i} and obtain

$$\begin{aligned}\mathbb{E}[|Z_i(\theta)|] &\le \mathbb{E}\Big[w_i \mathbb{E}\Big[\Big|d(y_i, g_i(\theta))^2 - d(y_i, g_i(\theta_0))^2 - d(Y_{x_i}, g_i(\theta))^2 - d(Y_{x_i}, g_i(\theta_0))^2\Big| \,\Big|\, y_i\Big]\Big] \\ &\le w_i d(g_i(\theta), g_i(\theta_0)) \mathbb{E}[\mathfrak{a}(y_i, Y_{x_i})]\,.\end{aligned}$$

As $\sup_{x\in\mathcal{X}} \mathbb{E}[\mathfrak{a}(Y_x, Y'_x)] \le \sigma_\kappa < \infty$, the processes Z_i are integrable. The stochastic processes $Z_1, \dots, Z_n$ with index set Θ are independent and integrable. Furthermore, $\mathbb{E}[Z_i(\theta)] = 0$ for all $\theta \in \Theta$, and $Z_i(\theta_0) = 0$. They fulfill the following quadruple property: Let Z'_i be independent copies of Z_i with y_i replaced by the independent copy y'_i. Then, for $\theta, \theta' \in \Theta$,

$$\left|Z_i(\theta) - Z_i(\theta') - Z'_i(\theta) + Z'_i(\theta')\right| \le w_i \mathfrak{a}(y_i, y'_i) d(g_i(\theta), g_i(\theta'))\,.$$

As $w_i = 0$ for $\left|\frac{x-t}{h}\right| > 1$, we have $w_i d(g_{t,h}(x,\theta), g_{t,h}(x,\theta')) \le w_i \mathfrak{b}(\theta,\theta')$. Thus, Theorem 5.56 implies

$$\mathbb{E}\left[\sup_{\theta\in\mathcal{B}} \left|\sum_{i=1}^n Z_i(\theta)\right|^\kappa\right] \le c_\kappa \gamma_2(\mathcal{B}, \mathfrak{b})^\kappa \mathbb{E}\left[\left(\sum_{i=1}^n w_i^2 \mathfrak{a}(y_i, y'_i)^2\right)^{\frac{\kappa}{2}}\right].$$

Define $W := \sum_{i=1}^n w_i^2$ and $v_i := w_i^2/W$. We obtain, using Jensen's inequality,

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{i=1}^n w_i^2 \mathfrak{a}(y_i, y'_i)^2\right)^{\frac{\kappa}{2}}\right] &= \mathbb{E}\left[\left(W \sum_{i=1}^n v_i \mathfrak{a}(y_i, y'_i)^2\right)^{\frac{\kappa}{2}}\right] \\ &\le W^{\frac{\kappa}{2}} \sum_{i=1}^n v_i \mathbb{E}\big[\mathfrak{a}(y_i, y'_i)^\kappa\big]\,.\end{aligned}$$

Lemma 5.31 shows that $\mathfrak{a}$ and $\mathfrak{b}$ are semi-metrics. Thus, we have

$$\mathbb{E}\big[\mathfrak{a}(y_i, y'_i)^\kappa\big] \le 2^\kappa \mathbb{E}[\mathfrak{a}(y_i, m_{x_i})^\kappa] \le 2^\kappa C_{\mathsf{Mom}}^\kappa$$

and, by Lemma 5.33, $W \le \frac{6 C_{\mathsf{Kmi}} C_{\mathsf{Kma}}}{nh}$, we obtain

$$\mathbb{E}\left[\sup_{\theta\in\mathcal{B}} \left|\bar{F}_t(\theta,\theta_0) - \hat{F}_t(\theta,\theta_0)\right|^\kappa\right] \le C_\kappa \left(2\,(6 C_{\mathsf{Kmi}} C_{\mathsf{Kma}})^{\frac{1}{2}}\, C_{\mathsf{Mom}} \gamma_2(\mathcal{B}, \mathfrak{b}) (nh)^{-\frac{1}{2}}\right)^\kappa . \qquad \square$$

The bias term can be bounded because of the smoothness assumption again.

Lemma 5.36. Assume Smoothness, R–Variance, and Kernel. Then

$$\left|\sum_{i=1}^n w_i \mathbb{E}[d(Y_{x_i}, g_i(\theta_{t,h}))^2 - d(Y_{x_i}, m_{x_i})^2]\right| \le C_{\mathsf{Vup}} C_{\mathsf{Smo}}^2 h^{2\beta}\,.$$

Proof. By R–Variance and Smoothness

$$\mathbb{E}\left[d\Big(Y_x, g\Big(\frac{x_i - t}{h}, \theta_{t,h}\Big)\Big)^2 - d(Y_x, m_x)^2\right] \leq C_{\mathsf{Vup}} d\Big(g\Big(\frac{x_i - t}{h}, \theta_{t,h}\Big), m_x\Big)^2$$
$$\leq C_{\mathsf{Vup}} C_{\mathsf{Smo}}^2 \left|x - t\right|^{2\beta} .$$

for all $x, t \in \mathbb{R}$. Hence, Kernel implies

$$\left|\sum_{i=1}^n w_i \mathbb{E}[d(Y_{x_i}, g_i(\theta_{t,h}))^2 - d(Y_{x_i}, m_{x_i})^2]\right| \leq C_{\mathsf{Vup}} C_{\mathsf{Smo}}^2 \sum_{i=1}^n w_i \left|x_i - t\right|^{2\beta}$$
$$\leq C_{\mathsf{Vup}} C_{\mathsf{Smo}}^2 h^{2\beta} . \qquad \square$$

A major step for obtaining a bound on the objects of interest instead of their objective function consists in using a peeling device (also called slicing). This technique is applied in the next 3 lemmata. Recall $U(\theta) = \sum_{i=1}^n w_i d(g_i(\theta), m_{x_i})^2$.

Lemma 5.37. Assume Variance, Smoothness, R–Variance, Moment, Kernel, Entropy, and Lipschitz. Then

$$\mathbb{E}[U(\hat{\theta}_{t,h})] \leq \frac{C_1}{nh} + C_2 h^{2\beta} ,$$

where

$$C_1 := c_\kappa C_{\mathsf{Lip}} C_{\mathsf{Vlo}}^2 C_{\mathsf{Mom}}^2 C_{\mathsf{Ent}}^2 C_{\mathsf{Kmi}}^2 C_{\mathsf{Kma}}^2$$
$$C_2 := c'_\kappa C_{\mathsf{Vlo}} C_{\mathsf{Vup}} C_{\mathsf{Lip}}^2 C_{\mathsf{Smo}}^2 .$$

Proof. Recall

$$\bar{F}_t(\theta) = \frac{1}{n} \sum_{i=1}^n w_i \mathbb{E}\Big[d(Y_{x_i}, g_i(\theta))^2\Big] .$$

Assume $U(\hat{\theta}_{t,h}) \in [a, b]$. Then by Variance

$$C_{\mathsf{Vlo}}^{-1} a \leq C_{\mathsf{Vlo}}^{-1} U(\hat{\theta}_{t,h})$$
$$\leq \sum_{i=1}^n w_i \mathbb{E}[d(Y_{x_i}, g_i(\hat{\theta}_{t,h}))^2 - d(Y_{x_i}, m_{x_i})^2]$$
$$\leq \bar{F}_t(\hat{\theta}_{t,h}, \theta_{t,h}) + \sum_{i=1}^n w_i \mathbb{E}[d(Y_{x_i}, g_i(\theta_{t,h}))^2 - d(Y_{x_i}, m_{x_i})^2] .$$

By Lemma 5.36

$$\sum_{i=1}^{n} w_i \mathbb{E}[d(Y_{x_i}, g_i(\theta_{t,h}))^2 - d(Y_{x_i}, m_{x_i})^2] \leq C_{\mathsf{Vup}} C_{\mathsf{Smo}}^2 h^{2\beta} .$$

For $b > 0$, let

$$\tilde{\mathcal{B}}_b := \left\{ \theta \in \Theta \colon \sum_{i=1}^{n} w_i d(g_i(\theta), m_{x_i})^2 \leq b \right\} .$$

By the minimizing property of $\hat{\theta}_{t,h}$,

$$\begin{aligned} \bar{F}_t(\hat{\theta}_{t,h}, \theta_{t,h}) &\leq \bar{F}_t(\hat{\theta}_{t,h}, \theta_{t,h}) - \hat{F}_t(\hat{\theta}_{t,h}, \theta_{t,h}) \\ &\leq \sup_{\theta \in \tilde{\mathcal{B}}_b} \left| \bar{F}_t(\theta, \theta_{t,h}) - \hat{F}_t(\theta, \theta_{t,h}) \right| . \end{aligned}$$

Thus,

$$C_{\mathsf{Vlo}}^{-1} a \leq C_{\mathsf{Vup}} C_{\mathsf{Smo}}^2 h^{2\beta} + \sup_{\theta \in \tilde{\mathcal{B}}_b} \left| \bar{F}_t(\theta, \theta_{t,h}) - \hat{F}_t(\theta, \theta_{t,h}) \right| .$$

Using Markov's inequality,

$$\begin{aligned} \mathbb{P}\Big(U(\hat{\theta}_{t,h}) \in [a,b]\Big) &\leq \mathbb{P}\left(C_{\mathsf{Vup}} C_{\mathsf{Smo}}^2 C_{\mathsf{Vlo}} h^{2\beta} + C_{\mathsf{Vlo}} \sup_{\theta \in \tilde{\mathcal{B}}_b} \left| \bar{F}_t(\theta, \theta_{t,h}) - \hat{F}_t(\theta, \theta_{t,h}) \right| \geq a \right) \\ &\leq 2^{\kappa-1} C_{\mathsf{Vlo}}^{\kappa} \frac{C_{\mathsf{Vup}}^{\kappa} C_{\mathsf{Smo}}^{2\kappa} h^{2\beta\kappa} + \mathbb{E}\Big[\sup_{\theta \in \tilde{\mathcal{B}}_b} \left| \bar{F}_t(\theta, \theta_{t,h}) - \hat{F}_t(\theta, \theta_{t,h}) \right|^{\kappa} \Big]}{a^{\kappa}} . \end{aligned}$$

By Lemma 5.35, with $\theta_0 = \theta_{t,h}$, with Lemma 5.39 below and Entropy

$$\begin{aligned} &\mathbb{E}\left[\sup_{\theta \in \tilde{\mathcal{B}}_b} \left| \bar{F}_t(\theta, \theta_{t,h}) - \hat{F}_t(\theta, \theta_{t,h}) \right|^{\kappa} \right] \\ &\leq c_{\kappa} \left(2 \left(6 C_{\mathsf{Kmi}} C_{\mathsf{Kma}} \right)^{\frac{1}{2}} C_{\mathsf{Mom}} C_{\mathsf{Ent}} \left(12 C_{\mathsf{Kmi}} C_{\mathsf{Kma}} b \right)^{\frac{1}{2}} (nh)^{-\frac{1}{2}} \right)^{\kappa} , \end{aligned}$$

for $b \geq 5 C_{\mathsf{Lip}}^2 h^{2\beta} + \frac{2 C_{\mathsf{Lip}}}{nh}$. Thus,

$$\mathbb{P}\Big(U(\hat{\theta}_{t,h}) \in [a,b]\Big) \leq 2^{\kappa-1} \frac{c_1^{\kappa} h^{2\beta\kappa} + \left(c_2 b^{\frac{1}{2}} (nh)^{-\frac{1}{2}} \right)^{\kappa}}{a^{\kappa}} ,$$

where

$$c_1 := C_{\mathsf{Vlo}} C_{\mathsf{Vup}} C_{\mathsf{Smo}}^2 \,,$$
$$c_2 := 18 c_\kappa^{\frac{1}{\kappa}} C_{\mathsf{Vlo}} C_{\mathsf{Mom}} C_{\mathsf{Ent}} C_{\mathsf{Kmi}} C_{\mathsf{Kma}} \,.$$

Thus, by Lemma 5.38 below

$$\mathbb{E}[U(\hat{\theta}_{t,h})] \leq 5 C_{\mathsf{Lip}}^2 h^{2\beta} + \frac{2 C_{\mathsf{Lip}}}{nh} + c'_\kappa \left(c_1 h^{2\beta} + \frac{c_2^2}{nh} \right) .$$

As all constants are chosen to be in $[1, \infty)$, we obtain the desired result. □

Lemma 5.38. Let V be a nonnegative random variable. Assume that for $0 \leq a_0 < a < b < \infty$, it holds

$$\mathbb{P}(V \in [a, b]) \leq c \frac{u^\kappa + \left(v b^{\frac{1}{2}} \right)^\kappa}{a^\kappa} \,.$$

where $c \geq 1, u, v > 0$, $\kappa > 2$. Then

$$\mathbb{E}[V] \leq a_0 + c_\kappa c^{\frac{2}{\kappa}} \left(u + v^2 \right) .$$

Proof. For $s > a_0$,

$$\begin{aligned}
\mathbb{P}(V > s) &\leq \sum_{k=0}^{\infty} \mathbb{P}\Big(V \in [s 2^k, s 2^{k+1}] \Big) \\
&\leq \sum_{k=0}^{\infty} c \frac{u^\kappa + v^\kappa s^{\frac{1}{2}\kappa} 2^{\frac{1}{2}\kappa} 2^{\frac{1}{2}k\kappa}}{s^\kappa 2^{k\kappa}} \\
&\leq c \left(u^\kappa s^{-\kappa} \sum_{k=0}^{\infty} 2^{-k\kappa} + 2^{\frac{1}{2}\kappa} v^\kappa s^{-\frac{1}{2}\kappa} \sum_{k=0}^{\infty} 2^{-\frac{1}{2}k\kappa} \right) \\
&\leq c'_\kappa c \left(u^\kappa s^{-\kappa} + v^\kappa s^{-\frac{1}{2}\kappa} \right) .
\end{aligned}$$

We integrate the tail to bound the expectation,

$$\mathbb{E}[V] \leq a_0 + \int_{a_0}^{\infty} \mathbb{P}(V > s) \mathrm{d}s \,.$$

For $A, B \geq 0$, it holds

$$\int_0^\infty \min(1, A s^{-\kappa}) \leq \frac{\kappa}{1-\kappa} A^{\frac{1}{\kappa}} \,, \qquad \int_0^\infty \min(1, B s^{-\frac{1}{2}\kappa}) \leq \frac{\kappa}{2-\kappa} B^{\frac{2}{\kappa}} \,.$$

Applying these inequalities to the tail bound above, we obtain

$$\mathbb{E}[V] \le a_0 + c_\kappa c^{\frac{2}{\kappa}} \left(u + v^2\right) . \qquad \square$$

Lemma 5.39. For $b > 0$, let

$$\mathcal{B}_b := \left\{ \theta \in \Theta \colon \int_{-\frac{1}{2}}^{\frac{1}{2}} d(g(x,\theta), g(x,\theta_{t,h}))^2 \, \mathrm{d}x \le b \right\}$$
$$\bar{\mathcal{B}}_b := \left\{ \theta \in \Theta \colon \sum_{i=1}^n w_i d(g_i(\theta), g_i(\theta_{t,h}))^2 \le b \right\}$$
$$\tilde{\mathcal{B}}_b := \left\{ \theta \in \Theta \colon \sum_{i=1}^n w_i d(g_i(\theta), m_{x_i})^2 \le b \right\}$$

Assume Smoothness, Kernel, and Lipschitz. Then, for all $b, s > 0$,

$$\tilde{\mathcal{B}}_b \subseteq \bar{\mathcal{B}}_{b+r} \qquad \text{and} \qquad \bar{\mathcal{B}}_s \subseteq \mathcal{B}_{s'}$$

where

$$r = 2C_{\mathsf{Lip}} h^\beta b^{\frac{1}{2}} + C_{\mathsf{Lip}}^2 h^{2\beta} \qquad\qquad s' = 6C_{\mathsf{Kmi}} C_{\mathsf{Kma}} \left(s + \frac{2C_{\mathsf{Lip}}}{nh} \right) .$$

Proof. For $\theta \in \Theta$, we obtain using the triangle inequality

$$\begin{aligned}
&d(g_i(\theta), m_{x_i})^2 - d(g_i(\theta), g_i(\theta_{t,h}))^2 \\
&\le d(m_{x_i}, g_i(\theta_{t,h})) \left(2d(g_i(\theta), m_{x_i}) + d(m_{x_i}, g_i(\theta_{t,h}))\right) \\
&\le 2C_{\mathsf{Smo}} \left|x_i - t\right|^\beta d(m_{x_i}, g_i(\theta)) + C_{\mathsf{Smo}}^2 \left|x_i - t\right|^{2\beta}
\end{aligned}$$

because of Smoothness. Thus, Kernel implies

$$\begin{aligned}
&\left| \sum_{i=1}^n w_i \left(d(g_i(\theta), m_{x_i})^2 - d(g_i(\theta), g_i(\theta_{t,h}))^2 \right) \right| \\
&\le 2C_{\mathsf{Smo}} \sum_{i=1}^n w_i \left|x_i - t\right|^\beta d(m_{x_i}, g_i(\theta)) + C_{\mathsf{Smo}}^2 \sum_{i=1}^n w_i \left|x_i - t\right|^{2\beta} \\
&\le 2C_{\mathsf{Smo}} h^\beta \sum_{i=1}^n w_i d(m_{x_i}, g_i(\theta)) + C_{\mathsf{Smo}}^2 h^{2\beta} .
\end{aligned}$$

Now assume $\theta \in \tilde{\mathcal{B}}_b$. Then $\sum_{i=1}^n w_i d(g_i(\theta), m_{x_i})^2 \le b$. We obtain, via Jensen's inequality,

$$\sum_{i=1}^n w_i d(m_{x_i}, g_i(\theta_{t,h})) \le \left(\sum_{i=1}^n w_i d(m_{x_i}, g_i(\theta_{t,h}))^2 \right)^{\frac{1}{2}} \le b^{\frac{1}{2}} .$$

Together, we get

$$\left|\sum_{i=1}^{n} w_i \left(d(g_i(\theta), m_{x_i})^2 - d(g_i(\theta), g_i(\theta_{t,h}))^2\right)\right| \leq 2C_{\mathsf{Smo}} h^{\beta} b^{\frac{1}{2}} + C_{\mathsf{Smo}}^2 h^{2\beta} =: r\,.$$

Thus,

$$\sum_{i=1}^{n} w_i d(g_i(\theta), g_i(\theta_{t,h}))^2 \leq \sum_{i=1}^{n} w_i d(g_i(\theta), m_{x_i})^2 + r \leq b + r$$

which shows $\tilde{\mathcal{B}}_b \subseteq \bar{\mathcal{B}}_{b+r}$. The relation $\bar{\mathcal{B}}_s \subseteq \mathcal{B}_{s'}$ follows from Lemma 5.32 by

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} d\Big(g(x,\theta), g(x,\tilde{\theta})\Big)^2 \,\mathrm{d}x \leq 6C_{\mathsf{Kmi}} C_{\mathsf{Kma}} \left(\sum_{i=1}^{n} w_i d\Big(g_i(\theta), g_i(\tilde{\theta})\Big)^2 + \frac{2C_{\mathsf{Lip}}}{nh}\right). \qquad \square$$

Finally, we can put together the results obtained so far to finish the proof of the main theorem.

Proof of Theorem 5.12. By Lemma 5.34,

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} d\Big(g(x,\hat{\theta}_{t,h}), g(x,\theta_{t,h})\Big)^2 \,\mathrm{d}x \leq 6C_{\mathsf{Kmi}} C_{\mathsf{Kma}} \left(U(\hat{\theta}_{t,h}) + 6C_{\mathsf{Kmi}} C_{\mathsf{Kma}} C_{\mathsf{Smo}}^2 h^{2\beta} + \frac{2C_{\mathsf{Lip}}}{nh}\right).$$

By Lemma 5.37,

$$\mathbb{E}[U(\hat{\theta}_{t,h})] \leq \frac{C_1'}{nh} + C_2' h^{2\beta}\,,$$

where

$$\begin{aligned} C_1' &:= c_\kappa C_{\mathsf{Lip}} C_{\mathsf{Vlo}}^2 C_{\mathsf{Mom}}^2 C_{\mathsf{Ent}}^2 C_{\mathsf{Kmi}}^2 C_{\mathsf{Kma}}^2\,, \\ C_2' &:= c_\kappa' C_{\mathsf{Vlo}} C_{\mathsf{Vup}} C_{\mathsf{Lip}}^2 C_{\mathsf{Smo}}^2\,. \end{aligned}$$

Thus, we obtain

$$\mathbb{E}\left[\int_{-\frac{1}{2}}^{\frac{1}{2}} d\Big(g(x,\hat{\theta}_{t,h}), g(x,\theta_{t,h})\Big)^2\right]\mathrm{d}x \leq \frac{C_1}{nh} + C_2 h^{2\beta}\,,$$

where

$$\begin{aligned} C_1 &:= 6C_{\mathsf{Kmi}} C_{\mathsf{Kma}} \left(c_\kappa C_{\mathsf{Lip}} C_{\mathsf{Vlo}}^2 C_{\mathsf{Mom}}^2 C_{\mathsf{Ent}}^2 C_{\mathsf{Kmi}}^2 C_{\mathsf{Kma}}^2 + 2C_{\mathsf{Lip}}\right), \\ C_2 &:= 6C_{\mathsf{Kmi}} C_{\mathsf{Kma}} \left(c_\kappa' C_{\mathsf{Vlo}} C_{\mathsf{Vup}} C_{\mathsf{Lip}}^2 C_{\mathsf{Smo}}^2 + 6C_{\mathsf{Kmi}} C_{\mathsf{Kma}} C_{\mathsf{Smo}}^2\right). \end{aligned} \qquad \square$$

5.A.3.2 Corollaries

Corollary 5.14 and Corollary 5.15 are direct implications of Theorem 5.12. We want to prove Corollary 5.11. It is a consequence of Corollary 5.14 with $\mathcal{Q} = \mathbb{S}^k$ and $g(x,(q,v)) = \mathsf{Exp}(q,xv)$ for $(q,v) \in \Theta \subseteq \mathsf{T}\mathbb{S}^k$. To apply this corollary, we need to show Entropy,

Lipschitz, and Connection for the sphere, as Variance, Smoothness and Kernel are assumed.

- Connection: As $(q,v),(p,u) \in \Theta$, it holds $|u|,|v| \leq \pi$. Lemma 5.29 implies

$$\begin{aligned}\int_{-\frac{1}{2}}^{\frac{1}{2}} d(\mathsf{Exp}(q,xu),\mathsf{Exp}(p,xv))^2 \,\mathrm{d}x &= \frac{1}{2}\int_{-1}^{1} d\Big(\mathsf{Exp}\Big(q,\frac{1}{2}xu\Big),\mathsf{Exp}\Big(p,\frac{1}{2}xv\Big)\Big)^2 \,\mathrm{d}x \\ &\geq \frac{1}{\pi}\|p-q\|^2 \\ &\geq \frac{1}{\pi^2} d_{\mathbb{S}^k}(p,q)^2\,.\end{aligned}$$

  Thus, we can choose $C_{\mathsf{Con}} := \pi^2$.

- Lipschitz: Let $\gamma_1(x) = \mathsf{Exp}(q,xv)$ and $\gamma_2(x) = \mathsf{Exp}(p,xu)$ be two geodesics. The squared distance $d(\gamma_1(x),\gamma_2(x))^2$ can be bounded by π-times the Euclidean distance. Furthermore, it changes not more than the distance of straight lines in $\mathbb{R}^{k+1}$ moving in opposite directions. Without loss of generality $d(\gamma_1(x),\gamma_2(x)) \leq d(\gamma_1(y),\gamma_2(y))$. Then

$$\begin{aligned}&\Big|d(\gamma_1(x),\gamma_2(x))^2 - d(\gamma_1(y),\gamma_2(y))^2\Big| \\ &\leq \left(d(\gamma_1(x),\gamma_2(x)) + \pi\,|x-y|\,(|u|+|v|)\right)^2 - d(\gamma_1(x),\gamma_2(x))^2 \\ &\leq |x-y|\left(\pi^2\,|x-y|\,(|u|+|v|)^2 + \pi d(\gamma_1(x),\gamma_2(x))\,(|u|+|v|)\right) \\ &\leq C_{\mathsf{Lip}}\,|x-y|\end{aligned}$$

  with $C_{\mathsf{Lip}} := 8\pi^4 + 4\pi$, where we used $|u|,|v| \leq \pi$ for $(q,v),(p,u) \in \Theta$.

- Entropy: Lemma 5.29 implies

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} d(\mathsf{Exp}(q,xu),\mathsf{Exp}(p,xv))^2 \mathrm{d}x \geq \frac{1}{\pi}\left(|p-q|^2 + |u-v|^2\right).$$

  Thus, $\mathcal{B}_b(\theta_0) \subseteq \left\{x \in \mathbb{R}^{2k+2} \colon |x| \leq \sqrt{\pi b}\right\}$. By Lemma 5.28, it holds $\mathfrak{b}((q,v),(p,u)) \leq 2\pi\,(|q-p| + |v-u|)$, yielding

$$\gamma_2(\mathcal{B}_b,\mathfrak{b}) \leq c\gamma_2(B_{\sqrt{\pi b}},|\cdot|) \leq c'\sqrt{kb}\,.$$

  Thus, we can choose $C_{\mathsf{Ent}} := c'\sqrt{kb}$.

5.A.4 Section 5.5: `LocFre`

First we state some properties of the weights w_i to be used later.

Lemma 5.40 ([Tsy08, Lemma 1.3])**.** Assume KERNEL. Then there is $C_{\mathsf{Ker}} \in [1, \infty)$ such that

$$\sum_{i=1}^{n} w_i = 1\,, \qquad w_i = 0 \text{ if } |x_i - t| > h\,, \qquad w_i \le \frac{C_{\mathsf{Ker}}}{nh}\,,$$
$$\sum_{i=1}^{n} |w_i| \le C_{\mathsf{Ker}}\,, \qquad \sum_{i=1}^{n} w_i^2 \le \frac{C_{\mathsf{Ker}}}{nh}\,.$$

for all $t \in [0,1]$, $n \ge n_0$.

5.A.4.1 Theorem

We prove Theorem 5.17. We first apply the variance inequality to relate a bound on the objective functions to a bound on the minimizers. The required uniform bound on the objective functions can be split into a bias and a variance part, which are bounded separately thereafter. Then, these results are put together in the application of a peeling device, which is used to bound the tail probabilities of the error. Integrating the tails leads to the required bounds in expectation.

Variance Inequality and Split. We define following notation for the objective functions

$$\hat{F}_t(q) := \sum_{i=1}^{n} w_i \mathfrak{c}(y_i, q) \qquad \hat{F}_t(q,p) := \hat{F}_t(q) - \hat{F}_t(p)\,,$$
$$\bar{F}_t(q) := \sum_{i=1}^{n} w_i \mathbb{E}[\mathfrak{c}(y_i, q)] \qquad \bar{F}_t(q,p) := \bar{F}_t(q) - \bar{F}_t(p)\,,$$
$$F_t(q) := \mathbb{E}[\mathfrak{c}(Y_t, q)] \qquad F_t(q,p) := F_t(q) - F_t(p)\,.$$

Using VARIANCE and the minimizing property of $\hat{m}_t$ we obtain

$$\begin{aligned} C_{\mathsf{Vlo}}^{-1} d(\hat{m}_t, m_t)^\alpha &\le F_t(\hat{m}_t, m_t) \\ &\le F_t(\hat{m}_t, m_t) - \hat{F}_t(\hat{m}_t, m_t) \\ &= \Big(F_t(\hat{m}_t, m_t) - \bar{F}_t(\hat{m}_t, m_t)\Big) + \Big(\bar{F}_t(\hat{m}_t, m_t) - \hat{F}_t(\hat{m}_t, m_t)\Big) \end{aligned}$$

The first parenthesis represents the bias part, the second one the variance part. We will bound the former using SMOOTHNESS, the later by an empirical process argument.

Variance. Define

$$Z_i(q) := w_i \left(\mathfrak{c}(y_i, q) - \mathfrak{c}(y_i, m_t)\right) - \mathbb{E}[w_i \left(\mathfrak{c}(y_i, q) - \mathfrak{c}(y_i, m_t)\right)]\,.$$

Then $Z_1, \dots, Z_n$ are independent and centered processes with $Z_i(m_t) = 0$. They are integrable due to MOMENT. By the definition of $\mathfrak{a}$,

$$|Z_i(q) - Z_i(p) - Z_i'(q) + Z_i'(p)| \le |w_i|\, \mathfrak{a}(y_i, y_i') d(q,p)\,,$$

where $Z_i(q)'$ and y_i' are independent copies of $Z_i(q)$ and y_i, respectively. Theorem 5.56 yields

$$\mathbb{E}\left[\sup_{q\in \mathrm{B}(m_t,d,\delta)}\left|\bar{F}_t(q,m_t)-\hat{F}_t(q,m_t)\right|^\kappa\right]=\mathbb{E}\left[\sup_{q\in \mathrm{B}(m_t,d,\delta)}\left|\sum_{i=1}^n Z_i(q)\right|^\kappa\right]$$
$$\leq c_\kappa\left(\mathbb{E}\left[\left(\sum_{i=1}^n w_i^2\mathfrak{a}(y_i,y_i')^2\right)^{\frac{\kappa}{2}}\right]^{\frac{1}{\kappa}}\gamma_2(\mathrm{B}(m_t,d,\delta),d)\right)^\kappa$$

for a constant c_κ depending only on κ. Define $W:=\sum_{i=1}^n w_i^2$ and $v_i:=w_i^2/W$. We apply Moment,

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{i=1}^n w_i^2\mathfrak{a}(y_i,y_i')^2\right)^{\frac{\kappa}{2}}\right]&=\mathbb{E}\left[\left(W\sum_{i=1}^n v_i\mathfrak{a}(y_i,y_i')^2\right)^{\frac{\kappa}{2}}\right]\\&\leq\mathbb{E}\left[W^{\frac{\kappa}{2}}\sum_{i=1}^n v_i\mathfrak{a}(y_i,y_i')^\kappa\right]\\&=W^{\frac{\kappa}{2}}\sum_{i=1}^n v_i\mathbb{E}[\mathfrak{a}(y_i,y_i')^\kappa]\\&\leq W^{\frac{\kappa}{2}}C_{\mathsf{Mom}}^\kappa\,.\end{aligned}$$

By Lemma 5.40, $W\leq C_{\mathsf{Ker}}(nh)^{-1}$. By Entropy, $\gamma_2(\mathrm{B}(m_t,d,\delta),d)\leq C_{\mathsf{Ent}}\delta$. Thus,

$$\mathbb{E}\left[\sup_{q\in \mathrm{B}(m_t,d,\delta)}\left|\bar{F}_t(q,m_t)-\hat{F}_t(q,m_t)\right|^\kappa\right]\leq c_\kappa\left(C_{\mathsf{Mom}}C_{\mathsf{Ent}}C_{\mathsf{Ker}}\delta(nh)^{-\frac{1}{2}}\right)^\kappa\,.$$

Bias. As $\sum_{i=1}^n w_i=1$ (Lemma 5.40), we have

$$F_t(q,m_t)-\bar{F}_t(q,m_t)=\sum_{i=1}^n w_i\mathbb{E}[\Diamond(Y_t,y_i,q,m_t)]\,.$$

Set $f(t):=\mathbb{E}[\mathfrak{c}(Y_t,q)-\mathfrak{c}(Y_t,p)]$. Applying Smoothness, a Taylor expansion, and the property that the weights annihilate polynomials [Tsy08, equation (1.68)], we obtain

$$\begin{aligned}\sum_{i=1}^n w_i\mathbb{E}[\Diamond(Y_t,y_i,q,p)]&=\sum_{i=1}^n w_i\left(R_i+\sum_{k=1}^\ell\frac{f'(t)}{k!}(x_i-t)^k\right)\\&=\sum_{i=1}^n w_iR_i\\&\leq\sum_{i=1}^n|w_i|\,|R_i|\,,\end{aligned}$$

for values $R_i \in \mathbb{R}$ with $|R_i| \leq d(q,p)L(q,p)\,|t - x_i|^\beta$. Thus,

$$\sum_{i=1}^{n} w_i \mathbb{E}[\Diamond(Y_t, y_i, q, m_t)] \leq C_{\mathsf{Ker}} d(q,p) L(q,p) h^\beta \,, \tag{5.2}$$

see Lemma 5.40. Finally we obtain

$$\begin{aligned}
&\mathbb{E}\Big[\Big|F_t(\hat{m}_t, m_t) - \bar{F}_t(\hat{m}_t, m_t)\Big|^\kappa \mathbb{1}_{[0,\delta]}(d(\hat{m}_t, m_t))\Big]^{\frac{1}{\kappa}} \\
&\leq \mathbb{E}\Big[\Big|C_{\mathsf{Ker}} d(\hat{m}_t, m_t) L(\hat{m}_t, m_t) h^\beta\Big|^\kappa \mathbb{1}_{[0,\delta]}(d(\hat{m}_t, m_t))\Big]^{\frac{1}{\kappa}} \\
&\leq C_{\mathsf{Ker}} C_{\mathsf{Smo}} \delta h^\beta \,.
\end{aligned}$$

Peeling. For $\delta > 0$ define

$$\Delta_\delta(q,p) = \Big(\Big|F_t(q,p) - \bar{F}_t(q,p)\Big| + \Big|\bar{F}_t(q,p) - \hat{F}_t(q,p)\Big|\Big)\, \mathbb{1}_{[0,\delta]}(d(q,p)) \,.$$

Recall that the variance inequality implies

$$C_{\mathsf{Vlo}}^{-1} d(\hat{m}_t, m_t)^\alpha \leq \Big(F_t(\hat{m}_t, m_t) - \bar{F}_t(\hat{m}_t, m_t)\Big) + \Big(\bar{F}_t(\hat{m}_t, m_t) - \hat{F}_t(\hat{m}_t, m_t)\Big) \,.$$

Let $0 < a < b < \infty$. The inequality above and Markov's inequality yield

$$\mathbb{P}(d(\hat{m}_t, m_t) \in [a,b]) \leq \mathbb{P}(a^\alpha \leq C_{\mathsf{Vlo}} \Delta_b(\hat{m}_t, m_t)) \leq \frac{C_{\mathsf{Vlo}}^\kappa \mathbb{E}[\Delta_b(\hat{m}_t, m_t)^\kappa]}{a^{\alpha\kappa}} \,.$$

Our previous consideration allow us the bound the expectation by a variance and a bias term:

$$\begin{aligned}
\mathbb{E}[\Delta_\delta(\hat{m}_t, m_t)^\kappa] \leq 2^{\kappa-1} \Bigg(&\mathbb{E}\Big[\Big|F_t(\hat{m}_t, m_t) - \bar{F}_t(\hat{m}_t, m_t)\Big|^\kappa \mathbb{1}_{[0,\delta]}(d(\hat{m}_t, m_t))\Big] \\
&+ \mathbb{E}\left[\sup_{q \in \mathrm{B}(m_t, d, \delta)} \Big|\bar{F}_t(q, m_t) - \hat{F}_t(q, m_t)\Big|^\kappa\right]\Bigg) \\
&\leq c_\kappa \Big(C_{\mathsf{Ker}} C_{\mathsf{Smo}} h^\beta + C_{\mathsf{Mom}} C_{\mathsf{Ent}} C_{\mathsf{Ker}} (nh)^{-\frac{1}{2}}\Big)^\kappa \delta^\kappa \,.
\end{aligned}$$

We are now prepared to apply peeling (also called slicing): Let $s > 0$. Set $A := C_{\mathsf{Vlo}} C_{\mathsf{Ker}} C_{\mathsf{Smo}} h^\beta + C_{\mathsf{Vlo}} C_{\mathsf{Mom}} C_{\mathsf{Ent}} C_{\mathsf{Ker}} (nh)^{-\frac{1}{2}}$. It holds

$$\begin{aligned}
\mathbb{P}(d(\hat{m}_t, m_t) > s) &\leq \sum_{k=0}^{\infty} \mathbb{P}\Big(d(\hat{m}_t, m_t) \in [2^k s, 2^{k+1} s]\Big) \\
&\leq \sum_{k=0}^{\infty} \frac{c_\kappa A^\kappa (2^{k+1} s)^\kappa}{(2^k s)^{\alpha\kappa}} \\
&\leq 2^\kappa c_\kappa A^\kappa s^{\kappa(1-\alpha)} \sum_{k=0}^{\infty} 2^{k\kappa(1-\alpha)} \\
&\leq \frac{2^\kappa}{1 - 2^{\kappa(1-\alpha)}} c_\kappa A^\kappa s^{\kappa(1-\alpha)} \,.
\end{aligned}$$

We integrate this tail bound to bound the expectation. For this we require $\kappa > \alpha/(\alpha-1)$. Set $B := \frac{2^\kappa}{1-2^{\kappa(1-\alpha)}} c_\kappa A^\kappa$, then

$$\begin{aligned}
\mathbb{E}[d(\hat{m}_t, m_t)^\alpha] &= \alpha \int_0^\infty s^{\alpha-1} \mathbb{P}(d(\hat{m}_t, m_t) > s) \mathrm{d}s \\
&\leq \alpha \int_0^\infty s^{\alpha-1} \min(1, B s^{\kappa(1-\alpha)}) \mathrm{d}s \\
&= \frac{1}{\alpha} b^{\frac{\alpha}{\kappa(\alpha-1)}} + \frac{1}{\kappa(\alpha-1)-\alpha} b^{\frac{\alpha-\kappa(\alpha-1)}{\kappa(\alpha-1)}} \\
&= \left(\frac{1}{\alpha} + \frac{1}{\kappa(\alpha-1)-\alpha} \right) b^{\frac{\alpha}{\kappa(\alpha-1)}} \\
&= c_{\kappa,\alpha} A^{\frac{\alpha}{\alpha-1}} .
\end{aligned}$$

Thus,

$$\mathbb{E}[d(\hat{m}_t, m_t)^\alpha] \leq c_{\kappa,\alpha} \left(C_{\mathsf{Vlo}} C_{\mathsf{Ker}} C_{\mathsf{Smo}} h^\beta + C_{\mathsf{Vlo}} C_{\mathsf{Mom}} C_{\mathsf{Ent}} C_{\mathsf{Ker}} (nh)^{-\frac{1}{2}} \right)^{\frac{\alpha}{\alpha-1}} .$$

5.A.4.2 Corollaries

Lemma 5.41. In Theorem 5.17 SMOOTHNESS can be replaced by SMOOTHDENSITY and BIASMOMENT when we replace C_{Smo} by $C_{\mathsf{Bom}} C_{\mathsf{SmD}}$.

Proof. Using the μ-density $y \mapsto \rho(y|t)$ of Y_t, we can write $\mathbb{E}[\overline{Y_t,q}^2 - \overline{Y_t,p}^2] = \int (\overline{y,q}^2 - \overline{y,p}^2) \rho(y|t) \mathrm{d}\mu(y)$. By SMOOTHDENSITY, $t \mapsto \rho(y|t) \in \Sigma(\beta, L(y))$. Thus, there are $a_k(y)$ such that $\rho(y|x) = R_y(x, x_0) + \sum_{k=0}^{\ell} a_k(y)(x - x_0)^k$ with $|R_y(x, x_0)| \leq L(y) |x - x_0|^\beta$. Using that the weights annihilate polynomials of order ℓ, we obtain

$$\begin{aligned}
\sum_{i=1}^n w_i \mathbb{E}[\Diamond(Y_t, y_i, q, p)] &= \int \sum_{i=1}^n w_i \left(\overline{y,q}^2 - \overline{y,p}^2 \right) (\rho(y|t) - \rho(y|x_i)) \, \mathrm{d}\mu(y) \\
&= \int \sum_{i=1}^n w_i \left(\overline{y,q}^2 - \overline{y,p}^2 \right) R_y(t, x_i) \mathrm{d}\mu(y) \\
&\leq \int \sum_{i=1}^n |w_i| \left| \overline{y,q}^2 - \overline{y,p}^2 \right| |R_y(t, x_i)| \, \mathrm{d}\mu(y) .
\end{aligned}$$

It holds

$$\left| \overline{y,q}^2 - \overline{y,p}^2 \right| |R_y(x, x_0)| \leq \overline{q,p} \, |x - x_0|^\beta \, (\overline{y,q} + \overline{y,p}) \, L(y) .$$

Together with $\sum_{i=1}^n |w_i| \leq C_{\mathsf{Ker}}$ from Lemma 5.40, we obtain

$$\left| \sum_{i=1}^n w_i \mathbb{E}[\Diamond(Y_t, y_i, q, p)] \right| \leq C_{\mathsf{Ker}} \, \overline{q,p} \, h^\beta \int (\overline{y,q} + \overline{y,p}) \, L(y) \mathrm{d}\mu(y) .$$

This replaces equation (5.2) in the proof of Theorem 5.17 with

$$L(q,p) = \int (\overline{y,q} + \overline{y,p})\, L(y)\mathrm{d}\mu(y)\,.$$

To make the replacement valid we have to ensure $\mathbb{E}[L(m_t,\hat{m}_t)^\kappa]^{\frac{1}{\kappa}} \le C_{\mathsf{Smo}}$. By Cauchy–Schwarz inequality,

$$\int (\overline{y,q} + \overline{y,p})\, L(y)\mathrm{d}\mu(y) \le H(q,p)\left(\int L(y)^2\mathrm{d}\mu(y)\right)^{\frac{1}{2}} \le H(q,p)C_{\mathsf{SmD}}\,.$$

BiasMoment states $\mathbb{E}[H(\hat{m}_t,m_t)^\kappa]^{\frac{1}{\kappa}} \le C_{\mathsf{Bom}}$. Thus, set $C_{\mathsf{Smo}} := C_{\mathsf{Bom}}C_{\mathsf{SmD}}$. □

Recall $H(q,p) = \left(\int (\overline{y,q} + \overline{y,p})^2\,\mu(\mathrm{d}y)\right)^{\frac{1}{2}}$.

Proposition 5.42. Assume BomBound, Variance, Kernel, Moment. To fulfill $\mathbb{E}[H(\hat{m}_t,m_t)^\kappa]^{\frac{1}{\kappa}} \le C_{\mathsf{Bom}}$ in BiasMoment, we can choose

$$C_{\mathsf{Bom}} := c_\kappa C_{\mathsf{Vlo}} C_{\mathsf{Len}} C_{\mathsf{Mom}} C_{\mathsf{Int}} C_{\mathsf{Ker}}\,.$$

Proof of Proposition 5.42. Using the triangle inequality

$$\begin{aligned} H(q,p)^2 &= \int (\overline{y,q} + \overline{y,p})^2\,\mu(\mathrm{d}y) \\ &\le \int (\overline{q,p} + 2\overline{y,p})^2\,\mu(\mathrm{d}y) \\ &\le 2\int \overline{q,p}^2 + 4\overline{y,p}^2\mu(\mathrm{d}y) \\ &\le 2\overline{q,p}^2 + 8\int \overline{y,p}^2\mu(\mathrm{d}y) \end{aligned}$$

as μ is a probability measure.

$$\begin{aligned} \mathbb{E}[H(\hat{m}_t,m_t)^\kappa]^{\frac{1}{\kappa}} &\le \mathbb{E}\left[\left(2\overline{\hat{m}_t,m_t}^2 + 8\int \overline{y,m_t}^2\mu(\mathrm{d}y)\right)^{\frac{\kappa}{2}}\right]^{\frac{1}{\kappa}} \\ &\le c_\kappa\left(\mathbb{E}\left[\overline{\hat{m}_t,m_t}^\kappa\right]^{\frac{1}{\kappa}} + \left(\int \overline{y,m_t}^2\mu(\mathrm{d}y)\right)^{\frac{1}{2}}\right)\,. \end{aligned}$$

Next, we will bound $\mathbb{E}[\overline{m_t,\hat{m}_t}^\kappa]$. Let $W := \sum_{i=1}^n |w_i|$. First, by Variance and the

minimizing property of $\hat{m}_t$,

$$\begin{aligned} C_{\mathsf{Vlo}}^{-1}\overline{m_t,\hat{m}_t}^2 &\leq F_t(\hat{m}_t, m_t) \\ &\leq F_t(\hat{m}_t, m_t) - \hat{F}_t(\hat{m}_t, m_t) \\ &= \sum_{i=1}^{n} w_i \mathbb{E}[\Diamond(Y_t, y_i, m_t, \hat{m}_t) \mid y_{1\dots n}] \\ &\leq \sum_{i=1}^{n} |w_i| \, \overline{\hat{m}_t,m_t} \, \mathbb{E}[\mathfrak{a}(Y_t, y_i) \mid y_i] \,. \end{aligned}$$

Thus,

$$C_{\mathsf{Vlo}}^{-1}\overline{m_t,\hat{m}_t} \leq \sum_{i=1}^{n} |w_i| \, \mathbb{E}[\mathfrak{a}(Y_t, y_i) \mid y_i] \,.$$

With Jensen's inequality

$$\begin{aligned} C_{\mathsf{Vlo}}^{-\kappa}\mathbb{E}[\overline{m_t,\hat{m}_t}^{\kappa}] &\leq \mathbb{E}\left[\left(\sum_{i=1}^{n} |w_i| \, \mathbb{E}[\mathfrak{a}(Y_t, y_i) \mid y_i]\right)^{\kappa}\right] \\ &= W^{\kappa}\mathbb{E}\left[\left(\sum_{i=1}^{n} \frac{|w_i|}{W}\mathbb{E}[\mathfrak{a}(Y_t, y_i) \mid y_i]\right)^{\kappa}\right] \\ &\leq W^{\kappa}\sum_{i=1}^{n} \frac{|w_i|}{W}\mathbb{E}[\mathbb{E}[\mathfrak{a}(Y_t, y_i) \mid y_i]^{\kappa}] \\ &\leq W^{\kappa}\sum_{i=1}^{n} \frac{|w_i|}{W}\mathbb{E}[\mathfrak{a}(Y_t, y_i)^{\kappa}] \\ &\leq W^{\kappa} \sup_{s,t\in[0,1]} \mathbb{E}\big[\mathfrak{a}(Y_t, Y'_s)^{\kappa}\big] \,. \end{aligned}$$

As $\mathfrak{a}$ is a semi-metric,

$$\begin{aligned} \mathbb{E}\big[\mathfrak{a}(Y_t, Y'_s)^{\kappa}\big] &\leq \mathbb{E}\big[\big(\mathfrak{a}(Y_t, m_t) + \mathfrak{a}(m_t, m_s) + \mathfrak{a}(m_s, Y'_s)\big)^{\kappa}\big] \\ &\leq 3^{\kappa}\left(2 \sup_{t\in[0,1]} \mathbb{E}[\mathfrak{a}(Y_t, m_t)^{\kappa}] + \mathfrak{a}(m_t, m_s)^{\kappa}\right) \\ &\leq c_{\kappa}\left(C_{\mathsf{Mom}}^{\kappa} + C_{\mathsf{Len}}^{\kappa}\right) \,. \end{aligned}$$

Lemma 5.40 shows $W \leq C_{\mathsf{Ker}}$. This completes the proof. □

Proof of Corollary 5.19. If $\mathsf{diam}(\mathcal{Q}, d) < \infty$, then

$$H(q,p) \leq \left(\int \left(2 \, \mathsf{diam}(\mathcal{Q}, d)\right)^2 \mu(\mathrm{d}y)\right)^{\frac{1}{2}} = 2 \, \mathsf{diam}(\mathcal{Q}, d)$$

Thus, we can choose $C_{\mathsf{Bom}} := 2 \, \mathsf{diam}(\mathcal{Q}, d)$. Using the triangle inequality we get $\overline{y,q}^2 -$

$\overline{y,p}^2 - \overline{z,q}^2 + \overline{z,p}^2 \leq 4\overline{q,p}\,\mathsf{diam}(\mathcal{Q},d)$. Thus, $\mathfrak{a}(y,z) \leq 4\,\mathsf{diam}(\mathcal{Q},d)$ and we can choose $C_{\mathsf{Mom}} := 4\,\mathsf{diam}(\mathcal{Q},d)$. To summarize,

$$\begin{aligned} C_1 &= c_{\kappa,\alpha} C_{\mathsf{Vlo}} C_{\mathsf{Ker}} C_{\mathsf{Smo}} \\ C_2 &= c_{\kappa,\alpha} C_{\mathsf{Vlo}} C_{\mathsf{Mom}} C_{\mathsf{Ent}} C_{\mathsf{Ker}} \\ C_{\mathsf{Mom}} &= 4\,\mathsf{diam}(\mathcal{Q},d) \\ C_{\mathsf{Smo}} &= C_{\mathsf{Bom}} C_{\mathsf{SmD}} \\ C_{\mathsf{Bom}} &= 2\ \mathsf{diam}(\mathcal{Q},d)\,. \end{aligned}$$

□

Proof of Corollary 5.20. VARIANCE holds in Hadamard spaces with $C_{\mathsf{Vlo}} := 1$. We bound $\mathbb{E}[H(\hat{m}_t, m_t)^\kappa]^{\frac{1}{\kappa}} \leq C_{\mathsf{Bom}}$ using

$$C_{\mathsf{Bom}} := c_\kappa C_{\mathsf{Vlo}} C_{\mathsf{Len}} C_{\mathsf{Mom}} C_{\mathsf{Int}} C_{\mathsf{Ker}}$$

see Proposition 5.42. To summarize,

$$\begin{aligned} C_1 &= c_{\kappa,\alpha} C_{\mathsf{Vlo}} C_{\mathsf{Ker}} C_{\mathsf{Smo}} \\ C_2 &= c_{\kappa,\alpha} C_{\mathsf{Vlo}} C_{\mathsf{Mom}} C_{\mathsf{Ent}} C_{\mathsf{Ker}} \\ C_{\mathsf{Vlo}} &= 1 \\ C_{\mathsf{Smo}} &= C_{\mathsf{Bom}} C_{\mathsf{SmD}} \\ C_{\mathsf{Bom}} &= c_\kappa C_{\mathsf{Vlo}} C_{\mathsf{Len}} C_{\mathsf{Mom}} C_{\mathsf{Int}} C_{\mathsf{Ker}} \\ (C_1 + C_2)^2 &\leq c'_\kappa \left(C_{\mathsf{Ker}}^2 C_{\mathsf{Len}} C_{\mathsf{Mom}} C_{\mathsf{Int}} C_{\mathsf{Ker}} C_{\mathsf{SmD}} + C_{\mathsf{Mom}} C_{\mathsf{Ent}} C_{\mathsf{Ker}} \right)^2 \\ &\leq c''_\kappa \left(C_{\mathsf{Ker}}^2 C_{\mathsf{Len}} C_{\mathsf{Mom}} C_{\mathsf{Int}} C_{\mathsf{SmD}} C_{\mathsf{Ent}} \right)^2 . \end{aligned}$$

□

Proof of Corollary 5.16. By Lemma 5.57, we can choose $C_{\mathsf{Ent}} := 2\sqrt{k}$. The diameter is $\mathsf{diam}(\mathbb{S}^k, d) = 2\pi$. □

5.A.4.3 Smoothness of regression function

Proposition 5.43 (Smoothness of regression function)**.** Let $(\mathcal{Q}, d)$ be a Hadamard space. Assume $t \to \rho(y|t) \in \Sigma(1, L(y))$. Assume there are $C_{\mathsf{Int}}, C_{\mathsf{SmD}} \in (0,\infty)$ with $\int \overline{y,m_t}^2 \mathrm{d}\mu(y) \leq C_{\mathsf{Int}}^2$ and $\int L(y)^2 \mathrm{d}\mu(y) \leq C_{\mathsf{SmD}}$. Then $t \mapsto m_t$ is Lipschitz continuous with constant $C_{\mathsf{Int}} C_{\mathsf{SmD}}$. In particular, we can choose $C_{\mathsf{Len}} = C_{\mathsf{Int}} C_{\mathsf{SmD}}$.

Proof of Proposition 5.43. Using the variance inequality twice, we have

$$\begin{aligned}
2\overline{m_s,m_t}^2 &\le (F_s(m_t,m_s) + F_t(m_s,m_t)) \\
&= \int \left(\overline{y,m_t}^2 - \overline{y,m_s}^2\right)(p(y|t) - p(y|s))\,\mathrm{d}\mu(y) \\
&\le \overline{m_s,m_t} \int (\overline{y,m_t} + \overline{y,m_s})\,|p(y|t) - p(y|s)|\,\mathrm{d}\mu(y)\,.
\end{aligned}$$

Thus, with the Lipschitz assumption on the density,

$$\begin{aligned}
\overline{m_s,m_t} &\le \frac{1}{2}\int (\overline{y,m_t} + \overline{y,m_s})\,|p(y|t) - p(y|s)|\,\mathrm{d}\mu(y) \\
&\le \frac{1}{2}\,|s-t| \int (\overline{y,m_t} + \overline{y,m_s})\,L(y)\mathrm{d}\mu(y) \\
&\le |s-t| \sup_{t\in[0,1]} \left(\int \overline{y,m_t}^2 \mathrm{d}\mu(y) \int L(y)^2 \mathrm{d}\mu(y)\right)^{\frac{1}{2}} \\
&\le |s-t|\, C_{\mathsf{SmD}} C_{\mathsf{Int}}\,.
\end{aligned}$$

□

5.A.5 Section 5.7: `TriFre`

5.A.5.1 Theorem

We prove Theorem 5.23. The difference of the objective functions is split into three parts in Lemma 5.44. In Lemma 5.45, we use a peeling device and the variance inequality to relate this difference to the distance between the minimizers $\hat{m}_t$ and m_t, which is the quantity to be bounded in the theorem. Of the three parts, two bias related quantities are bounded in Lemma 5.46 and Lemma 5.47 with an auxiliary result in Lemma 5.48. The third part, a variance term, is bounded in Lemma 5.49 via chaining. The bounds on the three parts are summarized in Lemma 5.50. In the end, the integral over t is applied to calculate the mean integrated squared error. Here, the auxiliary result Lemma 5.51 is applied.

For shorter notation define $F_t(q,p) := F_t(q) - F_t(p)$ and $\hat{F}_t(q,p) := \hat{F}_t(q) - \hat{F}_t(p)$. We introduce the Fourier coefficients $\vartheta_k(q,p)$ of $t \mapsto F_t(q,p)$ with respect to the trigonometric basis

$$\vartheta_k(q,p) := \int_0^1 \psi_k(x) F_x(q,p)\mathrm{d}x$$

such that $F_t(q,p) = \sum_{k=1}^{\infty} \vartheta_k(q,p)\psi_k(t)$ due to SmoothDensity. Define

$$r_t(q,p) := \sum_{k=N+1}^{\infty} \vartheta_k(q,p)\psi_k(t)\,, \qquad F_t^r(q,p) := \Psi_N(t)^\top \frac{1}{n}\sum_{i=1}^n \Psi_N(x_i) r_{x_i}(q,p)\,,$$

$$\varepsilon_t(y,q,p) := F_t(q,p) - \left(\overline{y,q}^2 - \overline{y,p}^2\right)\,, \quad F_t^\varepsilon(q,p) := \Psi_N(t)^\top \frac{1}{n}\sum_{i=1}^n \Psi_N(x_i)\varepsilon_{x_i}(y_i,q,p)\,.$$

Lemma 5.44. If $N < n$, then

$$F_t(q,p) - \hat{F}_t(q,p) = r_t(q,p) + F_t^{\varepsilon}(q,p) - F_t^{r}(q,p)\,.$$

Proof of Lemma 5.44. It holds $\frac{1}{n}\sum_{i=1}^{n}\psi_k(x_i)\psi_\ell(x_i) = \delta_{k\ell}$ for $k,\ell \in \{1,\dots,n-1\}$, see [Tsy08, Lemma 1.7]. Set

$$F_t^N(q,p) := \sum_{k=1}^{N}\vartheta_k(q,p)\psi_k(t)\,.$$

Then $\frac{1}{n}\sum_{i=1}^{n}\psi_k(x_i)F_{x_i}^N(q,p) = \vartheta_k(q,p)$ for $k \le N < n$. Thus,

$$F_t^N(q,p) = \Psi_N(t)^\top \frac{1}{n}\sum_{i=1}^{n}\Psi_N(x_i)F_{x_i}^N(q,p)\,.$$

As $F_t(q,p) - r_t(q,p) = F_t^N(q,p)$, we obtain

$$\begin{aligned}
&F_t(q,p) - \hat{F}_t(q,p) - r_t(q,p)\\
&= \Psi_N(t)^\top \frac{1}{n}\sum_{i=1}^{n}\Psi_N(x_i)F_{x_i}^N(q,p) - \Psi_N(t)\frac{1}{n}\sum_{i=1}^{n}\Psi_N(x_i)\left(\overline{y_i,q}^2 - \overline{y_i,p}^2\right)\\
&= \Psi_N(t)^\top \frac{1}{n}\sum_{i=1}^{n}\Psi_N(x_i)\left(F_{x_i}^N(q,p) - F_{x_i}(q,p) + F_{x_i}(q,p) - \left(\overline{y_i,q}^2 - \overline{y_i,p}^2\right)\right)\\
&= \Psi_N(t)^\top \frac{1}{n}\sum_{i=1}^{n}\Psi_N(x_i)\left(-r_{x_i}(q,p) + \varepsilon_{x_i}(y_i,q,p)\right)\\
&= F_t^{\varepsilon}(q,p) - F_t^{r}(q,p)\,.
\end{aligned}$$

□

Next, we apply the peeling device.

Lemma 5.45. For $b > 0$, define

$$U_{t,b} := \sup_{q\in \mathrm{B}(m_t,b,d)} F_t^{\epsilon}(q,m_t) + \left(r_t(\hat{m}_t,m_t) - F_t^{r}(\hat{m}_t,m_t)\right)\mathbb{1}_{[0,b]}(\overline{\hat{m}_t,m_t})\,.$$

Let $\kappa > 2$. Define

$$h(t) := \sup_{b>0}\left(\frac{\mathbb{E}[U_{t,b}^{\kappa}]}{b^{\kappa}}\right)^{\frac{1}{\kappa}}$$

Assume VARIANCE. Then

$$\mathbb{E}\left[\overline{\hat{m}_t,m_t}^2\right] \le \frac{4\kappa}{\kappa-2}C_{\mathsf{Vlo}}^2 h(t)^2\,.$$

Proof of Lemma 5.45. For a function $h(t) > 0$, we have

$$\mathbb{E}\left[\frac{\overline{\hat{m}_t,m_t}^2}{h(t)^2}\right] = \int_0^\infty 2s\mathbb{P}\Big(\overline{\hat{m}_t,m_t} > sh(t)\Big)\mathrm{d}s\,.$$

By Variance, the minimizing property of $\hat{m}_t$, and Lemma 5.44, we obtain

$$\begin{aligned} C_{\mathsf{Vlo}}^{-1}\,\overline{\hat{m}_t,m_t}^2 &\le F_t(\hat{m}_t, m_t) \\ &\le F_t(\hat{m}_t, m_t) - \hat{F}_t(\hat{m}_t, m_t) \\ &= r_t(\hat{m}_t, m_t) + \hat{F}_t^\epsilon(\hat{m}_t, m_t) - F_t^r(\hat{m}_t, m_t)\,. \end{aligned}$$

If $\overline{\hat{m}_t,m_t} \in [a,b]$ for $0 < a < b$, then

$$\begin{aligned} C_{\mathsf{Vlo}}^{-1}a^2 &\le C_{\mathsf{Vlo}}^{-1}\,\overline{\hat{m}_t,m_t}^2 \\ &\le F_t^\epsilon(\hat{m}_t, m_t) + r_t(\hat{m}_t, m_t) - F_t^r(\hat{m}_t, m_t) \\ &\le \sup_{q\in\mathrm{B}(m_t,b,d)} F_t^\epsilon(q, m_t) + (r_t(\hat{m}_t, m_t) - F_t^r(\hat{m}_t, m_t))\,\mathbb{1}_{[0,b]}(\overline{\hat{m}_t,m_t}) \\ &= U_{t,b}\,. \end{aligned}$$

Thus, by Markov's inequality

$$\mathbb{P}\Big(\overline{\hat{m}_t,m_t} \in [a,b]\Big) \le \mathbb{P}\Big(a^2 \le C_{\mathsf{Vlo}}U_{t,b}\Big) \le \frac{C_{\mathsf{Vlo}}^\kappa\mathbb{E}[U_{t,b}^\kappa]}{a^{2\kappa}}\,.$$

Let $a_k(s) = 2^k sh(t)$. As $\mathbb{E}[U_{t,b}^\kappa] \le b^\kappa h(t)^\kappa$, we have

$$\begin{aligned} \mathbb{P}\Big(\overline{\hat{m}_t,m_t} > sh(t)\Big) &\le \min\left(1, \sum_{k=0}^\infty \mathbb{P}\Big(\overline{\hat{m}_t,m_t} \in [a_k, a_{k+1})\Big)\right) \\ &\le \min\left(1, C_{\mathsf{Vlo}}^\kappa \sum_{k=0}^\infty \frac{a_{k+1}^\kappa h(t)^\kappa}{a_k^{2\kappa}}\right)\,. \end{aligned}$$

We obtain

$$\frac{a_{k+1}^\kappa h(t)^\kappa}{a_k^{2\kappa}} = \frac{\Big(2^{k+1}sh(t)\Big)^\kappa h(t)^\kappa}{\big(2^k sh(t)\big)^{2\kappa}} = \left(\frac{2\cdot 2^k sh(t)h(t)}{2^{2k}s^2h(t)^2}\right)^\kappa = \Big(2\cdot 2^{-k}s^{-1}\Big)^\kappa$$

and thus

$$\sum_{k=0}^\infty \frac{a_{k+1}^\kappa h(t)^\kappa}{a_k^{2\kappa}} = 2^\kappa s^{-\kappa}\sum_{k=0}^\infty 2^{-k\kappa} = \frac{2^\kappa}{1-2^{-\kappa}}s^{-\kappa}\,.$$

Putting everything together with $c_\kappa := \frac{2^\kappa}{1-2^{-\kappa}} C_{\mathsf{Vlo}}^\kappa$ yields

$$\begin{aligned}
h(t)^{-2}\mathbb{E}\Big[\overline{\hat{m}_t,m_t}^2\Big] &= 2\int_0^\infty s\mathbb{P}\Big(\overline{\hat{m}_t,m_t} > sh(t)\Big)\mathrm{d}s \\
&\leq 2\int_0^\infty s\min(1, c_\kappa s^{-\kappa})\,\mathrm{d}s \\
&= \int_0^{c_\kappa^{\frac{1}{\kappa}}} 2s\mathrm{d}s + 2c_\kappa \int_{c_\kappa^{\frac{1}{\kappa}}}^\infty s^{1-\kappa}\mathrm{d}s \\
&= c_\kappa^{\frac{2}{\kappa}} + 2c_\kappa \frac{1}{\kappa-2}\left(c_\kappa^{\frac{1}{\kappa}}\right)^{2-\kappa} \\
&= c_\kappa^{\frac{2}{\kappa}}\left(1+\frac{2}{\kappa-2}\right) \\
&\leq \frac{4\kappa}{\kappa-2} C_{\mathsf{Vlo}}^2\,.
\end{aligned}$$

□

Using the smoothness assumption, we are able to bound the r-term.

Lemma 5.46 (Bound on r)**.** Assume SmoothDensity. Then

$$\mathbb{E}[|r_t(\hat{m}_t, m_t)|^\kappa \mathbb{1}_{[0,b]}(\overline{\hat{m}_t,m_t})] \leq b^\kappa h_N(t)^\kappa C_{\mathsf{Bom}}^\kappa\,,$$

where

$$\begin{aligned}
h_N(t) &:= \left(\int\left(\sum_{\ell=N+1}^\infty \xi_\ell(y)\psi_\ell(t)\right)^2 \mu(\mathrm{d}y)\right)^{\frac{1}{2}} \\
H(q,p) &:= \left(\int (\overline{y,q} + \overline{y,p})^2\,\mu(\mathrm{d}y)\right)^{\frac{1}{2}}.
\end{aligned}$$

Proof. It holds

$$\begin{aligned}
\vartheta_k(q,p) &= \int_0^1 \psi_k(x)F_x(q,p)\mathrm{d}x \\
&= \int_0^1 \int \psi_k(x)\left(\overline{y,q}^2 - \overline{y,p}^2\right)\rho(y|x)\mathrm{d}\mu(y)\mathrm{d}x \\
&= \int \left(\overline{y,q}^2 - \overline{y,p}^2\right)\int_0^1 \psi_k(x)\rho(y|x)\mathrm{d}x\mathrm{d}\mu(y) \\
&= \int \left(\overline{y,q}^2 - \overline{y,p}^2\right)\xi(y)\mathrm{d}\mu(y)\,.
\end{aligned}$$

Thus,

$$\begin{aligned} r_t(q,p) &= \int \left(\overline{y,q}^2 - \overline{y,p}^2\right) \sum_{\ell=N+1}^{\infty} \xi_\ell(y)\psi_\ell(t)\mu(\mathrm{d}y) \\ &\leq \left(\int \left(\overline{y,q}^2 - \overline{y,p}^2\right)^2 \mu(\mathrm{d}y)\right)^{\frac{1}{2}} \left(\int \left(\sum_{\ell=N+1}^{\infty} \xi_\ell(y)\psi_\ell(t)\right)^2 \mu(\mathrm{d}y)\right)^{\frac{1}{2}} \\ &\leq \overline{q,p}H(q,p)h_N(t)\,. \end{aligned}$$

Finally, we obtain

$$\mathbb{E}[|r_t(\hat{m}_t, m_t)|^\kappa \mathbb{1}_{[0,b]}(\overline{\hat{m}_t, m_t})] \leq b^\kappa h_N(t)^\kappa \mathbb{E}[H(\hat{m}_t, m_t)^\kappa]\,. \qquad \square$$

Using the previous result, we can also establish a bound on F^r.

Lemma 5.47 (Bound on F^r)**.**

$$\mathbb{E}[F_t^r(\hat{m}_t, m_t)^\kappa \mathbb{1}_{[0,b]}(\overline{\hat{m}_t, m_t})] \leq c_\kappa \left(N n^{1-2\beta} C_{\mathsf{SmD}}\right)^\kappa b^\kappa C_{\mathsf{Bom}}^\kappa$$

where $c_\kappa \in [1, \infty)$ depends only on κ.

Proof. We will show that asymptotically $F_t^r(q,p) \lesssim r_t(q,p)$. Recall

$$\begin{aligned} F_t^r(q,p) &= \Psi_N(t)^\top \frac{1}{n} \sum_{i=1}^{n} \Psi_N(x_i) r_{x_i}(q,p) \\ r_t(q,p) &= \sum_{k=N+1}^{\infty} \vartheta_k(q,p)\psi_k(t) \end{aligned}$$

and define

$$r_{n,t}(q,p) = \sum_{\ell=n}^{\infty} \vartheta_\ell(q,p)\psi_\ell(t)$$

It holds

$$F_t^r(q,p) \leq \|\Psi_N(t)\| \, \|\frac{1}{n} \sum_{i=1}^{n} \Psi_N(x_i) r_{x_i}(q,p)\|$$

By Lemma 5.48 below, to be shown below,

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \Psi_N(x_i) r_{x_i}(q,p) \right\|^2 \leq \frac{1}{n} \sum_{i=1}^{n} r_{x_i}(q,p)^2$$

As in the proof of Lemma 5.46, we have

$$|r_{n,t}(q,p)| \le \overline{q,p} h_n(t)^\kappa H(q,p)\,,$$

where

$$h_n(t)^2 = \int \left(\sum_{\ell=n}^{\infty} \xi_\ell(y)\psi_\ell(t) \right)^2 \mu(\mathrm{d}y)\,.$$

Thus,

$$F_t^r(q,p)^2 \le \overline{q,p}^2 H(q,p)^2 \|\Psi_N(t)\|^2 \frac{1}{n} \sum_{i=1}^{n} h_n(x_i)^2$$

with $\|\Psi_N(t)\|^2 \le 2N$. As $\xi(y) \in \Theta(\beta, L(y))$, we have $\sum_{k=1}^{\infty} \xi_k(y)^2 w_k^{-2} \le L(y)^2$ with $w_{2k+1} = w_{2k} = (2k)^{-\beta}$.

$$\sum_{k=n}^{\infty} w_k^2 \le c n^{1-2\beta}\,.$$

Thus,

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} \left(\sum_{k=n}^{\infty} \xi_k(y)\psi_k(x_i) \right)^2 &\le \frac{1}{n} \sum_{i=1}^{n} \sum_{k=n}^{\infty} w_k^{-2} \xi_k(y)^2 \sum_{k=n}^{\infty} w_k^2 \psi_k(x_i)^2 \\
&\le 2 \sum_{k=n}^{\infty} w_k^{-2} \xi_k(y)^2 \sum_{k=n}^{\infty} w_k^2 \\
&\le c_0 L(y)^2 n^{1-2\beta}\,.
\end{aligned}$$

We obtain

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} h_n(x_i)^2 &\le \frac{1}{n} \sum_{i=1}^{n} \int \left(\sum_{\ell=n}^{\infty} \xi_\ell(y)\psi_\ell(x_i) \right)^2 \mu(\mathrm{d}y) \\
&\le c_0 n^{1-2\beta} \int L(y)^2 \mu(\mathrm{d}y)
\end{aligned}$$

and can bound

$$F_t^r(q,p)^2 \le 2c_0 \overline{q,p}^2 H(q,p)^2 N n^{1-2\beta} \int L(y)^2 \mu(\mathrm{d}y)\,.$$

Finally, the inequalities above yield

$$\mathbb{E}[F_t^r(\hat{m}_t, m_t)^\kappa \mathbb{1}_{[0,b]}(\overline{\hat{m}_t, m_t})] \le \left(2c_0 N n^{1-2\beta} \int L(y)^2 \mu(\mathrm{d}y) \right)^{\frac{\kappa}{2}} b^\kappa \mathbb{E}[H(\hat{m}_t, m_t)^\kappa]\,. \quad \square$$

We still have to prove following lemma, which was used in the previous proof.

Lemma 5.48. Let $f\colon [0,1] \to \mathbb{R}$ be any function and $N < n$. Then

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \Psi_N(x_i) f(x_i) \right\|^2 \leq \frac{1}{n} \sum_{i=1}^{n} f(x_i)^2 \,.$$

Proof of Lemma 5.48. Let $a_\ell := \frac{1}{n} \sum_{i=1}^{n} \psi_\ell(x_i) f(x_i)$ and $s(t) := f(t) - \sum_{\ell=1}^{N} a_\ell \psi_\ell(t)$. Then

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} s(x_i)\psi_k(x_i) &= \frac{1}{n} \sum_{i=1}^{n} \left(f(x_i) - \sum_{\ell=1}^{N} a_\ell \psi_\ell(x_i) \right) \psi_k(x_i) \\
&= \frac{1}{n} \sum_{i=1}^{n} f(x_i)\psi_k(x_i) - \sum_{\ell=1}^{N} a_\ell \frac{1}{n} \sum_{i=1}^{n} \psi_\ell(x_i)\psi_k(x_i) \\
&= a_k - a_k \\
&= 0
\end{aligned}$$

and thus

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} f(x_i)^2 &= \frac{1}{n} \sum_{i=1}^{n} \left(s(x_i) + \sum_{\ell=1}^{N} a_\ell \psi_\ell(x_i) \right)^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(s(x_i)^2 + s(x_i) \sum_{\ell=1}^{N} a_\ell \psi_\ell(x_i) + \sum_{\ell,k=1}^{N} a_\ell a_k \psi_\ell(x_i)\psi_k(x_i) \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} s(x_i)^2 + \sum_{\ell=1}^{N} a_\ell \frac{1}{n} \sum_{i=1}^{n} s(x_i)\psi_\ell(x_i) + \sum_{\ell,k=1}^{N} a_\ell a_k \frac{1}{n} \sum_{i=1}^{n} \psi_\ell(x_i)\psi_k(x_i) \\
&= \frac{1}{n} \sum_{i=1}^{n} s(x_i)^2 + \sum_{\ell}^{N} a_\ell^2 \,.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\left\| \frac{1}{n} \sum_{i=1}^{n} \Psi_N(x_i) f(x_i) \right\|^2 &= \sum_{\ell=1}^{N} \left(\psi_\ell(x_i) f(x_i) \right)^2 \\
&= \sum_{\ell=1}^{N} a_\ell^2 \,.
\end{aligned}$$

As $\frac{1}{n} \sum_{i=1}^{n} s(x_i)^2 \geq 0$ we have proved the claim. □

Next, we tackle the variance term.

Lemma 5.49 (Bound on F^ε)**.** Assume Moment, Entropy. Then there is an constant $c > 0$ depending only on κ such that

$$\mathbb{E}\left[\sup_{q\in\mathcal{B}} F_t^\varepsilon(q,p)^\kappa\right] \le c_\kappa C_{\mathsf{Mom}}^\kappa n^{-\frac{\kappa}{2}} C_{\mathsf{Ent}}^\kappa b^\kappa \left(\Psi_N(t)^\top \Psi_N(t)\right)^{\frac{\kappa}{2}} .$$

Proof of Lemma 5.49. Recall $F_t^\varepsilon(q,p) = \Psi_N(t)^\top \frac{1}{n}\sum_{i=1}^n \Psi_N(x_i)\varepsilon_{x_i}(y_i,q,p)$. Define $\alpha_i := \frac{1}{n}\Psi_N(t)^\top\Psi_N(x_i)$ and $\varepsilon_i(q,p) := \varepsilon_{x_i}(y_i,q,p)$. Then

$$F_t^\varepsilon(q,p) = \sum_{i=1}^n \alpha_i \varepsilon_i(q,p)\,,$$

where $\varepsilon_1,\dots,\varepsilon_n$ are independent and $\mathbb{E}[\varepsilon_i(q,p)] = 0$. We want to apply Theorem 5.56 with Z_i such that $Z_i(q) - Z_i(p) = \alpha_i\varepsilon_i(q,p)$ and $A_i := \alpha_i \mathfrak{a}(y_i,y_i')$. We need to show

$$|Z_i(q) - Z_i(p) - Z_i'(q) + Z_i'(p)| \le A_i\, \overline{q,p}$$

to obtain

$$\mathbb{E}\left[\sup_{q\in\mathcal{B}}\left|\sum_{i=1}^n Z_i(q)\right|^\kappa\right] \le C\,\mathbb{E}[\|A\|_2^\kappa]\,\gamma_2(\mathcal{B},d)^\kappa\,.$$

Using the quadruple property, we obtain

$$\begin{aligned}\varepsilon_i(q,p) - \varepsilon_i'(q,p) &= \left(F(q,p,x_i) - \left(\overline{y_i,q}^2 - \overline{y_i,p}^2\right)\right) - \left(F(q,p,x_i) - \left(\overline{y_i,q}^2 - \overline{y_i,p}^2\right)\right)\\ &\le \mathfrak{a}(y_i,y_i')\,\overline{q,p}\,.\end{aligned}$$

Thus, Theorem 5.56 yields

$$\mathbb{E}\left[\sup_{q\in\mathcal{B}} F_t^\varepsilon(q,p)^\kappa\right] \le C\gamma_2(\mathcal{B},d)^\kappa \mathbb{E}\left[\left(\sum_{i=1}^n \alpha_i^2 \mathfrak{a}(y_i,y_i')^2\right)^{\frac{\kappa}{2}}\right].$$

Let $a_i := \frac{\alpha_i^2}{\sum_{i=1}^n \alpha_i^2}$. Then

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_{i=1}^n \alpha_i^2 \mathfrak{a}(y_i, y_i')^2\right)^{\frac{\kappa}{2}}\right] &= \left(\sum_{i=1}^n \alpha_i^2\right)^{\frac{\kappa}{2}} \mathbb{E}\left[\left(\sum_{i=1}^n a_i \mathfrak{a}(y_i, y_i')^2\right)^{\frac{\kappa}{2}}\right] \\
&\leq \left(\sum_{i=1}^n \alpha_i^2\right)^{\frac{\kappa}{2}} \mathbb{E}\left[\sum_{i=1}^n a_i \mathfrak{a}(y_i, y_i')^\kappa\right] \\
&= \left(\sum_{i=1}^n \alpha_i^2\right)^{\frac{\kappa}{2}} \sum_{i=1}^n a_i \mathbb{E}\left[\mathfrak{a}(y_i, y_i')^\kappa\right] \\
&\leq \left(\sum_{i=1}^n \alpha_i^2\right)^{\frac{\kappa}{2}} \sup_t \mathbb{E}\left[\mathfrak{a}(Y_t, Y_t')^\kappa\right].
\end{aligned}$$

As $\mathfrak{a}$ is a semi metric, we have, using MOMENT,

$$\mathbb{E}\left[\mathfrak{a}(Y_t, Y_t')^\kappa\right] \leq 2^\kappa C_{\mathsf{Mom}}^\kappa .$$

Furthermore, it holds

$$\sum_{i=1}^n \alpha_i^2 = \frac{1}{n^2} \sum_{i=1}^n \Psi_N(t)^\top \Psi_N(x_i) \Psi_N(x_i)^\top \Psi_N(t) = \frac{1}{n} \Psi_N(t)^\top \Psi_N(t) .$$

Together we get

$$\mathbb{E}\left[\sup_{q \in \mathcal{B}} F_t^\varepsilon(q,p)^\kappa\right] \leq c_\kappa C_{\mathsf{Mom}}^\kappa n^{-\frac{\kappa}{2}} \gamma_2(\mathcal{B}, d)^\kappa \left(\Psi_N(t)^\top \Psi_N(t)\right)^{\frac{\kappa}{2}} . \qquad \square$$

Finally, we put the previous results together to proof our main theorem of this section.

Lemma 5.50. There is a constant $c > 0$ depending only on κ such that

$$h(t)^\kappa \leq c_\kappa \left(h_N(t)^\kappa C_{\mathsf{Bom}}^\kappa + \left(N n^{1-2\beta} C_{\mathsf{SmD}}\right)^\kappa C_{\mathsf{Bom}}^\kappa + C_{\mathsf{Mom}}^\kappa n^{-\frac{\kappa}{2}} C_{\mathsf{Ent}}^\kappa \|\Psi_N(t)\|^\kappa\right) .$$

Proof of Lemma 5.50. Lemma 5.46, Lemma 5.47, and Lemma 5.49. $\square$

Lemma 5.51. For the function h_N defined in Lemma 5.46, it holds

$$\int_0^1 h_N(t)^2 \mathrm{d}t \leq c\beta N^{-2\beta} C_{\mathsf{SmD}}^2 .$$

Proof of Lemma 5.51. We use Fubini's theorem and the weights $w_{2k+1} = w_{2k} = (2k)^{-\beta}$ from the definition of the ellipsoid $\Theta(\beta, L)$ and obtain

$$\begin{aligned}
\int_0^1 h_N(t)^2 \mathrm{d}t &= \int \int_0^1 \left(\sum_{\ell=N+1}^{\infty} \xi_\ell(y) \psi_\ell(t) \right)^2 \mathrm{d}t \mathrm{d}\mu(y) \\
&= \int_0^1 \int \left(\sum_{\ell=N+1}^{\infty} \xi_\ell(y) \psi_\ell(t) \right)^2 \mathrm{d}\mu(y) \mathrm{d}t \\
&= \int \sum_{\ell=N+1}^{\infty} \xi_\ell(y)^2 \mathrm{d}\mu(y) \\
&\leq \int w_{N+1}^2 \sum_{\ell=N+1}^{\infty} \xi_\ell(y)^2 w_\ell^{-2} \mathrm{d}\mu(y) \\
&\leq c\beta N^{-2\beta} \int L(y)^2 \mathrm{d}\mu(y) \,.
\end{aligned}$$

□

Proof of Theorem 5.23. We apply Lemma 5.45, Lemma 5.50, and Lemma 5.51 together with

$$\int_0^1 \|\Psi_N(t)\|^2 \,\mathrm{d}t = \int_0^1 \sum_{\ell=1}^{N} \psi_\ell(t)^2 \mathrm{d}t = N$$

to finally obtain

$$\begin{aligned}
\int_0^1 \mathbb{E}\left[\overline{\hat{m}_t, m_t}^2\right] \mathrm{d}t &\leq \frac{4\kappa}{\kappa - 2} C_{\mathsf{Vlo}}^2 \int_0^1 h(t)^2 \mathrm{d}t \\
&\leq c_\kappa C_{\mathsf{Vlo}}^2 \Bigg(C_{\mathsf{Bom}}^2 \int_0^1 h_N(t)^2 \mathrm{d}t + N n^{1-2\beta} C_{\mathsf{SmD}}^2 C_{\mathsf{Bom}}^2 + \\
&\qquad C_{\mathsf{Mom}}^2 n^{-1} C_{\mathsf{Ent}}^2 \int_0^1 \|\Psi_N(t)\|^2 \mathrm{d}t \Bigg) \\
&\leq c_{\kappa,\beta} C_{\mathsf{Vlo}}^2 \Bigg(C_{\mathsf{Bom}}^2 C_{\mathsf{SmD}}^2 N^{-2\beta} + C_{\mathsf{SmD}}^2 C_{\mathsf{Bom}}^2 N n^{1-2\beta} + \\
&\qquad C_{\mathsf{Mom}}^2 C_{\mathsf{Ent}}^2 \frac{N}{n} \Bigg) .
\end{aligned}$$

□

5.A.5.2 Corollaries

We first need to prove an auxiliary results before we can tackle the corollaries themselves. Recall $H(q,p) = \left(\int \left(\overline{y,q} + \overline{y,p} \right)^2 \mu(\mathrm{d}y) \right)^{\frac{1}{2}}$.

Proposition 5.52. Assume BomBound. To fulfill $\mathbb{E}[H(\hat{m}_t, m_t)^\kappa]^{\frac{1}{\kappa}} \leq C_{\mathsf{Bom}}$, we can choose

$$C_{\mathsf{Bom}} := c_\kappa C_{\mathsf{Len}} C_{\mathsf{Mom}} C_{\mathsf{Int}} \left(1 + \log(N) + \frac{N^2}{n}\right)$$

where $c_\kappa > 0$ depends only on κ.

This proposition is proven in two steps: Lemma 5.53 and Lemma 5.54. Let $w_i := \frac{1}{n}\left|\Psi_N(t)^\top \Psi_N(x_i)\right|$ and $W := \sum_{i=1}^n |w_i|$.

Lemma 5.53. Assume Variance. There is a constant $c_\kappa \in [1, \infty)$ depending only on κ such that

$$\mathbb{E}[H(\hat{m}_t, m_t)^\kappa]^{\frac{1}{\kappa}} \leq c_\kappa \left(C_{\mathsf{Vlo}} W \left(C_{\mathsf{Len}} + C_{\mathsf{Mom}}\right) + C_{\mathsf{Int}}\right) .$$

Proof of Lemma 5.53. Using the triangle inequality

$$\begin{aligned} H(q,p)^2 &= \int (\overline{y,q} + \overline{y,p})^2 \, \mu(\mathrm{d}y) \\ &\leq \int (\overline{q,p} + 2\overline{y,p})^2 \, \mu(\mathrm{d}y) \\ &\leq 2 \int \overline{q,p}^2 + 4\overline{y,p}^2 \mu(\mathrm{d}y) \\ &\leq 2\overline{q,p}^2 + 8 \int \overline{y,p}^2 \mu(\mathrm{d}y) \end{aligned}$$

as μ is a probability measure.

$$\begin{aligned} \mathbb{E}[H(\hat{m}_t, m_t)^\kappa]^{\frac{1}{\kappa}} &\leq \mathbb{E}\left[\left(2\overline{\hat{m}_t, m_t}^2 + 8 \int \overline{y, m_t}^2 \mu(\mathrm{d}y)\right)^{\frac{\kappa}{2}}\right]^{\frac{1}{\kappa}} \\ &\leq c_\kappa \left(\mathbb{E}\left[\overline{\hat{m}_t, m_t}^\kappa\right]^{\frac{1}{\kappa}} + \left(\int \overline{y, m_t}^2 \mu(\mathrm{d}y)\right)^{\frac{1}{2}}\right) . \end{aligned}$$

Next, we will bound $\mathbb{E}[\overline{m_t, \hat{m}_t}^\kappa]$. First, by Variance and the minimizing property of $\hat{m}_t$,

$$\begin{aligned} C_{\mathsf{Vlo}}^{-1} \overline{m_t, \hat{m}_t}^2 &\leq F_t(\hat{m}_t, m_t) \\ &\leq F_t(\hat{m}_t, m_t) - \hat{F}_t(\hat{m}_t, m_t) \\ &\leq \sum_{i=1}^n |w_i| \; \overline{\hat{m}_t, m_t} \, \mathbb{E}[\mathfrak{a}(Y_t, y_i) \mid y_i] . \end{aligned}$$

Thus,

$$C_{\mathsf{Vlo}}^{-1}\overline{m_t,\hat{m}_t} \leq \sum_{i=1}^{n} |w_i|\, \mathbb{E}[\mathfrak{a}(Y_t, y_i) \mid y_i]\,.$$

With Jensen's inequality

$$\begin{aligned}
C_{\mathsf{Vlo}}^{-\kappa}\mathbb{E}[\overline{m_t,\hat{m}_t}^{\kappa}] &\leq \mathbb{E}\left[\left(\sum_{i=1}^{n} |w_i|\, \mathbb{E}[\mathfrak{a}(Y_t, y_i) \mid y_i]\right)^{\kappa}\right] \\
&= W^{\kappa}\mathbb{E}\left[\left(\sum_{i=1}^{n} \frac{|w_i|}{W}\mathbb{E}[\mathfrak{a}(Y_t, y_i) \mid y_i]\right)^{\kappa}\right] \\
&\leq W^{\kappa}\sum_{i=1}^{n} \frac{|w_i|}{W}\mathbb{E}[\mathbb{E}[\mathfrak{a}(Y_t, y_i) \mid y_i]^{\kappa}] \\
&\leq W^{\kappa}\sum_{i=1}^{n} \frac{|w_i|}{W}\mathbb{E}[\mathfrak{a}(Y_t, y_i)^{\kappa}] \\
&\leq W^{\kappa}\sup_{s,t\in[0,1]} \mathbb{E}\big[\mathfrak{a}(Y_t, Y_s')^{\kappa}\big]\,.
\end{aligned}$$

As $\mathfrak{a}$ is a semi-metric,

$$\begin{aligned}
\mathbb{E}\big[\mathfrak{a}(Y_t, Y_s')^{\kappa}\big] &\leq \mathbb{E}\big[\big(\mathfrak{a}(Y_t, m_t) + \mathfrak{a}(m_t, m_s) + \mathfrak{a}(m_s, Y_s')\big)^{\kappa}\big] \\
&\leq 3^{\kappa}\left(2\sup_{t\in[0,1]} \mathbb{E}[\mathfrak{a}(Y_t, m_t)^{\kappa}] + \mathfrak{a}(m_t, m_s)^{\kappa}\right) \\
&\leq c_{\kappa}\left(C_{\mathsf{Mom}}^{\kappa} + C_{\mathsf{Len}}^{\kappa}\right)\,. \qquad \square
\end{aligned}$$

Lemma 5.54. There is an universal constant $c \in (0, \infty)$ such that

$$W \leq c\left(1 + \log(N) + \frac{N^2}{n}\right)\,.$$

Proof of Lemma 5.54. Let $g_t(s) := \left|\sum_{\ell=1}^{N} \psi_\ell(t)\psi_\ell(s)\right|$. Then

$$W = \sum_{i=1}^{n} |w_i| = \frac{1}{n}\sum_{i=1}^{n} \left|\Psi_N(t)^{\top}\Psi_N(x_i)\right| = \frac{1}{n}\sum_{i=1}^{n} g_t(x_i)\,.$$

By the standard comparison between an integral of a Lipschitz–continuous function an

the corresponding Riemann sum, we obtain

$$\left|\int_0^1 g_t(s)\mathrm{d}s - \frac{1}{n}\sum_{i=1}^{n} g_t(x_i)\right| \le \sup_{s\in[0,1]} \frac{|g_t'(s)|}{n} \le 4\pi\frac{N^2}{n}\,.$$

This bound is quite rough and could be improved. But we will choose $N_n \le n^{\frac{1}{3}}$ and thus $\frac{N_n^2}{n} \to 0$. For $x \in \mathbb{R}$ denote $[x]$ the fractional part of x, i.e., the number $[x] \in [0,1)$ that fulfills $[x] = x - k$ for a $k \in \mathbb{Z}$. For $\ell \ge 2$,

$$\psi_\ell(t)\psi_\ell(s) = \frac{1}{2}\Big((-1)^\ell \cos(2\pi\ell[t+s]) + \cos(2\pi\ell[t-s])\Big)\,.$$

The function $(s,t) \mapsto \sum_{\ell=1}^{N} \psi_\ell(t)\psi_\ell(s)$ only depends on $[s+t]$ and $[s-t]$. When integrating s from 0 to 1, $[s+t]$ and $[s-t]$ run through every value in $[0,1)$. Thus

$$\begin{aligned}
&\sup_{t\in[0,1]}\int_0^1 \left|1 + \sum_{\ell=2}^{N}\psi_\ell(t)\psi_\ell(s)\right|\mathrm{d}s \\
&= \sup_{t\in[0,1]}\int_0^1 \left|1 + \frac{1}{2}\sum_{\ell=2}^{N}\Big((-1)^\ell\cos(2\pi\ell[t+s]) + \cos(2\pi\ell[t-s])\Big)\right|\mathrm{d}s \\
&\le 1 + \frac{1}{2}\sup_{t\in[0,1]}\int_0^1 \left|\sum_{\ell=2}^{N}\Big((-1)^\ell\cos(2\pi\ell[t+s])\Big)\right|\mathrm{d}s \\
&\quad + \frac{1}{2}\sup_{t\in[0,1]}\int_0^1 \left|\sum_{\ell=2}^{N}\cos(2\pi\ell[t-s])\right|\mathrm{d}s \\
&= 1 + \frac{1}{2}\int_0^1 \left|\sum_{\ell=2}^{N}(-1)^\ell\cos(2\pi\ell s)\right|\mathrm{d}s + \frac{1}{2}\int_0^1 \left|\sum_{\ell=2}^{N}\cos(2\pi\ell s)\right|\mathrm{d}s\,.
\end{aligned}$$

Lagrange's trigonometric identities state

$$\begin{aligned}
2\sum_{\ell=1}^{L}\cos(\ell x) &= -1 + \frac{\sin\Big((L+\frac{1}{2})x\Big)}{\sin\big(\frac{x}{2}\big)}\,, \\
2\sum_{\ell=1}^{L}(-1)^\ell\cos(\ell x) &= -1 + \frac{(-1)^{L+1}\sin\Big((L+\frac{1}{2})x\Big)}{-\sin\big(\frac{x}{2}\big)}\,.
\end{aligned}$$

Thus, we have to bound the integral

$$\int_0^1 \left|\frac{\sin((2L+1)\pi s)}{\sin(\pi s)}\right|\mathrm{d}s\,.$$

It holds $|\sin(\pi x)| \geq \frac{1}{2}\pi \min(x, 1-x)$ for $x \in [0,1]$. Let $a = k\pi$ for $k \in \mathbb{N}$. Then

$$\begin{aligned}\int_0^1 \left|\frac{\sin(as)}{\sin(\pi s)}\right| \mathrm{d}s &\leq \frac{2}{\pi}\int_0^1 \frac{|\sin(as)|}{\min(s,1-s)}\mathrm{d}s \\ &= \frac{4}{\pi}\int_0^{\frac{1}{2}} \frac{|\sin(as)|}{s}\mathrm{d}s \\ &= \frac{4}{\pi}\int_0^{\frac{1}{2}a} \frac{|\sin(t)|}{t}\mathrm{d}t\,.\end{aligned}$$

We bound this integral as follows,

$$\begin{aligned}\int_0^{\frac{1}{2}k\pi} \frac{|\sin(t)|}{t}\mathrm{d}t &= \int_0^{\pi} \frac{|\sin(t)|}{t}\mathrm{d}t + \int_{\pi}^{\frac{1}{2}k\pi} \frac{|\sin(t)|}{t}\mathrm{d}t \\ &\leq \int_0^{\pi} \frac{\sin(t)}{t}\mathrm{d}t + \int_{\pi}^{\frac{1}{2}k\pi} \frac{1}{t}\mathrm{d}t \\ &\leq 2 + \log(\frac{1}{2}k\pi) - \log(\pi) \\ &= 2 + \log(\frac{1}{2}k)\,.\end{aligned}$$

Thus, we obtain

$$\int_0^1 \left|\frac{\sin(2k\pi s)}{\sin(\pi s)}\right| \mathrm{d}s \leq \frac{8}{\pi} + \frac{4}{\pi}\log(\frac{1}{2}k)\,,$$

which yields

$$\sup_{t\in[0,1]} \int_0^1 \left|1 + \sum_{\ell=2}^{N} \psi_\ell(t)\psi_\ell(s)\right| \mathrm{d}s \leq c_0 + c_1 \log(N)\,. \qquad \square$$

Proof of Corollary 5.24. If $\mathsf{diam}(\mathcal{Q}, d) < \infty$, then

$$H(q,p) \leq \left(\int (2\ \mathsf{diam}(\mathcal{Q},d))^2\, \mu(\mathrm{d}y)\right)^{\frac{1}{2}} = 2\ \mathsf{diam}(\mathcal{Q},d)\,.$$

Thus, we can choose $C_{\mathsf{Bom}} := 2\ \mathsf{diam}(\mathcal{Q}, d)$. Using the triangle inequality we get $\overline{y,q}^2 - \overline{y,p}^2 - \overline{z,q}^2 + \overline{z,p}^2 \leq 4\overline{q,p}\,\mathsf{diam}(\mathcal{Q},d)$. Thus, $\mathfrak{a}(y,z) \leq 4\,\mathsf{diam}(\mathcal{Q},d)$ and we can choose $C_{\mathsf{Mom}} = 4\,\mathsf{diam}(\mathcal{Q},d)$. $\square$

Proof of Corollary 5.25. VARIANCE holds in Hadamard spaces with $C_{\mathsf{Vlo}} := 1$. We bound $\mathbb{E}[H(\hat{m}_t, m_t)^\kappa]^{\frac{1}{\kappa}} \leq C_{\mathsf{Bom}}$ using

$$C_{\mathsf{Bom}} := c_\kappa C_{\mathsf{Len}} C_{\mathsf{Mom}} C_{\mathsf{Int}} \left(1 + \log(N) + \frac{N^2}{n}\right)\,,$$

see Proposition 5.52. □

5.B Chaining

Definition 5.55 (Measures of Entropy [Tal14])**.**

(i) Given a set $\mathcal{Q}$ an *admissible sequence* is an increasing sequence $(\mathcal{A}_k)_{k\in\mathbb{N}_0}$ of partitions of $\mathcal{Q}$ such that $\mathcal{A}_0 = \mathcal{Q}$ and $\mathsf{card}(\mathcal{A}_k) \leq 2^{2^k}$ for $k \geq 1$.

By an increasing sequence of partitions we mean that every set of $\mathcal{A}_{k+1}$ is contained in a set of $\mathcal{A}_k$. We denote by $A_k(q)$ the unique element of $\mathcal{A}_k$ which contains $q \in \mathcal{Q}$.

(ii) Let $(\mathcal{Q}, d)$ be a pseudo-metric space. Define

$$\gamma_2(\mathcal{Q}, d) := \inf \sup_{q\in\mathcal{Q}} \sum_{k=0}^{\infty} 2^{\frac{k}{2}} \,\mathsf{diam}(A_k(q), d)\,,$$

where the infimum is taken over all admissible sequences in $\mathcal{Q}$ and

$$\mathsf{diam}(A, d) := \sup_{q,p\in A} d(q, p)$$

for $A \subseteq \mathcal{Q}$.

Theorem 5.56 (Empirical process bound)**.** Let $(\mathcal{Q}, d)$ be a separable pseudo-metric space and $\mathcal{B} \subseteq \mathcal{Q}$. Let $Z_1, \dots, Z_n$ be centered, independent, and integrable stochastic processes indexed by $\mathcal{Q}$ with a $q_0 \in \mathcal{B}$ such that $Z_i(q_0) = 0$ for $i = 1, \dots, n$. Let $(Z_1', \dots, Z_n')$ be an independent copy of $(Z_1, \dots, Z_n)$. Assume the following Lipschitz-property: There is a random vector A with values in $\mathbb{R}^n$ such that

$$\left|Z_i(q) - Z_i(p) - Z_i'(q) + Z_i'(p)\right| \leq A_i d(q, p)$$

for $i = 1, \dots, n$ and all $q, p \in \mathcal{B}$. Let $\kappa \geq 1$. Then

$$\mathbb{E}\left[\sup_{q\in\mathcal{B}} \left|\sum_{i=1}^{n} Z_i(q)\right|^{\kappa}\right] \leq c_\kappa \,\mathbb{E}[\|A\|_2^\kappa]\,\gamma_2(\mathcal{B}, d)^\kappa\,,$$

where $c_\kappa \in (0, \infty)$ depends only on κ.

Proof. See [Sch19b, Theorem 6]. □

Lemma 5.57. In the Euclidean space $\mathbb{R}^k$ with the metric induced by the Euclidean norm $|\cdot|$, it holds $\gamma_2(\mathrm{B}(x, r, |\cdot|), |\cdot|) \leq 2r\sqrt{k}$ for any point $x \in \mathbb{R}^k$ and radius $r > 0$.

Proof. See [Pol90, section 4] and comparison to the entropy integral as in Remark 5.3. □

Bibliography

[Aba78] T. J. Abatzoglou. "The minimum norm projection on C^2-manifolds in $\mathbf{R}^n$". In: *Trans. Amer. Math. Soc.* 243 (1978), pp. 115–122.

[ABY13] M. Arnaudon, F. Barbaresco, and L. Yang. "Medians and means in Riemannian geometry: existence, uniqueness and computation". In: *Matrix information geometry*. Springer, Heidelberg, 2013, pp. 169–197.

[AC11] M. Agueh and G. Carlier. "Barycenters in the Wasserstein Space." In: *SIAM J. Math. Analysis* 43.2 (2011), pp. 904–924.

[AD89] J. Aczél and J. Dhombres. *Functional equations in several variables*. Vol. 31. Encyclopedia of Mathematics and its Applications. With applications to mathematics, information theory and to the natural and social sciences. Cambridge University Press, Cambridge, 1989, pp. xiv+462.

[ALP20] A. Ahidar-Coutrix, T. Le Gouic, and Q. Paris. "Convergence rates for empirical barycenters in metric spaces: curvature, convexity and extendable geodesics". In: *Probab. Theory Related Fields* 177.1-2 (2020), pp. 323–368.

[AW95] Z. Artstein and R. J.-B. Wets. "Consistency of minimizers and the SLLN for stochastic programs". In: *J. Convex Anal.* 2.1-2 (1995), pp. 1–17.

[Bač14a] M. Bačák. "Computing medians and means in Hadamard spaces". In: *SIAM J. Optim.* 24.3 (2014), pp. 1542–1566.

[Bač14b] M. Bačák. *Convex analysis and optimization in Hadamard spaces*. Vol. 22. De Gruyter Series in Nonlinear Analysis and Applications. De Gruyter, Berlin, 2014, pp. viii+185.

[Bac18] M. Bacak. *Old and new challenges in Hadamard spaces*. 2018. arXiv: `1807.01355 [math.FA]`.

[BBI01] D. Burago, Y. Burago, and S. Ivanov. *A course in metric geometry*. Vol. 33. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2001, pp. xiv+415.

[BD17] L. Birbrair and M. P. Denkowski. "Medial axis and singularities". In: *J. Geom. Anal.* 27.3 (2017), pp. 2339–2380.

[BDG07] E. del Barrio, P. Deheuvels, and S. van de Geer. *Lectures on empirical processes*. EMS Series of Lectures in Mathematics. Theory and statistical applications. European Mathematical Society (EMS), Zürich, 2007, pp. x+254.

[BFW19] D. Banholzer, J. Fliege, and R. Werner. "On rates of convergence for sample average approximations in the almost sure sense and in mean". In: *Mathematical Programming* (May 2019).

[BGW05] A. Banerjee, X. Guo, and H. Wang. “On the optimality of conditional expectation as a Bregman predictor”. In: *IEEE Trans. Inform. Theory* 51.7 (2005), pp. 2664–2669.

[BHV01] L. J. Billera, S. P. Holmes, and K. Vogtmann. “Geometry of the space of phylogenetic trees”. In: *Adv. in Appl. Math.* 27.4 (2001), pp. 733–767.

[Big+17] J. Bigot, R. Gouet, T. Klein, and A. López. “Geodesic PCA in the Wasserstein space by convex PCA”. In: *Ann. Inst. H. Poincaré Probab. Statist.* 53.1 (Feb. 2017), pp. 1–26.

[BL14] W. Bednorz and R. Latała. “On the boundedness of Bernoulli processes”. In: *Ann. of Math. (2)* 180.3 (2014), pp. 1167–1203.

[BLO18] D. Barden, H. Le, and M. Owen. “Limiting behaviour of Fréchet means in the space of phylogenetic trees”. In: *Ann. Inst. Statist. Math.* 70.1 (2018), pp. 99–129.

[BN08] I. D. Berg and I. G. Nikolaev. “Quasilinearization and curvature of Aleksandrov spaces”. In: *Geom. Dedicata* 133 (2008), pp. 195–218.

[Bou02] O. Bousquet. “A Bennett concentration inequality and its application to suprema of empirical processes”. In: *C. R. Math. Acad. Sci. Paris* 334.6 (2002), pp. 495–500.

[BP03] R. Bhattacharya and V. Patrangenaru. “Large sample theory of intrinsic and extrinsic sample means on manifolds. I”. In: *Ann. Statist.* 31.1 (2003), pp. 1–29.

[BP05] R. Bhattacharya and V. Patrangenaru. “Large sample theory of intrinsic and extrinsic sample means on manifolds. II”. In: *Ann. Statist.* 33.3 (2005), pp. 1225–1259.

[Car16] M. de Carvalho. “Mean, what do you mean?” In: *Amer. Statist.* 70.3 (2016), pp. 270–274.

[CCM97] H. I. Choi, S. W. Choi, and H. P. Moon. “Mathematical theory of medial axis transform”. In: *Pacific J. Math.* 181.1 (1997), pp. 57–88.

[CHS03] C. Choirat, C. Hess, and R. Seri. “A functional version of the Birkhoff ergodic theorem for a normal integrand: A variational approach”. In: *Ann. Probab.* 31.1 (Jan. 2003), pp. 63–92.

[CLM20] Y. Chen, Z. Lin, and H.-G. Müller. *Wasserstein Regression*. 2020. arXiv: `2006.09660 [stat.ME]`.

[CM20] Y. Chen and H.-G. Müller. *Uniform convergence of local Fréchet regression, with applications to locating extrema and time warping for metric-space valued trajectories*. 2020. arXiv: `2006.13548 [stat.ME]`.

[Cor+17] E. Cornea, H. Zhu, P. Kim, and J. G. Ibrahim. “Regression models on Riemannian symmetric spaces”. In: *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 79.2 (2017), pp. 463–482.

[Dav+10] B. C. Davis, P. T. Fletcher, E. Bullitt, and S. Joshi. “Population Shape Regression from Random Design Data”. In: *International Journal of Computer Vision* 90.2 (Nov. 2010), pp. 255–266.

[DD16] M. M. Deza and E. Deza. *Encyclopedia of distances*. Fourth. Springer, Berlin, 2016, pp. xxii+756.

[DKZ09] I. L. Dryden, A. Koloydenko, and D. Zhou. “Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging”. In: *Ann. Appl. Stat.* 3.3 (2009), pp. 1102–1123.

[DM16] I. L. Dryden and K. V. Mardia. *Statistical shape analysis with applications in R*. Second. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 2016, pp. xxiii+454.

[DM19a] P. Dubey and H.-G. Müller. “Fréchet analysis of variance for random objects”. In: *Biometrika* 106.4 (2019), pp. 803–821.

[DM19b] P. Dubey and H.-G. Müller. *Fréchet Change Point Detection*. 2019. arXiv: `1911.11864 [math.ST]`.

[DN03] H. A. David and H. N. Nagaraja. *Order statistics*. Third. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2003, pp. xvi+458.

[Edw95] R. E. Edwards. *Functional analysis*. Theory and applications, Corrected reprint of the 1965 original. Dover Publications, Inc., New York, 1995, pp. xvi+783.

[EH19] B. Eltzner and S. F. Huckemann. “A smeary central limit theorem for manifolds with application to high-dimensional spheres”. In: *Ann. Statist.* 47.6 (2019), pp. 3360–3381.

[EHW19] G. Eichfelder, T. Hotz, and J. Wieditz. “An algorithm for computing Fréchet means on the sphere”. In: *Optim. Lett.* 13.7 (2019), pp. 1523–1533.

[EJ20] S. N. Evans and A. Q. Jaffe. *Strong laws of large numbers for Fréchet means*. 2020. arXiv: `2012.12859 [math.PR]`.

[EPR13] L. Ellingson, V. Patrangenaru, and F. Ruymgaart. “Nonparametric estimation of means on Hilbert manifolds and extrinsic analysis of mean shapes of contours”. In: *J. Multivariate Anal.* 122 (2013), pp. 317–333.

[Fed59] H. Federer. “Curvature measures”. In: *Trans. Amer. Math. Soc.* 93 (1959), pp. 418–491.

[FHR21] T. Fissler, J. Hlavinová, and B. Rudloff. “Elicitability and identifiability of set-valued measures of systemic risk”. In: *Finance Stoch.* 25.1 (2021), pp. 133–165.

[Fle13] P. T. Fletcher. “Geodesic regression and the theory of least squares on Riemannian manifolds”. In: *Int. J. Comput. Vis.* 105.2 (2013), pp. 171–185.

[Fré48] M. Fréchet. “Les éléments aléatoires de nature quelconque dans un espace distancié”. In: *Ann. Inst. H. Poincaré* 10 (1948), pp. 215–310.

[FVJ09] P. T. Fletcher, S. Venkatasubramanian, and S. Joshi. “The geometric median on Riemannian manifolds with application to robust atlas estimation”. In: *Neuroimage* 45.1 Suppl (Mar. 2009), S143–152.

[FZ16] T. Fissler and J. F. Ziegel. “Higher order elicitability and Osband’s principle”. In: *Ann. Statist.* 44.4 (2016), pp. 1680–1707.

[Gau09] C. F. Gauß. *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*. Hamburgi sumptibus Frid. Perthes et I.H.Besser, 1809.

[Gee00] S. A. van de Geer. *Applications of empirical process theory*. Vol. 6. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2000, pp. xii+286.

[GGR18] S. Gadat, I. Gavra, and L. Risser. “How to calculate the barycenter of a weighted graph”. In: *Math. Oper. Res.* 43.4 (2018), pp. 1085–1118.

[Gne11] T. Gneiting. “Making and evaluating point forecasts”. In: *J. Amer. Statist. Assoc.* 106.494 (2011), pp. 746–762.

[Gou+19] T. L. Gouic, Q. Paris, P. Rigollet, and A. J. Stromme. *Fast convergence of empirical barycenters in Alexandrov spaces and the Wasserstein space*. 2019. arXiv: `1908.00828 [math.ST]`.

[Gow75] J. C. Gower. “Generalized Procrustes analysis”. In: *Psychometrika* 40 (1975), pp. 33–51.

[GSK12] C. E. Ginestet, A. Simmons, and E. D. Kolaczyk. “Weighted Frechet means as convex combinations in metric spaces: properties and generalized median inequalities”. In: *Statist. Probab. Lett.* 82.10 (2012), pp. 1859–1863.

[HE20] S. F. Huckemann and B. Eltzner. “Data Analysis on Non-Standard Spaces”. In: *WIREs Computational Statistics* (2020). early access.

[Hei09] M. Hein. “Robust Nonparametric Regression with Metric-Space Valued Output”. In: *Advances in Neural Information Processing Systems*. Ed. by Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta. Vol. 22. Curran Associates, Inc., 2009, pp. 718–726.

[Hei14] C. Heinrich. “The mode functional is not elicitable”. In: *Biometrika* 101.1 (2014), pp. 245–251.

[HF06] L. de Haan and A. Ferreira. *Extreme value theory*. Springer Series in Operations Research and Financial Engineering. An introduction. Springer, New York, 2006, pp. xviii+417.

[HH15] T. Hotz and S. F. Huckemann. “Intrinsic means on the circle: uniqueness, locus and asymptotics”. In: *Ann. Inst. Statist. Math.* 67.1 (2015), pp. 177–193.

[HHM10] S. F. Huckemann, T. Hotz, and A. Munk. “Intrinsic shape analysis: geodesic PCA for Riemannian manifolds modulo isometric Lie group actions”. In: *Statist. Sinica* 20.1 (2010), pp. 1–58.

[HL98] H. Hendriks and Z. Landsman. “Mean location and sample mean location on manifolds: asymptotics, tests, confidence regions”. In: *J. Multivariate Anal.* 67.2 (1998), pp. 227–243.

[HLP52] G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. 2d ed. Cambridge, at the University Press, 1952, pp. xii+324.

[Hot+13] T. Hotz, S. F. Huckemann, H. Le, J. S. Marron, J. C. Mattingly, E. Miller, J. Nolen, M. Owen, V. Patrangenaru, and S. Skwerer. “Sticky central limit theorems on open books”. In: *Ann. Appl. Probab.* 23.6 (2013), pp. 2238–2258.

[Hub64] P. J. Huber. “Robust estimation of a location parameter”. In: *Ann. Math. Statist.* 35 (1964), pp. 73–101.

[Huc+15] S. F. Huckemann, J. C. Mattingly, E. Miller, and J. Nolen. “Sticky central limit theorems at isolated hyperbolic planar singularities”. In: *Electron. J. Probab.* 20 (2015), no. 78, 34.

[Huc11] S. F. Huckemann. “Intrinsic inference on the mean geodesic of planar shapes and tree discrimination by leaf growth”. In: *Ann. Statist.* 39.2 (2011), pp. 1098–1124.

[Kar14] H. Karcher. *Riemannian Center of Mass and so called karcher mean*. 2014. arXiv: `1407.2087 [math.HO]`.

[Kar77] H. Karcher. “Riemannian center of mass and mollifier smoothing”. In: *Comm. Pure Appl. Math.* 30.5 (1977), pp. 509–541.

[Kem87] J. H. B. Kemperman. “The median of a finite measure on a Banach space”. In: *Statistical data analysis based on the L_1-norm and related methods (Neuchâtel, 1987)*. North-Holland, Amsterdam, 1987, pp. 217–230.

[Ken84] D. G. Kendall. “Shape manifolds, Procrustean metrics, and complex projective spaces”. In: *Bull. London Math. Soc.* 16.2 (1984), pp. 81–121.

[Kle14] A. Klenke. *Probability theory*. Second. Universitext. A comprehensive course. Springer, London, 2014, pp. xii+638.

[Koe05] R. Koenker. *Quantile regression*. Vol. 38. Econometric Society Monographs. Cambridge University Press, Cambridge, 2005, pp. xvi+349.

[Kol30] A. N. Kolmogorov. “Sur la notion de la moyenne”. In: *Atti Accad. Naz. Lincei* 12 (1930), pp. 388–391.

[KP17] Y.-H. Kim and B. Pass. “Wasserstein barycenters over Riemannian manifolds”. In: *Adv. Math.* 307 (2017), pp. 640–683.

[KW01] L. A. Korf and R. J.-B. Wets. “Random LSC functions: An ergodic theorem”. In: *Mathematics of Operations Research* 26.2 (2001), pp. 421–445.

[LL15] T. Le Gouic and J.-M. Loubes. “Barycenter in Wasserstein spaces: existence and consistency”. In: *Geometric science of information*. Vol. 9389. Lecture Notes in Comput. Sci. Springer, Cham, 2015, pp. 104–108.

[Llo82] S. P. Lloyd. “Least squares quantization in PCM”. In: *IEEE Trans. Inform. Theory* 28.2 (1982), pp. 129–137.

[LM19] Z. Lin and H.-G. Müller. *Total Variation Regularized Fréchet Regression for Metric-Space Valued Data*. 2019. arXiv: `1904.09647 [math.ME]`.

[LMP20] Z. Lin, H.-G. Müller, and B. U. Park. *Additive Models for Symmetric Positive-Definite Matrices, Riemannian Manifolds and Lie groups*. 2020. arXiv: `2009.08789 [stat.ME]`.

[LPS08] N. S. Lambert, D. M. Pennock, and Y. Shoham. “Eliciting Properties of Probability Distributions”. In: *Proceedings of the 9th ACM Conference on Electronic Commerce*. EC ’08. Chicago, Il, USA: Association for Computing Machinery, 2008, pp. 129–138.

[LV09] J. Lott and C. Villani. “Ricci curvature for metric-measure spaces via optimal transport”. In: *Ann. of Math. (2)* 169.3 (2009), pp. 903–991.

[MA14] J. S. Marron and A. M. Alonso. “Overview of object oriented data analysis”. In: *Biom. J.* 56.5 (2014), pp. 732–753.

[Mac67] J. MacQueen. “Some methods for classification and analysis of multivariate observations”. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. Berkeley, Calif.: University of California Press, 1967, pp. 281–297.

[MD87] P. Milasevic and G. R. Ducharme. “Uniqueness of the spatial median”. In: *Ann. Statist.* 15.3 (1987), pp. 1332–1333.

[NK13] S. Nandi and D. Kundu. “Estimation of parameters of partially sinusoidal frequency model”. In: *Statistics* 47.1 (2013), pp. 45–60.

[Nye11] T. M. W. Nye. “Principal components analysis in the space of phylogenetic trees”. In: *Ann. Statist.* 39.5 (2011), pp. 2716–2739.

[Oht12] S.-i. Ohta. “Barycenters in Alexandrov spaces of curvature bounded below”. In: *Adv. Geom.* 12.4 (2012), pp. 571–587.

[Pat98] V. Patrangenaru. *Asymptotic statistics on manifolds and their applications*. Thesis (Ph.D.)–Indiana University. ProQuest LLC, Ann Arbor, MI, 1998, p. 185.

[PDM19] A. Petersen, S. Deoni, and H.-G. Müller. “Fréchet estimation of time-varying covariance matrices from sparse data, with application to the regional coevolution of myelination in the developing brain”. In: *Ann. Appl. Stat.* 13.1 (2019), pp. 393–419.

[PFA06] X. Pennec, P. Fillard, and N. Ayache. “A Riemannian Framework for Tensor Computing”. In: *International Journal of Computer Vision* 66.1 (Jan. 2006), pp. 41–66.

[PM19a] A. Petersen and H.-G. Müller. “Fréchet regression for random objects with Euclidean predictors”. In: *Ann. Statist.* 47.2 (2019), pp. 691–719.

[PM19b] A. Petersen and H.-G. Müller. “Wasserstein covariance for multiple random densities”. In: *Biometrika* 106.2 (2019), pp. 339–351.

[Pol90] D. Pollard. *Empirical processes: theory and applications*. Vol. 2. NSF-CBMS Regional Conference Series in Probability and Statistics. Institute of Mathematical Statistics, Hayward, CA; American Statistical Association, Alexandria, VA, 1990, pp. viii+86.

[R D08] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2008.

[RW98] R. Rockafellar and R. J.-B. Wets. *Variational Analysis*. Heidelberg, Berlin, New York: Springer Verlag, 1998.

[Sch19a] C. Schötz. *Arbitrary Rates of Convergence for Projected and Extrinsic Means*. 2019. arXiv: `1910.11223 [math.ST]`.

[Sch19b] C. Schötz. “Convergence rates for the generalized Fréchet mean via the quadruple inequality”. In: *Electron. J. Stat.* 13.2 (2019), pp. 4280–4345.

[Sch20a] C. Schötz. *Regression in Nonstandard Spaces with Fréchet and Geodesic Approaches*. 2020. arXiv: `2012.13332 [math.ST]`.

[Sch20b] C. Schötz. *Strong Laws of Large Numbers for Generalizations of Fréchet Mean Sets*. 2020. arXiv: `2012.12762 [math.PR]`.

[SHS10] F. Steinke, M. Hein, and B. Schölkopf. “Nonparametric regression between general Riemannian manifolds”. In: *SIAM J. Imaging Sci.* 3.3 (2010), pp. 527–563.

[SO20] H.-Y. Shin and H.-S. Oh. *Robust Geodesic Regression*. 2020. arXiv: `2007.04518 [stat.ML]`.

[Ste56] H. Steinhaus. “Sur la division des corps matériels en parties”. In: *Bull. Acad. Polon. Sci. Cl. III.* 4 (1956), 801–804 (1957).

[Stu02] K.-T. Sturm. “Nonlinear martingale theory for processes with values in metric spaces of nonpositive curvature”. In: *Ann. Probab.* 30.3 (2002), pp. 1195–1222.

[Stu03] K.-T. Sturm. “Probability measures on metric spaces of nonpositive curvature”. In: *Heat kernels and analysis on manifolds, graphs, and metric spaces (Paris, 2002)*. Vol. 338. Contemp. Math. Amer. Math. Soc., Providence, RI, 2003, pp. 357–390.

[Sve81] H. Sverdrup-Thygeson. “Strong law of large numbers for measures of central tendency and dispersion of random variables in compact metric spaces”. In: *Ann. Statist.* 9.1 (1981), pp. 141–145.

[Tal14] M. Talagrand. *Upper and lower bounds for stochastic processes*. Vol. 60. Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics]. Modern methods and classical problems. Springer, Heidelberg, 2014, pp. xvi+626.

[Tsy08] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. 1st. Springer Publishing Company, Incorporated, 2008.

[Vil03] C. Villani. *Topics in optimal transportation*. Vol. 58. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2003, pp. xvi +370.

[Vil09] C. Villani. *Optimal transport*. Vol. 338. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Old and new. Springer-Verlag, Berlin, 2009, pp. xxii+973.

[VW96] A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. With applications to statistics. Springer-Verlag, New York, 1996, pp. xvi+508.

[Wan+12] Y. Wang, J. S. Marron, B. Aydin, A. Ladha, E. Bullitt, and H. Wang. "A nonparametric regression model with tree-structured response". In: *J. Amer. Statist. Assoc.* 107.500 (2012), pp. 1272–1285.

[WJ87] R. Williamson and L. Janos. "Constructing metrics with the Heine-Borel property". In: *Proc. Amer. Math. Soc.* 100.3 (1987), pp. 567–573.

[WM07] H. Wang and J. S. Marron. "Object oriented data analysis: sets of trees". In: *Ann. Statist.* 35.5 (2007), pp. 1849–1873.

[Yok17] T. Yokota. "Convex functions and p-barycenter on CAT(1)-spaces of small radii". In: *Tsukuba J. Math.* 41.1 (2017), pp. 43–80.

[Zie77] H. Ziezold. "On expected figures and a strong law of large numbers for random elements in quasi-metric spaces". In: *Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes and of the Eighth European Meeting of Statisticians (Tech. Univ. Prague, Prague, 1974), Vol. A*. Springer Netherlands, 1977, pp. 591–602.

[ZP19] Y. Zemel and V. M. Panaretos. "Fréchet means and Procrustes analysis in Wasserstein space". In: *Bernoulli* 25.2 (2019), pp. 932–976.