Exponential Concentration for Median-of-means Estimators on NPC spaces

by

Ho Yun

A Dissertation
submitted in fulfillment of the requirement
for the degree of
Master of Science
in
Statistics

The Department of Statistics
College of Natural Sciences
Seoul National University
February, 2022

윤호 석사학위논문

Exponential Concentration for Median-of-means Estimators on NPC Spaces

2022년 2월

서울대학교 대학원
통계학과 통계학전공
윤 호

Abstract

Ho Yun
The Department of Statistics
The Graduate School
Seoul National University

In Euclidean spaces, the empirical mean vector as a mean estimator has polynomial concentration unless a strong tail assumption is imposed. The idea of median-of-means tournament has been considered as a way of robustification for the empirical mean vector. In this paper, to address the sub-optimal performance of the empirical mean in a more general setting, we consider general Polish spaces with a general metric, which are allowed to be non-compact and of infinite-dimension. We discuss the estimation of the associated population Fréchet mean, and for this we extend the existing notion of median-of-means to this general setting. We devise several new notions and inequalities associated with the geometry of the underlying metric, and using them we show that the new estimators achieve exponential concentration under only second moment condition on the underlying distribution, while the empirical Fréchet mean has polynomial concentration. We focus our study on spaces with non-positive Alexandrov curvature since they afford slower rates of convergence than spaces with positive curvature.

Keywords : concentration inequalities, NPC spaces, non-Euclidean geometry, median-of-means, Fréchet mean, power transform metric.

Student Number : 2018-28485

Contents

1 Introduction **1**

2 Assumptions **6**

3 Empirical Fréchet Means **12**

4 Consideration of Assumptions **15**

4.1 Common choice $\eta = d^2$ 16

4.2 Cases with $\eta = d^\alpha$ 17

4.3 Metric entropy 21

4.4 Wasserstein space 24

5 Geometric-Median-of-Means **25**

5.1 Common choice $\eta = d^2$ 29

5.2 Cases with $\eta = d^\alpha$ 32

6 Discussion **37**

7 Proofs **38**

7.1 Proofs of theorems in Chapter 3 38

7.2 Proofs of propositions in Chapter 4 44

7.3 Proofs of theorems in Chapter 5 47
7.4 Additional proposition 50

Reference **52**

Abstract in Korean **58**

Chapter 1

Introduction

The notion of Fréchet mean extends the definition of mean, as a center of probability distribution, to metric space settings. Given a Borel probability measure P on a metric space $(\mathcal{M}, d)$ and a functional $\eta : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$, the *Fréchet mean* (or the barycenter) [Fré48] of P is any x^* such that

$$x^* \in \operatorname*{arg\,min}_{x \in \mathcal{M}} \int_{\mathcal{M}} \eta(x, y) \, \mathrm{d}P(y). \tag{1.1}$$

This accords with the usual definition of the Euclidean mean for $\mathcal{M} = \mathbb{R}^D$ when $\eta(x, y) = d(x, y)^2 = |x - y|^2$. In this paper, we consider the estimation of the Fréchet mean driven from a heavy-tailed distribution. Our problem is to find estimators that have better non-asymptotic accuracy than the *empirical Fréchet mean*

$$x_n \in \operatorname*{arg\,min}_{x \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} \eta(x, X_i) \tag{1.2}$$

when P is heavy-tailed on $\mathcal{M}$. Our main results assertively answer this question for global *non-positive curvature* (NPC) spaces, also called CAT(0) or Hadamard spaces, that are of finite- or infinite-dimension.

Our coverage with NPC spaces is genuinely broad enough. It includes

Hilbert spaces with Euclidean spaces as a special case, and various other types of metric spaces, some of which are listed below.

- A hyperbolic space $\mathcal{H}_D$ has constant non-positive sectional curvature, which results in rich geometrical features with explicit expression for the log and exp maps. The deviation of two geodesics in a hyperbolic space accelerates while drifting away from the origin, which allows a natural hierarchical structure in neural networks [GBH18; TBG18].

- The space $\mathcal{S}_D^+$ of symmetric positive definite matrices has non-constant and non-positive sectional curvature, which appears frequently in diffusion tensor imaging [Fil+05; Fil+07]. The space $\mathcal{S}_D^+$ is not only a Riemannian manifold, but also an Abelian Lie group with additional algebraic structure [Ars+07; Lin19; PSF19]. Thus, structural modeling is allowed for random elements taking values in $\mathcal{S}_D^+$ [LMP21].

- The Wasserstein space $\mathcal{P}_2(\mathbb{R})$ over $\mathbb{R}$ has vanishing Alexandrov curvature [Klo10] and plays a fundamental role in optimal transport [Vil09]. The Wasserstein space has rich applications in modern theories, such as change point detection [HKW21], and Wasserstein regression [CLM21; ZKP22; GP21].

Apart from the above-mentioned examples, there are other NPC spaces that are of great importance in application, such as phylogenetic trees [PSF19; BHV01] and Euclidean buildings [Rou04].

A great deal of statistical inference is fundamentally based on the estimation of the Fréchet mean x^*. While classical statistics leaned toward asymptotic behavior of estimators, the derivation of non-asymptotic probability bounds, called *concentration or tail inequalities*, has drawn increasing attention recently. For an estimator $\hat{x} = \hat{x}(X_1, \ldots, X_n)$ of x^*, concentration inequalities

for $\hat{x}$ are given in the form of

$$\mathbb{P}\left(d(\hat{x}, x^*) \leq r(n, \Delta)\right) \geq 1 - \Delta, \tag{1.3}$$

where $r(n, \Delta)$ is the radius of concentration corresponding to a tail probability level Δ whose dependence on n is typically determined by the metric-entropy of $\mathcal{M}$. There have been only a few attempts to establish such concentration inequalities when $(\mathcal{M}, d)$ is not Euclidean, and all of them have been restricted to the empirical Fréchet mean $\hat{x} = x_n$, to the best of our knowledge. For $\mathcal{M} = \mathbb{R}^D$, it is widely known that the empirical mean x_n is sub-optimal achieving only *polynomial concentration* for heavy-tailed P in the sense that $\Delta^{-1} = f(n, r(n, \Delta))$ for some f with $f(n, r)$ for fixed n being a polynomial function of r.

A solution to alleviating the sub-optimality of the empirical mean x_n is to partition $\{X_1, \ldots, X_n\}$ into a certain number of blocks and then take a 'median' of the within-block sample means. The above idea of robustification against heavy-tailed distribution, while inheriting the efficiency of the empirical mean for light-tailed distribution, was first introduced by Nemirovsky and Yudin [NY83]. When $\mathcal{M} = \mathbb{R}$, the resulting estimator, termed as *median-of-means*, achieves the concentration inequality (1.3) with $r(n, \Delta) = C \times n^{-1/2}\sqrt{\log(1/\Delta)}$ for some constant $C > 0$ [Cat12; Dev+16]. Lugosi [LM19] extended the result to $\mathcal{M} = \mathbb{R}^D$ by developing the idea of 'median-of-means tournament'. Both results establish *exponential concentration* for the median-of-means estimator $\hat{x}$ in the sense that $\Delta^{-1} = f(n, r(n, \Delta))$ with $f(n, r)$ for fixed n being an exponential function of r. All of the above-mentioned works, however, treated Euclidean spaces for $\eta = d^2$ with extensive use of the associated inner product. Hsu [HS16] treated arbitrary metric spaces for $\eta = d^2$. However, the latter work does not use the geometric features of the underlying

metric space but assumes certain high-level conditions. The conditions include the existence of an estimator $\hat{x}$ and a random distance $DIST$ on $(\mathcal{M}, d)$ such that $\mathbb{P}\left(d(\hat{x}, x^*) \leq \varepsilon\right) \geq 2/3$ for some $\varepsilon > 0$ and $\mathbb{P}(d(x,y)/2 \leq DIST(x,y) \leq 2d(x,y)) \geq 8/9$ for all $x, y \in \mathcal{M}$.

In this paper, we first extend the notion of median-of-means to general metric spaces $\mathcal{M}$. Then, we address the problem of robust estimation by taking into account the *metric geometry* of the underlying space. To this end, we use the *CN, quadruple and variance inequalities*, which are not well known in statistics, instead of inner product. We show that, when $\mathcal{M}$ is an NPC space and $\eta(x,y) = d(x,y)^\alpha$, the corresponding *geometric-median-of-means* estimator achieves exponential concentration for all $\alpha \in (1,2]$, under only the second moment condition $\mathbb{E}\, d(x^*, X)^2 < +\infty$. In particular, for the treatment of the 'bridging' case where $\alpha \in (1,2)$, we introduce a further extended notion of the geometric-median-of-means, for which we devise generalized versions of the CN and variance inequalities. Our work is the first that provides with concentration inequalities for median-of-means type estimators with explicit constants, when η is not necessarily d^2 or $\mathcal{M}$ is a possibly infinite-dimensional non-Euclidean space.

We work with (possibly non-compact) NPC spaces for the geometric-median-of-means estimators since the Fréchet mean x_n has poor performance in such spaces. In fact, the concentration properties of x_n depend heavily on the compactness and curvature of $\mathcal{M}$. For general Polish spaces, an exponential concentration inequality may be established with x_n if the space is compact [ACLGP20]. For non-compact geodesic spaces, however, only polynomial concentration is possible with x_n unless a strong assumption on the tail of P is imposed. The latter was proved for Euclidean spaces, a special case of non-compact spaces [Cat12]. As for the curvature of the underlying space, x_n has a

poorer rate of convergence for $\mathcal{M}$ with non-positive curvature than with positive curvature (Chapter 3 and Section 4.3). Curvature and compactness are related in case $\mathcal{M}$ is a Riemannian manifold. The Bonnet-Myers theorem states that, if the sectional curvature of a Riemannian manifold is bounded from below by $\kappa > 0$, then $\operatorname{diam}(\mathcal{M}) \leq \pi/\sqrt{\kappa}$ so that it is compact. To complement the existing works for x_n, we demonstrate the polynomial concentration of x_n, as well, for general Polish spaces in Chapter 3, and for NPC spaces as a specialization of the latter in Section 4.2. We note that there have been few works on non-asymptotic theory of x_n for non-Euclidean $\mathcal{M}$, although its asymptotic theory has been widely studied [BP03; BP05; LGL17; SCG03]. The work in Chapter 3 for the empirical Fréchet mean x_n paves our way for developing the main results in Chapter 5 for the geometric-median-of-means estimators.

Our treatment of NPC spaces relies on the metric geometry of the underlying space $\mathcal{M}$, rather than on the differential geometry of $\mathcal{M}$. Consequently, the radius of concentration $r(n, \Delta)$ in the exponential inequalities in Chapter 5 does not involve any term related to the structure of the tangential vector space of $\mathcal{M}$, which corresponds to Σ_X in Lugosi [LM19] when $\mathcal{M} = \mathbb{R}^D$. We find that assuming $\mathbb{E}\, d(x^*, X)^2 < +\infty$ is enough to deduce the exponential concentration. The flexibility inherent in our framework thus allows our work to serve as the basic constituent for a wide range of principal methods for non-Euclidean data. In particular, the theoretical development achieved in this paper may be adapted to the robustification of various recent Fréchet regression techniques [LMP21; CLM21; ZKP22; GP21; PM19].

Chapter 2

Assumptions

In this chapter, we present main structures of the underlying metric, where we base our theory, and key assumptions on the entropy of the underlying space. The validity of the assumptions will be discussed in Chapter 4.

Let $(\mathcal{M}, d)$ be a separable and complete metric space (Polish space). Consider the set of all probability measures on $\mathcal{M}$ denoted by $\mathcal{P}(\mathcal{M})$. Let P be a probability measure with finite second moment, i.e.

$$P \in \mathcal{P}_2(\mathcal{M}) := \Big\{ P \in \mathcal{P}(\mathcal{M}) : \int_{\mathcal{M}} d(x,y)^2 \, \mathrm{d}P(y) < +\infty \ \text{ for some } \ x \in \mathcal{M} \Big\}.$$

We note that, if $\int_{\mathcal{M}} d(x,y)^2 \, \mathrm{d}P(y) < +\infty$ for some $x \in \mathcal{M}$, then it holds for all $x \in \mathcal{M}$. Let $\eta : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ be a measurable function. Throughout this paper, we assume that there exists $x^* \in \mathcal{M}$ at (1.1) for the underlying measure P. Let $X_1, X_2, \ldots, X_n$ be the i.i.d. observations of a random element X governed by a probability measure P, and P_n be its empirical probability measure. Then, the empirical Fréchet mean x_n at (1.2) can be written as

$$x_n \in \operatorname*{arg\,min}_{x \in \mathcal{M}} \int_{\mathcal{M}} \eta(x,y) \, \mathrm{d}P_n(y).$$

To analyze the deviation of x_n from x^* by making use of the difference of their η-functional values, we introduce two assumptions on $P \in \mathcal{P}_2(\mathcal{M})$:

(A1) *Quadruple inequality*: There is a nonnegative function $l : \mathcal{M} \times \mathcal{M} \to [0, +\infty)$, called *growth function*, such that, for any $y, z, p, q \in \mathcal{M}$,

$$l(y, z) = 0 \;\Leftrightarrow\; y = z,$$
$$\eta(y, p) - \eta(y, q) - \eta(z, p) + \eta(z, q) \leq 2l(y, z) \cdot d(p, q).$$

(A2) *Variance inequality*: There exist constants $K > 0$ and $\beta \in (0, 2)$ such that, for all $x \in \mathcal{M}$,

$$l\left(x, x^*\right)^2 \leq K \left(\int_{\mathcal{M}} \left(\eta(x, y) - \eta\left(x^*, y\right)\right) \mathrm{d}P(y) \right)^{\beta}.$$

We note that (A1) and (A2) together imply the *uniqueness* of the Fréchet mean x^*.

Example 1. Consider the case where $\mathcal{M}$ is a Hilbert space $\mathcal{H}$ with an inner product $\langle \cdot, \cdot \rangle$ and $d(x, y) = \|x - y\|$ for the induced norm $\| \cdot \|$ of $\langle \cdot, \cdot \rangle$. Let $\eta = d^2$. If X has finite second moment, i.e. $\mathbb{E}\, d(x^*, X)^2 < +\infty$, then $x^* = \mathbb{E}X$ is the *unique* barycenter of X in the sense of Bochner integration. Also, it holds that

$$\begin{aligned}
&\eta(y, p) - \eta(y, q) - \eta(z, p) + \eta(z, q) \\
&\quad = (2\langle y - q, q - p\rangle + \|q - p\|^2) - (2\langle z - q, q - p\rangle + \|q - p\|^2) \\
&\quad = 2\langle y - z, q - p\rangle \leq 2\|y - z\| \cdot \|p - q\|.
\end{aligned}$$

Thus, (A1) holds with $l = d$. Moreover, (A2) is satisfied with equality holding always for all $x \in \mathcal{M}$ with $K = \beta = 1$:

$$\begin{aligned}
\mathbb{E}\left(\eta(x, X) - \eta\left(x^*, X\right)\right) &= \mathbb{E}\left(2\langle x^* - X, x - x^*\rangle + \|x - x^*\|^2\right) \\
&= 2\langle x^* - \mathbb{E}X, x - x^*\rangle + \|x - x^*\|^2 = \|x - x^*\|^2. \qquad \square
\end{aligned}$$

For curved spaces, the inequality in (A2) may be satisfied, but with equality not holding always for all $x \in \mathcal{M}$ in general, contrary to the Hilbetian case. Moreover, both x_n and x^*, defined in the format of M-estimation, do not have a closed form expression for curved metric spaces. Therefore, in order to derive a concentration inequality for x_n, we need an inequality that gives an upper bound to the discrepancy $l\left(x_n, x^*\right)$ between x_n and x^*. The variance inequality (A2) implies that $l\left(x_n, x^*\right)$ can be controlled by the positive function $\eta(x_n, \cdot) - \eta\left(x^*, \cdot\right)$, called the *empirical excess risk* of η:

$$l\left(x_n, x^*\right)^2 \leq K \left(\int_{\mathcal{M}} \left(\eta(x_n, y) - \eta\left(x^*, y\right)\right) \mathrm{d}P(y) \right)^{\beta}. \tag{2.1}$$

For the usual choice $\eta = d^2$, it turns out that (A1) and (A2) hold with $l = d, K = \beta = 1$ for general NPC spaces $\mathcal{M}$, see Section 4.1 for details.

Bounding the right hand side of (2.1) with a high probability depends on the geometric properties of the class of functions $\eta(x, \cdot) - \eta\left(x^*, \cdot\right)$ for $x \in \mathcal{M}$. It turns out that the dependence is through the *centered functional* η_c defined by $\eta_c(x, \cdot) = \eta(x, \cdot) - \int_{\mathcal{M}} \eta(x, y) \mathrm{d}P(y)$. Put $f_\eta(x, \cdot) = \eta_c(x, \cdot) - \eta_c(x^*, \cdot)$, $x \in \mathcal{M}$.

Definition 1. *For $\delta \geq 0$,*

$$\begin{aligned}
&\mathcal{M}_\eta(\delta) = \left\{ x \in \mathcal{M} : \int_{\mathcal{M}} \left(\eta(x, y) - \eta\left(x^*, y\right)\right) \mathrm{d}P(y) \leq \delta \right\}, \\
&\mathcal{F}_\eta(\delta) = \{ f_\eta(x, \cdot) : x \in \mathcal{M}_\eta(\delta) \}, \\
&\sigma_\eta^2(\delta) = \sup \left\{ \int_{\mathcal{M}} f_\eta(x, y)^2 \, \mathrm{d}P(y) : x \in \mathcal{M}_\eta(\delta) \right\}.
\end{aligned}$$

Example 2. Consider the η and X in Example 1. Let $\Sigma_X : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ be the covariance operator of X defined by $\Sigma_X(x, y) = \mathbb{E}\left(\langle x, X - x^* \rangle \langle y, X - x^* \rangle\right)$ and λ_{max} be its largest eigenvalue. From Example 1, it is straightforward to

see that

$$
\begin{aligned}
\mathcal{M}_\eta(\delta) &= B(x^*, \sqrt{\delta}), \\
\mathbb{E}\,\eta(x, X) &= \mathbb{E}\,\eta(x^*, X) + \|x - x^*\|^2 = \mathrm{tr}(\Sigma_X) + \|x - x^*\|^2, \\
\eta_c(x, y) &= \eta(x, y) - \mathbb{E}\eta(x, X) = \|x - y\|^2 - \|x - x^*\|^2 - \mathrm{tr}(\Sigma_X), \\
f_\eta(x, y) &= \eta_c(x, y) - \eta_c(x^*, y) \\
&= \|x - y\|^2 - \|x^* - y\|^2 - \|x - x^*\|^2 = 2\langle x - x^*, x^* - y\rangle, \\
\|f_\eta(x, \cdot) - f_\eta(y, \cdot)\|_{2,P}^2 &= 4\,\mathbb{E}\,\left(\langle x - y, X - x^*\rangle^2\right) = 4\,\Sigma_X(x - y, x - y),
\end{aligned}
$$

where $B(x, r)$ denotes the ball centered at x with radius r, and $\|f\|_{2,P}^2 = \mathbb{E}\,f(X)^2$. Note that $f_\eta(x, \cdot) : \mathcal{H} \to \mathbb{R}$ is an affine function and $f_\eta(x^*, \cdot) \equiv 0 \equiv f_\eta(\cdot, x^*)$. Also, from the Cauchy-Schwarz inequality, we have

$$
\begin{aligned}
\sigma_\eta^2(\delta) &= \sup\left\{4\mathbb{E}(\langle x - x^*, X - x^*\rangle^2) : x \in B(x^*, \sqrt{\delta})\right\} \\
&= \sup\left\{4\Sigma_X(x - x^*, x - x^*) : x \in B(x^*, \sqrt{\delta})\right\} \\
&= 4\delta \cdot \lambda_{max}. \qquad \square
\end{aligned}
$$

Under the assumptions (A1) and (A2), it holds that

$$
\begin{aligned}
&\sup_{x \in \mathcal{M}_\eta(\delta)} f_\eta(x, y) \\
&= \sup_{x \in \mathcal{M}_\eta(\delta)} \int_{\mathcal{M}} \left(\eta(x, y) - \eta(x^*, y) - \eta(x, z) + \eta(x^*, z)\right)\, \mathrm{d}P(z) \\
&\le 2 \sup_{x \in \mathcal{M}_\eta(\delta)} \int_{\mathcal{M}} l(x, x^*) d(y, z)\, \mathrm{d}P(z) \\
&\le 2\sqrt{K\delta^\beta} \int_{\mathcal{M}} d(y, z)\, \mathrm{d}P(z) =: H_{\delta,\eta}(y).
\end{aligned} \tag{2.2}
$$

By definition $H_{\delta,\eta}$ *envelops* the class $\mathcal{F}_\eta(\delta)$ of functions under the assumptions (A1) and (A2). Let $\|\cdot\|_{2,P_n}$ be defined by

$$
\|f\|_{2,P_n}^2 = n^{-1} \sum_{i=1}^n f(X_i)^2, \quad f : \mathcal{M} \to \mathbb{R}.
$$

Note that $\|\cdot\|_{2,P_n}$ is a pseudo metric. To analyze high probability concentration, toward zero, of the right hand side of (2.1), we consider the following assumption on the $\|\cdot\|_{2,P_n}$-metric entropy of $\mathcal{M}$.

(B1) *Finite-dimensional* $\mathcal{M}$: There are some constants $A, D > 0$ such that, for any $\delta > 0$ and $n \in \mathbb{N}$,

$$N\left(\tau\|H_{\delta,\eta}\|_{2,P_n}, \mathcal{F}_\eta(\delta), \|\cdot\|_{2,P_n}\right)) \leq \left(\frac{A}{\tau}\right)^D, \quad 0 < \tau \leq 1.$$

The constant D in the assumption (B1) is related to the index of VC(Vapnik-Červonenkis)-type class of functions, which appears frequently in M-estimation. According to the common definition [GN21], $\mathcal{F}_\eta(\delta)$ is of VC-type with respect to $H_{\delta,\eta}$ if

$$\sup_{Q\in\mathcal{P}(\mathcal{M})} N\left(\tau\|H_{\delta,\eta}\|_{2,Q}, \mathcal{F}_\eta(\delta), \|\cdot\|_{2,Q}\right)) \leq \left(\frac{A}{\tau}\right)^{D_{\mathrm{vc}}} \tag{2.3}$$

for some constants $A, D_{\mathrm{vc}} > 0$. The constant D_{vc}, termed as *VC index*, may not equal the dimension of $\mathcal{M}$ in general, but is usually larger, and (2.3) implies (B1) with $D = D_{\mathrm{vc}}$, the latter being what we actually need in our framework. Because of the implication, $\mathcal{F}_\eta(\delta)$ with (B1) may be regarded as a *weak* VC-type class of functions, and D as a *weak* VC index. In Proposition 3 given later in Chapter 4 we show that (B1) holds with $D = \dim(\mathcal{M})$ in case $\mathcal{M}$ is an NPC space with $\dim(\mathcal{M}) < +\infty$ and $\eta = d^2$.

For infinite-dimensional scenarios, we make the following assumption on the geometric complexity of $\mathcal{M}$.

(B2) *Infinite-dimensional* $\mathcal{M}$: There are some constants $A, \gamma > 0$ such that, for any $\delta > 0$ and $n \in \mathbb{N}$,

$$\log N\left(\tau\|H_{\delta,\eta}\|_{2,P_n}, \mathcal{F}_\eta(\delta), \|\cdot\|_{2,P_n}\right)) \leq \frac{A}{\tau^{2\gamma}}, \quad 0 < \tau \leq 1.$$

The constant γ describes how quickly the covering number grows as τ decreases. For probability measures P with non-compact support, the complexity constant depends largely on the curvature of $\mathcal{M}$. Here and throughout the paper, 'curvature' means sectional curvature for Riemannian manifolds, and Alexandrov curvature for general metric spaces. In case $\eta = d^2$, we get that $\gamma = 1$ for Hilbert spaces $\mathcal{M}$, $\gamma \leq 1$ for geodesic spaces with positive curvature, and $\gamma \geq 1$ for geodesic spaces with non-positive curvature, see Section 4.3. Based on this, we call γ the *curvature complexity* of $\mathcal{M}$.

Chapter 3

Empirical Fréchet Means

In this chapter, we present two theorems that establish polynomial concentration for empirical Fréchet means under the assumptions (A1), (A2), (B1) and (B2) in the case where $\mathcal{M}$ is a general Polish space. The theorems are used in developing exponential concentration for geometric-median-of-means estimators to be introduced in Chapter 5. Throughout this chapter, we assume that P has finite second moment, i.e., $\sigma_X^2 := \mathbb{E}\, d(x^*, X)^2 < +\infty$.

Theorem 1. *Assume (A1), (A2) and (B1), and let (K, β) and (A, D) be the constant pairs that appear in (A2) and (B1), respectively. Then, for all $n \in \mathbb{N}$ and $\Delta \in (0, 1)$,*

$$l(x_n, x^*) \leq C_\Delta \cdot \left(\frac{\sigma_X}{\sqrt{n}}\right)^{\frac{\beta}{2-\beta}}$$

with probability at least $1 - \Delta$, where C_Δ is given by

$$C_\Delta = K^{\frac{1}{2-\beta}} \left\{ 32 \left(24\sqrt{AD} + \sqrt{\frac{2}{\Delta}} \right) \right\}^{\frac{\beta}{2-\beta}}.$$

In the case where $\mathcal{M}$ is an NPC space to be introduced in the next chapter, choosing $\eta = d^2$ gives $l = d$ and $K = \beta = 1$, see Section 4.1. In this case,

Theorem 1 provides an upper bound of order $\sigma_X/\sqrt{n\Delta}$ for $d(x_n, x^*)$. Note that, in the trivial case where $\mathcal{M} = \mathbb{R}^D$ with $d(x, y) = |x - y|$, an application of the Chebyshev inequality gives

$$\mathbb{P}\left(|x_n - x^*| \leq \frac{\sigma_X}{\sqrt{n\Delta}}\right) \geq 1 - \Delta.$$

Here and throughout this paper, $|\cdot|$ denotes the Euclidean norm. We pay extra factors in C_Δ related to the complexity of $\mathcal{M}$ to deal with general metric spaces. The following theorem is for infinite-dimensional scenarios with the assumption (B2).

Theorem 2. *Assume (A1), (A2) and (B2), and let (K, β) and (A, γ) be the constant pairs that appear in (A2) and (B2), respectively. Then, there is a universal constant $C_{A,\gamma}$ depending only on $A > 0$ and $\gamma > 0$ such that, for all $n \in \mathbb{N}$ and $\Delta \in (0, 1)$,*

$$l(x_n, x^*) \leq \begin{cases} K^{\frac{1}{2-\beta}} \left(C_{A,\gamma} \cdot \dfrac{1}{n^{1/2}} \cdot \dfrac{\sigma_X}{\sqrt{\Delta}}\right)^{\frac{\beta}{2-\beta}}, & \text{if } 0 < \gamma < 1 \\ K^{\frac{1}{2-\beta}} \left(C_{A,1} \cdot \dfrac{\log n}{n^{1/2}} \cdot \dfrac{\sigma_X}{\sqrt{\Delta}}\right)^{\frac{\beta}{2-\beta}}, & \text{if } \gamma = 1 \\ K^{\frac{1}{2-\beta}} \left(C_{A,\gamma} \cdot \dfrac{1}{n^{1/2\gamma}} \cdot \dfrac{\sigma_X}{\sqrt{\Delta}}\right)^{\frac{\beta}{2-\beta}}, & \text{if } \gamma > 1 \end{cases}$$

with probability at least $1 - \Delta$.

An explicit form of the constant $C_{A,\gamma}$ in Theorem 2 may be found in the proof of the theorem in Section 7.1. The theorem demonstrates that the consistency of the empirical Fréchet mean x_n remains to hold for infinite-dimensional $(\mathcal{M}, d)$, but with slower rates of convergence to x^* for increasing n when $\gamma \geq 1$, compared to the finite-dimensional case in Theorem 1. It tells that, for infinite-dimensional geodesic spaces $\mathcal{M}$, decreasing the curvature of $\mathcal{M}$ results in slowing down the rate of convergence of x_n to x^* since the curvature complexity γ

gets larger as the curvature decreases. This implies that the rate is slower for $\mathcal{M}$ with non-positive curvature than with positive curvature. We note that, for the finite-dimensional case, the rate of convergence of x_n does not depend on the curvature, as is shown in Theorem 1. The constant A in C_Δ, however, gets larger as the curvature of $\mathcal{M}$ decreases in case $\mathcal{M}$ is a Riemannian manifold and $\eta = d^2$, see Section 4.3.

Theorems 1 and 2 reveal that, for fixed n, the empirical Fréchet mean achieves only polynomial concentration speeds. In Chapter 5 we discuss in depth alternative estimators that afford exponential speeds, basically replacing $1/\Delta$ by $\log(1/\Delta)$ in the concentration inequalities.

Chapter 4

Consideration of Assumptions

In this section, we discuss the validity of the assumptions (A1), (A2), (B1) and (B2) for non-positive curvature (NPC) spaces. We also derive generalized versions of the CN and variance inequalities.

Definition 2. *A Polish space $(\mathcal{M}, d)$ is called an (global) NPC space if for any $x_0, x_1 \in \mathcal{M}$, there exists $y \in \mathcal{M}$ such that*

$$d(z,y)^2 \leq \frac{1}{2}d(z,x_0)^2 + \frac{1}{2}d(z,x_1)^2 - \frac{1}{4}d(x_0,x_1)^2, \quad z \in \mathcal{M}.$$

Example 3. Any Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ is an NPC space: for any $x_0, x_1, z \in \mathcal{M}$

$$\begin{aligned}
&\frac{1}{2}d(z,x_0)^2 + \frac{1}{2}d(z,x_1)^2 - \frac{1}{4}d(x_0,x_1)^2 \\
&\qquad = \frac{1}{4}\left(2\|z-x_0\|^2 + 2\|z-x_1\|^2 - \|(z-x_0)-(z-x_1)\|^2\right) \\
&\qquad = \frac{1}{4}\|(z-x_0)+(z-x_1)\|^2 \\
&\qquad = d\left(z, \frac{x_0+x_1}{2}\right)^2. \qquad \square
\end{aligned}$$

Throughout this section, $\mathcal{M}$ is an NPC space. Also, when there is no confusion, with an abuse of terminology, 'Riemannian manifold' means a smooth,

complete and connected finite-dimensional Riemannian manifold. By the Hopf-Rinow Theorem, such Riemannian manifold is geodesically complete.

4.1 Common choice $\eta = d^2$

Let us first discuss some properties of NPC spaces when $\eta(x, y) = d(x, y)^2$. The geometry of metric measure spaces with non-positive curvature is mainly developed by Sturm [SCG03]. Note that the existence and uniqueness of the Fréchet mean for any probability measure are guaranteed for such spaces.

We have seen in Example 1 that, for Hilbert spaces, the inner product structure allows us to easily verify (A1) and the equality in (A2) with $l = d$, $K = \beta = 1$. For curved spaces, however, $d(x, y)^2 - d(x^*, y)^2$ cannot be expressed nicely, thus our assumptions (A1) and (A2) may not be easy to check. For example, for Riemannian manifolds $\mathcal{M}$, the relationship between the embedded distance $\|\log_p x - \log_p y\|$ for $p, x, y \in \mathcal{M}$ and the original distance $d(x, y)$ depends considerably on the curvature, see Remark 1 below. Nevertheless, using the fact that the geodesic deviation accelerates as two geodesics move further away from the origin, one may prove the following inequalities for global NPC spaces $\mathcal{M}$, see [SCG03] for details.

CN inequality: For any $y \in \mathcal{M}$ and for any geodesic $\gamma : [0, 1] \to \mathcal{M}$,

$$d(\gamma_t, y)^2 \leq (1 - t)d(\gamma_0, y)^2 + t\, d(\gamma_1, y)^2 - t(1 - t)d(\gamma_0, \gamma_1)^2, \quad t \in [0, 1].$$

Quadruple inequality: For any $y, z, p, q \in \mathcal{M}$,

$$d(y, p)^2 - d(z, p)^2 - d(y, q)^2 + d(z, q)^2 \leq 2d(y, z)d(p, q).$$

Variance inequality: For any $x \in \mathcal{M}$ and for any $P \in \mathcal{P}_2(\mathcal{M})$,

$$d\left(x, x^*\right)^2 \leq \int_{\mathcal{M}} \left(d(x, y)^2 - d\left(x^*, y\right)^2\right) \mathrm{d}P(y).$$

Here, 'CN' stands for Courbure Négative in French. Therefore, not only for Hilbert spaces but also for NPC spaces, our assumptions (A1) and (A2) are satisfied with $K = \beta = 1$, $l = d$ for the usual choice $\eta(x, y) = d(x, y)^2$.

Remark 1. *We note that $\eta = d^2$ satisfies the Hamilton-Jacobi equation, see (14.29) in [Vil09], and the homogeneous Taylor polynomial of order 4 for η gives the following formula: for any $p \in \mathcal{M}$ and $v, w \in T_p\mathcal{M}$,*

$$d\left(\exp_p(tv), \exp_p(tw)\right)^2 = \|v - w\|^2 \cdot t^2 - \frac{1}{3}\,\mathrm{Riem}(v, w, w, v) \cdot t^4 + O(t^5),$$

where 'Riem' stands for the Riemannian curvature tensor.

4.2 Cases with $\eta = d^\alpha$

Here, we consider the choice $\eta = d^\alpha$, or equivalently $\eta = d_\alpha^2$ with $d_\alpha = d^{\alpha/2}$, for $\alpha \in (1, 2]$. We note that the Fréchet mean x^* corresponding to $\alpha = 1$ is analogous to the conventional median for $\mathcal{M} = \mathbb{R}$, thus is often called *Fréchet median*. We exclude the case $\alpha = 1$ in our discussion, however, for the reason to be given shortly. We also note that d_α is a metric for $\alpha \in (1, 2]$, and is often called *power transform metric*. The associated Fréchet mean is called *α-power Fréchet mean*. With a slight abuse of notation we continue to denote it by x^* throughout this paper.

Fig. 4.1 illustrates the α-power Fréchet means for several $\alpha \in [1, 2]$ when $\mathcal{M} = \mathbb{R}^2$, $d(x, y) = |x - y|$ and P has the equal probability mass $1/3$ at three points $a_1 = (0, h)$, $a_2 = (-\sqrt{3}, 0)$, $a_3 = (\sqrt{3}, 0)$. The right panel depicts t in $x^* = (0, t)$ as a function of h. For $\alpha = 2$, $x^* = (a_1 + a_2 + a_3)/3 = (0, h/3)$ becomes most sensitive to the change of $a_1 = (0, h)$ from a certain point on the scale of h. For $\alpha = 1$, $x^* = \arg\min_{x \in \mathbb{R}^2} \overline{xa_1} + \overline{xa_2} + \overline{xa_3}$, known as the *Fermat point*, is invariant for $h \geq 1$. As the cases $\alpha = 1.1$ and $\alpha = 1.5$ demonstrate, x^*

for $\alpha \in (1,2)$ is resistent to outlying $a_1 = (0,h)$ to a certain extent depending on α: the smaller α is, the more it resists.

Fig. 4.1 also indicates that all α-power Fréchet means for different values of α meet at $(0,1)$ when $a_1 = (0,3)$. This is not a coincidence. Proposition 5 in the Supplement shows that, if the underlying probability measure P is invariant under rotation around a point z, then z is the unique α-power Fréchet mean for all $\alpha \geq 1$.

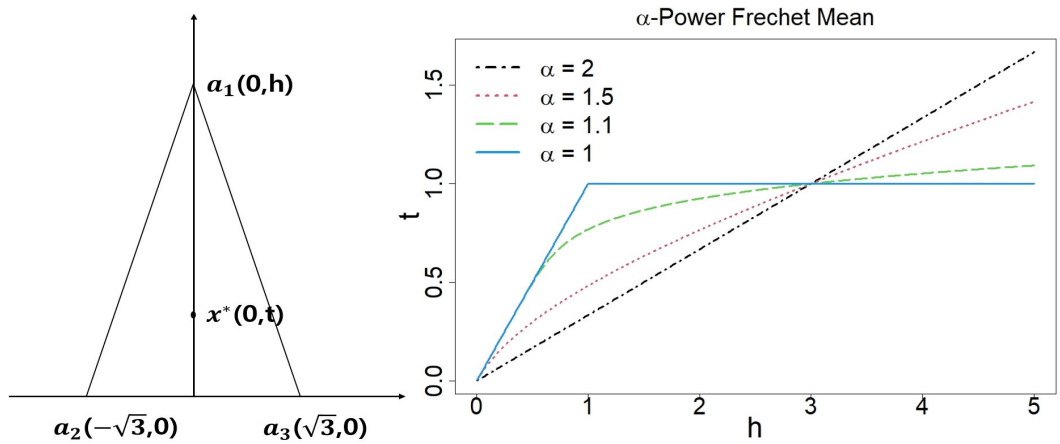

Figure 4.1: The left panel depicts the positions of the α-power Fréchet mean x^* and the three points a_1, a_2, a_3 having equal mass. The right panel shows the change of x^* as a_1 moves with a_2 and a_3 staying fixed, for $\alpha = 1/1.1/1.5/2$ (solid/dashed/dotted/dot-dashed).

The rates of convergence for α-power Fréchet means are studied for NPC spaces with $\alpha \in [1,2]$ in Schötz [Sch19]. In the latter work it is proved that the assumption (A1) holds with $l(\cdot,\cdot) = \alpha 2^{-\alpha+1} d(\cdot,\cdot)^{\alpha-1}$: for any $y, z, p, q \in \mathcal{M}$,

$$d(y,p)^\alpha - d(z,p)^\alpha - d(y,q)^\alpha + d(z,q)^\alpha \leq \alpha 2^{-\alpha+2} d(y,z)^{\alpha-1} d(p,q), \quad \alpha \in [1,2]. \tag{4.1}$$

Moreover, according to Appendix E in [Sch19], no growth function satisfying (A1) exists for $\alpha > 2$ and $0 < \alpha < 1$. For $\alpha = 1$, (4.1) implies (A1) with the growth function $l(y, z) = I(y \neq z)$, but with this the assumption (A2) makes no sense, so that Theorems 1 and 2 are not meaningful for $\eta = d$. For the case where $\alpha = 1$, [Bač14a] provided some results analogous to Theorems 1 and 2. [Bač14b], [Bač18] also introduced stochastic proximal point algorithms (PPA) to compute Fréchet medians in NPC spaces.

We are not aware of any types of CN or variance inequalities for the power transform metric in the literature. In the next two propositions we derive generalized CN and variance inequalities for $\alpha \in (1, 2]$. Thus, the theorems in Chapter 3 remain valid for α-power Fréchet means as well.

Proposition 1 (Power transform CN inequality)**.** *Let $\gamma : [0, 1] \to \mathcal{M}$ be a geodesic and $\alpha \in [1, 2]$. Then, it holds that, for any $\delta \geq 0$, $t \in [0, 1]$ and $z \in \mathcal{M}$,*

$$d(\gamma_t, z)^\alpha \leq (1+\delta)^{1-\alpha/2} \left[(1-t)^{\alpha/2} d(\gamma_0, z)^\alpha + t^{\alpha/2} d(\gamma_1, z)^\alpha\right] - \delta^{1-\alpha/2} \left[t(1-t) d(\gamma_0, \gamma_1)^2\right]^{\alpha/2}.$$

Our result in Proposition 1 reduces to the CN inequality in Section 4.1 when $\alpha = 2$. It is believed to be a sharp generalization since it is derived from the CN inequality in Section 4.1 and a version of Hölder's inequality, both of which are sharp. With given three points $x, y, z \in \mathcal{M}$, Proposition 1 enables us to get an upper bound for the power transform metric $\eta(\cdot, z) = d(\cdot, z)^\alpha$ along the geodesic from x to y, which does not seem to be feasible for general η. We will illustrate how to use this inequality in a concrete way in the proof of the following proposition, and also in the proofs of the concentration inequalities given in Theorems 5 and 6 later in Section 5.2.

To state the second proposition, for $\alpha > 0$ we let

$$\mathcal{P}_\alpha(\mathcal{M}) := \Big\{ P \in \mathcal{P} : \int_{\mathcal{M}} d(x,y)^\alpha \mathrm{d}P(y) < +\infty \ \text{ for some } \ x \in \mathcal{M} \Big\}.$$

For $P \in \mathcal{P}_\alpha(\mathcal{M})$, define $F_\alpha(\cdot) = \int_{\mathcal{M}} d(\cdot, y)^\alpha \mathrm{d}P(y)$ and

$$b_\alpha(x) = \sup_{t \in (0,1]} \frac{F_\alpha(\gamma_t^x) - \left\{ t^{\alpha/2} + (1-t)^{\alpha/2} \right\} F_\alpha(x^*)}{t^{\alpha/2} d(x,x^*)^\alpha}, \quad x \in \mathcal{M} \setminus \{x^*\},$$

where $\gamma^x : [0,1] \to \mathcal{M}$ is the geodesic from x^* to x.

Proposition 2 (Power transform variance inequality)**.** *Let $\alpha \in [1,2]$ and $P \in \mathcal{P}_\alpha(\mathcal{M})$. If $b_\alpha(x) > 0$, then*

$$d\left(x, x^*\right)^\alpha \leq \frac{1}{b_\alpha(x)} \int_{\mathcal{M}} \left(d(x,y)^\alpha - d\left(x^*, y\right)^\alpha \right) \mathrm{d}P(y), \quad x \in \mathcal{M} \setminus \{x^*\}.$$

Therefore, if

$$B_\alpha := \inf_{x \in \mathcal{M} \setminus \{x^*\}} b_\alpha(x) > 0, \tag{4.2}$$

then for any $x \in \mathcal{M}$,

$$d\left(x, x^*\right)^\alpha \leq \frac{1}{B_\alpha} \int_{\mathcal{M}} \left(d(x,y)^\alpha - d\left(x^*, y\right)^\alpha \right) \mathrm{d}P(y).$$

Proposition 2 tells that, in order to establish the power transform variance inequality, it suffices to check that, for all $x \in \mathcal{M} \setminus \{x^*\}$, $F_\alpha(\gamma_t^x)$ gets apart from $\left(t^{\alpha/2} + (1-t)^{\alpha/2}\right) F_\alpha(x^*)$ by more than a positive constant multiple of $t^{\frac{\alpha}{2}} d(x,x^*)^\alpha$, at some point γ_t^x along the geodesic from x^* to x. Note that $F_\alpha(x^*) = \inf_{x \in \mathcal{M}} F_\alpha(x)$ and $t^{\alpha/2} + (1-t)^{\alpha/2} \geq 1$ for all $t \in [0,1]$. For the common choice $\eta = d^2$, i.e. $\alpha = 2$, it follows from the (power transform) CN inequality that, for any $x \in \mathcal{M} \setminus \{x^*\}$,

$$b_2(x) = \sup_{t \in (0,1]} \frac{F_2(\gamma_t^x) - F_2(x^*)}{t \cdot d(x,x^*)^2} \geq \sup_{t \in (0,1]} \frac{t^2 \cdot d(x,x^*)^2}{t \cdot d(x,x^*)^2} = 1.$$

Thus, we may take $B_2 = 1$ in this case and the proposition gives the usual variance inequality in Section 4.1. For $\eta = d^\alpha$ with $\alpha \in (1,2]$ in general, if $P \in \mathcal{P}_\alpha(\mathcal{M})$ satisfies the condition (4.2), then (A1) and (A2) hold with $l(y,z) = \alpha 2^{-\alpha+1} d(y,z)^{\alpha-1}, K = \alpha^2 2^{-2\alpha+2} B_\alpha^{-2+2/\alpha}$ and $\beta = 2 - 2/\alpha \in (0,1]$. Thus, in this general case as well, Theorems 1 and 2 hold under the entropy conditions (B1) and (B2), respectively. The theorems give that

$$\mathbb{P}\left(d(x_n, x^*) \leq 64 \Big(\frac{K_\alpha^{\alpha/2}}{\alpha} \Big)^{1/(\alpha-1)} \cdot \Big(24\sqrt{AD} + \sqrt{\frac{2}{\Delta}} \Big) \cdot \frac{\sigma_X}{\sqrt{n}} \right) \geq 1 - \Delta \quad (4.3)$$

for finite-dimensional NPC spaces $\mathcal{M}$ and

$$\mathbb{P}\left(d(x_n, x^*) \leq 2 \Big(\frac{K_\alpha^{\alpha/2}}{\alpha} \Big)^{1/(\alpha-1)} \cdot C_{A,\gamma} \cdot \rho_n \cdot \frac{\sigma_X}{\sqrt{\Delta}} \right) \geq 1 - \Delta \quad (4.4)$$

for infinite-dimensional cases, where $K_\alpha = \alpha^2 2^{-2\alpha+2} B_\alpha^{-2+2/\alpha}$ and

$$\rho_n = \begin{cases} n^{-1/2} & \textit{if} \;\; 0 < \gamma < 1 \\ n^{-1/2} \cdot \log n & \textit{if} \;\; \gamma = 1 \\ n^{-1/2\gamma} & \textit{if} \;\; \gamma > 1. \end{cases}$$

Note that the concentration rates in terms of Δ and n in (4.3) and (4.4) do not depend on $\alpha \in (1,2]$.

4.3 Metric entropy

VC-type classes appear frequently in the study of empirical processes. Our assumption (B1) on the complexity of $\mathcal{M}$ in terms of the random entropy is crucial for the derivation of non-asymptotic concentration properties of x_n. It gives universal non-stochastic bounds to the random entropies $N(\tau, \mathcal{F}_\eta(\delta), \|\cdot\|_{2,P_n})$. The calculation of the (weak) VC index D in (B1), i.e. the uniform control of

the random covering numbers, is difficult in many cases (see Section 7.2 in [VH14]). A common technique to obtain D is to exploit the combinatorial structure of the class of functions, provided that it is a VC subgraph class of functions, see [BLM13; GN21; VH14] and references therein. However, with a more explicit assumption (B1′) given below, which essentially characterizes the dimension of the underlying spaces, we may calculate directly the (weak) VC index without combinatorial notions of complexity such as shattering.

(B1′) There are some constants $A_1, D_1 > 0$ such that, for any $\tau \in (0, r]$,

$$N(\tau, B(x^*, r), d) \leq \left(\frac{A_1 r}{\tau}\right)^{D_1}.$$

Proposition 3. *Let $\eta = d^\alpha$ with $1 < \alpha \leq 2$. Assume (A2) and (B1′). Then (B1) holds with $A = A_1^{\alpha-1}$ and $D = D_1/(\alpha - 1)$:*

$$N\left(\tau \|H_{\delta,\eta}\|_{2,P_n}, \mathcal{F}_\eta(\delta), \|\cdot\|_{2,P_n}\right) \leq \left(\frac{A_1}{\tau^{1/(\alpha-1)}}\right)^{D_1}, \quad 0 < \tau \leq 1.$$

In particular, when $\eta = d^2$ where (A2) is satisfied, (B1′) alone implies (B1) with $A = A_1$ and $D = D_1$.

Considering that the VC index D_{vc} introduced in Chapter 2 is usually larger than the dimension D_1 of the underlying space $\mathcal{M}$, the second result in Proposition 3 is striking as it states that the (weak) VC index D equals D_1 in our framework when $\eta = d^2$. It is noteworthy that the right hand side of the inequality in Proposition 3 does not involve any term related to δ. This can be interpreted as that the growth of $\|H_{\delta,\eta}\|_{2,P_n}$ counterbalances the increasing complexity of the class $\mathcal{F}_\eta(\delta)$ as δ gets larger.

When $\mathcal{M}$ is a Riemannian manifold and $\eta = d^\alpha$ with $\alpha \in (1, 2]$, the constant A in (B1) is indispensably related to the *volume control problem*, which is one

of the fundamental problems in geometry. Indeed, the constant A_1 in (B1′) for a Riemannian manifold depends on how fast the volume of a ball grows as its radius increases, which relies on the sectional (or Ricci) curvature of $\mathcal{M}$. The Bishop-Günther inequality gives an upper bound to the volume change in terms of the sectional curvature, see Theorem 3.101 (ii) in [GHL90]. For the reversed inequality, named as the Bishop-Gromov inequality, see [Vil09]. Because of these inequalities, A_1 thus A in (B1) becomes smaller as the curvature of $\mathcal{M}$ increases when $\eta = d^\alpha$ with $\alpha \in (1, 2]$.

To encompass infinite-dimensional cases, we made another complexity assumption (B2) in Chapter 2. The following proposition demonstrates that, when $(\mathcal{M}, d)$ is a Hilbert space and $\eta = d^2$, the constants A and γ in (B2) turn out to be calculated explicitly as $A = 1/32$ and $\gamma = 1$.

Proposition 4. *Let $\mathcal{M}$ be a Hilbert space and $\eta = d^2$ with $d(x, y) = \|x - y\|$. Then, for any probability measure $P \in \mathcal{P}_2(\mathcal{M})$,*

$$\log N\left(\tau \|H_{\delta,\eta}\|_{2,P}, \mathcal{F}_\eta(\delta), \|\cdot\|_{2,P}\right) \leq \frac{1}{32\tau^2}, \quad 0 < \tau \leq 1.$$

Furthermore, for the empirical measure P_n, it holds that

$$\log N\left(\tau \|H_{\delta,\eta}\|_{2,P_n}, \mathcal{F}_\eta(\delta), \|\cdot\|_{2,P_n}\right) \leq \frac{1}{32\tau^2}, \quad 0 < \tau \leq 1.$$

Proposition 4 may be used to verify (B2) with $\eta = d^2$ for Riemannian manifolds $(\mathcal{M}, d)$. Note that $d(x, y) \leq \|\log_p x - \log_p y\|$ for $\mathcal{M}$ with non-negative curvature, while $d(x, y) \geq \|\log_p x - \log_p y\|$ for $\mathcal{M}$ with non-positive curvature, i.e. for Hadamard manifolds. By embedding $\mathcal{M}$ into the tangent space $T_{x^*}\mathcal{M}$ and applying Proposition 4 to $T_{x^*}\mathcal{M}$, one may argue that (B2) is satisfied with some $\gamma \leq 1$ for Riemannian manifolds with non-negative curvature, and with some $\gamma \geq 1$ for Hadamard manifolds. In fact, γ in (B2), termed as curvature complexity, can be made smaller as the curvature of $\mathcal{M}$

gets larger. The latter follows from the *Toponogov comparison theorem*: the larger is the sectional curvature of an underlying space $\mathcal{M}$, the slower becomes the acceleration of the deviation between two geodesics emanating from a single point.

4.4 Wasserstein space

For a separable Banach space $(\mathcal{X}, \|\cdot\|)$, $\mathcal{P}_2(\mathcal{X})$ is called *Wasserstein space* and can be written as

$$\mathcal{P}_2(\mathcal{X}) = \{\mu \in \mathcal{P}(\mathcal{X}) : \int_{\mathcal{X}} \|x\|^2 \mathrm{d}\mu(x) < \infty\},$$

where $\mathcal{P}(\mathcal{X})$ denotes the set of all probability measures on $\mathcal{X}$. The Wasserstein space $\mathcal{P}_2(\mathcal{X})$ is equipped with the *Wasserstein distance*

$$W_2(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} \|x - y\|^2 \mathrm{d}\pi(x, y) \right)^{1/2}, \quad \mu, \nu \in \mathcal{P}_2(\mathcal{X})$$

where $\Pi(\mu, \nu)$ denotes the family of all probability measures on $\mathcal{M} \times \mathcal{M}$ with marginals μ and ν.

The Wasserstein space $\mathcal{P}_2(\mathcal{X})$ for a general Banach space $\mathcal{X}$ has non-negative Alexandrov curvature at any probability measure $\mu \in \mathcal{P}_2(\mathcal{X})$ that is absolutely continuous with respect to all non-degenerate Gaussian measures [ACLGP20; PZ20]. For $\mathcal{X} = \mathbb{R}$, however, $\mathcal{P}_2(\mathbb{R})$ has vanishing Alexandrov curvature [Klo10]. Thus, the latter is an NPC space, and (A1) and (A2) are satisfied with $K = \beta = 1$ and $l = W_2$ for the usual choice $\eta(\mu, \nu) = W_2(\mu, \nu)^2$, see Section 4.1. For the metric entropy, $\mathcal{P}_2(\mathbb{R})$ satisfies (B2) for any $\gamma > 1/2$ (see Example 2.6 of [ACLGP20]), thus the conclusion of Theorem 2 is valid for $\mathcal{P}_2(\mathbb{R})$.

Chapter 5

Geometric-Median-of-Means

For empirical Fréchet means in non-compact metric spaces, polynomial concentration, as we derived in Chapter 3, is the best one can achieve. In this section we introduce new estimators, for which we show that they have exponential concentration in general NPC spaces. The definitions of the estimators are for general metric spaces $(\mathcal{M}, d)$ and functionals η.

Let the random sample $\{X_1, \ldots, X_n\}$ be partitioned into k disjoint and independent blocks $\mathcal{B}_1, \ldots, \mathcal{B}_k$ of size $m \geq n/k$. For each $1 \leq j \leq k$, define

$$F_{n,j}(x) = \frac{1}{m} \sum_{X_i \in \mathcal{B}_j} \eta(x, X_i). \tag{5.1}$$

For two points $a, b \in \mathcal{M}$, one may interpret $F_{n,j}(a) < F_{n,j}(b)$ as that a is 'closer' than b to the 'center' of the jth block $\mathcal{B}_j$. Indeed, in case $\mathcal{M} = \mathbb{R}^D$ and $\eta(x, y) = |x - y|^2$,

$$F_{n,j}(a) < F_{n,j}(b) \quad \text{if and only if} \quad |a - Z_j| < |b - Z_j|, \tag{5.2}$$

where Z_j in general is the sample Fréchet mean of the block $\mathcal{B}_j$ defined by

$$Z_j \in \arg\min_{x \in \mathcal{M}} F_{n,j}(x).$$

More generally, when $\mathcal{M}$ is a Hilbert space and $\eta(x,y) = \|x-y\|^2$, then $F_{n,j}(a) < F_{n,j}(b)$ is equivalent to $\|a - Z_j\| < \|b - Z_j\|$. This follows from $F_{n,j}(x) = F_{n,j}(Z_j) + \|x - Z_j\|^2$.

Definition 3. *For $a, b \in \mathcal{M}$, we say that 'a defeats b' if $F_{n,j}(a) \leq F_{n,j}(b)$ for more than $k/2$ blocks $\mathcal{B}_j$. For $x \in \mathcal{M}$, let*

$$S_x = \{a \in \mathcal{M} : a \text{ defeats } x\}, \quad r_x = \arg\min\{r > 0 : S_x \subset B(x,r)\}.$$

We call S_x the 'x-defeating region' and r_x the 'x-defeating radius'. The new estimator x_{MM} of x^* is then defined by

$$x_{MM} \in \arg\min_{x \in \mathcal{M}} r_x.$$

We call it 'geometric-median-of-means', or simply 'median-of-means' when there is no confusion.

Remark 2. *By definition, x defeats itself so that $x \in S_x$ for all $x \in \mathcal{M}$. Also, 'a defeats b' does not always imply 'b does not defeat a'. Both a and b can defeat each other, and if it happens then there exists at least one j such that $F_{n,j}(a) = F_{n,j}(b)$. Furthermore, $r_x \leq r$ if and only if any point a with $d(x,a) > r$ cannot defeat x since*

$$r_x = \max\{d(x,a) : a \in \mathcal{M} \text{ defeats } x\}.$$

In case $\mathcal{M}$ is a Euclidean space, the median-of-means may be interpreted in terms of Tukey depth, see [Hop20].

In view of (5.2), our definition of 'defeat' is a natural extension of the 'median-of-means tournament' introduced in [LM19] for $\mathcal{M} = \mathbb{R}^D$: '$a$ defeats b' if $|a - Z_j| \leq |b - Z_j|$ for more than $k/2$ blocks $\mathcal{B}_j$. We note that, for curved metric spaces, the equivalence between $F_{n,j}(a) \leq F_{n,j}(b)$ and $d(a, Z_j) \leq d(b, Z_j)$ is no

longer valid in general. Our definition in terms of $F_{n,j}(x)$ is preferable to the one based on $d(x, Z_j)$ since the latter needs the much more onerous computation of sample Fréchet means Z_j for curved spaces. Our definition dispenses with the calculation of Z_j in all competitions between two points in $\mathcal{M}$. In case $\eta : \mathcal{M} \times \mathcal{M} \to [0, +\infty)$ is continuous, the x-defeating region S_x for any $x \in \mathcal{M}$ is a closed subset of $\mathcal{M}$ containing x. This would entail that $x \mapsto r_x$ is a continuous function, from which one may argue that the minimum of r_x over $x \in \mathcal{M}$ is attained at some point in $\mathcal{M}$.

From the discussion immediately before Definition 3, one may interpret 'a defeats b' as that a is closer than b to the centers of more than half of the k blocks. The idea of minimizing the radius of defeating region is that, if x is far away from x^*, and thus from the block centers Z_j, then it is more likely that x would be defeated by some point located far from x, i.e. r_x would be large. Since x_{MM} is determined by the ordering relation based on $F_{n,j}$ rather than by the magnitudes of $F_{n,j}$ themselves, it reflects the geometric structure of η and inherits the characteristics of the Euclidean median of $Z_1, \ldots, Z_k$. Indeed, when $\mathcal{M} = \mathbb{R}$ and $\eta(x, y) = |x - y|^2$, x_{MM} in Definition 3 coincides with the usual sample median of $Z_1, \ldots, Z_k$.

To illustrate how x_{MM} works, we simulated $n = 10,000$ data points from a bivariate distribution and chose $k = 5$ for the number of blocks. In Figure 5.1 we depicted them on $[-1, 1]^2$ and also Z_j ($\bullet$) for $1 \leq j \leq 5$. The figure demonstrates that r_x, which is the radius of the smallest ball centered at $x = \blacktriangle$ covering the 'violet/sky-blue/blue' regions, tends to decrease as $x \in \mathcal{M}$ gets closer to the Fréchet mean $x^* = \blacklozenge$. To see how sensitive x_{MM} is to the change of data points, imagine that the data points in a single block changes completely to arbitrary values. This would change only one $F_{n,j}(\cdot)$ among the five, regardless how extreme the change of the data points is. Since the points a in

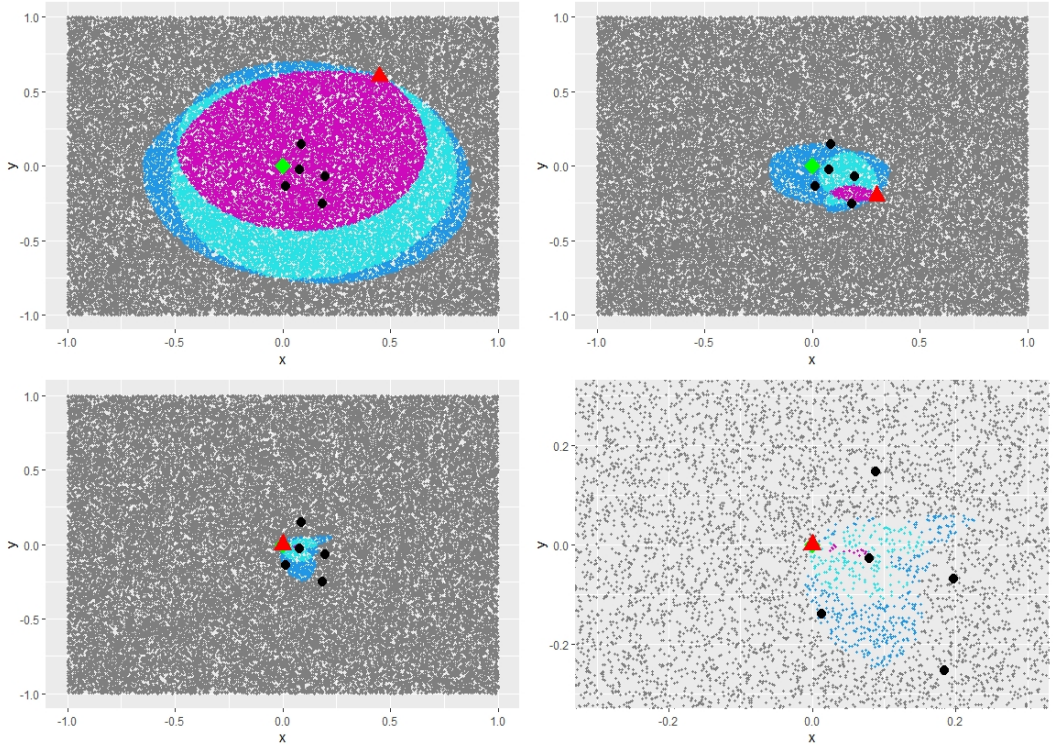

Figure 5.1: A dataset of size $n = 10,000$ was generated from a bivariate distribution on $\mathbb{R}^2$ with mean $\blacklozenge = (0, 0)$, and partitioned into $k = 5$ blocks randomly. Their sample means are depicted as $\bullet$ points. The color of each region indicates for how many blocks the points a have $F_{n,j}(a) \leq F_{n,j}(\blacktriangle)$ (5:violet, 4:sky blue, 3:blue, $\leq$2:gray). Thus, the union of violet/sky-blue/blue colored regions is the x-defeating region $S_{\blacktriangle}$. The coordinates of $\blacktriangle$ in the top-left and top-right panels, respectively, are (0.45,0.6) and (0.3,-0.2). The bottom-right panel is the zoomed-in picture of the bottom-left with $\blacktriangle = (0, 0)$, which is identical to the true mean $\blacklozenge$.

the violet and sky-blue regions, respectively, have $F_{n,j}(a) \leq F_{n,j}(\blacktriangle)$ for 5 and 4 blocks with the original dataset, they still defeat $x = \blacktriangle$ with the modified dataset. From this one may infer that there would be no significant change in the ordering of r_x across $x \in \mathcal{M}$. This consideration suggests that x_{MM} is more robust than x_n to large deviation of a few blocks, which results in x_{MM} having stronger concentration than x_n, provided that the number of blocks (k) is sufficiently large. The latter has been evidenced for $\mathcal{M} = \mathbb{R}$ by [Cat12; Dev+16] and for $\mathcal{M} = \mathbb{R}^D$ by [LM19].

In the next two subsections, we make precise the above heuristic discussion for NPC spaces with $\eta = d^\alpha$ for $\alpha \in (1, 2]$.

5.1 Common choice $\eta = d^2$

Let $X_1, \ldots, X_n$ be i.i.d. random elements taking values in an NPC space $(\mathcal{M}, d)$ with finite second moment. Here, we focus on the case $\eta = d^2$. The following theorem is essential for deriving an exponential concentration for x_{MM} when $\mathcal{M}$ is of finite dimension.

Theorem 3. *Assume (B1) with some constants $A, D > 0$. Let $\Delta \in (0, 1)$ and $q \in (0, 1/2)$. Let k denote the number of blocks $\mathcal{B}_j$. If $k = \lceil 1/(2q^2) \log(1/\Delta) \rceil$, then it holds that, with probability at least $1 - \Delta$, x^* defeats all $x \in \mathcal{M}$ with $d(x, x^*) > R_q$ but any such x does not defeat x^*, where*

$$R_q = C_q \sigma_X \sqrt{\frac{\log(1/\Delta)}{n}}, \quad C_q = \frac{32\sqrt{2}}{q} \left(24\sqrt{AD} + \frac{2}{\sqrt{1 - 2q}} \right). \tag{5.3}$$

Let $\mathcal{E}$ denote an event where, for all x with $d(x, x^*) > R_q$, x^* defeats x but x does not defeat x^*. On $\mathcal{E} \cap \{d(x_{MM}, x^*) > R_q\}$, one has $x^* \in S_{x_{MM}}$, which implies $S_{x_{MM}} \nsubseteq B(x_{MM}, R_q)$ so that $r_{x_{MM}} > R_q$. On $\mathcal{E}$, one also gets that

$x \notin S_{x^*}$ for all x with $d(x, x^*) > R_q$, which implies $S_{x^*} \subset B(x^*, R_q)$ so that $r_{x^*} \leq R_q$ on $\mathcal{E}$. By the definition of x_{MM}, it holds that $r_{x_{MM}} \leq r_{x^*}$, however. This means that

$$\mathbb{P}\big(\mathcal{E} \cap \{d(x_{MM}, x^*) > R_q\}\big) = 0.$$

The foregoing arguments gives the following corollary of Theorem 3.

Corollary 1. *Assume (B1) with some constants $A, D > 0$. Let $\Delta \in (0,1)$ and $q \in (0, 1/2)$. Let k denote the number of blocks $\mathcal{B}_j$. If $k = \lceil 1/(2q^2)\log(1/\Delta)\rceil$, then it holds that $d(x_{MM}, x^*) \leq R_q$ with probability at least $1 - \Delta$, where R_q is the constant defined at* (5.3).

The constant factor C_q in the radius of concentration R_q depends on $q \in (0, 1/2)$. Taking too small (large) q close to 0 (1/2) leads to too large (small) number of blocks k, which results in inflating the constant C_q and impairing the concentration property of x_{MM}. There is an optimal q in the interval $(0, 1/2)$ that minimizes C_q since C_q is a smooth function of $q \in (0, 1/2)$ and diverges to $+\infty$ as q approaches either to 0 or to 1/2. We note that x_{MM} with too small k is not much differentiated from the empirical Fréchet mean x_n, while with too large k the block Fréchet means Z_j would be scattered and thus there would be no guarantee that points x close to x^* have small x-defeating radius r_x.

The following theorem is for infinite-dimensional $\mathcal{M}$ and also gives an exponential concentration for x_{MM}.

Theorem 4. *Assume (B2) with some constants $A > 0$ and $\gamma \geq 1$. Let $\Delta \in (0,1)$ and $q \in (0, 1/2)$. Let k denote the number of blocks $\mathcal{B}_j$. If $k = \lceil 1/(2q^2)\log(1/\Delta)\rceil$, then it holds that, with probability at least $1 - \Delta$, x^* de-*

feats all $x \in \mathcal{M}$ with $d(x, x^*) > R_q$ but any such x does not defeat x^*, where

$$R_{q,\gamma} = \begin{cases} c_{q,1} \cdot \sigma_X \cdot \log n \cdot \sqrt{\dfrac{\log(1/\Delta)}{n}} & \text{if } \gamma = 1 \\ \\ c_{q,\gamma} \cdot \sigma_X \cdot \left(\dfrac{\log(1/\Delta)}{n}\right)^{1/2\gamma} & \text{if } \gamma > 1 \end{cases} \tag{5.4}$$

where $c_{q,\gamma} = \frac{2C_{A,\gamma}}{q\sqrt{1-2q}}$ with $C_{A,\gamma}$ appearing in Theorem 2.

Corollary 2. *Assume (B2) with some constants $A > 0$ and $\gamma \geq 1$. Let $\Delta \in (0,1)$ and $q \in (0, 1/2)$. Let k denote the number of blocks $\mathcal{B}_j$. If $k = \lceil 1/(2q^2)\log(1/\Delta)\rceil$, then it holds that $d(x_{MM}, x^*) \leq R_{q,\gamma}$ with probability at least $1 - \Delta$, where $R_{q,\gamma}$ is the constant defined at (5.4).*

As in the case of the empirical Fréchet mean x_n for infinite-dimensional $\mathcal{M}$, see (4.4), decreasing the curvature of $\mathcal{M}$ (increasing γ) results in slowing down the rate of convergence of x_{MM} to x^*. We can also make a similar remark for the dependence of the constant factor $c_{q,\gamma}$ on $q \in (0, 1/2)$ as in the discussion of Corollary 1. In the infinite-dimensional case, however, $c_{q,\gamma}$ is minimized at $q = 1/3$ regardless of the values of A and γ.

We note that the constants C_q and $c_{q,\gamma}$ in Theorems 3 and 4, respectively, may not be optimal. One might improve them by careful sharpening of various inequalities in the proofs of the theorems. Rather than optimizing the constants, we lay stress on *exponential* concentration. It is also noteworthy that our results do not involve terms such as $\mathrm{tr}(\Sigma_X)$, as opposed to the radius of concentration derived by Lugosi [LM19] for the case $\mathcal{M} = \mathbb{R}^D$, since we do not assume any differential structure for the underlying NPC space.

5.2 Cases with $\eta = d^\alpha$

Here, we consider a more general setting where $\eta = d^\alpha$ for $1 < \alpha \leq 2$. We note that the CN inequality (1) in Section 4.1 plays an important role in establishing Theorems 3 and 4. For the general case with $\eta = d^\alpha$, we use the power transform CN inequality established in Proposition 1.

The general estimators are built on the following notion of 'defeat by fraction'. The definition applies not only to $\eta = d^\alpha$ but also to a general measurable function $\eta : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$.

Definition 4. *Let ρ be a positive real number. For $a, b \in \mathcal{M}$, we say that 'a defeats b by fraction ρ' if $F_{n,j}(a) \leq \rho \cdot F_{n,j}(b)$ for more than $k/2$ blocks $\mathcal{B}_j$. For $x \in \mathcal{M}$, let*

$$\begin{aligned} S_{\rho,x} &= \{a \in \mathcal{M} : a \text{ defeats } x \text{ by fraction } \rho\}, \\ r_{\rho,x} &= \min\{r > 0 : S_{\rho,x} \subset B(x,r)\} \\ &= \max\{d(x,a) : a \in \mathcal{M} \text{ defeats } x \text{ by fraction } \rho\}. \end{aligned}$$

We call $S_{\rho,x}$ the 'x-defeating-by-ρ region' and r_x the 'x-defeating-by-ρ radius'. The estimator $x_{\rho,MM}$ of x^* is then defined by

$$x_{\rho,MM} \in \arg\min_{x \in \mathcal{M}} r_{\rho,x}.$$

We call it 'ρ-geometric-median-of-means', or simply 'ρ-median-of-means' if there is no confusion.

Clearly, the case $\rho = 1$ in the above definition coincides with Definition 3. By defintion, for any $0 < \rho_1 < \rho_2$, if a defeats b by fraction ρ_1, then a defeats b by fraction ρ_2. Therefore, for any fixed $x \in \mathcal{M}$, the x-defeating-by-ρ region $S_{\rho,x}$ increases as ρ increases, and $\rho \mapsto r_{\rho,x}$ is a monotone increasing function.

For $0 < \rho < 1$, the x-defeating-by-ρ region does not contain x since $S_{\rho,x}$ collects those points in $\mathcal{M}$ that are 'strictly better' than x. If ρ is too small, $S_{\rho,x}$ can be an empty set for some $x \in \mathcal{M}$, in which case $r_{\rho,x} = 0$. We note that the two events 'a defeats b by fraction ρ' and 'b defeats a by fraction $1/\rho$' do not complement each other, but either of the two always occurs. Both can occur simultaneously, and if so then there exists at least one j such that $F_{n,j}(a) = \rho \cdot F_{n,j}(b)$. As in the case of $\rho = 1$, the minimum of $r_{\rho,x}$ over $x \in \mathcal{M}$ is attained at some point in $\mathcal{M}$ when $\eta : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ is continuous.

To state a generalization of Theorem 3 to the case $\eta = d^\alpha$, put

$$M_{\alpha,\rho} = \sup\left\{\delta^{1-\alpha/2} t^{\alpha/2}(1-t)^{\alpha/2} : 0 < t < 1, \delta > 0, \frac{1-(1+\delta)^{1-\alpha/2}(1-t)^{\alpha/2}}{(1+\delta)^{1-\alpha/2}\, t^{\alpha/2}} \geq \rho\right\}.$$

Note that $M_{\alpha,\rho} = 1/4$ for $\alpha = 2$ and $\rho \leq 1$ since for any $0 < t < 1$ and $\delta > 0$,

$$\frac{1-(1+\delta)^{1-2/2}(1-t)^{2/2}}{(1+\delta)^{1-2/2}\, t^{2/2}} = \frac{t}{t} = 1.$$

However, for $0 < \alpha < 2$, we note that $t^{\alpha/2} + (1-t)^{\alpha/2} > 1$ for all $0 < t < 1$ and thus

$$\frac{1-(1+\delta)^{1-\alpha/2}(1-t)^{\alpha/2}}{(1+\delta)^{1-\alpha/2}\, t^{\alpha/2}} < 1 \tag{5.5}$$

for all $0 < t < 1$ and $\delta > 0$. Hence, taking $\rho \geq 1$ when $\eta = d^\alpha$ for $0 < \alpha < 2$, as (5.5) shows, would give $M_{\alpha,\rho} = \sup \emptyset = -\infty$. In fact, we find that the derivation of exponential concentration is intractable for $x_{\rho,MM}$ with $\rho \geq 1$ when $1 < \alpha < 2$, which is why we introduce the new notions of 'defeat by fraction' and 'ρ-geometric-median-of-means estimator'. Fig. 5.2 demonstrates the shapes of $M_{\alpha,\rho}$ as a function of ρ for several choices of α. It also depicts $M_{\alpha,\rho}^{-1/\alpha}$ on the log scale that appears in the constant factors in the concentration inequalities in the following theorems and corollaries.

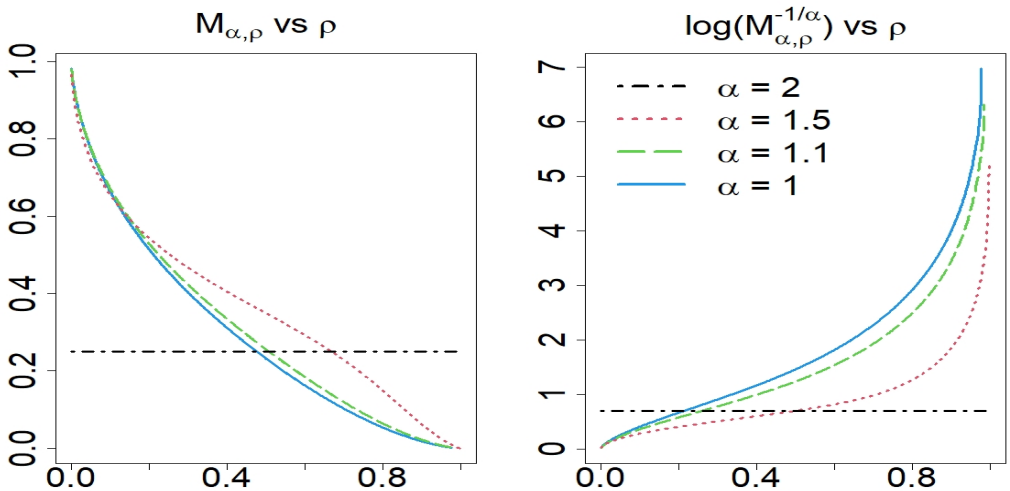

Figure 5.2: The shapes of $M_{\alpha,\rho}$ (left) and $\log M_{\alpha,\rho}^{-1/\alpha}$ (right) as functions of ρ for $\alpha = 1/1.1/1.5/2$ (solid/dashed/dotted/dot-dashed).

Theorem 5. *Assume (B1) with some constants $A, D > 0$ and that there exists a constant $B_\alpha > 0$ such that*

$$d\,(x, x^*)^\alpha \leq \frac{1}{B_\alpha} \int_M \left(d(x,y)^\alpha - d\,(x^*, y)^\alpha\right) \mathrm{d}P(y). \tag{5.6}$$

Let $\rho \in (0,1]$, $\Delta \in (0,1)$ and $q \in (0, 1/2)$. Put $K_\alpha = \alpha^2 2^{-2\alpha+2} B_\alpha^{-2+2/\alpha}$. Let k denote the number of blocks $\mathcal{B}_j$. If $k = \lceil 1/(2q^2)\log(1/\Delta)\rceil$, then it holds that, with probability at least $1-\Delta$, x^* defeats by fraction $1/\rho$ all $x \in \mathcal{M}$ with $d(x, x^*) > R_{q,\alpha,\rho}$ but any such x does not defeat x^* by fraction ρ, where

$$\begin{aligned} R_{q,\alpha,\rho} &= C_{q,\alpha,\rho}\, \sigma_X \sqrt{\frac{\log(1/\Delta)}{n}}, \\ C_{q,\alpha,\rho} &= M_{\alpha,\rho}^{-1/\alpha} \cdot \frac{16\sqrt{2K_\alpha}}{q} \left(24\sqrt{AD} + \frac{2}{\sqrt{1-2q}}\right). \end{aligned} \tag{5.7}$$

Recall that Proposition 2 gives a sufficient condition for the existence of $B_\alpha > 0$ such that (5.6) holds. Also, we note that (5.6) holds with $B_\alpha = 1$ when

$\alpha = 2$, see Section 4.1. Thus, when $\alpha = 2$ and $M_{\alpha,\rho} = 1/4$, we have $K_\alpha = 1$ so that Theorem 5 with $\rho = 1$ reduces to Theorem 3. The following corollary may be derived from Theorem 5 as Corollary 1 is from Theorem 3.

Corollary 3. *Assume the conditions in Theorem 5. Let $\rho \in (0, 1]$, $\Delta \in (0, 1)$ and $q \in (0, 1/2)$. Let k denote the number of blocks $\mathcal{B}_j$. If $k = \lceil 1/(2q^2) \log(1/\Delta) \rceil$, then it holds that $d(x_{\rho,MM}, x^*) \leq R_{q,\alpha,\rho}$ with probability at least $1 - \Delta$, where $R_{q,\alpha,\rho}$ is the constant defined at* (5.7).

The constant factor $C_{q,\alpha,\rho}$ depends on q and ρ. As in Corollary 1 for x_{MM}, it is minimized at some point $q \in (0, 1/2)$. The minimizing q depends on A and D, but is independent of α and ρ. As for the dependence on ρ, we note that $\rho \in (0, 1) \mapsto C_{q,\alpha,\rho} \in (0, +\infty)$ is an increasing function when $1 < \alpha < 2$, as is well illustrated by the right panel of Fig. 5.2. The increasing speed gets extremely fast as ρ approaches to 1. Since taking a smaller ρ shrinks the defeating regions $S_{\rho,x}$, it results in having $x_{\rho,MM}$ stay closer to x^*, which explains the result that the radius of concentration $R_{q,\alpha,\rho}$ gets smaller for smaller ρ.

Below, we present versions of Theorem 5 and Corollary 3 when $\mathcal{M}$ is of infinite-dimension satisfying the entropy condition (B2). Again, when $\alpha = 2$, we have $K_\alpha = 1$ and $M_{\alpha,\rho} = 1/4$ so that Theorem 6 with $\rho = 1$ reduces to Theorem 4.

Theorem 6. *Assume (B2) with some constants $A > 0$ and $\gamma \geq 1$ and that there exists a constant $B_\alpha > 0$ such that* (5.6) *holds. Let $\rho \in (0, 1]$, $\Delta \in (0, 1)$ and $q \in (0, 1/2)$. Put $K_\alpha = \alpha^2 2^{-2\alpha+2} B_\alpha^{-2+2/\alpha}$. Let k denote the number of blocks $\mathcal{B}_j$. If $k = \lceil 1/(2q^2) \log(1/\Delta) \rceil$, then it holds that, with probability at least $1 - \Delta$, x^* defeats by fraction $1/\rho$ all $x \in \mathcal{M}$ with $d(x, x^*) > R_{q,\alpha,\rho}$ but*

any such x does not defeat x^* by fraction ρ, where

$$R_{q,\alpha,\rho,\gamma} = \begin{cases} c_{q,\alpha,\rho,1} \cdot \dfrac{\log n}{n^{1/2}} \cdot \sigma_X \cdot \sqrt{\log \dfrac{1}{\Delta}} & \text{if } \gamma = 1 \\ \\ c_{q,\alpha,\rho,\gamma} \cdot \dfrac{1}{n^{1/2\gamma}} \cdot \sigma_X \cdot \left(\log \dfrac{1}{\Delta}\right)^{1/2\gamma} & \text{if } \gamma > 1, \end{cases} \tag{5.8}$$

$$c_{q,\alpha,\rho,\gamma} = K_\alpha^{1/2} M_{\alpha,\rho}^{-1/\alpha} \cdot \frac{C_{A,\gamma}}{q\sqrt{1-2q}}$$

and $C_{A,\gamma}$ is the constant that appears in Theorem 2.

Corollary 4. *Assume the conditions in Theorem 6. Let $\rho \in (0,1]$, $\Delta \in (0,1)$ and $q \in (0,1/2)$. Let k denote the number of blocks $\mathcal{B}_j$. If $k = \lceil 1/(2q^2)\log(1/\Delta)\rceil$, then it holds that $d(x_{\rho,MM}, x^*) \le R_{q,\alpha,\rho,\gamma}$ with probability at least $1-\Delta$, where $R_{q,\alpha,\rho,\gamma}$ is the constant defined at* (5.8).

From (4.3) and (4.4) in Section 4.2 we have observed that the concentration rates for x_n in terms of Δ and n do not depend on $\alpha \in (1,2]$. This is also the case with the geometric-median-of-means estimators x_{MM} and $x_{\rho,MM}$, which can be seen by comparing Corollaries 1 and 2 with Corollaries 3 and 4, respectively. The dependence pattern of the rate of convergence of $x_{\rho,MM}$ on γ is the same as x_n and x_{MM}. Also, the dependence of $c_{q,\alpha,\rho,\gamma}$ on ρ is the same as in the finite-dimensional case. For the dependence on q, as in the case of x_{MM}, the constant factor is minimized at $q = 1/3$ irrespective of A, γ and ρ.

Remark 3. *For NPC spaces $\mathcal{M}$ with $\eta = d^2$, the curvature complexity γ is greater than or equal to* 1*. However, γ may be $\gamma < 1$ when $1 < \alpha < 2$. In such case, one may prove that $R_{q,\alpha,\rho,\gamma}$ in Theorem 6 is given by*

$$R_{q,\alpha,\rho,\gamma} = c_{q,\alpha,\rho,\gamma} \cdot \frac{1}{n^{1/2}} \cdot \sigma_X \cdot \sqrt{\log \frac{1}{\Delta}}, \quad 0 < \gamma < 1$$

for the same constant $c_{q,\alpha,\rho,\gamma}$ given at (5.8).

Chapter 6

Discussion

Our results can be applied to any NPC spaces of finite or infinite dimension, such as Hilbert spaces, hyperbolic spaces, manifolds of SPD matrices, and the Wasserstein space $\mathcal{P}_2(\mathbb{R})$, etc. Our work is an extensive generalization of previous works on the methods of median-of-means. It is the first attempt that extends the notion of median-of-means to a general class of metric spaces with a rich class of metrics, and derives exponential concentration for the extended notions of median-of-means in such a general setting. As we discussed in this paper, we stress that the sample Fréchet mean has poor concentration for non-compact or negatively curved spaces. For such spaces, our geometric-median-of-means estimators are efficient antidotes to the sample Fréchet mean.

Chapter 7

Proofs

Throughout the Appendix, for a measurable space $(\mathcal{S}, \mathcal{B})$, a probability measure Q on $\mathcal{B}$ and a measurable function $f : \mathcal{S} \to \mathbb{R}$, we often denote $\int_{\mathcal{S}} f(y)\, \mathrm{d}Q(y)$ simply by Qf. For instance, $Pf = \mathbb{E}\left(f(X)\right)$ and $P_n f = n^{-1} \sum_{i=1}^{n} f(X_i)$. We also suppress the dependence on η of $\mathcal{M}_\eta(\delta)$ and other associated terms.

7.1 Proofs of theorems in Chapter 3

To provide an upper bound to the right hand side of (2.1) with high probability, we need a tail inequality for empirical processes. In our setup, $\|\eta(x, \cdot) - \eta\left(x^*, \cdot\right)\|_\infty$ may be unbounded as x moves. Under some strong condition on the tail of P, one may be able to obtain an exponential tail inequality, see [Ada08; GL13]. Since we assume only finite second moment of P, we use the following polynomial tail inequality.

Lemma 1 ([LVDG14])**.** *Let $X_1, \ldots, X_n$ be i.i.d. copies of X taking values in a measurable space $(\mathcal{S}, \mathcal{B})$ with probability measure P, and let $\mathcal{G}$ be a countable class of measurable functions $f : \mathcal{S} \to \mathbb{R}$ with $Pf = 0$. Put $Z =$*

$\sup_{f\in\mathcal{G}}(P-P_n)f$ *and* $\sigma^2=\sup_{f\in\mathcal{G}}Pf^2$. *Assume that the envelope* H *of the class* $\mathcal{G}$ *satisfies* $\mathbb{E}(H^p)\le M^p$ *for some* $p\ge 1$ *and* $M>0$. *Then, for any* $\varepsilon>0$, *it holds that*

$$\mathbb{P}\left(Z\ge 4\,\mathbb{E}(Z)+\varepsilon\right)\le \min_{1\le l\le p}\frac{l\cdot\Gamma(l/2)\left(\sqrt{32/n}M\right)^l}{\varepsilon^l}.$$

If $\mathbb{E}(H^2)\le M^2$, *in particular, we get that, for any* $\Delta\in(0,1)$,

$$\mathbb{P}\left(Z\le 4\,\mathbb{E}(Z)+\frac{8M}{\sqrt{n\Delta}}\right)\ge 1-\Delta.$$

Below, we present two more lemmas for the proof of the theorems. The proof of the following lemma is deferred to the Supplement. Recall the definition of $H_\delta\equiv H_{\delta,\eta}$ given at (2.2), which envelops $\mathcal{F}(\delta)\equiv\mathcal{F}_\eta(\delta)$.

Lemma 2. *Let* $\eta:\mathcal{M}\times\mathcal{M}\to\mathbb{R}$ *be a measurable function and* X *an* $\mathcal{M}$*-valued random element with Fréchet mean* x^* *and covariance* σ_X^2. *Let* $\delta>0$. *Then, under the assumptions (A1) and (A2),*

$$\sigma(\delta)\le\bar\sigma(\delta),\quad \mathbb{E}\big(H_\delta(X_1)^2\big)\le\bar\sigma(\delta)^2,\quad \mathbb{E}\left(\|H_\delta\|_{2,P_n}^2\right)\le\bar\sigma(\delta)^2,$$

where $\bar\sigma(\delta)=4\sqrt{K\sigma_X^2\delta^\beta}$.

Proof of Lemma 2. Recall the definition of H_δ at (2.2). Then,

$$\begin{aligned}\max\left\{\sigma^2(\delta),\mathbb{E}H_\delta(X_1)^2\right\}&\le 4K\delta^\beta\int_{\mathcal{M}}\left(\int_{\mathcal{M}}d(y,z)\,\mathrm{d}P(z)\right)^2\mathrm{d}P(y)\\&\le 4K\delta^\beta\int_{\mathcal{M}}\int_{\mathcal{M}}d(y,z)^2\,\mathrm{d}P(z)\,\mathrm{d}P(y)\\&\le 8K\delta^\beta\int_{\mathcal{M}}\int_{\mathcal{M}}\left(d(y,x^*)^2+d(z,x^*)^2\right)\mathrm{d}P(z)\,\mathrm{d}P(y)\\&=16K\sigma_X^2\delta^\beta.\end{aligned}$$

Now, let $X_1', \dots, X_n'$ be an independent copy of $X_1, \dots, X_n$. By the triangular inequality, it holds that

$$\begin{aligned}
\mathbb{E}\left(\|H_\delta\|_{2,P_n}^2\right) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n H_\delta(X_i)^2\right) = \frac{1}{n}\sum_{i=1}^n 4K\delta^\beta \mathbb{E}\left(\int_{\mathcal{M}} d(X_i, z)\,\mathrm{d}P(z)\right)^2 \\
&= \frac{4K\delta^\beta}{n}\sum_{i=1}^n \mathbb{E}\left(\int_{\mathcal{M}} d(X_i, z)^2\,\mathrm{d}P(z)\right) \\
&= \frac{4K\delta^\beta}{n}\sum_{i=1}^n \mathbb{E}\left(d(X_i, X_i')^2\right) = 4K\delta^\beta\,\mathbb{E}\left(d(X_1, X_1')^2\right) \\
&\le 8K\delta^\beta\,\mathbb{E}\left(d(X_1, x^*)^2 + d(X_1', x^*)^2\right) \le 16K\sigma_X^2\delta^\beta.
\end{aligned}$$

□

The following lemma provides an improved chaining bound for Gaussian processes. For a proof, see Theorem 5.31 in [VH14] or Lemma 5.1 in [ACLGP20].

Lemma 3. *Let $(X_t)_{t\in\mathcal{F}}$ be a real-valued process indexed by a pseudo metric space $(\mathcal{F}, d)$ with the following properties: (i) there exists a countable subset $\mathcal{F}' \subset \mathcal{F}$ such that $X_t = \lim_{s\to t, s\in\mathcal{F}'} X_s$ a.s. for any $t \in \mathcal{F}$; (ii) X_t is sub-Gaussian, i.e.*

$$\log \mathbb{E}\big(e^{\theta(X_s - X_t)}\big) \le \theta^2 d(s,t)^2/2$$

for any $s, t \in \mathcal{F}$ and $\theta \in \mathbb{R}$; (iii) there exists a random variable L such that $|X_s - X_t| \le L\,d(s,t)$ a.s. for all $s, t \in \mathcal{F}$. Then, for any $S \subset \mathcal{F}$ and any $\varepsilon \ge 0$, it holds that

$$\mathbb{E}\big(\sup_{t\in S} X_t\big) \le 2\,\varepsilon\,\mathbb{E}(L) + 12\int_\varepsilon^{+\infty} \sqrt{\log N(u, \mathcal{F}, d)}\,\mathrm{d}u.$$

Proof of Theorem 1. Define $\delta_n = P\left(\eta\left(x_n, .\right) - \eta\left(x^*, .\right)\right)$ and

$$\begin{aligned}
\phi_n(\delta) &= \sup\left\{(P - P_n)\left(\eta(x, .) - \eta\left(x^*, .\right)\right) : x \in \mathcal{M}(\delta)\right\} \\
&= \sup\left\{(P - P_n)\left(\eta_c(x, .) - \eta_c\left(x^*, .\right)\right) : x \in \mathcal{M}(\delta)\right\}
\end{aligned}$$

for $\delta \geq 0$. Since x_n is a minimizer of $P_n\eta(x,.)$, it follows from the definition of ϕ_n that

$$\delta_n \leq (P - P_n)\left(\eta\left(x_n,.\right) - \eta\left(x^*,.\right)\right) \leq \phi_n\left(\delta_n\right).$$

Applying Lemmas 1 and 2 we get that, with probability at least $1-(\Delta/2)$,

$$\phi_n(\delta) \leq 4\mathbb{E}\,\phi_n(\delta) + \frac{8\sqrt{2}\cdot\bar{\sigma}(\delta)}{\sqrt{n\Delta}}. \tag{7.1}$$

We first get an upper bound to $\mathbb{E}\,\phi_n(\delta)$. Let $\{\varepsilon_i\}$ be a Rademacher sequence, i.e. random signs independent of X_i's. Then, by the symmetrization of the associated empirical process (see [GN21]) we obtain

$$\begin{aligned}\mathbb{E}\,\phi_n(\delta) &\leq 2\,\mathbb{E}\left(\sup_{x\in\mathcal{M}(\delta)} n^{-1}\sum_{i=1}^{n}\varepsilon_i\left(\eta_c(x,X_i)-\eta_c(x^*,X_i)\right)\right)\\ &= 2\,\mathbb{E}\left(\sup_{x\in\mathcal{M}(\delta)} n^{-1}\sum_{i=1}^{n}\varepsilon_i\,\eta_c(x,X_i)\right).\end{aligned}$$

One can easily check that the Rademacher empirical process $\{Y_f : f \in (\mathcal{F}(\delta), \|\cdot\|_{2,P_n})\}$ for the pseudo metric space $(\mathcal{F}(\delta), \|\cdot\|_{2,P_n})$ given by

$$Y_f := \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i f(X_i)$$

is sub-Gaussian and $\sqrt{n}$-Lipschitz with respect to $\|\cdot\|_{2,P_n}$, conditionally on the X_i's. Thus, it satisfies the conditions of Lemma 3 (see [ACLGP20]). Applying Lemma 3 with (B1) and using the inequalities for H_δ given in Lemma 2, we

get

$$
\begin{aligned}
\mathbb{E}\,\phi_n(\delta) &\le 2\,\mathbb{E}\inf_{\varepsilon\ge 0}\left(2\varepsilon + \frac{12}{\sqrt{n}}\int_{\varepsilon}^{\infty}\sqrt{\log N(u,\mathcal{F}(\delta),\|\cdot\|_{2,P_n})}\mathrm{d}u\right)\\
&\le 2\,\mathbb{E}\inf_{\varepsilon\ge 0}\left(2\varepsilon + \frac{12}{\sqrt{n}}\int_{\varepsilon}^{\|H_\delta\|_{2,P_n}}\sqrt{D\log\left(\frac{A\|H_\delta\|_{2,P_n}}{u}\right)}\,\mathrm{d}u\right)\\
&= 2\,\mathbb{E}\left(\|H_\delta\|_{2,P_n}\right)\cdot\inf_{\varepsilon'\ge 0}\left(2\varepsilon' + \frac{12}{\sqrt{n}}\int_{\varepsilon'}^{1}\sqrt{D\log\left(\frac{A}{u}\right)}\,\mathrm{d}u\right) \qquad (7.2)\\
&\le 48\,\mathbb{E}\left(\|H_\delta\|_{2,P_n}\right)\cdot\sqrt{\frac{AD}{n}}\\
&\le 48\,\bar{\sigma}(\delta)\sqrt{\frac{AD}{n}},
\end{aligned}
$$

where in the third inequality we have used $\log x \le x-1 \le x$ for $x>0$.

The inequalities (7.1) and (7.2) imply that, with probability at least $1-(\Delta/2)$,

$$
\begin{aligned}
\phi_n(\delta) &\le \bar{\sigma}(\delta)\left(192\sqrt{\frac{AD}{n}} + \frac{8\sqrt{2}}{\sqrt{n\Delta}}\right)\\
&\le 32\sqrt{\frac{K\sigma_X^2\delta^\beta}{n}}\left(24\sqrt{AD} + \sqrt{\frac{2}{\Delta}}\right)\\
&=: b_n(\delta,\Delta).
\end{aligned}
$$

Since $\phi_n(\delta)$ is a increasing function and $b_n(\delta,\Delta)$ is decreasing in Δ for fixed δ, it follows from Theorem 4.3 in [Kol11] that

$$
\delta_n \le \phi_n(\delta_n) \le b_n(\Delta) := \inf\left\{\tau>0 : \sup_{\delta\ge\tau}\delta^{-1}b_n\left(\delta,\Delta^{\frac{\delta}{\tau}}\right)\le 1\right\} \qquad (7.3)
$$

with probability at least $1-\Delta$. Since $\bar{\sigma}(\delta)/\delta$ is decreasing in δ as $\beta\in(0,2)$,

$$
\sup_{\delta\ge\tau}\delta^{-1}b_n\left(\delta,\Delta^{\frac{\delta}{\tau}}\right) = \frac{b_n(\tau,\Delta)}{\tau} = 32\sqrt{\frac{K\sigma_X^2\tau^{-(2-\beta)}}{n}}\left(24\sqrt{AD} + \sqrt{\frac{2}{\Delta}}\right).
$$

This gives

$$b_n(\Delta) = \inf\left\{\tau > 0 : 32\sqrt{\frac{K\sigma_X^2\tau^{-(2-\beta)}}{n}}\left(24\sqrt{AD} + \sqrt{\frac{2}{\Delta}}\right) \le 1\right\} \tag{7.4}$$

$$= \left\{32\sqrt{\frac{K\sigma_X^2}{n}}\left(24\sqrt{AD} + \sqrt{\frac{2}{\Delta}}\right)\right\}^{\frac{2}{2-\beta}}$$

Applying (7.4) to (7.3), we obtain that, with probability at least $1 - \Delta$,

$$\begin{aligned} l(x_n, x^*) &\le \sqrt{K} \cdot \delta_n^{\beta/2} \\ &\le K^{\frac{1}{2-\beta}}\left\{32\left(24\sqrt{AD} + \sqrt{\frac{2}{\Delta}}\right)\frac{\sigma_X}{\sqrt{n}}\right\}^{\frac{\beta}{2-\beta}}. \end{aligned}$$

This completes the proof of Theorem 1. □

Proof of Theorem 2. The proof is similar to that of Theorem 1 for the case of finite-dimensional $\mathcal{M}$. The difference is in the covering number $N(u, \mathcal{F}(\delta), \|\cdot\|_{2,P_n})$. We get

$$\begin{aligned} \mathbb{E}\,\phi_n(\delta) &\le 2\,\mathbb{E}\inf_{\varepsilon\ge 0}\left(2\varepsilon + \frac{12}{\sqrt{n}}\int_\varepsilon^{\|H_\delta\|_{2,P_n}}\sqrt{\frac{A\|H_\delta\|_{2,P_n}^{2\gamma}}{u^{2\gamma}}}\,\mathrm{d}u\right) \\ &= 4\,\mathbb{E}\,(\|H_\delta\|_{2,P_n}) \cdot \inf_{\varepsilon\ge 0}\left(\varepsilon + 6\sqrt{\frac{A}{n}}\int_\varepsilon^1 u^{-\gamma}\,\mathrm{d}u\right) \\ &\le 4\,\mathbb{E}\,(\|H_\delta\|_{2,P_n}) \times \begin{cases} \dfrac{6}{1-\gamma}\sqrt{\dfrac{A}{n}} & \text{if } 0 < \gamma < 1 \\ 6\sqrt{\dfrac{A}{n}}\left(1 - \log\left(6\sqrt{\dfrac{A}{n}}\right)\right) & \text{if } \gamma = 1 \\ \dfrac{\gamma}{\gamma-1}\left(6\sqrt{\dfrac{A}{n}}\right)^{1/\gamma} & \text{if } \gamma > 1. \end{cases} \end{aligned}$$

Therefore, $\phi_n(\delta_n) \le b_n(\Delta)$ with probability at least $1-\Delta$, now with

$$b_n(\Delta) = \begin{cases} \left(32\sqrt{K\sigma_X^2}\left(\frac{12}{1-\gamma}\sqrt{\frac{A}{n}} + \sqrt{\frac{2}{n\Delta}}\right)\right)^{\frac{2}{2-\beta}} & \text{if } 0<\gamma<1 \\ \left(32\sqrt{K\sigma_X^2}\left(12\sqrt{\frac{A}{n}}\left(1-\log\left(6\sqrt{\frac{A}{n}}\right)\right) + \sqrt{\frac{2}{n\Delta}}\right)\right)^{\frac{2}{2-\beta}} & \text{if } \gamma=1 \\ \left(32\sqrt{K\sigma_X^2}\left(\frac{2\gamma}{\gamma-1}\left(6\sqrt{\frac{A}{n}}\right)^{1/\gamma} + \sqrt{\frac{2}{n\Delta}}\right)\right)^{\frac{2}{2-\beta}} & \text{if } \gamma>1. \end{cases}$$

This gives the theorem. □

7.2 Proofs of propositions in Chapter 4

Proof of Proposition 1. Since $1/2 \le \alpha/2 \le 1$, we have $a^{\alpha/2} + b^{\alpha/2} \ge (a+b)^{\alpha/2}$ for any $a, b \ge 0$, so that

$$\begin{aligned} &(1+\delta)^{1-\alpha/2}\left\{(1-t)^{\alpha/2}d(z,\gamma_0)^\alpha + t^{\alpha/2}d(z,\gamma_1)^\alpha\right\} - d(z,\gamma_t)^\alpha \\ &\quad \ge (1+\delta)^{1-\alpha/2}\left\{(1-t)d(z,\gamma_0)^2 + td(z,\gamma_1)^2\right\}^{\alpha/2} - d(z,\gamma_t)^\alpha. \end{aligned} \tag{7.5}$$

An application of Hölder's inequality gives

$$(\delta_1+\delta_2)^{1-\alpha/2}(a_1+a_2)^{\alpha/2} \ \ge\ \delta_1^{1-\alpha/2}a_1^{\alpha/2} + \delta_2^{1-\alpha/2}a_2^{\alpha/2}$$

for all $\delta_i, a_i \ge 0$ and $\alpha \in (0,2)$. The above inequality also holds for $\alpha = 0$ and 2. Applying the inequality with $\delta_1 = 1, \delta_2 = \delta, a_1 = d(z,\gamma_t)^2, a_1 + a_2 = (1-t)d(z,\gamma_0)^2 + td(z,\gamma_1)^2$ to the right hand side of the inequality at (7.5), we get

$$\begin{aligned} &(1+\delta)^{1-\alpha/2}\left((1-t)^{\alpha/2}d(z,\gamma_0)^\alpha + t^{\alpha/2}d(z,\gamma_1)^\alpha\right) - d(z,\gamma_t)^\alpha \\ &\quad \ge \delta^{1-\alpha/2}\left((1-t)d(z,\gamma_0)^2 + td(z,\gamma_1)^2 - d(z,\gamma_t)^2\right)^{\alpha/2} \\ &\quad \ge \delta^{1-\alpha/2}\left(t(1-t)d(\gamma_0,\gamma_1)^2\right)^{\alpha/2}. \end{aligned}$$

We note that $a_2 \ge 0$ and the last inequality follows from the CN inequality. □

Proof of Proposition 2. For $x \in \mathcal{M} \setminus \{x^*\}$, we apply Proposition 1 to $\delta = 0$ and γ being the geodesic $\gamma^x : [0,1] \to \mathcal{M}$ with $\gamma_0^x = x^*$ and $\gamma_1^x = x$, and then integrate both sides of the inequality with respect to z. This gives

$$(1-t)^{\alpha/2} F_\alpha(x^*) + t^{\alpha/2} F_\alpha(x) - F_\alpha(\gamma_t^x) \geq 0 \tag{7.6}$$

for any $0 \leq t \leq 1$. Take an arbitrary $\varepsilon > 0$. By the definition of $b_\alpha(x)$, it holds that, for any $x \in \mathcal{M} \setminus \{x^*\}$, there exists $t \equiv t(x) > 0$ such that

$$F_\alpha(\gamma_t^x) - \left(t^{\alpha/2} + (1-t)^{\alpha/2}\right) F_\alpha(x^*) \geq (b_\alpha(x) - \varepsilon)\, t^{\frac{\alpha}{2}}\, d(x, x^*)^\alpha. \tag{7.7}$$

From (7.6) and (7.7), it follows that

$$t^{\alpha/2} \left(F_\alpha(x) - F_\alpha(x^*)\right) \geq (b_\alpha(x) - \varepsilon) t^{\alpha/2} d(x, x^*)^\alpha,$$

so that $F_\alpha(x) - F_\alpha(x^*) \geq (b_\alpha(x) - \varepsilon) d(x, x^*)^\alpha$. Since $\varepsilon > 0$ was arbitrarily chosen, we have

$$F_\alpha(x) - F_\alpha(x^*) \geq b_\alpha(x) \cdot d(x, x^*)^\alpha,$$

which completes the proof of the proposition. □

Proof of Proposition 3. Recall $H_\delta(y) = 2\sqrt{K\delta^\beta} \int_{\mathcal{M}} d(y,z)\, \mathrm{d}P(z)$ from (2.2). Now, for $x, y \in \mathcal{M}(\delta)$,

$$\begin{aligned}
&\|f(x,\cdot) - f(y,\cdot)\|_{2,P_n}^2 \\
&\quad = \frac{1}{n} \sum_{i=1}^n \left(\eta_c(x, X_i) - \eta_c(y, X_i)\right)^2 \\
&\quad = \frac{1}{n} \sum_{i=1}^n \left(\int_{\mathcal{M}} \left(d(x, X_i)^\alpha - d(y, X_i)^\alpha - d(x,z)^\alpha + d(y,z)^\alpha\right) \mathrm{d}P(z) \right)^2 \\
&\quad \leq \frac{\alpha^2 2^{-2\alpha+4}}{n} \cdot d(x,y)^{2\alpha-2} \sum_{i=1}^n \left(\int_{\mathcal{M}} d(X_i, z) \mathrm{d}P(z) \right)^2 \\
&\quad = \frac{\alpha^2 2^{-2\alpha+2}}{K\delta^\beta} \cdot d(x,y)^{2\alpha-2} \|H_\delta\|_{2,P_n}^2,
\end{aligned}$$

where the inequality follows from (4.1). Since $l(\cdot,\cdot) = \alpha 2^{-\alpha+1} d(\cdot,\cdot)^{\alpha-1}$ and $\beta = 2 - 2/\alpha$,

$$\mathcal{M}(\delta) \subset B\left(x^*, \left(\frac{K\delta^\beta}{\alpha^2 2^{-2\alpha+2}}\right)^{\frac{1}{2(\alpha-1)}}\right).$$

Thus, it holds that

$$\begin{aligned}
&N\left(\tau\|H_\delta\|_{2,P_n}, \mathcal{F}(\delta), \|\cdot\|_{2,P_n}\right) \\
&\quad\le N\left(\left(\frac{\tau^2 K\delta^\beta}{\alpha^2 2^{-2\alpha+2}}\right)^{\frac{1}{2(\alpha-1)}}, \mathcal{M}(\delta), d\right) \\
&\quad\le N\left(\left(\frac{\tau^2 K\delta^\beta}{\alpha^2 2^{-2\alpha+2}}\right)^{\frac{1}{2(\alpha-1)}}, B\left(x^*, \left(\frac{K\delta^\beta}{\alpha^2 2^{-2\alpha+2}}\right)^{\frac{1}{2(\alpha-1)}}\right), d\right) \\
&\quad\le \left(\frac{A_1}{\tau^{1/(\alpha-1)}}\right)^{D_1}.
\end{aligned}$$

□

Proof of Proposition 4. Recall from Example 2 that $\mathcal{M}(\delta) = B(x^*, \sqrt{\delta})$, $\|f(x,\cdot) - f(y,\cdot)\|_{2,P}^2 = 4\,\Sigma_X(x-y, x-y)$. Also, $\mathcal{F}(\delta) = \left\{2\langle x - x^*, x^* - \cdot\rangle : x \in B(x^*, \sqrt{\delta})\right\}$ and $\sup_{x\in\mathcal{F}(\delta)} f(x,\cdot) = 2\sqrt{\delta}\|\cdot - x^*\|$ is the envelope of the class $\mathcal{F}(\delta)$. By Sudakov's minorisation (see Theorem 2.4.12. in [GN21] and also [Fer75] for the specified constant),

$$\log N\left(\tau, \mathcal{F}(\delta), \|\cdot\|_{2,P}\right) \le \frac{1}{8}\left(\frac{\mathbb{E}_g\left(\Sigma_X^{1/2}(g,g)\right)}{\tau/\sqrt{\delta}}\right)^2 \le \frac{\operatorname{tr}(\Sigma_X)\delta}{8\tau^2}$$

where g is a standard Gaussian random element taking values in $(\mathcal{H}, d)$. Since $\|H_\delta\|_{2,P}^2 = 4\,\delta\,\mathbb{E}\left(\langle X - x^*, X - x^*\rangle\right) = 4\,\delta\operatorname{tr}(\Sigma_X)$, we have

$$\log N\left(\tau\|H_\delta\|_{2,P}, \mathcal{F}(\delta), \|\cdot\|_{2,P}\right) \le \frac{1}{32\tau^2}$$

With the same machinery, one can also deduce the same result for the empirical measure P_n. □

7.3 Proofs of theorems in Chapter 5

Without loss of generality, we assume that $n = m \cdot k$, where k is the number of blocks in splitting the sample and m is the size of each block.

Proof of Theorem 3. Let $F(x) = \int_{\mathcal{M}} \eta(x, y)\, \mathrm{d}P(y)$. By the definition of x^* it holds that, for each block $\mathcal{B}_j$,

$$F_{n,j}(x^*) - F_{n,j}(Z_j) \leq F_{n,j}(x^*) - F_{n,j}(Z_j) - F(x^*) + F(Z_j).$$

The right hand side has an upper bound that is analogous to $\phi_n(\delta_n)$ in the proof of Theorem 1, which is obtained by substituting the empirical measure corresponding to $\mathcal{B}_j$ for P_n and Z_j for x_n. Thus, replacing Δ by $(1-2q)/2$ (so that $1-\Delta$ by $q+1/2$) and n by $m = n/k$ with $K = \beta = 1$, we get from (7.3) and (7.4) that

$$\mathbb{P}\left(F_{n,j}(x^*) - F_{n,j}(Z_j) \leq \varepsilon_{k,q}^2\right) \geq q + \frac{1}{2}, \tag{7.8}$$

where

$$\varepsilon_{k,q} = 32\sqrt{\frac{k\sigma_X^2}{n}}\left(24\sqrt{AD} + \frac{2}{\sqrt{1-2q}}\right). \tag{7.9}$$

By the CN inequality in Section 4.1, we have

$$\begin{aligned} F_{n,j}(Z_j) \leq F_{n,j}(\gamma_{1/2}^x) &\leq \frac{F_{n,j}(x)}{2} + \frac{F_{n,j}(x^*)}{2} - \frac{d(x^*, x)^2}{4} \\ \Leftrightarrow F_{n,j}(x) - F_{n,j}(Z_j) &\geq -\left(F_{n,j}(x^*) - F_{n,j}(Z_j)\right) + \frac{d(x^*, x)^2}{2}, \end{aligned}$$

where $\gamma^x : [0,1] \to \mathcal{M}$ is the geodesic with $\gamma_0^x = x^*$ and $\gamma_1^x = x$. Thus, denoting by $\mathcal{E}_{n,j}$ the event

$$F_{n,j}(x) > F_{n,j}(x^*) \quad \text{for all } \ x \in \mathcal{M} \text{ with } d(x, x^*) > 2\varepsilon_{k,q},$$

we get from (7.8) that $\mathbb{P}\left(\mathcal{E}_{n,j}\right) \geq q + 1/2$ since $F_{n,j}(x^*) - F_{n,j}(Z_j) \leq \varepsilon_{k,q}^2$ implies

$$F_{n,j}(x) - F_{n,j}(Z_j) > -\left(F_{n,j}(x^*) - F_{n,j}(Z_j)\right) + 2\varepsilon_{k,q}^2 \ \geq \ F_{n,j}(x^*) - F_{n,j}(Z_j)$$

for all x with $d(x, x^*) > 2\varepsilon_{k,q}$. By applying Høffding's inequality to $\sum_{j=1}^{k} I(\mathcal{E}_{n,j})$, we obtain

$$\begin{aligned}
1 - \Delta &\leq 1 - e^{-2q^2 k} \\
&\leq \mathbb{P}\left(\sum_{j=1}^{k} I(\mathcal{E}_{n,j}) > k/2\right) \\
&\leq \mathbb{P}\left(\sum_{j=1}^{k} I\left(F_{n,j}(x) > F_{n,j}(x^*)\right) > k/2 \ \text{ for all } x \in \mathcal{M} \text{ with } d(x, x^*) > 2\varepsilon_{k,q}\right).
\end{aligned}$$

This completes the proof of the theorem. □

Proof of Theorem 4. The proof is essentially the same as that of Theorem 3 except that we use Theorem 2 instead of Theorem 1. We obtain (7.8) now with

$$\varepsilon_{k,q} = \begin{cases} C_{A,1} \cdot \dfrac{\log(n/k)}{\sqrt{n/k}} \cdot \dfrac{\sigma_X}{\sqrt{(1-2q)/2}}, & \text{if } \gamma = 1 \\ \\ C_{A,\gamma} \cdot (k/n)^{1/2\gamma} \cdot \dfrac{\sigma_X}{\sqrt{(1-2q)/2}}, & \text{if } \gamma > 1. \end{cases} \tag{7.10}$$

Since

$$\begin{aligned}
&\frac{1}{\sqrt{2}q} \frac{\log n \sqrt{\log(1/\Delta)}}{\sqrt{n}} = \frac{\sqrt{k} \log n}{\sqrt{n}} \geq \frac{\log(n/k)}{\sqrt{n/k}}, \\
&\frac{1}{\sqrt{2}q} n^{-1/2\gamma} \left(\log \frac{1}{\Delta}\right)^{1/2\gamma} \geq \left(\frac{2q^2 n}{\log(1/\Delta)}\right)^{-1/2\gamma} = (k/n)^{1/2\gamma},
\end{aligned} \tag{7.11}$$

we get $\varepsilon_{k,q} \leq R_q/2$. The rest of the proof is the same as in the proof of Theorem 3. □

Proof of Theorem 5. First, we follow the lines leading to (7.8), now using (7.4) with $K = K_\alpha$ and $\beta = 2 - 2/\alpha$ instead of $K = \beta = 1$. We may prove

$$\mathbb{P}\left(F_{n,j}(x^*) - F_{n,j}(Z_j) \leq K_\alpha^{\alpha/2} \varepsilon_{k,q}^{\alpha}\right) \geq q + \frac{1}{2}. \tag{7.12}$$

By integrating both sides of the inequality in Proposition 1 with respect to z for $\gamma = \gamma^x : [0,1] \to \mathcal{M}$, we obtain that, for all $0 \le t \le 1$ and $\delta > 0$,

$$\begin{aligned}(1+\delta)^{1-\alpha/2}\left((1-t)^{\alpha/2}F_{n,j}(x^*) + t^{\alpha/2}F_{n,j}(x)\right) - F_{n,j}(\gamma_t^x) \\ \ge \delta^{1-\alpha/2}\left(t(1-t)d(x,x^*)^2\right)^{\alpha/2}.\end{aligned}$$

From the definition of Z_j and the above inequality, we get

$$\begin{aligned}F_{n,j}(Z_j) \le F_{n,j}(\gamma_t) \le (1+\delta)^{1-\alpha/2}\left((1-t)^{\alpha/2}F_{n,j}(x^*) + t^{\alpha/2}F_{n,j}(x)\right) \\ - \delta^{1-\alpha/2}\left(t(1-t)d(x,x^*)^2\right)^{\alpha/2}.\end{aligned}$$

This gives that, on the event where $F_{n,j}(x^*) - F_{n,j}(Z_j) \le K_\alpha^{\alpha/2}\varepsilon_{k,q}^\alpha$,

$$\begin{aligned}&(1+\delta)^{1-\alpha/2}t^{\alpha/2}F_{n,j}(x) \\ &\quad > \left(1-(1+\delta)^{1-\alpha/2}(1-t)^{\alpha/2}\right)F_{n,j}(x^*) + \left(\delta^{1-\alpha/2}t^{\alpha/2}(1-t)^{\alpha/2} - M_{\alpha,\rho}\right)\cdot\frac{K_\alpha^{\alpha/2}\varepsilon_{k,q}^\alpha}{M_{\alpha,\rho}}\end{aligned}$$

or equivalently

$$F_{n,j}(x) > \frac{1-(1+\delta)^{1-\frac{\alpha}{2}}(1-t)^{\frac{\alpha}{2}}}{(1+\delta)^{1-\frac{\alpha}{2}}t^{\frac{\alpha}{2}}}\cdot F_{n,j}(x^*) + \frac{\delta^{1-\frac{\alpha}{2}}t^{\frac{\alpha}{2}}(1-t)^{\frac{\alpha}{2}} - M_{\alpha,\rho}}{(1+\delta)^{1-\frac{\alpha}{2}}t^{\frac{\alpha}{2}}}\cdot\frac{K_\alpha^{\alpha/2}\varepsilon_{k,q}^\alpha}{M_{\alpha,\rho}}$$

for all $x \in \mathcal{M}$ with $d(x,x^*) > K_\alpha^{1/2}M_{\alpha,\rho}^{-1/\alpha}\varepsilon_{k,q}$. Thus, from (7.12) and the definition of $M_{\alpha,\rho}$ it follows that

$$\mathbb{P}\left(F_{n,j}(x) > \rho\cdot F_{n,j}(x^*) \text{ for all } x \in \mathcal{M} \text{ with } d(x,x^*) > \frac{K_\alpha^{1/2}\varepsilon_{k,q}}{M_{\alpha,\rho}^{1/\alpha}}\right) \ge q + \frac{1}{2}. \tag{7.13}$$

Applying Høffding's inequality as in the proof of Theorem 3 with (7.13), we may complete the proof of the theorem. □

Proof of Theorem 6. The proof is essentially the same as that of Theorem 5 except that we use the definition of $\varepsilon_{k,q}$ at (7.10) instead of the one at (7.9). Using (7.11) we get $K_\alpha^{1/2}M_{\alpha,\rho}^{-1/\alpha}\varepsilon_{k,q} \le R_{q,\alpha,\rho}$. □

7.4 Additional proposition

Proposition 5. *Let P be a probability measure in $\mathbb{R}^D$ and $\eta(x,y) = \psi(|x-y|)$ where $\psi : [0,+\infty) \to \mathbb{R}$ is strictly increasing and convex. Assume that there exists $z \in \mathbb{R}^D$ such that*

$$P(A) = P\left(Q(A-z)+z\right), \quad A \in \mathscr{B}(\mathbb{R}^D)$$

for an orthogonal matrix $Q \in \mathbb{R}^{D\times D}$ with $I+Q+\cdots+Q^{m-1} = 0$ for some integer $m \geq 2$. Then, z is the unique Fréchet mean with respect to $\eta : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$.

Proof of Proposition 5. We write $d(x,y) = |x-y|$. Since $Q \in \mathbb{R}^{D\times D}$ is an orthogonal matrix, it holds that, for any $x, y \in \mathbb{R}^D$,

$$\sum_{j=0}^{m-1} d(x, Q^j y) = \sum_{j=0}^{m-1} d(Qx, Q^j y) = \cdots = \sum_{j=0}^{m-1} d(Q^{m-1}x, Q^j y). \tag{7.14}$$

By (7.14) and the subadditivity of the Euclidean norm, we get

$$\begin{aligned}\sum_{j=0}^{m-1} d(x, Q^j y) &= \frac{1}{m}\sum_{l=0}^{m-1}\sum_{j=0}^{m-1} d(Q^l x, Q^j y) \\ &\geq \sum_{j=0}^{m-1} d\left(\frac{1}{m}\sum_{l=0}^{m-1} Q^l x, Q^j y\right) = \sum_{j=0}^{m-1} d(0, Q^j y).\end{aligned}$$

Now, by Jensen's inequality,

$$\frac{1}{m}\sum_{j=0}^{m-1} \psi(d(x, Q^j y)) \geq \psi\left(\frac{1}{m}\sum_{j=0}^{m-1} d(x, Q^j y)\right) \geq \psi\left(\frac{1}{m}\sum_{j=0}^{m-1} d(0, Q^j y)\right)$$

and the equality holds if and only if $x = 0$. Considering translation, for any $x, y \in \mathbb{R}^D$,

$$\sum_{j=0}^{m-1} \psi\left(d(x, Q^j(y-z)+z)\right) \geq \sum_{j=0}^{m-1} \psi\left(d(z, Q^j(y-z)+z)\right)$$

and the equality holds if and only if $x = z$. Therefore,

$$\begin{aligned}
\int_{\mathbb{R}^D} \eta(x,y)\mathrm{d}P(y) &= \frac{1}{m}\int_{\mathbb{R}^D} \sum_{j=0}^{m-1} \psi\left(d(x, Q^j(y-z)+z)\right)\mathrm{d}P(y) \\
&\geq \frac{1}{m}\int_{\mathbb{R}^D} \sum_{j=0}^{m-1} \psi\left(d(z, Q^j(y-z)+z)\right)\mathrm{d}P(y) \\
&= \int_{\mathbb{R}^D} \eta(d(z,y))\mathrm{d}P(y)
\end{aligned}$$

and the equality holds if and only if $x = z$. $\square$

References

[ACLGP20] Adil Ahidar-Coutrix, Thibaut Le Gouic, and Quentin Paris. "Convergence rates for empirical barycenters in metric spaces: curvature, convexity and extendable geodesics". In: *Probability Theory and Related Fields* 177.1 (2020), pp. 323–368.

[Ada08] Radoslaw Adamczak. "A tail inequality for suprema of unbounded empirical processes with applications to Markov chains". In: *Electronic Journal of Probability* 13 (2008), pp. 1000–1034.

[Ars+07] V. Arsigny et al. "Geometric means in a novel vector space structure on symmetric positive-definite matrices". In: *SIAM Journal of Matrix Analysis and Applications* 29 (2007), pp. 328–347.

[Bač14a] Miroslav Bačák. "Computing medians and means in Hadamard spaces". In: *SIAM Journal on Optimization* 24.3 (2014), pp. 1542–1566.

[Bač14b] Miroslav Bačák. *Convex Analysis and Optimization in Hadamard Spaces*. de Gruyter, 2014.

[Bač18] Miroslav Bačák. "Old and new challenges in Hadamard spaces". In: *arXiv preprint arXiv:1807.01355* (2018).

[BHV01] Louis J Billera, Susan P Holmes, and Karen Vogtmann. "Geometry of the space of phylogenetic trees". In: *Advances in Applied Mathematics* 27.4 (2001), pp. 733–767.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[BP03] Rabi Bhattacharya and Vic Patrangenaru. "Large sample theory of intrinsic and extrinsic sample means on manifolds". In: *The Annals of Statistics* 31.1 (2003), pp. 1–29.

[BP05] Rabi Bhattacharya and Vic Patrangenaru. "Large sample theory of intrinsic and extrinsic sample means on manifolds: II". In: *The Annals of Statistics* 33 (2005), pp. 1225–1259.

[Cat12] Olivier Catoni. "Challenging the empirical mean and empirical variance: a deviation study". In: *Annales de l'IHP Probabilités et statistiques* 48.4 (2012), pp. 1148–1185.

[CLM21] Yaqing Chen, Zhenhua Lin, and Hans-Georg Müller. "Wasserstein regression". In: *Journal of the American Statistical Association* 116 (2021), pp. 1–14.

[Dev+16] Luc Devroye et al. "Sub-Gaussian mean estimators". In: *The Annals of Statistics* 44.6 (2016), pp. 2695–2725.

[Fer75] Xavier Fernique. "Regularité des trajectoires des fonctions aléatoires gaussiennes". In: *Ecole d'Eté de Probabilités de Saint-Flour IV—1974*. Springer, 1975, pp. 1–96.

[Fil+05] Pierre Fillard et al. "Extrapolation of sparse tensor fields: Application to the modeling of brain variability". In: *Biennial International Conference on Information Processing in Medical Imaging*. Springer. 2005, pp. 27–38.

[Fil+07] Pierre Fillard et al. "Measuring brain variability by extrapolating sparse tensor fields measured on sulcal lines". In: *Neuroimage* 34.2 (2007), pp. 639–650.

[Fré48] Maurice Fréchet. "Les éléments aléatoires de nature quelconque dans un espace distancié". In: *Annales de l'institut Henri Poincaré* 10.4 (1948), pp. 215–310.

[GBH18] Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. "Hyperbolic neural networks". In: *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*. 2018.

[GHL90] Sylvestre Gallot, Dominique Hulin, and Jacques Lafontaine. *Riemannian Geometry*. Vol. 2. Springer, 1990.

[GL13] Sara van de Geer and Johannes Lederer. "The Bernstein–Orlicz norm and deviation inequalities". In: *Probability Theory and Related Fields* 157.1-2 (2013), pp. 225–250.

[GN21] Evarist Giné and Richard Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, 2021.

[GP21] Laya Ghodrati and Victor M. Panaretos. "Distribution-on-distribution regression via optimal transport maps". In: *arXiv:2104.09418* (2021).

[HKW21] Lajos Horváth, Piotr Kokoszka, and Shixuan Wang. "Monitoring for a change point in a sequence of distributions". In: *The Annals of Statistics* 49.4 (2021), pp. 2271–2291.

[Hop20] Samuel B Hopkins. "Mean estimation with sub-Gaussian rates in polynomial time". In: *The Annals of Statistics* 48.2 (2020), pp. 1193–1213.

[HS16] Daniel Hsu and Sivan Sabato. "Loss minimization and parameter estimation with heavy tails". In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 543–582.

[Klo10] Benoît Kloeckner. "A geometric study of Wasserstein spaces: Euclidean spaces". In: *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze* 9.2 (2010), pp. 297–323.

[Kol11] Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*. Vol. 2033. Springer Science & Business Media, 2011.

[LGL17] Thibaut Le Gouic and Jean-Michel Loubes. "Existence and consistency of Wasserstein barycenters". In: *Probability Theory and Related Fields* 168.3 (2017), pp. 901–917.

[Lin19] Zhenhua Lin. "Riemannian geometry of symmetric positive definite matrices via Cholesky decomposition". In: *SIAM Journal of Matrix Analysis and Applications* 40 (2019), pp. 1353–1370.

[LM19] Gábor Lugosi and Shahar Mendelson. "Sub-Gaussian estimators of the mean of a random vector". In: *The Annals of Statistics* 47.2 (2019), pp. 783–794.

[LMP21] Zhenhua Lin, Hans-Georg Müller, and Byeong U. Park. "Additive models for symmetric positive-definite matrices, Riemannian manifolds and Lie groups". In: *arXiv preprint arXiv:2009.08789* (2021).

[LVDG14] Johannes Lederer and Sara Van De Geer. "New concentration inequalities for suprema of empirical processes". In: *Bernoulli* 20.4 (2014), pp. 2020–2038.

[NY83] Arkadij Semenovič Nemirovskij and David Borisovich Yudin. "Problem Complexity and Method Efficiency in Optimization". In: (1983).

[PM19] Alexander Petersen and Hans-Georg Müller. "Fréchet regression for random objects with Euclidean predictors". In: *The Annals of Statistics* 47.2 (2019), pp. 691–719.

[PSF19] Xavier Pennec, Stefan Sommer, and Tom Fletcher. *Riemannian Geometric Statistics in Medical Image Analysis*. Academic Press, 2019.

[PZ20] Victor M. Panaretos and Yoav Zemel. *An Invitation to Statistics in Wasserstein Space*. Springer Nature, 2020.

[Rou04] Guy Rousseau. "Euclidean buildings". In: *Géométries à Courbure Négative ou Nulle, Groupes Discrets et Rigidité*. Vol. 18. Société Mathématique de France. 2004, pp. 77–116.

[SCG03] Karl-Theodor Sturm, T Coulhon, and A Grigor'yan. "Probability measures on metric spaces of nonpositive curvature". In: *Heat Kernels and Analysis on Manifolds, Graphs, and Metric Spaces, Contemporary Mathematics* **358**. American Mathematical Society (2003).

[Sch19] Christof Schötz. "Convergence rates for the generalized Fréchet mean via the quadruple inequality". In: *Electronic Journal of Statistics* 13.2 (2019), pp. 4280–4345.

[TBG18] Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. "Poincaré glove: Hyperbolic word embeddings". In: *arXiv preprint arXiv:1810.06546* (2018).

[VH14] Ramon Van Handel. *Probability in high dimension*. Tech. rep. Princeton Univ NJ, 2014.

[Vil09] Cédric Villani. *Optimal Transport: Old and New, Grundlehren der mathematischen Wissenschaften* **338**. Springer, 2009.

[ZKP22] Chao Zhang, Piotr Kokoszka, and Alexander Petersen. "Wasserstein autoregressive models for density time series". In: *Journal of Time Series Analysis* 43 (2022), pp. 30–52.

국문초록

유클리드공간에서 표본평균 벡터는 평균 추정량에 대해 오직 다항적 집중만을 가진다. 이에, 중앙값-평균(median-of-means) 추정량이 표본평균 벡터의 강건화의 일환으로 제시되었다. 본 논문에서는 비옹골(non-compact) 혹은 무한 차원 폴란드 공간(Polish space)에서 일반적인 거리가 주어졌을 때의 프레셰(Fréchet) 모평균 추정에 관한 문제를 다룬다. 이를 위해, 기존의 중앙값-평균의 정의를 확장하였고, 주어진 공간의 기하학적 성질을 반영하기 위해 고안된 개념들과 부등식을 이용하여 표본 프레셰 평균은 다항적 집중만을 가지는 반면, 본문에서 제시한 새로운 추정량은 지수적 집중을 가짐을 보인다. 특히, 본 연구는 양곡률 공간보다 수렴속도가 더 느린 음곡률 공간에 초점을 두고 있다.

주요어 : 집중 부등식, 음곡률 공간, 비유클리드 기하, 중앙값-평균, 프레셰 평균, 멱 평균 거리.

학　번 : 2018-28485

Exponential Concentration for Median-of-means Estimators on NPC Spaces

지도교수 박 병 욱

이 논문을 이학석사학위논문으로 제출함

2021년 10월

서울대학교 대학원

통계학과 통계학전공

윤 호

윤호의 석사학위논문을 인준함

2022년 1월

위 원 장 오 희 석

부 위 원 장 박 병 욱

위 원 장 원 철