Bernoulli **29**(4), 2023, 2927–2960
https://doi.org/10.3150/22-BEJ1569

Exponential concentration for geometric-median-of-means in non-positive curvature spaces

HO YUN[a] and BYEONG U. PARK[b]

Department of Statistics, Seoul National University, Seoul, South Korea, [a]*zrose0921@gmail.com,*
[b]*bupark@snu.ac.kr*

In Euclidean spaces, the empirical mean vector as an estimator of the population mean is known to have polynomial concentration unless a strong tail assumption is imposed on the underlying probability measure. The idea of median-of-means tournament has been considered as a way of overcoming the sub-optimality of 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 study the concentration properties of the extended notions of median-of-means as the estimators of the population Fréchet mean. We show that the new estimators achieve exponential concentration under only a 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. We note that this is the first work that derives non-asymptotic concentration inequalities for extended notions of the median-of-means in non-vector spaces with a general metric.

Keywords: Concentration inequalities; Fréchet mean; median-of-means estimators; non-Euclidean geometry; NPC spaces; power transform metric

1. Introduction

The notion of a 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) [21] of P is any x^* such that

$$x^* \in \operatorname*{arg\,min}_{x \in \mathcal{M}} \int_{\mathcal{M}} \eta(x, y) \, \mathrm{d}P(y). \tag{1}$$

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 of a heavy-tailed distribution. Our goal is to find estimators that have better non-asymptotic accuracy than the *empirical Fréchet mean*,

$$x_n \in \operatorname*{arg\,min}_{x \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} \eta(x, X_i) \tag{2}$$

when P is heavy-tailed on $\mathcal{M}$. The x_n is an M-estimator in a broad sense. The present work is an achievement of this goal for global *non-positive curvature* (NPC) spaces, also called CAT(0) or Hadamard spaces, that are of finite- or infinite-dimension.

1350-7265 © 2023 ISI/BS

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 due to explicit expressions 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 [23,49].
- 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 [19,20]. The space $\mathcal{S}_D^+$ is not only a Riemannian manifold, but also an Abelian Lie group with additional algebraic structure [3,37,45]. Thus, additive regression modeling is allowed for random elements taking values in $\mathcal{S}_D^+$ [38].
- The Wasserstein space $\mathcal{P}_2(\mathbb{R})$ over $\mathbb{R}$ has vanishing Alexandrov curvature [29] and plays a fundamental role in optimal transport [52]. The Wasserstein space has rich applications in modern theories, such as change point detection [27], and Wasserstein regression [14,24,54].

Apart from the above-mentioned examples, there are other NPC spaces, such as phylogenetic trees [9,45], that are of great importance in applications.

A great deal of statistical inference is fundamentally based on the estimation of the Fréchet mean x^*. While classical statistics leaned toward the 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

$$\mathbb{P}\left(d(\hat{x}, x^*) \leq r(n, \Delta)\right) \geq 1 - \Delta, \tag{3}$$

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 a linear space, 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. This robustifies the empirical mean against heavy-tailed distribution while it inherits its efficiency for light-tailed distribution. The idea was first introduced by [43]. When $\mathcal{M} = \mathbb{R}$, the resulting estimator, termed as *median-of-means*, achieves the concentration inequality (3) with $r(n,\Delta) = C \times n^{-1/2}\sqrt{\log(1/\Delta)}$ for some constant $C > 0$ [18]. The one-dimensional result was extended to $\mathcal{M} = \mathbb{R}^D$ by [39] developing the idea of 'median-of-means tournament'. The resulting estimator $\hat{x}$, also termed as median-of-means, was found to achieve a *sub-Gaussian performance*:

$$\mathbb{P}\left(\|\hat{x} - x^*\| \leq C_1\sqrt{\frac{\mathrm{tr}(\Sigma_X)}{n}} + C_2\sqrt{\frac{\|\Sigma_X\|\log(1/\Delta)}{n}}\right) \geq 1 - \Delta \tag{4}$$

for some constants $C_1, C_2 > 0$, where Σ_X is the covariance matrix and $\|\cdot\|$ is the operator norm. The concentration property at (4) is what the empirical mean achieves when X has a multivariate sub-Gaussian distribution, so the name sub-Gaussian performance. Both results establish *exponential concentration* in the sense that $\Delta^{-1} = f(n, r(n,\Delta))$ with $f(n,r)$ for fixed n being an exponential function of r. There have been also proposed several other mean estimators satisfying (4) that can be computed in linear time $(nD\log(1/\Delta))^{O(1)}$ by using the median-of-means principle, see [15,17,26,35], and other related works on robust mean estimation, e.g. [12,41].

All the aforementioned works, however, treated Euclidean spaces for $\eta = d^2$ with extensive use of the associated inner product. Apart from the Euclidean cases, there are few works for infinite-dimensional $\mathcal{M}$, e.g. [36] for a kernel-enriched domain and [42] for a Banach space, both of which considered $\eta = d^2$. We are also aware of [28] that studied the case of arbitrary metric spaces. 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 $\mathbb{P}(d(\hat{x}, x^*) \le \varepsilon) \ge 2/3$ for some $\varepsilon > 0$ and $\mathbb{P}(d(x,y)/2 \le DIST(x,y) \le 2d(x,y)) \ge 8/9$ for all $x, y \in \mathcal{M}$. We highlight that the present work is the first to use the median-of-means principle without imposing strong assumptions such as in [28] when $\mathcal{M}$ is a non-vector space. Our technical development is significantly different from the existing works in the literature. We note that there is no distribution on non-vector spaces corresponding to Gaussian or sub-Gaussian distribution on $\mathbb{R}^D$, neither are available the notions of trace and operator norm, so that an analogue of the sub-Gaussian performance as at (4) for non-vector spaces does not seem to be possible. Nevertheless, we establish for our estimators exponential concentration in the sense that the inverse of ‘probability regret’ Δ at (3) is an exponential function of the radius of concentration.

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 (‘Courbure Négative’ in French), quadruple and variance inequalities*, which are not well known in statistics, instead of the 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 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-vector 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 [2]. 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 [12]. 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 (Sections 3 and 4.3). Curvature and compactness are related in the case where $\mathcal{M}$ 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 $\operatorname{diam}(\mathcal{M}) \le \pi/\sqrt{\kappa}$ 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 Section 3, and for NPC spaces as a specialization of the latter in Section 4. 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 [7,8,32,48]. The work in Section 3 for the empirical Fréchet mean x_n paves our way for developing the main results in Section 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 Section 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 [39] 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 [14,24,38,46,54].

2. Assumptions

In this section, we present the main structures of the underlying metric on which we base our theory, and key assumptions on the entropy of the underlying space. The validity of the assumptions will be discussed in Section 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.

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

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 $x^* \in \mathcal{M}$ for which (1) holds and 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 (2) can be written as

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

To analyze the deviation of x_n from x^* by making use of the difference of their η-functional values, we introduce two assumptions, the first on $\eta : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ and the second 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 $y,z,p,q \in \mathcal{M}$,

$$\begin{aligned} &l(y,z) = 0 \iff y = z, \\ &\eta(y,p) - \eta(y,q) - \eta(z,p) + \eta(z,q) \le 2l(y,z) \cdot d(p,q). \end{aligned}$$

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

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

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

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

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

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

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 Hilbertian case. Moreover, both x_n and x^* do not have a closed form expression for curved metric spaces although x_n has for Hilbert spaces. Therefore, in order to derive a concentration inequality for x_n, we need an inequality that gives an upper bound to the discrepancy $l(x_n, x^*)$ between x_n and x^*. The variance inequality (A2) implies that $l(x_n, x^*)$ can be controlled by the positive function $\eta(x_n, \cdot) - \eta(x^*, \cdot)$, called the *empirical excess risk* of η:

$$l\left(x_n, x^*\right)^2 \le K\left(\int_{\mathcal{M}} \left(\eta(x_n, y) - \eta\left(x^*, y\right)\right) \mathrm{d}P(y)\right)^{\beta}. \tag{5}$$

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 (5) with a high probability depends on the geometric properties of the class of functions $\eta(x,\cdot) - \eta(x^*,\cdot)$ 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 $\delta \ge 0$,

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

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 $\mathcal{M}_\eta(\delta) = B(x^*, \sqrt{\delta})$ and $\mathbb{E}\,\eta(x,X) = \mathrm{tr}(\Sigma_X) + \|x - x^*\|^2$, where $B(x,r)$ denotes the ball centered at x with radius r, $\mathrm{tr}(\Sigma_X) = \sum_k \Sigma_X(e_k, e_k)$ and $\{e_k : k \ge 1\}$ is an arbitrary orthonormal basis of $\mathcal{H}$. Let $\|\cdot\|_{2,P}$ be defined by $\|f\|_{2,P}^2 = \mathbb{E}\, f(X)^2$. Then,

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

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

$$\begin{aligned}\sup_{x\in\mathcal{M}_\eta(\delta)} f_\eta(x,y) &= \sup_{x\in\mathcal{M}_\eta(\delta)}\int_{\mathcal{M}} (\eta(x,y)-\eta(x^*,y)-\eta(x,z)+\eta(x^*,z))\,\mathrm{d}P(z)\\ &\le 2\sup_{x\in\mathcal{M}_\eta(\delta)}\int_{\mathcal{M}} l(x,x^*)d(y,z)\,\mathrm{d}P(z) \qquad (6)\\ &\le 2\sqrt{K\delta^\beta}\int_{\mathcal{M}} d(y,z)\,\mathrm{d}P(z) =: H_{\delta,\eta}(y).\end{aligned}$$

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 (5), we consider the following assumption on the $\|\cdot\|_{2,P_n}$-metric entropy of $\mathcal{M}$. For a totally bounded subset $\mathcal{S}$ of a metric space $(\mathcal{F},d)$, we let $N(\tau,\mathcal{S},d)$ denote the minimal number of balls with radius τ that cover $\mathcal{S}$, and call it τ-covering number.

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

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

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 [25], $\mathcal{F}_\eta(\delta)$ is of VC-type with respect to $H_{\delta,\eta}$ if

$$\sup_{Q\in\mathcal{P}(\mathcal{M})} N\left(\tau\|H_{\delta,\eta}\|_{2,Q},\mathcal{F}_\eta(\delta),\|\cdot\|_{2,Q}\right) \le \left(\frac{A}{\tau}\right)^{D_{\mathrm{vc}}} \qquad (7)$$

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 (7) 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 Section 4 we show that (B1) holds with $D=\dim(\mathcal{M})$ in the case where $\mathcal{M}$ 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 $A,\zeta>0$ such that, for any $\delta>0$ and $n\in\mathbb{N}$,

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

The constant ζ 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. When $\eta = d^2$, we get that $\zeta = 1$ for Hilbert spaces $\mathcal{M}$, $\zeta \le 1$ for geodesic spaces with positive curvature, and $\zeta \ge 1$ for geodesic spaces with non-positive curvature, see Section 4.3. Based on this, we call ζ the *curvature complexity* of $\mathcal{M}$.

3. Empirical Fréchet means

In this section, 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 Section 5. Throughout this section, 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 $\Delta \in (0,1)$,*

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

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 section, 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

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

Here and throughout this paper, $|\cdot|$ denotes the Euclidean norm. The extra factor C_Δ in Theorem 1 is a price we pay for 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, ζ) be the constant pairs that appear in (A2) and (B2), respectively. Then, there is a universal constant $C_{A,\zeta}$ depending only on $A > 0$ and $\zeta > 0$ such that, for all $n \in \mathbb{N}$ and $\Delta \in (0,1)$,*

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

with probability at least $1 - \Delta$.

An explicit form of the constant $C_{A,\zeta}$ in Theorem 2 may be found in the proof of the theorem in Appendix A.2. The theorem demonstrates that the consistency of the empirical Fréchet mean x_n continues to hold for infinite-dimensional $(\mathcal{M}, d)$, but with slower rates of convergence to x^* for increasing n when $\zeta \geq 1$, compared to the finite-dimensional case in Theorem 1. It shows 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 ζ *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 the case where $\mathcal{M}$ is a Riemannian manifold and $\eta = d^2$, see Section 4.3.

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

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 $x_0, x_1, z \in \mathcal{M}$

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

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 a 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 [48]. 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 [48] for details.

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

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

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

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

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

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

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 [52], 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. 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 $h \ge 1$. 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. 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 S.1 in the Supplementary Material shows that, if the

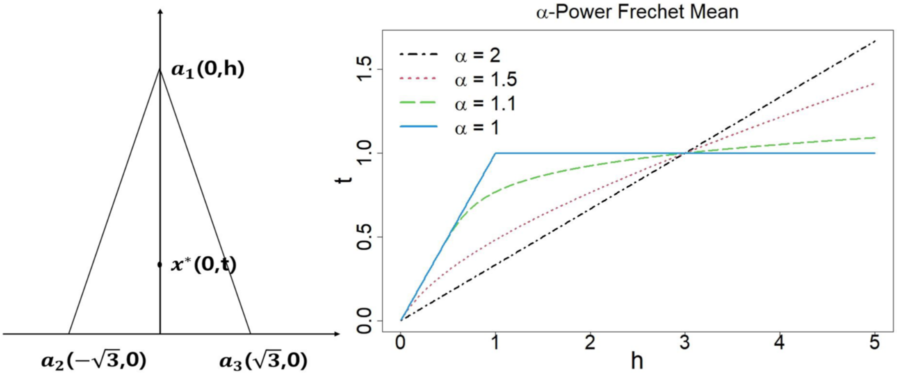

Figure 1. The left panel depicts the positions of the α-power Fréchet mean x^* and the three points a_1, a_2, a_3 having equal mass. The right panel shows the change of x^* as a_1 moves with a_2 and a_3 staying fixed, for $\alpha = 1/1.1/1.5/2$ (solid/dashed/dotted/dot-dashed).

underlying probability measure P is invariant under rotation around a point z, then z is the unique α-power Fréchet mean for all $\alpha \geq 1$.

The rates of convergence for α-power Fréchet means are studied for NPC spaces with $\alpha \in [1,2]$ in [47]. 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 $y,z,p,q \in \mathcal{M}$,

$$d(y,p)^\alpha - d(z,p)^\alpha - d(y,q)^\alpha + d(z,q)^\alpha \leq \alpha 2^{-\alpha+2} d(y,z)^{\alpha-1} d(p,q), \quad \alpha \in [1,2]. \tag{8}$$

Moreover, according to Appendix E in [47], no growth function satisfying (A1) exists for $\alpha > 2$ and $0 < \alpha < 1$. For $\alpha = 1$, (8) 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 $\alpha = 1$, some results analogous to Theorems 1 and 2 were provided in [4]. Stochastic proximal point algorithms (PPA) to compute Fréchet medians in NPC spaces were also introduced in [5,6].

In the next two propositions we derive generalized CN and variance inequalities for $\alpha \in (1,2]$. Thus, the theorems in Section 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 $z \in \mathcal{M}$,*

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

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. When 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 6 and 7 later in Section 5.2.

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

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

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*

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

Therefore, if $B_\alpha := \inf_{x\in\mathcal{M}\setminus\{x^*\}} b_\alpha(x) > 0$, then for any $x \in \mathcal{M}$,

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

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 $(t^{\alpha/2} + (1-t)^{\alpha/2})F_\alpha(x^*)$ by more than a positive constant multiple of $t^{\alpha/2} d(x,x^*)^\alpha$, 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 $t^{\alpha/2} + (1-t)^{\alpha/2} \ge 1$ 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 $x \in \mathcal{M} \setminus \{x^*\}$,

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

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 $B_\alpha > 0$, 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

$$\mathbb{P}\left(d(x_n,x^*) \le 64\left(\frac{K_\alpha^{\alpha/2}}{\alpha}\right)^{1/(\alpha-1)} \cdot \left(24\sqrt{AD} + \sqrt{\frac{2}{\Delta}}\right) \cdot \frac{\sigma_X}{\sqrt{n}} \right) \ge 1 - \Delta \tag{9}$$

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

$$\mathbb{P}\left(d(x_n,x^*) \le 2\left(\frac{K_\alpha^{\alpha/2}}{\alpha}\right)^{1/(\alpha-1)} \cdot C_{A,\zeta} \cdot \rho_n \cdot \frac{\sigma_X}{\sqrt{\Delta}} \right) \ge 1 - \Delta \tag{10}$$

for infinite-dimensional cases, where $K_\alpha = \alpha^2 2^{-2\alpha+2} B_\alpha^{-2+2/\alpha}$ and $\rho_n = n^{-1/2}$ if $0 < \zeta < 1$; $n^{-1/2} \cdot \log n$ if $\zeta = 1$; $n^{-1/2\zeta}$ if $\zeta > 1$. Note that the concentration rates in terms of Δ and n in (9) and (10) do not depend on $\alpha \in (1,2]$.

Remark 2. There are other choices of η, not of the form d^α, that may be of interest in statistics. For example, one may be interested in $\eta(x,y) = L_\delta(d(x,y))$, where L_δ with $\delta > 0$ is the Huber loss defined by

$$L_\delta : [0,+\infty) \to [0,+\infty), \quad x \mapsto \begin{cases} x^2/2, & 0 \le x \le \delta \\ \delta \cdot (x - \delta/2), & x > \delta. \end{cases}$$

This choice shares with $\eta = d^\alpha$ for $\alpha \in [1,2]$ the idea of combining the squared and absolute losses. Another example that may be of practical interest is the Kullback-Leibler divergence [31], $\eta(\mu,\nu) = D_{KL}(\nu||\mu)$, when $(\mathcal{M},d) = (\mathcal{P}_2(\mathbb{R}),W_2)$. The latter is an example of asymmetric functional. However, it seems difficult to prove the basic inequalities in (A1) and (A2) for general η. In particular, we are not aware of any type of bound for $\eta(\gamma_t,z)$ along a geodesic $\gamma : [0,1] \to \mathcal{M}$ for general η, which we need for the proof of (A2). To the best of our knowledge, even the results for $\eta = d^\alpha$ we present in Propositions 1 and 2 are the first.

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 [51]). 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 [11,25,51] 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 $\tau \in (0,r]$,

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

For a finite-dimensional normed space $\mathcal{M}$, one may take $D_1 = \dim(\mathcal{M})$ irrespective of the underlying norm, since all norms in such a space are equivalent. On the contrary, A_1 depends on the choice of a metric d and $A_1 = 3$ for the Euclidean norm when $\mathcal{M} = \mathbb{R}^D$. In any case, (B1′) is for finite-dimensional $\mathcal{M}$ and thus the dependence of A_1 and D_1 on the metric d does not need to be made explicit because the values of A_1 and D_1 do not affect the *convergence rates* in Theorem 1 of Section 3 and in Theorems 3 and 6 of Section 5 that are for finite-dimensional cases.

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

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

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 Section 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 [22]. For the reversed inequality, named as the Bishop-Gromov inequality, see [52]. Because of these inequalities, A_1 thus A in (B1) becomes smaller as the curvature of $\mathcal{M}$ increases when $\eta = d^\alpha$ with $\alpha \in (1,2]$.

Contrary to the case of finite-dimensional $\mathcal{M}$, a version of (B1′) is not true in many cases of infinite-dimensional $\mathcal{M}$. If $\mathcal{M}$ is an infinite-dimensional normed space, then any closed ball is non-compact, so that there is some $\tau_0 > 0$ such that $\log N(\tau, B(x^*, r), d) = \infty$ for any $\tau < \tau_0$. Therefore, the approach that mimics the finite-dimensional case does not work for infinite-dimensional $\mathcal{M}$ in general. However, for separable Hilbert spaces we may calculate directly the explicit constants in the assumption (B2), $A = 1/32$ and $\zeta = 1$ as demonstrated in the following proposition.

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

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

Furthermore, for the empirical measure P_n, it holds that

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

Proposition 4 may be used to verify (B2) with $\eta = d^2$ for Riemannian manifolds $(\mathcal{M}, d)$. Note that $d(x,y) \le \|\log_p x - \log_p y\|$ for $\mathcal{M}$ with non-negative curvature, while $d(x,y) \ge \|\log_p x - \log_p y\|$ 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 $\zeta \le 1$ for Riemannian manifolds with non-negative curvature, and with some $\zeta \ge 1$ for Hadamard manifolds. In fact, ζ 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 the sectional curvature of an underlying space $\mathcal{M}$ is, the slower 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 [2,44]. For $\mathcal{X}=\mathbb{R}$, however, $\mathcal{P}_2(\mathbb{R})$ has vanishing Alexandrov curvature [29]. 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. Even though $\mathcal{P}_2(\mathbb{R})$ is not compact, if we restrict ourselves to $\mathcal{M}=\mathcal{P}_2([-B,B])\subset\mathcal{P}_2(\mathbb{R})$ for $0<B<\infty$, then $\mathcal{M}$ is compact in $\mathcal{P}_2(\mathbb{R})$ (see Corollary 2.2.5 in [44]) with a finite diameter: $W_2(\mu,\nu)\le 2B$ for all $\mu,\nu\in\mathcal{P}_2([-B,B])$. This implies that the Wasserstein ball $B(\mu^*,r)\subset\mathcal{P}_2(\mathbb{R})$ is way larger than $\mathcal{P}_2([-r/2,r/2])$, since the former set includes probability measures with non-compact support and there is no hope that one can prove (B2) via a version of (B1′) when $\mathcal{M}=\mathcal{P}_2(\mathbb{R})$. Nonetheless, for $\mathcal{M}=\mathcal{P}_2([-B,B])$ for some $B>0$, we may obtain a version of (B1′) for any $D_1>1$ due to Theorem A.1 in [10]:

$$N\left(\tau,\mathcal{P}_2([-B,B]),W_2\right)\le\left(\frac{\sqrt{16e}B}{\tau}\right)^{8B/\tau}.$$

This would give (B2) for some $A,\zeta>1/2$ that do not depend on n as in the finite-dimensional case, see Subsection 2.2.4 of [44] or Appendix A of [10] for the explicit constants.

5. Geometric-median-of-means

For empirical Fréchet means in non-compact metric spaces, polynomial concentration, as we derived in Section 3, is the best one can achieve. In this section we introduce new estimators and 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 $m\ge n/k$. For each $1\le j\le k$, define

$$F_{n,j}(x)=\frac{1}{m}\sum_{X_i\in\mathcal{B}_j}\eta(x,X_i).\tag{11}$$

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

$$F_{n,j}(a)<F_{n,j}(b)\quad\text{if and only if}\quad |a-Z_j|<|b-Z_j|,\tag{12}$$

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

$$Z_j\in\operatorname*{arg\,min}_{x\in\mathcal{M}}F_{n,j}(x).$$

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 $F_{n,j}(a) \le F_{n,j}(b)$ for more than $k/2$ blocks $\mathcal{B}_j$. For $x \in \mathcal{M}$, let

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

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

$$x_{MM} \in \arg\min_{x \in \mathcal{M}} r_x. \tag{13}$$

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

Remark 3. We note that 'a defeats b' if and only if median$\{F_{n,j}(a) - F_{n,j}(b) : 1 \le j \le k\} \le 0$, see [33]. The minimum in (13) is always achieved, provided that $\eta : \mathcal{M} \times \mathcal{M} \to [0, +\infty)$ is continuous and for any $x \in \mathcal{M}$, $\eta(x,y) \to \infty$ as $d(x,y) \to \infty$. For any $x \in \mathcal{M}$, the x-defeating region S_x is a closed and bounded subset of $\mathcal{M}$ containing x, thus $r_x < +\infty$. This would entail that $x \mapsto r_x$ is a continuous function, and with the fact that $r_x \to \infty$ as $\min\{d(x,X_1), \dots, d(x,X_n)\} \to \infty$, one may argue that the minimum of r_x over $x \in \mathcal{M}$ is attained at some point in $\mathcal{M}$. 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, $r_x \le r$ 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 the case where $\mathcal{M}$ is a Euclidean space, the median-of-means may be interpreted in terms of Tukey depth, see [26].

In view of (12), our definition of 'defeat' is a natural extension of the notion introduced in [39] for $\mathcal{M} = \mathbb{R}^D$: '$a$ defeats b' if $|a - Z_j| \le |b - Z_j|$ for more than $k/2$ blocks $\mathcal{B}_j$. We note that, for curved metric spaces, the equivalence between $F_{n,j}(a) \le F_{n,j}(b)$ and $d(a,Z_j) \le d(b,Z_j)$ 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 $\mathcal{M}$.

Although $d(a,Z_j) \le d(b,Z_j)$ is not equivalent to $F_{n,j}(a) \le F_{n,j}(b)$ for curved spaces, one may roughly interpret 'a defeats b' as that a is closer than b to the centers of more than half of the k blocks, for $\eta = d^\alpha$ with $\alpha \in (1,2]$. 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 $Z_1, \dots, Z_k$. 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 $Z_1, \dots, Z_k$.

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 2 we depicted them on $[-1,1]^2$ and also Z_j (●) for $1 \le j \le 5$. The figure demonstrates that r_x, which is the radius of the smallest ball centered at $x =$ ▲ covering the 'violet/sky-blue/blue' regions, tends to decrease as $x \in \mathcal{M}$ gets closer to the Fréchet mean $x^* =$ ♦. 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 the violet and sky-blue regions, respectively, have $F_{n,j}(a) \le F_{n,j}($▲$)$ for 5 and 4 blocks with the original dataset, they

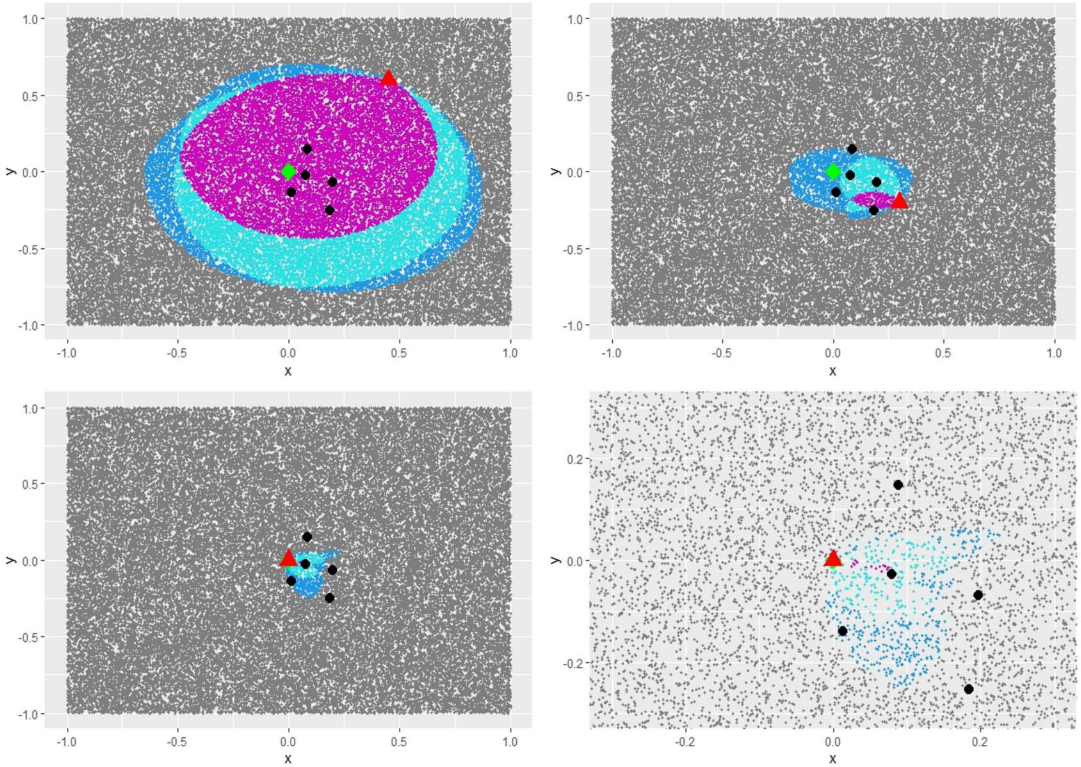

Figure 2. Illustration of the x-defeating region S_x for three choices of x = ▲: ▲ = $(0.45, 0.6)$ in the top-left, ▲ = $(0.3, -0.2)$ in the top-right and ▲ = $(0,0)$ in the bottom-left panel. For this a dataset $\{X_i : 1 \le i \le n\}$ of size $n = 10{,}000$ was generated from a bivariate distribution on $\mathbb{R}^2$ with mean ◆ = $(0,0)$, and it was partitioned into $k = 5$ blocks randomly. The block sample means $Z_1, \dots, Z_5$ are depicted as • points. The bottom-right panel is the zoomed-in picture of the bottom-left. In each panel with a given ▲, the color of each region indicates the 'defeating score', against ▲, of the points in the region, where the 'defeating score' of a point a against ▲ equals the number of blocks $\mathcal{B}_j$ such that $F_{n,j}(a) \le F_{n,j}$(▲). The violet region is for the score 5, the sky blue for 4, the blue for 3 and the gray for the scores ≤ 2. Thus, the union of violet/sky-blue/blue colored regions is the x-defeating region $S_▲$ in each panel.

still defeat x = ▲ 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 [12,18] and for $\mathcal{M} = \mathbb{R}^D$ by [39].

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 $X_1, \dots, X_n$ 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*

$$R_q = C_q \sigma_X \sqrt{\frac{\log(1/\Delta)}{n}}, \quad C_q = \frac{32\sqrt{2}}{q}\left(24\sqrt{AD} + \frac{2}{\sqrt{1-2q}}\right). \tag{14}$$

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 $S_{x_{MM}} \not\subseteq B(x_{MM}, R_q)$ 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 $r_{x^*} \le R_q$ on $\mathcal{E}$. By the definition of x_{MM}, it holds that $r_{x_{MM}} \le r_{x^*}$, however. This means that

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

The foregoing arguments give 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 $d(x_{MM},x^*) \le R_q$ with probability at least $1-\Delta$, where R_q is the constant defined at (14).*

Remark 4. Note that the condition $\Delta \in [e^{-2q^2 n}, 1)$ is latent in Theorem 3 and also in Theorems 4, 6 and 7 and Corollaries 1 to 4, since the number of blocks $k = \lceil 1/(2q^2)\log(1/\Delta)\rceil \le n$. When $\mathcal{M} = \mathbb{R}$, it is known that one should impose $\Delta \in [\Delta_{min}, 1)$ for some $\Delta_{min} > 0$ to achieve a sub-Gaussian performance, see [18].

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 $\zeta \ge 1$. 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^* defeats all $x \in \mathcal{M}$ with $d(x,x^*) > R_q$ but any such x does not defeat x^*, where*

$$R_{q,\zeta} = \begin{cases} c_{q,1} \cdot \sigma_X \cdot \log n \cdot \sqrt{\dfrac{\log(1/\Delta)}{n}} & \text{if } \zeta = 1 \\ c_{q,\zeta} \cdot \sigma_X \cdot \left(\dfrac{\log(1/\Delta)}{n}\right)^{1/2\zeta} & \text{if } \zeta > 1 \end{cases} \tag{15}$$

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

Corollary 2. *Assume (B2) with some constants $A > 0$ and $\zeta \geq 1$. 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 $d(x_{MM}, x^*) \leq R_{q,\zeta}$ with probability at least $1-\Delta$, where $R_{q,\zeta}$ is the constant defined at (15).*

As in the case of the empirical Fréchet mean x_n for infinite-dimensional $\mathcal{M}$, see (10), decreasing the curvature of $\mathcal{M}$ (increasing ζ) 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 $c_{q,\zeta}$ on $q \in (0,1/2)$ as in the discussion of Corollary 1. In the infinite-dimensional case, however, $c_{q,\zeta}$ is minimized at some point $q \in (0,1/2)$.

We note that the constants C_q and $c_{q,\zeta}$ in Theorems 3 and 4, respectively, may not be optimal. One might improve them by carefully sharpening of various inequalities in the proofs of the theorems. Rather than optimizing the constants, we focus on deriving *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 [39] for the case $\mathcal{M} = \mathbb{R}^D$, since we do not assume any differential structure for the underlying NPC space. The rates of concentration in Corollaries 1 and 2 are not optimal when $\mathcal{M}$ is a Hilbert space unless $\mathcal{M} = \mathbb{R}$. In the latter case, the optimal rate of concentration is known to be $O(\sqrt{\mathrm{tr}(\Sigma_X)/n} + \sqrt{\|\Sigma_X\|\log(1/\Delta)/n})$ as in (4). It is noteworthy that $\sigma_X^2 = \mathrm{tr}(\Sigma_X)$ when $\mathcal{M}$ is a Hilbert space. However, metric spaces without a differential structure do not have an equivalent of the covariance matrix Σ_X in general. Moreover, $\mathrm{tr}(\Sigma_X)$ in [39] arises from the *dual Sudakov inequality*, which accounts for the covering number of a sphere $r \cdot S^{D-1}$ with respect to the norm $\|\cdot\|_{2,P}$ in terms of r and $\mathrm{tr}(\Sigma_X)$. The inequality is based on the linear structure of $\mathbb{R}^D$ and the fact that $\|\cdot\|_{2,P}$ is translation invariant, therefore it is no longer valid for non-vector spaces. Hence, even for Hadamard manifolds where a differential structure is available, it seems intractable to obtain an inequality that corresponds to the dual Sudakov inequality.

Now, we present a theorem that gives the *breakdown point* of x_{MM}. The breakdown point of an estimator is the smallest proportion of data corruption that can upset the estimator completely. It tells the level of resistance by an estimator against data corruption and is a popular measure of robustness in statistics. Let $\mathcal{X}_n = \{X_1, \ldots, X_n\}$. For a configuration $\{i(1), \ldots, i(\ell)\} \subset \{1, 2, \ldots, n\}$, let $\tilde{\mathcal{X}}_n(i(1), \ldots, i(\ell))$ denote the modification of $\mathcal{X}_n$ for which $X_{i(j)}$ for $1 \leq j \leq \ell$ in $\mathcal{X}_n$ are replaced by $\tilde{X}_{i(j)}$, respectively. For an estimator $\hat{x}$, the breakdown point of $\hat{x}$ is defined as

$$\varepsilon_n^* := \frac{1}{n}\min\Big\{\ell : \text{there exists a dataset } \mathcal{X}_n \text{ and a configuration } \{i(1),\ldots,i(\ell)\} \text{ such that} \\ \sup_{\tilde{X}_{i(1)},\ldots,\tilde{X}_{i(\ell)}} d\big(\hat{x}(\mathcal{X}_n), \hat{x}\big(\tilde{\mathcal{X}}_n(i(1),\ldots,i(\ell))\big)\big) = \infty\Big\}.$$

For the above definition to make sense, we consider the case where $\mathrm{diam}(\mathcal{M}) = \infty$. The following theorem demonstrates that the breakdown point ε_n^* of x_{MM} for an NPC space $(\mathcal{M}, d)$ equals that of the median-of-means tournament for $\mathcal{M} = \mathbb{R}^D$.

Theorem 5. *Let $(\mathcal{M}, d)$ be an NPC space where $X_1, \ldots, X_n$ take values. Let k denote the number of blocks $\mathcal{B}_j$. Then, the breakdown point of x_{MM} associated with $\eta = d^2$ is independent of partition $\{\mathcal{B}_j : 1 \leq j \leq k\}$ and equals $\varepsilon_n^* = n^{-1} \cdot \lceil (k+1)/2 \rceil$.*

One may be interested in studying the concentration properties of geometric-median-of-means when some portion of the dataset are corrupted. This has been done by [16] for $\mathcal{M} = \mathbb{R}^D$. Its extension to NPC spaces is a challenging topic for future study.

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 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

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

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

$$\frac{1 - (1+\delta)^{1-\alpha/2}(1-t)^{\alpha/2}}{(1+\delta)^{1-\alpha/2}\, t^{\alpha/2}} < 1 \tag{16}$$

for all $0 < t < 1$ and $\delta > 0$. Hence, taking $\rho \geq 1$ when $\eta = d^\alpha$ for $0 < \alpha < 2$, as (16) shows, would give $M_{\alpha,\rho} = \sup \varnothing = -\infty$. In fact, we find that the derivation of exponential concentration is intractable for

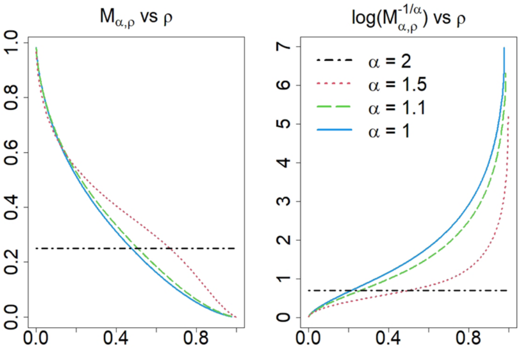

Figure 3. The shapes of $M_{\alpha,\rho}$ (left) and $\log M_{\alpha,\rho}^{-1/\alpha}$ (right) as functions of ρ for $\alpha = 1/1.1/1.5/2$ (solid/dashed/dotted/dot-dashed).

$x_{\rho,MM}$ 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. 3 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.

Theorem 6. *Assume (B1) with some constants $A, D > 0$ and that there exists a constant $B_\alpha > 0$ such that*

$$d(x, x^*)^\alpha \leq \frac{1}{B_\alpha} \int_M \left(d(x,y)^\alpha - d(x^*, y)^\alpha \right) \mathrm{d}P(y). \tag{17}$$

Let $\rho \in (0,1)$ for $\alpha \in (1,2)$ or $\rho = 1$ when $\alpha = 2$. Also, let $\Delta \in (0,1)$ 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

$$\begin{aligned} R_{q,\alpha,\rho} &= C_{q,\alpha,\rho}\, \sigma_X \sqrt{\frac{\log(1/\Delta)}{n}}, \\ C_{q,\alpha,\rho} &= M_{\alpha,\rho}^{-1/\alpha} \cdot \frac{16\sqrt{2K_\alpha}}{q} \left(24\sqrt{AD} + \frac{2}{\sqrt{1-2q}} \right). \end{aligned} \tag{18}$$

Recall that Proposition 2 gives a sufficient condition for the existence of $B_\alpha > 0$ such that (17) holds. Also, we note that (17) 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 6 with $\rho = 1$ reduces to Theorem 3. The following corollary may be derived from Theorem 6 as Corollary 1 is from Theorem 3.

Corollary 3. *Assume the conditions and consider the ranges of (ρ, α), Δ and q in Theorem 6. 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 (18).*

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. 3. 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 6 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 7 with $\rho = 1$ reduces to Theorem 4.

Theorem 7. *Assume (B2) with some constants $A > 0$ and $\zeta \geq 1$ and that there exists a constant $B_\alpha > 0$ such that (17) holds. Consider the ranges of (ρ,α), Δ and q in Theorem 6. 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*

$$R_{q,\alpha,\rho,\zeta} = \begin{cases} c_{q,\alpha,\rho,1} \cdot \dfrac{\log n}{n^{1/2}} \cdot \sigma_X \cdot \sqrt{\log \dfrac{1}{\Delta}} & \text{if } \zeta = 1 \\[2ex] c_{q,\alpha,\rho,\zeta} \cdot \dfrac{1}{n^{1/2\zeta}} \cdot \sigma_X \cdot \left(\log \dfrac{1}{\Delta}\right)^{1/2\zeta} & \text{if } \zeta > 1, \end{cases} \tag{19}$$

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

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

Corollary 4. *Assume the conditions and consider the ranges of (ρ,α), Δ and q in Theorem 7. Let k denote the number of blocks $\mathcal{B}_j$. If $k = \lceil 1/(2q^2)\log(1/\Delta)\rceil$, then it holds that $d(x_{\rho,MM},x^*) \leq R_{q,\alpha,\rho,\zeta}$ with probability at least $1-\Delta$, where $R_{q,\alpha,\rho,\zeta}$ is the constant defined at (19).*

From (9) and (10) in Section 4.2 we have observed that the concentration rates for the empirical Fréchet mean 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 the curvature complexity ζ is the same as x_n and x_{MM}. Also, the dependence of $c_{q,\alpha,\rho,\zeta}$ 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 some point $q \in (0,1/2)$.

Remark 5. For NPC spaces $\mathcal{M}$ with $\eta = d^2$, the curvature complexity ζ is greater than or equal to 1 (Proposition 4). However, ζ may be strictly less than 1 when $\eta = d^\alpha$ with $1 < \alpha < 2$. In the latter case, one may prove that the radius of concentration $R_{q,\alpha,\rho,\zeta}$ in Theorem 7 is given by

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

for the same constant $c_{q,\alpha,\rho,\zeta}$ given at (19).

6. Concluding remarks

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 median-of-means method. 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.

For Euclidean or Hilbertian spaces $\mathcal{M}$, there is a large body of works that study sub-Gaussian mean estimators under only a second moment condition, see [26,39,40] and references therein. For general metric spaces, however, the definition of sub-Gaussianity itself is not available. It is a challenging future topic to generalize the notion of sub-Gaussian performance to more general metric spaces and investigate the concentration properties of the corresponding empirical Fréchet means with the extended notion of sub-Gaussianity.

We admit that there is an issue of algorithmic feasibility with the geometric-median-of-means estimator studied in our paper. The computational issue is also present in the Euclidean case for the median-of-means tournament estimator, see [39]. There are some alternative proposals that are equipped with an efficient algorithm. These include the geometric median of [42], Catoni-Giulini estimator of [13], the Hopkins' estimator [26] and those in the follow-up studies by [15,17,35]. However, all these estimators are for the case where $\mathcal{M}$ is a Euclidean or a Banach space. In particular, the estimators studied in [15,17,26,35] combine the idea of semi-definite programming (SDP) and *r-centrality* (see [26] for definition), which requires an inner product structure for the underlying space. For some spaces that admit a tangential structure equipped with a bi-invariant metric, one may borrow the idea of Hopkins [26] to find a robust Fréchet mean estimator equipped with an efficient algorithm. For instance, if $\mathcal{M} = \mathcal{S}_D^+$, the space of symmetric positive-definite matrices, and it is endowed with the log-Euclidean metric, one might project the dataset onto the tangent at the identity I_D via the logarithmic map, compute the Hopkins' estimator from the projected data, and then transform the result back to $\mathcal{S}_D^+$ via the exponential map. It is straightforward to show that the resulting estimator of the Fréchet mean is consistent. However, this is not an estimator of our interest in this paper, and the special treatment would restrict the study to Riemannian manifolds. It is a challenging topic of study to develop an efficient algorithm for the geometric-median-of-means estimator in the general setting of NPC spaces.

Appendix A: Proofs of theorems

In the Appendix, we give the proofs of Theorems 1–7. The proofs of the propositions in Section 4 can be found in the Supplementary Material [53]. Throughout the Appendix and the Supplementary Material, we often denote $\int_{\mathcal{S}} f(y)\,\mathrm{d}Q(y)$ simply by Qf for a measurable space $(\mathcal{S},\mathcal{B})$, a probability measure Q on $\mathcal{B}$ and a measurable function $f : \mathcal{S} \to \mathbb{R}$. For instance, $Pf = \mathbb{E}(f(X))$ 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.

A.1. Some lemmas

Here we present three lemmas that are used in the proofs of the main theorems. Our first lemma is a tail inequality for empirical processes. In our setup, $\|\eta(x,\cdot) - \eta(x^*,\cdot)\|_\infty$ may be unbounded as x varies.

Under some strong condition on the tail of P, one may be able to obtain an exponential tail inequality, see [1,50]. Since we assume only a finite second moment of P, we use the following polynomial tail inequality.

Lemma 1 ([34]). *Let $X_1,\dots,X_n$ 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 $\Delta\in(0,1)$,

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

For the statement of the second lemma, recall the definition of $H_\delta\equiv H_{\delta,\eta}$ given at (6), 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 $\bar\sigma(\delta)=4\sqrt{K\sigma_X^2\delta^\beta}$.

The following lemma provides an improved chaining bound for Gaussian processes. For a proof, see Theorem 5.31 in [51] or Lemma 5.1 in [2].

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

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

A.2. Proofs of theorems in Section 3

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

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

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

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

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

$$\phi_n(\delta) \le 4\mathbb{E}\,\phi_n(\delta) + \frac{8\sqrt{2}\cdot\bar{\sigma}(\delta)}{\sqrt{n\Delta}}. \tag{20}$$

We first get an upper bound on $\mathbb{E}\,\phi_n(\delta)$. 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 [25]) we obtain

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

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 $\sqrt{n}$-Lipschitz with respect to $\|\cdot\|_{2,P_n}$, conditionally on the X_i's. It is also sub-Gaussian. To see this, we note that, for any $a_1,\ldots,a_n \in \mathbb{R}$,

$$\mathbb{E}\left(\exp\left(\sum_{i=1}^{n}a_i\varepsilon_i\right)\right) = \prod_{i=1}^{n}\mathbb{E}e^{a_i\varepsilon_i} = \prod_{i=1}^{n}\frac{e^{a_i}+e^{-a_i}}{2} \le \prod_{i=1}^{n}e^{a_i^2/2} = \exp\left(\sum_{i=1}^{n}\frac{a_i^2}{2}\right),$$

where the inequality follows from Taylor's expansion. From this we get that, for any $f, g \in \mathcal{F}(\delta)$ and $\theta \in \mathbb{R}$,

$$\begin{aligned}\mathbb{E}\left(e^{\theta(Y_f - Y_g)}\,|X_1,\ldots,X_n\right) &= \mathbb{E}\left(\exp\left(\frac{\theta}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i(f-g)(X_i)\right)\Big|X_1,\ldots,X_n\right)\\ &\le \exp\left(\frac{\theta^2}{2n}\sum_{i=1}^{n}(f(X_i)-g(X_i))^2\right)\\ &= \exp\left(\frac{\theta^2}{2}\|f-g\|_{2,P_n}^2\right).\end{aligned}$$

Thus, Y_f satisfies the conditions of Lemma 3 (see [2]). Applying Lemma 3 with (B1) and using the inequalities for H_δ given in Lemma 2, we get

$$\begin{aligned}
\mathbb{E}\,\phi_n(\delta) &\le 2\,\mathbb{E}\inf_{\varepsilon\ge 0}\left(2\varepsilon + \frac{12}{\sqrt{n}}\int_\varepsilon^\infty \sqrt{\log N(u,\mathcal{F}(\delta),\|\cdot\|_{2,P_n})}\mathrm{d}u\right)\\
&\le 2\,\mathbb{E}\inf_{\varepsilon\ge 0}\left(2\varepsilon + \frac{12}{\sqrt{n}}\int_\varepsilon^{\|H_\delta\|_{2,P_n}} \sqrt{D\log\left(\frac{A\|H_\delta\|_{2,P_n}}{u}\right)}\,\mathrm{d}u\right)\\
&= 2\,\mathbb{E}\left(\|H_\delta\|_{2,P_n}\right)\cdot\inf_{\varepsilon'\ge 0}\left(2\varepsilon' + \frac{12}{\sqrt{n}}\int_{\varepsilon'}^{1} \sqrt{D\log\left(\frac{A}{u}\right)}\,\mathrm{d}u\right) \qquad (21)\\
&\le 48\,\mathbb{E}\left(\|H_\delta\|_{2,P_n}\right)\cdot\sqrt{\frac{AD}{n}}\\
&\le 48\,\bar{\sigma}(\delta)\sqrt{\frac{AD}{n}},
\end{aligned}$$

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

The inequalities (20) and (21) 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 an increasing function and $b_n(\delta,\Delta)$ is decreasing in Δ for fixed δ, it follows from Theorem 4.3 in [30] that

$$\delta_n \le \phi_n(\delta_n) \le b_n(\Delta) := \inf\left\{\tau>0 : \sup_{\delta\ge\tau}\delta^{-1}b_n\left(\delta,\Delta\cdot\frac{\delta}{\tau}\right)\le 1\right\} \qquad (22)$$

with probability at least $1-\Delta$. Since $b_n(\delta,\Delta\cdot\delta/\tau)$ is decreasing in δ for $\beta\in(0,2)$,

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

This gives

$$\begin{aligned}
b_n(\Delta) &= \inf\left\{\tau>0 : 32\sqrt{\frac{K\sigma_X^2\tau^{-(2-\beta)}}{n}}\left(24\sqrt{AD}+\sqrt{\frac{2}{\Delta}}\right)\le 1\right\} \qquad (23)\\
&= \left\{32\sqrt{\frac{K\sigma_X^2}{n}}\left(24\sqrt{AD}+\sqrt{\frac{2}{\Delta}}\right)\right\}^{\frac{2}{2-\beta}}.
\end{aligned}$$

Applying (23) to (22), we obtain that, with probability at least $1-\Delta$,

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

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

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

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

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

This gives the theorem. □

A.3. Proofs of theorems in Section 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 $\mathcal{B}_j$,

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

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 (22) and (23) that

$$\mathbb{P}\left(F_{n,j}(x^*) - F_{n,j}(Z_j) \le \varepsilon_q^2\right) \ge q + \frac{1}{2}, \tag{24}$$

where

$$\varepsilon_q = 32\sqrt{\frac{k\sigma_X^2}{n}}\left(24\sqrt{AD} + \frac{2}{\sqrt{1-2q}}\right). \tag{25}$$

By the CN inequality in Section 4.1, we have

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

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

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

we get from (24) that $\mathbb{P}(\mathcal{E}_{n,j}) \ge q + 1/2$ since $F_{n,j}(x^*) - F_{n,j}(Z_j) \le \varepsilon_q^2$ implies

$$F_{n,j}(x) - F_{n,j}(Z_j) > -\left(F_{n,j}(x^*) - F_{n,j}(Z_j)\right) + 2\varepsilon_q^2 \;\ge\; F_{n,j}(x^*) - F_{n,j}(Z_j)$$

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

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

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 (24) now with ε_q at (25) being replaced by

$$\varepsilon_{q,\zeta} = \begin{cases} C_{A,1} \cdot \dfrac{\log(n/k)}{\sqrt{n/k}} \cdot \dfrac{\sigma_X}{\sqrt{(1-2q)/2}}, & \text{if } \zeta = 1 \\ C_{A,\zeta} \cdot (k/n)^{1/2\zeta} \cdot \dfrac{\sigma_X}{\sqrt{(1-2q)/2}}, & \text{if } \zeta > 1. \end{cases} \tag{26}$$

Since

$$\begin{aligned} &\frac{1}{\sqrt{2}q} \frac{\log n \sqrt{\log(1/\Delta)}}{\sqrt{n}} = \frac{\sqrt{k}\log n}{\sqrt{n}} \geq \frac{\log(n/k)}{\sqrt{n/k}}, \\ &\frac{1}{\sqrt{2}q} n^{-1/2\zeta} \left(\log \frac{1}{\Delta}\right)^{1/2\zeta} \geq \left(\frac{2q^2 n}{\log(1/\Delta)}\right)^{-1/2\zeta} = (k/n)^{1/2\zeta}, \end{aligned} \tag{27}$$

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

Proof of Theorem 5. Fix a partition $\{\mathcal{B}_j : 1 \leq j \leq k\}$ of the original dataset $\mathcal{X}_n = \{X_1, \ldots, X_n\}$. Let $\mathcal{B}_j = \{X_{1,j}, \ldots, X_{m,j}\}$ for $1 \leq j \leq k$. Choose a point $O \in \mathcal{M}$ and let $D_O = \max_{1\leq i\leq n} d(O, X_i)$. We let $\tilde{X}_i$ denote a corrupted value corresponding to X_i, and we write $\tilde{A}$ instead of A if a term A involves corrupted values. For example, we write $\tilde{F}_{n,j}(x)$ instead of $F_{n,j}(x)$ when the jth block contains a corrupted value. In particular, we write in this proof $\tilde{S}_x$ and $\tilde{r}_x$ for each point $x \in \mathcal{M}$ rather than S_x and r_x, respectively, since the defeating region and radius always depend on corrupted values. Put $L := \lceil (k+1)/2 \rceil$.

We first show that $\varepsilon_n^* \leq L/n$ by contradiction. Suppose that it is false, i.e., $\varepsilon_n^* > L/n$. Then, for an arbitrary configuration $\{j(1), \ldots, j(L)\} \subset \{1, 2, \ldots, k\}$ with the corruption $\tilde{X}_{1,1} = \tilde{X}_{1,2} = \cdots = \tilde{X}_{1,L} = \tilde{x}$, there exists $R > 0$ such that $\sup_{\tilde{x}\in\mathcal{M}} d(O, \tilde{x}_{MM}) < R$. We may assume $R > \sqrt{m-1} \cdot D_O$. Now, let $\tilde{\gamma} : [0,1] \to \mathcal{M}$ be the geodesic connecting $\tilde{\gamma}_0 = O$ and $\tilde{\gamma}_1 = \tilde{x}$ with the length $\tilde{D} := d(\tilde{x}, O)$ larger than R. For $j = 1, \ldots, L$, by the triangular inequality,

$$\begin{aligned} m \cdot (\tilde{F}_{n,j}(\tilde{x}_{MM}) - \tilde{F}_{n,j}(\tilde{\gamma}_t)) &\geq d(\tilde{x}, \tilde{x}_{MM})^2 - (1-t)^2\tilde{D}^2 - \sum_{i=2}^{m} d(\tilde{\gamma}_t, X_{ij})^2 \\ &\geq (\tilde{D} - R)^2 - (1-t)^2\tilde{D}^2 - (m-1)(t\tilde{D} + D_O)^2 \\ &= (2t - mt^2)\tilde{D}^2 - 2(R + D_O(m-1)t)\tilde{D} + (R^2 - (m-1) \cdot D_O^2) \\ &\geq (2t - mt^2)\tilde{D}^2 - 2(R + D_O(m-1)t)\tilde{D}. \end{aligned}$$

This implies that for all $t < 2/m$, the point $\tilde{\gamma}_t$ defeats $\tilde{x}_{MM}$ whenever

$$\tilde{D} \geq \frac{2(R + D_O(m-1)t)}{2t - mt^2},$$

so the defeating radius of $\tilde{x}_{MM}$ satisfies $\tilde{r}_{\tilde{x}_{MM}} \geq d(\tilde{\gamma}_t, \tilde{x}_{MM})$. Therefore,

$$\liminf_{\tilde{D}\to\infty} \frac{\tilde{r}_{\tilde{x}_{MM}}}{\tilde{D}} \geq \liminf_{\tilde{D}\to\infty} \sup_{0<t<2/m} \frac{d(\tilde{\gamma}_t, \tilde{x}_{MM})}{\tilde{D}} = \liminf_{\tilde{D}\to\infty} \sup_{0<t<2/m} \frac{d(\tilde{\gamma}_t, O)}{\tilde{D}} = \frac{2}{m}. \tag{28}$$

Now, choose any $x \in \tilde{S}_{\tilde{\gamma}_{1/m}} (\neq \emptyset)$, i.e. x defeating $\tilde{\gamma}_{1/m}$. Then, ∃ at least one $j_0 \in \{j(1), \ldots, j(L)\}$ such that $\tilde{F}_{n,j_0}(x) \leq \tilde{F}_{n,j_0}(\tilde{\gamma}_{1/m})$. Due to the CN inequality,

$$\begin{aligned} 0 &\leq m \cdot (\tilde{F}_{n,j_0}(\tilde{\gamma}_{1/m}) - \tilde{F}_{n,j_0}(x)) \qquad (29) \\ &= d(\tilde{\gamma}_{1/m}, \tilde{x})^2 - d(x, \tilde{x})^2 + \sum_{i=2}^{m} \left(d(\tilde{\gamma}_{1/m}, X_{i,j_0})^2 - d(x, X_{i,j_0})^2 \right) \\ &\leq \left(\frac{m-1}{m} \right)^2 \tilde{D}^2 - d(x, \tilde{x})^2 \\ &\quad + \sum_{i=2}^{m} \left\{ \frac{m-1}{m} d(O, X_{i,j_0})^2 + \frac{1}{m} d(\tilde{x}, X_{i,j_0})^2 - \frac{m-1}{m^2} \tilde{D}^2 - d(x, X_{i,j_0})^2 \right\}. \end{aligned}$$

Note that again by the CN inequality,

$$d(x, \tilde{\gamma}_{1/m})^2 \leq \frac{m-1}{m} d(O, x)^2 + \frac{1}{m} d(x, \tilde{x})^2 - \frac{m-1}{m^2} \tilde{D}^2.$$

Plugging this inequality into (29) and using the triangular inequality, we get

$$\begin{aligned} 0 &\leq \frac{m-1}{m^2} \tilde{D}^2 - m \cdot d(x, \tilde{\gamma}_{1/m})^2 \\ &\quad + \sum_{i=2}^{m} \left\{ \frac{m-1}{m} d(O, X_{i,j_0})^2 + \frac{1}{m} d(\tilde{x}, X_{i,j_0})^2 - \frac{m-1}{m^2} \tilde{D}^2 + d(O, x)^2 - d(x, X_{i,j_0})^2 \right\} \\ &\leq \frac{m-1}{m^2} \tilde{D}^2 - m \cdot d(x, \tilde{\gamma}_{1/m})^2 \\ &\quad + (m-1) \left\{ \frac{m-1}{m} D_O^2 + \frac{1}{m} (\tilde{D} + D_O)^2 - \frac{m-1}{m^2} \tilde{D}^2 + 2D_O \cdot d(O, x) - D_O^2 \right\} \\ &\leq \frac{m-1}{m^2} \tilde{D}^2 - m \cdot d(x, \tilde{\gamma}_{1/m})^2 \\ &\quad + (m-1) \left\{ \frac{m-1}{m} D_O^2 + \frac{1}{m} (\tilde{D} + D_O)^2 - \frac{m-1}{m^2} \tilde{D}^2 + 2D_O \left(\frac{\tilde{D}}{m} + d(x, \tilde{\gamma}_{1/m}) \right) - D_O^2 \right\} \\ &= -m \cdot d(x, \tilde{\gamma}_{1/m})^2 + 2(m-1) D_O \cdot d(x, \tilde{\gamma}_{1/m}) + \frac{2(m-1)\tilde{D}(\tilde{D} + 2mD_O)}{m^2}. \end{aligned}$$

Therefore,

$$d(x, \tilde{\gamma}_{1/m}) \leq \frac{m(m-1)D_O + \sqrt{m^2(m-1)^2 D_O^2 + 2m(m-1)\tilde{D}(\tilde{D} + 2mD_O)}}{m^2}.$$

Since $x \in \tilde{S}_{\tilde{\gamma}_{1/m}}$ was chosen arbitrarily, we have

$$\limsup_{\tilde{D} \to \infty} \frac{\tilde{r}_{\tilde{\gamma}_{1/m}}}{\tilde{D}} \leq \limsup_{\tilde{D} \to \infty} \frac{m(m-1)D_O + \sqrt{m^2(m-1)^2 D_O^2 + 2m(m-1)\tilde{D}(\tilde{D} + 2mD_O)}}{m^2 \tilde{D}}$$

$$= \frac{\sqrt{2}}{m} < \frac{2}{m}.$$

In view of (28), the above strict inequality is contradictory to the fact that $\tilde{r}_{\tilde{x}_{MM}} \le \tilde{r}_{\tilde{y}_{1/m}}$ for all $\tilde{D}$ from the definition of geometric-median-of-means.

Next, we show that

$$\sup_{*} d(O, \tilde{x}_{MM}) < \infty, \tag{30}$$

where $\sup_*$ denotes the supremum over all configurations of $s \le (L-1)$ arbitrary corruptions among $\{X_1, \dots, X_n\}$. We note that (30) implies $\varepsilon_n^* \ge L/n$. To prove (30), let $s \le L-1$ be the number of corrupted $\tilde{X}_i$ and think of a configuration of the indices of $\tilde{X}_i$, say $\{i(1), \dots, i(s)\} \subset \{1, 2, \dots, n\}$. The corrupted $\tilde{X}_{i(1)}, \dots, \tilde{X}_{i(s)}$ are scattered across the k blocks $\mathcal{B}_j$, $1 \le j \le k$. Without loss of generality, let $\mathcal{B}_1, \dots, \mathcal{B}_J$ denote those blocks that do not contain any of the corrupted values. We note that $J > k/2$ since $s \le L-1$. We claim

$$\sup_{\tilde{X}_{i(1)}, \dots, \tilde{X}_{i(s)}} \max_{1 \le j \le J} \tilde{r}_{Z_j} < \infty. \tag{31}$$

Then, by the definition of the geometric-median-of-means we get

$$\sup_{\tilde{X}_{i(1)}, \dots, \tilde{X}_{i(s)}} \tilde{r}_{\tilde{x}_{MM}} < \infty. \tag{32}$$

Also, by the definition of x-defeating radius and since $J > k/2$, it holds that

$$\begin{aligned} \tilde{r}_x &\ge \mathrm{rad}_x \left(\bigcap_{j=1}^{J} \{ y \in \mathcal{M} : F_{n,j}(y) \le F_{n,j}(x) \} \right) \\ &\ge \mathrm{rad}_x \left(\bigcap_{j=1}^{k} \{ y \in \mathcal{M} : F_{n,j}(y) \le F_{n,j}(x) \} \right) \end{aligned} \tag{33}$$

for all $x \in \mathcal{M}$, where $\mathrm{rad}_x(A)$ stands for the radius of the smallest ball centered at x that covers A. The right hand side of the second inequality in (33) depends solely on the original dataset $\{X_1, \dots, X_n\}$, independent of data corruption. Now, suppose that there exists $s \le L-1$ and a configuration $\{i(1), \dots, i(s)\}$ such that

$$\sup_{\tilde{X}_{i(1)}, \dots, \tilde{X}_{i(s)}} d(O, \tilde{x}_{MM}) = \infty.$$

Then, since the right hand side of the second inequality in (33) diverges to infinity as $d(O, x) \to \infty$, we would obtain

$$\sup_{\tilde{X}_{i(1)}, \dots, \tilde{X}_{i(s)}} \tilde{r}_{\tilde{x}_{MM}} = \infty,$$

which contradicts (32). This proves (30).

It remains to prove (31). Let $1 \le j \le J$ be fixed. Then, for any x that defeats Z_j, there exists at least one un-corrupted block $\mathcal{B}_l$ $(1 \le l \le J)$ such that $F_{n,l}(x) \le F_{n,l}(Z_j)$, since $J > k/2$ and the number of indices $l : 1 \le l \le k$ such that $F_{n,l}(x) \le F_{n,l}(Z_j)$ or $\tilde{F}_{n,l}(x) \le \tilde{F}_{n,l}(Z_j)$ is greater than $k/2$. This implies that

$$\begin{aligned}
\tilde{r}_{Z_j} &\le \max_{x \in \mathcal{M}} \{ d(x, Z_j) : F_{n,l}(x) \le F_{n,l}(Z_j) \text{ for some } 1 \le l \le J \} \\
&\le \max_{x \in \mathcal{M}} \left\{ \sqrt{F_{n,l}(x)} + \max_{X_i \in \mathcal{B}_l} d(X_i, Z_j) : F_{n,l}(x) \le F_{n,l}(Z_j) \text{ for some } 1 \le l \le J \right\} \\
&\le \max_{1 \le l \le J} \left(\sqrt{F_{n,l}(Z_j)} + \max_{X_i \in \mathcal{B}_l} d(X_i, Z_j) \right) \\
&\le \max_{1 \le l \le k} \left(\sqrt{F_{n,l}(Z_j)} + \max_{X_i \in \mathcal{B}_l} d(X_i, Z_j) \right).
\end{aligned} \tag{34}$$

In (34), the second inequality follows from

$$d(x, Z_j) \le \frac{1}{m} \sum_{X_i \in \mathcal{B}_l} \big(d(x, X_i) + d(X_i, Z_j) \big) \le \sqrt{F_{n,l}(x)} + \max_{X_i \in \mathcal{B}_l} d(X_i, Z_j).$$

The right hand side of the last inequality in (34) depends solely on the original dataset $\{X_1, \dots, X_n\}$, independent of the configuration of $\{i(1), \dots, i(s)\}$ and the corrupted values $\tilde{X}_{i(1)}, \dots, \tilde{X}_{i(s)}$. This gives (31). □

Proof of Theorem 6. First, we follow the lines leading to (24), now using (23) with $K = K_\alpha$ and $\beta = 2 - 2/\alpha$ instead of $K = \beta = 1$. We may prove

$$\mathbb{P}\left(F_{n,j}(x^*) - F_{n,j}(Z_j) \le K_\alpha^{\alpha/2} \varepsilon_q^\alpha \right) \ge q + \frac{1}{2}. \tag{35}$$

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 $\delta > 0$,

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

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

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

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

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

or equivalently

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

for all $x \in \mathcal{M}$ with $d(x,x^*) > K_\alpha^{1/2} M_{\alpha,\rho}^{-1/\alpha}\varepsilon_q$. Thus, from (35) and the definition of $M_{\alpha,\rho}$ it follows that

$$\mathbb{P}\left(F_{n,j}(x) > \rho \cdot F_{n,j}(x^*) \text{ for all } x \in \mathcal{M} \text{ with } d(x,x^*) > \frac{K_\alpha^{1/2}\varepsilon_q}{M_{\alpha,\rho}^{1/\alpha}}\right) \geq q + \frac{1}{2}. \tag{36}$$

Applying Høffding's inequality as in the proof of Theorem 3 with (36), we may complete the proof of the theorem. □

Proof of Theorem 7. The proof is essentially the same as that of Theorem 6 except that we use the definition of $\varepsilon_{q,\zeta}$ at (26) instead of ε_q at (25). Using (27) we get $K_\alpha^{1/2} M_{\alpha,\rho}^{-1/\alpha}\varepsilon_{q,\zeta} \leq R_{q,\alpha,\rho,\zeta}$. □

Acknowledgments

The authors would like to thank an associate editor and three referees for constructive comments.

Funding

Research of the authors was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2019R1A2C3007355).

Supplementary Material

Supplement to "Exponential concentration for geometric-median-of-means in non-positive curvature spaces" (DOI: 10.3150/22-BEJ1569SUPP; .pdf). The Supplementary Material contains an additional proposition and the proofs of Lemma 2 in Appendix A.1 and the propositions in Section 4.

References

[1] Adamczak, R. (2008). A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.* **13** 1000–1034. MR2424985 https://doi.org/10.1214/EJP.v13-521

[2] Ahidar-Coutrix, A., Le Gouic, T. and Paris, Q. (2020). Convergence rates for empirical barycenters in metric spaces: Curvature, convexity and extendable geodesics. *Probab. Theory Related Fields* **177** 323–368. MR4095017 https://doi.org/10.1007/s00440-019-00950-0

[3] Arsigny, V., Fillard, P., Pennec, X. and Ayache, N. (2007). Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM J. Matrix Anal. Appl.* **29** 328–347. MR2288028 https://doi.org/10.1137/050637996

[4] Bačák, M. (2014). Computing medians and means in Hadamard spaces. *SIAM J. Optim.* **24** 1542–1566. MR3264572 https://doi.org/10.1137/140953393

[5] Bačák, M. (2014). *Convex Analysis and Optimization in Hadamard Spaces*. *De Gruyter Series in Nonlinear Analysis and Applications* **22**. Berlin: de Gruyter. MR3241330 https://doi.org/10.1515/9783110361629

[6] Bačák, M. (2018). Old and new challenges in Hadamard spaces. ArXiv preprint. Available at arXiv:1807.01355.

[7] Bhattacharya, R. and Patrangenaru, V. (2003). Large sample theory of intrinsic and extrinsic sample means on manifolds. I. *Ann. Statist.* **31** 1–29. MR1962498 https://doi.org/10.1214/aos/1046294456

[8] Bhattacharya, R. and Patrangenaru, V. (2005). Large sample theory of intrinsic and extrinsic sample means on manifolds. II. *Ann. Statist.* **33** 1225–1259. MR2195634 https://doi.org/10.1214/009053605000000093

[9] Billera, L.J., Holmes, S.P. and Vogtmann, K. (2001). Geometry of the space of phylogenetic trees. *Adv. in Appl. Math.* **27** 733–767. MR1867931 https://doi.org/10.1006/aama.2001.0759

[10] Bolley, F., Guillin, A. and Villani, C. (2007). Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probab. Theory Related Fields* **137** 541–593. MR2280433 https://doi.org/10.1007/s00440-006-0004-7

[11] Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford: Oxford Univ. Press. MR3185193 https://doi.org/10.1093/acprof:oso/9780199535255.001.0001

[12] Catoni, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** 1148–1185. MR3052407 https://doi.org/10.1214/11-AIHP454

[13] Catoni, O. and Giulini, I. (2018). Dimension-free PAC-Bayesian bounds for the estimation of the mean of a random vector. ArXiv preprint. Available at arXiv:1802.04308.

[14] Chen, Y., Lin, Z. and Müller, H.-G. (2021). Wasserstein regression. *J. Amer. Statist. Assoc.* **116** 1–14.

[15] Cherapanamjeri, Y., Flammarion, N. and Bartlett, P.L. (2019). Fast mean estimation with sub-Gaussian rates. In *Conference on Learning Theory* 786–806. PMLR.

[16] Depersin, J. and Lecué, G. (2021). On the robustness to adversarial corruption and to heavy-tailed data of the Stahel-Donoho median of means. ArXiv preprint. Available at arXiv:2101.09117.

[17] Depersin, J. and Lecué, G. (2022). Robust sub-Gaussian estimation of a mean vector in nearly linear time. *Ann. Statist.* **50** 511–536. MR4382026 https://doi.org/10.1214/21-aos2118

[18] Devroye, L., Lerasle, M., Lugosi, G. and Oliveira, R.I. (2016). Sub-Gaussian mean estimators. *Ann. Statist.* **44** 2695–2725. MR3576558 https://doi.org/10.1214/16-AOS1440

[19] Fillard, P., Arsigny, V., Pennec, X., Hayashi, K.M., Thompson, P.M. and Ayache, N. (2007). Measuring brain variability by extrapolating sparse tensor fields measured on sulcal lines. *NeuroImage* **34** 639–650. https://doi.org/10.1016/j.neuroimage.2006.09.027

[20] Fillard, P., Arsigny, V., Pennec, X., Thompson, P.M. and Ayache, N. (2005). Extrapolation of sparse tensor fields: Application to the modeling of brain variability. In *Biennial International Conference on Information Processing in Medical Imaging* 27–38. Springer.

[21] Fréchet, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. Henri Poincaré* **10** 215–310. MR0027464

[22] Gallot, S., Hulin, D. and Lafontaine, J. (1990). *Riemannian Geometry*, 2nd ed. *Universitext*. Berlin: Springer. MR1083149 https://doi.org/10.1007/978-3-642-97242-3

[23] Ganea, O.-E., Bécigneul, G. and Hofmann, T. (2018). Hyperbolic neural networks. In *NeurIPS 2018*.

[24] Ghodrati, L. and Panaretos, V.M. (2022). Distribution-on-distribution regression via optimal transport maps. *Biometrika* **109** 957–974. MR4519110 https://doi.org/10.1093/biomet/asac005

[25] Giné, E. and Nickl, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. *Cambridge Series in Statistical and Probabilistic Mathematics, [40]*. New York: Cambridge Univ. Press. MR3588285 https://doi.org/10.1017/CBO9781107337862

[26] Hopkins, S.B. (2020). Mean estimation with sub-Gaussian rates in polynomial time. *Ann. Statist.* **48** 1193–1213. MR4102693 https://doi.org/10.1214/19-AOS1843

[27] Horváth, L., Kokoszka, P. and Wang, S. (2021). Monitoring for a change point in a sequence of distributions. *Ann. Statist.* **49** 2271–2291. MR4319250 https://doi.org/10.1214/20-aos2036

[28] Hsu, D. and Sabato, S. (2016). Loss minimization and parameter estimation with heavy tails. *J. Mach. Learn. Res.* **17** Paper No. 18. MR3491112

[29] Kloeckner, B. (2010). A geometric study of Wasserstein spaces: Euclidean spaces. *Ann. Sc. Norm. Super. Pisa Cl. Sci. (5)* **9** 297–323. MR2731158

[30] Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. *Lecture Notes in Math.* **2033**. Heidelberg: Springer. MR2829871 https://doi.org/10.1007/978-3-642-22147-7

[31] Le Cam, L. (2012). *Asymptotic Methods in Statistical Decision Theory*. Springer Science & Business Media.

[32] Le Gouic, T. and Loubes, J.-M. (2017). Existence and consistency of Wasserstein barycenters. *Probab. Theory Related Fields* **168** 901–917. MR3663634 https://doi.org/10.1007/s00440-016-0727-z

[33] Lecué, G. and Lerasle, M. (2020). Robust machine learning by median-of-means: Theory and practice. *Ann. Statist.* **48** 906–931. MR4102681 https://doi.org/10.1214/19-AOS1828

[34] Lederer, J. and van de Geer, S. (2014). New concentration inequalities for suprema of empirical processes. *Bernoulli* **20** 2020–2038. MR3263097 https://doi.org/10.3150/13-BEJ549

[35] Lei, Z., Luh, K., Venkat, P. and Zhang, F. (2020). A fast spectral algorithm for mean estimation with sub-Gaussian rates. In *Conference on Learning Theory* 2598–2612. PMLR.

[36] Lerasle, M., Szabó, Z., Mathieu, T. and Lecué, G. (2019). MONK outlier-robust mean embedding estimation by median-of-means. In *International Conference on Machine Learning* 3782–3793. PMLR.

[37] Lin, Z. (2019). Riemannian geometry of symmetric positive definite matrices via Cholesky decomposition. *SIAM J. Matrix Anal. Appl.* **40** 1353–1370. MR4032859 https://doi.org/10.1137/18M1221084

[38] Lin, Z., Müller, H.-G. and Park, B.U. (2021). Additive models for symmetric positive-definite matrices, Riemannian manifolds and Lie groups. ArXiv preprint. Available at arXiv:2009.08789.

[39] Lugosi, G. and Mendelson, S. (2019). Sub-Gaussian estimators of the mean of a random vector. *Ann. Statist.* **47** 783–794. MR3909950 https://doi.org/10.1214/17-AOS1639

[40] Lugosi, G. and Mendelson, S. (2019). Mean estimation and regression under heavy-tailed distributions: A survey. *Found. Comput. Math.* **19** 1145–1190. MR4017683 https://doi.org/10.1007/s10208-019-09427-x

[41] Lugosi, G. and Mendelson, S. (2021). Robust multivariate mean estimation: The optimality of trimmed mean. *Ann. Statist.* **49** 393–410. MR4206683 https://doi.org/10.1214/20-AOS1961

[42] Minsker, S. (2015). Geometric median and robust estimation in Banach spaces. *Bernoulli* **21** 2308–2335. MR3378468 https://doi.org/10.3150/14-BEJ645

[43] Nemirovsky, A.S. and Yudin, D.B. (1983). *Problem Complexity and Method Efficiency in Optimization*. *Wiley-Interscience Series in Discrete Mathematics*. New York: Wiley. MR0702836

[44] Panaretos, V.M. and Zemel, Y. (2020). *An Invitation to Statistics in Wasserstein Space*. *SpringerBriefs in Probability and Mathematical Statistics*. Cham: Springer. MR4350694 https://doi.org/10.1007/978-3-030-38438-8

[45] Pennec, X., Sommer, S. and Fletcher, T. (2019). *Riemannian Geometric Statistics in Medical Image Analysis*. Academic Press.

[46] Petersen, A. and Müller, H.-G. (2019). Fréchet regression for random objects with Euclidean predictors. *Ann. Statist.* **47** 691–719. MR3909947 https://doi.org/10.1214/17-AOS1624

[47] Schötz, C. (2019). Convergence rates for the generalized Fréchet mean via the quadruple inequality. *Electron. J. Stat.* **13** 4280–4345. MR4023955 https://doi.org/10.1214/19-EJS1618

[48] Sturm, K.-T., Coulhon, T. and Grigor'yan, A. (2003). Probability measures on metric spaces of nonpositive curvature. Heat kernels and analysis on manifolds, graphs, and metric spaces, contemporary mathematics. *Am. Math. Soc.* **358**.

[49] Tifrea, A., Bécigneul, G. and Ganea, O.-E. (2018). Poincaré glove: Hyperbolic word embeddings. ArXiv preprint. Available at arXiv:1810.06546.

[50] van de Geer, S. and Lederer, J. (2013). The Bernstein-Orlicz norm and deviation inequalities. *Probab. Theory Related Fields* **157** 225–250. MR3101846 https://doi.org/10.1007/s00440-012-0455-y

[51] Van Handel, R. (2014). Probability in high dimension Technical Report Princeton Univ NJ.

[52] Villani, C. (2009). *Optimal Transport: Old and New*. *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]* **338**. Berlin: Springer. MR2459454 https://doi.org/10.1007/978-3-540-71050-9

[53] Yun, H. and Park, B.U. (2023). Supplement to "Exponential Concentration for Geometric-Median-of-Means in Non-Positive Curvature Spaces." https://doi.org/10.3150/22-BEJ1569SUPP

[54] Zhang, C., Kokoszka, P. and Petersen, A. (2022). Wasserstein autoregressive models for density time series. *J. Time Series Anal.* **43** 30–52. MR4400283 https://doi.org/10.1111/jtsa.12590

Received March 2022 and revised November 2022