
Robust Estimation in metric spaces: Achieving Exponential Concentration with a Fréchet Median

Jakwang Kim*

University of British Columbia

Jiyoung Park*

Texas A&M University

Anirban Bhattacharya

Texas A&M University

Abstract

There is growing interest in developing statistical estimators that achieve exponential concentration around a population target even when the data distribution has heavier than exponential tails. More recent activity has focused on extending such ideas beyond Euclidean spaces to Hilbert spaces and Riemannian manifolds. In this work, we show that such exponential concentration in presence of heavy tails can be achieved over a broader class of parameter spaces called $\text{CAT}(\kappa)$ spaces, a very general metric space equipped with the minimal essential geometric structure for our purpose, while being sufficiently broad to encompass most typical examples encountered in statistics and machine learning. The key technique is to develop and exploit a general concentration bound for the Fréchet median in $\text{CAT}(\kappa)$ spaces. We illustrate our theory through a number of examples, and provide empirical support through simulation studies.

1 Introduction

A fundamental challenge in statistical estimation pertains to dealing with heavy-tailed data. Many estimators used in practice are built upon the assumption of light-tailed data, and their finite-sample (or non-asymptotic) statistical properties do not necessarily carry over when the data generating distribution has heavy tails. A standard way to characterize such non-asymptotic behavior of statistical estimators is via concentration inequalities; see Boucheron et al. (2013) for a book-level treatment. Given an estimator $\hat{\theta}_n$ based

on n samples for parameter θ , a concentration inequality provides a *non-asymptotic bound* to some distance $d(\hat{\theta}_n, \theta)$ being greater than a tolerance level $\varepsilon > 0$, as a function of n and ε . Inverting such a concentration inequality, one can obtain a growth rate on the sample size as a function of the tolerance level ε to provably achieve a desired level of confidence (say 95%). As a simple illustrative example, the sample average of i.i.d. real-valued observations concentrate exponentially fast (in terms of sample size) around the population mean if the true data distribution has sub-exponential tails by Bernstein’s inequality. However, if one weakens the tail assumption to expand the class of true distributions to those with finite second moment, then one can only establish polynomial concentration as dictated by Chebyshev’s inequality (Catoni, 2012). Exponential concentration is desirable not only to obtain logarithmic dependence of the sample size on the tolerance level for a single estimator, but also to combine multiple dependent estimators via a union bound. Accordingly, there has been extensive recent research towards constructing alternative ‘robust’ estimators that achieve exponential concentration rates in situations where standard M-estimators (or method of moment estimators) fail to provide one (Minsker, 2019; Audibert and Catoni, 2011; Oliveira and Rico, 2024).

Modern statistics and machine learning routinely encounter data beyond the classical Euclidean settings. Hyperspheres are used to model the directional data and spatial data (Hall et al., 1987; Jeong et al., 2017; Zhang and Chen, 2021); Hilbert spaces serve as a base space in functional data analysis (Petersen and Müller, 2016); hyperbolic spaces have become popular for hierarchical data (Nickel and Kiela, 2017); and graph and tree data are predominant in network data analysis (Fortunato, 2010; Abu-Ata and Dragan, 2016). Accordingly, several recent attempts have been made to address statistical problems in more general spaces; see, e.g., Arnaudon et al. (2013); Brunel and Serres (2024); Holmes (2003); Köstenberger and Stark (2024); Romon and Brunel (2023); Sturm (2000) for some representative examples.

*: **Equal contribution, alphabetically ordered.** Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

Motivated by the extensive literature on robust estimations and statistical methods beyond Euclidean spaces, this work proposes a method for robust estimation—boosting weakly concentrating estimators to strongly concentrate—in general metric spaces under minimal assumptions. Specifically, we focus on $\text{CAT}(\kappa)$ spaces, which is a general metric space with the minimal essential geometric structure necessary for our purposes, and introduce a procedure to boost weakly concentrating estimators in these spaces. $\text{CAT}(\kappa)$ spaces encompass not only widely studied spaces such as Hilbert spaces (Euclidean spaces) and Riemannian manifolds, but also other various spaces that have gained attention in data science and yet have been less examined in the context of robust estimation; see Section 2.1 for examples.

The contributions of this work are as follows:

- (1) We propose a method to boost estimators with polynomial concentration to strong (exponential) concentration in $\text{CAT}(\kappa)$ spaces; a general setting that encompasses most spaces concerned in statistics and machine learning; by leveraging properties of the Fréchet median (Section 3).
- (2) We show the applicability of our method across a wide range of statistical problems (Section 4).
- (3) We show the proposed method allows for a tractable algorithm, and verify its strength numerically (Section 5).

2 Preliminaries

2.1 $\text{CAT}(\kappa)$ space

In this section, we introduce $\text{CAT}(\kappa)$ spaces. A more rigorous definition is provided in Appendix A.1.

For each $\kappa \in \mathbb{R}$, the 2-dimensional model spaces M_κ^2 are defined as follows:

$$M_\kappa^2 = \begin{cases} \mathbb{R}^2 & \text{if } \kappa = 0, \\ \frac{1}{\sqrt{\kappa}} \mathbb{S}^2 & \text{if } \kappa > 0, \\ \frac{1}{\sqrt{|\kappa|}} \mathbb{H}^2 & \text{if } \kappa < 0 \end{cases}$$

where \mathbb{S}^2 and \mathbb{H}^2 are the 2-dimensional unit sphere and hyperbolic plane, respectively. A geodesic space (\mathcal{X}, d) is a $\text{CAT}(\kappa)$ space if every geodesic triangle in \mathcal{X} has a *comparison triangle* in M_κ^2 with the same side lengths, such that the original triangle is *thinner* (a precise mathematical formulation is given in Appendix A.1). This condition, known as the *$\text{CAT}(\kappa)$ inequality*, describes how much the space is curved, allowing curvature of spaces to be defined with minimal structure. See Figure 1. By definition, \mathbb{R}^2 , \mathbb{S}^2 , and \mathbb{H}^2 are $\text{CAT}(0)$, $\text{CAT}(1)$, and $\text{CAT}(-1)$ spaces, respectively.

A lot of spaces studied in practice belong to the category of $\text{CAT}(\kappa)$ spaces. Below are some examples.



Figure 1: Triangles in M_κ^2 spaces for different κ . Triangles in a $\text{CAT}(\kappa)$ space is thinner than triangles in M_κ^2 . **Left:** Euclidean triangle ($\kappa = 0$). **Middle:** Spherical triangle ($\kappa > 0$). **Right:** Hyperbolic triangle ($\kappa < 0$).

- Riemannian manifolds with sectional curvature upper bounded by κ : Hyperspheres, hyperbolic spaces, information geometry, Kendall shape spaces, space of Symmetric Positive Definite (SPD) matrices, to name a few.
- Infinite dimensional spaces: Hilbert spaces and infinite dimensional hyperbolic spaces are prominent examples of $\text{CAT}(0)$.
- Singular spaces and stratified spaces: These spaces have gained interest recently (Geiger et al., 2001; Mattingly et al., 2023a,b,c) but do not fall under classical spaces. On the other hand, many of these spaces are $\text{CAT}(\kappa)$ spaces (Burago et al., 2001a)[Theorem 9.1.21].
- Spaces of phylogenetic trees: Phylogenetic trees are widely studied objects in the field of biology and statistics. The space of phylogenetic trees can be endowed with a metric that makes it $\text{CAT}(0)$ (Billera et al., 2001).
- Metric graphs and trees: Any metric graph with cycles of length less than 2π is $\text{CAT}(1)$ (Brown, 2016)[Remark 2.12]. Specifically, metric trees are $\text{CAT}(0)$.

In particular, $\text{CAT}(0)$ spaces are often referred to as Hadamard spaces or Non-Positively Curved (NPC) spaces in the literature. They have drawn attention in statistics community due to their applicability to practical examples and favorable properties when incorporating a probability measure (Arnaudon et al., 2013; Brunel and Serres, 2024; Gouic et al., 2019; Köstemberger and Stark, 2024; Romon and Brunel, 2023; Sturm, 2000; Yun and Park, 2023). These favorable properties are discussed in Remark 2.2.

2.2 Fréchet mean and median

In this section, we introduce the notions of Fréchet mean and median, which are generalizations of classical mean and median to a general separable and complete metric space (\mathcal{X}, d) . Define $\mathcal{P}_p(\mathcal{X})$ as the set of Borel probability measures P on \mathcal{X} with a finite p^{th} moment,

i.e., $\int_{\mathcal{X}} d^p(x, y) dP(y) < \infty$ for some $x \in \mathcal{X}$.

Definition 2.1 (Fréchet mean and median). *Given $P \in \mathcal{P}_p(\mathcal{X})$, suppose $x^* \in \mathcal{X}$ satisfies*

$$x^* \in \operatorname{argmin}_{x \in \mathcal{X}} \int_{\mathcal{X}} d^p(x, y) dP(y).$$

Any such x^ with $p = 1$ is called a Fréchet median of P , and with $p = 2$ is called a Fréchet mean of P .*

Fréchet mean and median are also referred to as *barycenter* and *geometric median*, respectively, in some literature.

Remark 2.2 (Existence and uniqueness of the Fréchet mean and median). *For NPC spaces, Fréchet mean (resp. median) globally exists for any probability measure with a finite second (resp. first) moment (Bačák, 2014)[Lemma 2.3]. In addition, Fréchet mean is unique (Bačák, 2014)[Theorem 2.4]. Fréchet median may not be unique, but when it is not unique they form a single geodesic segment in the same manner medians behave in \mathbb{R} (Schötz, 2024)[Theorem 6.6]. On the other hand, for $\text{CAT}(\kappa)$ with $\kappa > 0$, a probability measure in addition needs to be supported within a ball of radius smaller than $D_\kappa/2$ to guarantee the (unique) existence of a Fréchet median (resp. mean) (Yokota, 2017).*

We lastly note that the Fréchet mean and median are not the only extensions of the classical mean and median to a general metric space. Sturm (2000)[Section 7] introduced a convex mean that coincides with the Fréchet mean in Euclidean space but not necessarily in a general metric space. Similarly, Lugosi and Mendelson (2016) suggested a tournament-based median, which matches the Fréchet median in \mathbb{R} but not beyond it. In addition, Dai and López-Pintado (2021) suggested to extend the idea of Tukey’s depth based median to general geodesic spaces.

2.3 Robust estimation and a median-of-means estimator

Catoni (2012) was one of the first to demonstrate the existence of an M-estimator that achieves exponential concentration only under moment conditions in \mathbb{R} . Since then, numerous studies have explored the existence of estimators with similar favorable properties in more general spaces. For instance, Lugosi and Mendelson (2021) generalized the notion of ‘trimmed mean’ to \mathbb{R}^d and showed its exponential concentration.

One of the most extensively studied methods for this pursuit is a median-of-means (MoM) estimator (Nemirovsky and Yudin, 1983; Alon et al., 1996). The construction of a MoM estimator is as follows: first split n data into k disjoint blocks, compute sample means for each block, and then take a median over

those means. A MoM estimator interpolates between the mean estimator (when $k = 1$) and the median estimator (when $k = n$). Therefore, choosing an appropriate k guarantees accurate estimation of the mean with a level of robustness similar to that of the median. Lerasle and Oliveira (2011) showed MoM estimators achieve exponential concentration over the real line. Beyond \mathbb{R} , different choices of medians lead to different MoM estimators. Minsker (2015); Lin et al. (2020) utilize the Fréchet median as a median and demonstrate the existence of MoM estimators in Banach spaces and certain families of Riemannian manifolds. Meanwhile, Lugosi and Mendelson (2019); Yun and Park (2023) define MoM via a “median-of-means tournament” and derive concentration bounds in \mathbb{R}^d and NPC spaces, although their MoM estimators face computational issues (Lugosi and Mendelson, 2019; Yun and Park, 2023)[Section 4, 6]. The philosophy of MoM has been extended to more general target quantities beyond the mean. For example, Minsker et al. (2017) proposed a robust posterior distribution using MoM-inspired ideas.

3 Main result: Boosting the weak estimators by Fréchet median

In this section, we adopt the Fréchet median robust estimation in $\text{CAT}(\kappa)$ spaces based on the method proposed by Minsker (2015). Minsker (2015)[Lemma 2.1] establishes a key link between the Fréchet median and robustness in Hilbert spaces, which is later extended to Riemannian manifolds with some assumptions by Lin et al. (2020)[Lemma 2.1]. The idea is that the empirical Fréchet median controls the geometric discrepancies between data points, which guarantees the desirable concentration. One of the main contributions of this work is the extension of this idea to general $\text{CAT}(\kappa)$ spaces.

For (\mathcal{X}, d) a $\text{CAT}(\kappa)$ space, an empirical Fréchet median of $x_1, \dots, x_k \in \mathcal{X}$ is defined by the Fréchet median of the empirical measure $\sum_{j=1}^k \delta_{x_j}/k$. We will use the notation $\text{med}(x_1, \dots, x_k)$ to denote the empirical Fréchet median of x_1, \dots, x_k .

Lemma 3.1 (Geometric discrepancy near the Fréchet median). *Let (\mathcal{X}, d) be a $\text{CAT}(\kappa)$ space, and fix $x_1, \dots, x_k \in \mathcal{X}$. Denote $x^* := \text{med}(x_1, \dots, x_k)$. Fix $\alpha \in (0, 0.5)$ and write $C_\alpha = (1 - \alpha)(1 - 2\alpha)^{-1/2}$. Suppose either (a) or (b) holds:*

- (a) *For $\kappa \leq 0$, assume there exists $z \in \mathcal{X}$ such that $d(x^*, z) > C_\alpha r$ for some $r > 0$.*
- (b) *For $\kappa > 0$, write $D_\kappa = \pi/\sqrt{\kappa}$. Assume x^* exists, $x_j \in B(x^*, D_\kappa/2)$ for all $j = 1, \dots, k$, and there exists $z \in \mathcal{X}$ such that $\frac{\pi}{2}C_\alpha r < d(x^*, z) \leq D_\kappa/2$ for some $0 < r < D_\kappa/(C_\alpha \pi)$.*

Under (a) or (b), there exists a subset $J \subseteq \{1, \dots, k\}$, with cardinality $|J| > \alpha k$, such that for all $j \in J$, $d(x_j, z) > r$.

Lemma 3.1 implies that if a point z is far away from a Fréchet median, then it also has to be far away from the bulk of the points, x_j 's. The proof of Lemma 3.1 relies on the behavior of the triangle $\triangle x_j x^* z$ in $\text{CAT}(\kappa)$ spaces. Utilizing the $\text{CAT}(\kappa)$ inequality, or the triangle comparison, it turns out that such z cannot be close to x_j 's. A complete proof will be provided in Appendix B.1.

Remark 3.2.

1. When $\kappa > 0$, the assumptions on the positions of points and the existence of x^* are required. As mentioned in Remark 2.2, this is possible whenever points are distributed in a ball of radius smaller than $D_\kappa/2$.
2. Lemma 3.1 covers the case of Hilbert spaces discussed in Minsker (2015), but does not extend to Banach spaces. This limitation arises from the fact that, unlike Hilbert triangles, Banach triangles may not satisfy inner-product-based inequalities such as the Cauchy-Schwarz inequality. For more details, see Khamsi and Shukri (2017).
3. Restricted to Riemannian manifolds, the curvature upper bound condition in Lemma 3.1 implies the Lipschitz logarithmic map condition proposed by Lin et al. (2020)[Lemma 2.1], with the Lipschitz constant being $\pi/2$. We conjecture that the converse is also true: the Lipschitz logarithmic map condition is valid only if the sectional curvature at x^* is upper bounded. If this conjecture holds, then Lemma 3.1 precisely encompasses the findings of Lin et al. (2020).

Now, we proceed to the main theorem.

Theorem 3.3 (Boosting a weak estimator). *Suppose the parameter space Θ is $\text{CAT}(\kappa)$ space. Let $\theta \in \Theta$ be a parameter of interest and $\hat{\theta}_j$, $j = 1, \dots, k$ be independent estimators of θ . Let $\hat{\theta}_{\text{FMoE}} := \text{med}(\hat{\theta}_1, \dots, \hat{\theta}_k)$ be a ‘Fréchet median of estimators’.*

Fix $\alpha \in (0, 1/2)$ and $p \in (0, \alpha)$. Write $\psi(\alpha, p) := (1 - \alpha) \log \frac{1-\alpha}{1-p} + \alpha \log \frac{\alpha}{p}$ and set C_α same as in Lemma 3.1.

- (a) For $\kappa \leq 0$, suppose there exists $\epsilon > 0$ such that $\mathbb{P}(d(\hat{\theta}_j, \theta) > \epsilon) \leq p$ for all $j = 1, \dots, k$. Then,

$$\mathbb{P}\left[d(\hat{\theta}_{\text{FMoE}}, \theta) > C_\alpha \epsilon\right] \leq \exp(-k\psi(\alpha, p)).$$

- (b) For $\kappa > 0$, suppose there exists $\epsilon \in (0, D_\kappa/(\pi C_\alpha))$ such that $\mathbb{P}(d(\hat{\theta}_j, \theta) > \epsilon) \leq p$ for all $j = 1, \dots, k$. Assume $\hat{\theta}_{\text{FMoE}}$ exists, $\hat{\theta}_j \in B(\hat{\theta}_{\text{FMoE}}, D_\kappa/2)$, and

$\hat{\theta}_{\text{FMoE}} \in B(\theta, D_\kappa/2)$ almost surely. Then,

$$\mathbb{P}\left[d(\hat{\theta}_{\text{FMoE}}, \theta) > \frac{\pi C_\alpha \epsilon}{2}\right] \leq \exp(-k\psi(\alpha, p)).$$

We provide the proof in Appendix B.1.

Remark 3.4 ($\kappa > 0$ case). *For $\kappa > 0$, a priori concentrations $d(\hat{\theta}_{\text{FMoE}}, \hat{\theta}_j) \leq D_\kappa/2$ and $d(\hat{\theta}_{\text{FMoE}}, \theta) \leq D_\kappa/2$ are required. The first condition is a common assumption in $\text{CAT}(\kappa)$ spaces (Brunel and Serres, 2024)[Theorem 18]. The second condition is new here. If one avoids using this condition, one only obtains a conditional form of concentration: if $\hat{\theta}_j \in B(\hat{\theta}_{\text{FMoE}}, D_\kappa/2 - \epsilon)$ almost surely, one can obtain*

$$\begin{aligned} \mathbb{P}\left[d(\hat{\theta}_{\text{FMoE}}, \theta) > \frac{\pi C_\alpha \epsilon}{2} \mid d(\hat{\theta}_{\text{FMoE}}, \theta) \leq D_\kappa/2\right] \\ \leq \frac{\exp(-k\psi(\alpha, p))}{1 - p^k}. \end{aligned}$$

When the context is clear, we will refer to $\hat{\theta}$ as the ‘original estimator’ throughout the paper. Theorem 3.3 states that by taking the Fréchet median, the original estimator can be boosted to achieve exponential concentration even when only weak, e.g., polynomial, concentration would be expected. Algorithm 1 outlines the procedure for constructing a Fréchet median of estimators (FMoE) based on the original estimator.

Algorithm 1 Boosting a preliminary estimator.

Require: Input data x_1, \dots, x_n , $\text{CAT}(\kappa)$ space (Θ, d) , the block size k .

- 1: Split the data x_i into k disjoint blocks, with each block consisting of $\lfloor n/k \rfloor$ data points.
 - 2: **for** $j \leftarrow 1$ to k **do**
 - 3: $\hat{\theta}_j \leftarrow$ The original estimator from j^{th} block data points.
 - 4: **return** $\hat{\theta}_{\text{FMoE}} \leftarrow$ Fréchet median of $\hat{\theta}_j$ with respect to the metric d_Θ .
-

Remark 3.5 (Time complexity of Algorithm 1). *To the best of our knowledge, the time complexity of Algorithm 1 in general setting is unknown, as that of computing the Fréchet median is unknown beyond Euclidean spaces. That said, for the time complexity with respect to n for a fixed d , one can expect the time complexity of FMoE matches the time complexity of the original estimator. A heuristic argument proceeds as follows: suppose the time complexities of an original estimator and Fréchet median given n elements are $O(n^\alpha)$ and $O(n^\beta)$ for some $\alpha, \beta > 0$. Then, given k , the time complexity of obtaining the original estimators over k blocks will be $O(n^\alpha k^{1-\alpha})$. In sum, the total time complexity will be $O(n^\alpha k^{1-\alpha} + k^\beta)$. Typically, both from*

theoretical observations and numerical simulations, k is relatively small compared to n (see Sections 4 and 5). Therefore, if one assumes $k = O(1)$, $O(n^\alpha k^{1-\alpha})$ term dominates as n gets larger, meaning the complexity matches the original estimator's time complexity $O(n^\alpha)$.

We conclude this section by highlighting the favorable properties of our method. First, it can boost any type of estimator. While most research on statistical estimation in general metric spaces has focused on Fréchet mean estimation (Ahidar-Coutrix et al., 2018; Brunel and Serres, 2024; Gouic et al., 2019; Sturm, 2000; Yun and Park, 2023), our method also applies to estimators that need not be Fréchet means. We illustrate the advantages of this broad applicability in Section 4.2.

Second, our method is nearly fully implementable. Given the original estimators, Algorithm 1 requires only the computation of the Fréchet median of them. This is always feasible in NPC spaces (Bačák, 2014)[Algorithm 4.3]. In $\text{CAT}(\kappa)$ spaces with $\kappa > 0$, while no universal algorithm for Fréchet median exists, algorithms tailored to specific domains can be utilized (Fletcher et al., 2009; Boria et al., 2019). This improves upon Yun and Park (2023), which considered the median-of-means estimators in NPC spaces using a tournament-based median but lacked computational tractability.

4 Statistical applications

This section illustrates how the proposed method boosts some widely used estimators to achieve exponential concentration without sub-Gaussian conditions on the data. While our method is generally applicable, we investigate the boosting of two important estimators as examples: (1) Fréchet mean estimators in NPC spaces and (2) the canonical sample covariance estimator.

4.1 Boosting Fréchet means: Fréchet median-of-means estimators

As the primary application, we focus on the widely studied problem of estimating the Fréchet mean in NPC spaces (Brunel and Serres, 2024; Gouic et al., 2019; Sturm, 2000; Yun and Park, 2023). This problem has received significant attention in recent years due to the existence guarantee of the Fréchet mean and median in these spaces, as discussed in Remark 2.2. For this problem, the proposed method becomes a Fréchet median-of-means (FMoM). Throughout this section we assume that (\mathcal{X}, d) is a NPC space with a curvature lower bound $\kappa_{\min}(\mathcal{X}) \in [-\infty, 0]$ and that $x_i \stackrel{i.i.d.}{\sim} P \in \mathcal{P}_2(\mathcal{X})$ for $i = 1, \dots, n$. We denote the

Fréchet mean and the second moment of P as x^* and $\sigma^2 := \mathbb{E}_{y \sim P} [d^2(x^*, y)]$ respectively.

There are two natural ways to construct a Fréchet mean estimator in NPC spaces: (1) Empirical Fréchet mean \hat{x}_{EM} and (2) inductive mean \hat{x}_{IM} . These serve as alternative estimators of the population mean though they differ in some aspects. The explanations of these estimators will be provided in Appendix A.1.1. Proposition 4.1 shows a weak concentration of these estimators under mild conditions.

Proposition 4.1. *Let \hat{x} be either \hat{x}_{EM} or \hat{x}_{IM} . If $\hat{x} = \hat{x}_{EM}$, in addition assume $\kappa_{\min}(\mathcal{X}) > -\infty$. Then,*

$$\mathbb{E} [d^2(\hat{x}, x^*)] \leq \frac{\sigma^2}{n}.$$

Furthermore, for any $\epsilon > 0$

$$\mathbb{P} [d(\hat{x}, x^*) > \epsilon] \leq \frac{\sigma^2}{n\epsilon^2}.$$

Proof. For the expected error bound, see Gouic et al. (2019)[Corollary 3.4] for \hat{x}_{EM} case and Sturm (2000)[Theorem 4.7] for \hat{x}_{IM} case.

The concentration inequality directly follows from Markov inequality. \square

On the contrary, achieving the exponential concentration requires sub-Gaussian type assumptions, which are stronger than the usual sub-Gaussian conditions in Euclidean space; see the discussion in Brunel and Serres (2024)[Definition 3].

Proposition 4.2. *Brunel and Serres (2024)[Corollary 11] Suppose P satisfies the following sub-Gaussian type assumption, i.e.,*

$$\sup_{f \in \mathcal{F}} \mathbb{E}_{X \sim P} \left[e^{\lambda(f(X) - \mathbb{E}[f(X)])} \right] \leq e^{\frac{\lambda^2 K^2}{2}} \quad \forall \lambda > 0$$

where $\mathcal{F} = \{f : \mathcal{X} \rightarrow \mathbb{R} \mid f \text{ is a 1-Lipschitz function}\}$. Let \hat{x} be either \hat{x}_{EM} or \hat{x}_{IM} . If $\hat{x} = \hat{x}_{EM}$, in addition assume $\kappa_{\min}(\mathcal{X}) > -\infty$. Then,

$$\mathbb{P} \left[d(\hat{x}, x^*) \geq \frac{\sigma}{\sqrt{n}} + K \sqrt{\frac{\log(1/\delta)}{n}} \right] \leq \delta.$$

Now, we turn our attention to FMoM. Theorem 4.3 shows a FMoM achieves the exponential concentration without the sub-Gaussian assumptions, just as in Hilbert space.

Theorem 4.3. *Fix $\delta > 0$. Let \hat{x}_{FMoM} be a Fréchet median-of-means of \hat{x}_j 's where \hat{x} is either \hat{x}_{EM} or \hat{x}_{IM} . Set $k = \lfloor \log(1/\delta)/\psi(7/18, 1/10) \rfloor + 1$. If $\hat{x} = \hat{x}_{EM}$, in addition assume $\kappa_{\min}(\mathcal{X}) > -\infty$. Then*

$$\mathbb{P} \left[d(\hat{x}_{FMoM}, x^*) \geq 11 \sqrt{\frac{\sigma^2 \log(1.4/\delta)}{n}} \right] \leq \delta.$$

The proof is provided in Appendix B.2.

Remark 4.4 (Fréchet mean estimation for $\kappa > 0$). *In the Fréchet mean estimation problem in $CAT(\kappa)$ with $\kappa > 0$, the bounded support condition, as mentioned in Remark 2.2, is necessary to ensure the unique existence of the Fréchet mean. One might wonder whether this assumption automatically implies exponential concentration of the empirical Fréchet mean, as in Euclidean space. However, even in this case, proper finiteness of the metric entropy must be imposed to guarantee exponential concentration (Brunel and Serres, 2024)[Theorem 18]. Conversely, in spaces with $\kappa_{\min} \geq 0$ that satisfy the so-called extendible geodesics condition, polynomial concentration is achieved (Gouic et al., 2019)[Theorem 3.3, 4.2]. Notably, this condition can be met even if the space lacks a strong metric entropy bound (Ahidar-Coutrix et al., 2018; Gouic et al., 2019)[Example 2.6, Corollary 4.4].*

4.2 Boosting a sample covariance estimator

As mentioned in Section 3, the main advantage of this work lies in its extendability to problems beyond the Fréchet mean estimation. The proposed method is applicable for inducing a robust estimator whenever (1) a parameter space can endow a $CAT(\kappa)$ structure, and (2) the original estimator achieves weak concentration (e.g. polynomial). In this section, its application to the estimation of the sample covariance matrix is provided.

Unlike the Fréchet mean problem discussed in Section 4.1, original estimators may not necessarily rely on the geometry of $CAT(\kappa)$ space. The main difficulty in this section lies in obtaining weak concentration of the original estimators with respect to the metric of a $CAT(\kappa)$ space. Fortunately, an estimator built under Euclidean geometry achieves the polynomial concentration with respect to the metrics of a $CAT(\kappa)$ space for this sample covariance problem. Once polynomial concentration for the original estimator is established, the procedure goes similarly to Theorem 4.3, resulting in exponential concentration for a FMoE estimator.

4.2.1 Geometry of symmetric positive definite matrices

Symmetric positive definite (SPD) matrices arise in many fields, hence their estimation and concentration are an important issue. While there are several geometries in matrix spaces, two metrics are specifically tailored to SPD matrices: affine invariant metric and Bures-Wasserstein metric.

First, the affine invariant metric is defined as follows:

$$d_{AI}(A, B) := \|\log A^{-1/2}BA^{-1/2}\|_F.$$

This metric coincides with the Fisher-Rao metric between multivariate Gaussian distributions with fixed mean and covariance matrices A and B (Nielsen, 2023). Additionally, the Fréchet mean of SPD matrices with respect to d_{AI} coincides with the geometric mean and plays an important role in diffusion tensor imaging (Fillard et al., 2005).

The Bures-Wasserstein metric is defined as follows:

$$d_{BW}^2(A, B) := \text{tr}(A) + \text{tr}(B) - 2\text{tr}(A^{1/2}BA^{1/2})^{1/2}.$$

This metric arises naturally in the fields of quantum information and optimal transport. Particularly, this metric is a Wasserstein distance between two multivariate Gaussian distributions with fixed mean and covariance matrices A and B (Bhatia et al., 2019).

While these two metrics both inherit the SPD constraint naturally, their geometries are quite different. The metric space (SPD, d_{AI}) forms a NPC space (Bhatia and Holbrook, 2006)[Proposition 5]. In contrast, (SPD, d_{BW}) forms a non-negative curvature spaces and is not in fact a $CAT(\kappa)$ space for any $\kappa > 0$ (Takatsu, 2009)[Theorem 1.1]. However, when restricted to the set of SPD matrices whose smallest eigenvalue is lower bounded by $\sqrt{3/(2\kappa)}$, it becomes a $CAT(\kappa)$ space (Masart et al., 2019)[Proposition 2].

4.2.2 Concentration analysis

One difficulty in the sample covariance matrix estimation problem lies in the SPD matrix constraint. Many analyses of covariance matrices utilizing matrix norms may potentially violate this constraint. For example, it is non-trivial whether taking a matrix norm Fréchet median of SPD matrices satisfies the SPD matrix constraint.

Conversely, since (SPD, d_{AI}) is a NPC space, the Fréchet median of SPD matrices with respect to d_{AI} resides within the SPD space as discussed in Section 2.2. For (SPD, d_{BW}) one requires additional assumptions that the eigenvalues of matrices should be lower bound by $\sqrt{3/(2\kappa)}$ and matrices are supported within a ball of radius smaller than $D_\kappa/2$. Under these additional conditions, one can also guarantee that their Fréchet median with respect to d_{BW} is SPD, maintaining the eigenvalue lower bound. In this regard, our method produces an estimator that concentrates exponentially with respect to the chosen metric while satisfying the SPD constraint. We set the original estimator as the canonical sample covariance matrix $\hat{\Sigma} = \sum_{i=1}^n X_i X_i^T / n$ (assuming the mean being 0 for simplicity). However, as noted in Section 3, our approach is applicable to any covariance estimator that exhibits weak concentration with respect to the chosen metric; e.g., the Fréchet mean of $X_i X_i^T$'s, which corresponds to FMoM

discussed in Section 4.1.

The following proposition shows the weak concentration of sample covariance matrix estimator w.r.t. both d_{AI} and d_{BW} .

Proposition 4.5 (Polynomial tail bound for covariance matrix estimator). *Let $X_i \stackrel{i.i.d.}{\sim} P \in \mathcal{P}_4(\mathbb{R}^d)$ a distribution with mean 0 and covariance matrix Σ with a fixed dimension d . Let $\hat{\Sigma} = \sum_i X_i X_i^T / n$ be a sample covariance estimator. Then, writing λ_{\min} the smallest eigenvalue of Σ ,*

$$\mathbb{P} \left[d_{AI}(\hat{\Sigma}, \Sigma) \geq \epsilon \right] \leq \frac{Cd^4}{n\lambda_{\min}^2 \left(1 - \exp \left(-\frac{\epsilon}{\sqrt{d}} \right) \right)^2}$$

for some constant $C > 0$ only depends on the moments of P .

For d_{BW} , in addition assume both $\hat{\Sigma}$ and Σ have the eigenvalue lower bound by $\lambda_0 > 0$. Then

$$\mathbb{P} \left[d_{BW}(\hat{\Sigma}, \Sigma) \geq \epsilon \right] \leq \frac{Cd^4}{4n\lambda_0\epsilon^2}$$

with the same C in the above.

We provide a proof in Appendix B.2.

Remark 4.6.

1. The above bound may not be optimal with respect to the dimension d . We do not pursue obtaining the optimal dimension bound, as our main goal is to verify the polynomial concentration with respect to n .
2. For the d_{BW} bound, the eigenvalue lower bound assumption on $\hat{\Sigma}$ may not sound plausible at glance, as $\lambda_{\min}(\hat{\Sigma})$ is a random quantity. However, its value will be in fact concentrated in $B(\lambda_{\min}(\Sigma), C/\sqrt{n})$ for some universal constant $C > 0$ whenever P has a finite fourth moment; see Appendix B.2. Accordingly, it is not too harmful to regard $\lambda_0 = \lambda_{\min}(\Sigma)/2$ in practical applications.

Given weak concentration of $\hat{\Sigma}$, we can proceed with the same technique in Theorem 4.3 to obtain the exponential concentration rate of $\hat{\Sigma}_{FMoE}$. This yields the following result:

Theorem 4.7 (Exponential concentration of median of sample covariance matrices). *Under the same setting in Proposition 4.5, we set $\hat{\Sigma}_{FMoE}$ as in Algorithm 1 with the original estimator being a sample covariance matrix and the metric d being either d_{AI} or d_{BW} . Set $k = \lceil \log(1/\delta)/\psi(0.4, 0.1) \rceil + 1$.*

(a) For $d = d_{AI}$, whenever $n \geq 2kCd^4/\lambda_{\min}^2$, we have

$$\mathbb{P} \left[d_{AI}(\hat{\Sigma}_{FMoE}, \Sigma) \geq -1.3\sqrt{d} \log \left(1 - \frac{9d^2}{\lambda_{\min}} \sqrt{\frac{C \log(1.4/\delta)}{n}} \right) \right] \leq \delta.$$

(b) For $d = d_{BW}$, again assume both $\hat{\Sigma}$ and Σ have the eigenvalue lower bound by $\lambda_0 > 0$. In addition assume conditions in Theorem 3.3(b) holds with $\kappa = 3/(2\lambda_0^2)$, $\theta = \Sigma$, and $\hat{\theta}_j = \hat{\Sigma}_j$. Then, whenever $n > 6\lambda_0 kCd^4$, we have

$$\mathbb{P} \left[d_{BW}(\hat{\Sigma}_{FMoE}, \Sigma) > 12d^2 \sqrt{\frac{C \log(1.4/\delta)}{2n\lambda_0}} \right] \leq \delta.$$

We provide a proof in Appendix B.2.

Remark 4.8.

1. Note that d can vary in Proposition 4.5; as long as $d^4 = o(n/\log n)$ Proposition 4.5 ensures polynomial concentration. Therefore, our method still achieves the exponential concentration in such high dimensional settings.
2. Our estimator is comparable to estimators studied in Abdalla and Zhivotovskiy (2023); Oliveira and Rico (2024), while their estimators require some additional (still weak) assumptions. For example, d_{BW} bound of the estimator in Oliveira and Rico (2024) $\tilde{\Sigma}$ (considering 0 contamination) will be

$$\mathbb{P} \left(d_{BW}(\tilde{\Sigma}, \Sigma) \geq \frac{C\lambda_{\max}\sqrt{d}}{2} \sqrt{\frac{r(\Sigma) + \log(1/\delta)}{n\lambda_0}} \right) \leq \delta.$$

Here, $r(\Sigma)$ denotes the effective rank of Σ . This bound is more desirable with respect to dimension d , but has the additional factor λ_{\max} , and the additive term $\frac{C\lambda_{\max}\sqrt{\lambda_{\max}r(\Sigma)}}{2\sqrt{n\lambda_0}}$. Furthermore, note the dimension dependencies of our bounds come from Proposition 4.5. If one uses a different original estimator, or obtains the tighter dimension bounds of $\hat{\Sigma}$, $\hat{\Sigma}_{FMoE}$ can exhibit a tighter concentration in terms of the dimension. A similar analysis can be conducted for d_{AI} metric bound.

The result of Theorem 4.7 shows that we can obtain the estimator with respect to d_{AI} and d_{BW} that achieves the exponential concentration only with the moment assumptions. On the other hand, the proof of Proposition 4.5 indicates that the original sample covariance matrix can achieve the exponential concentration with respect to d_{AI} and d_{BW} when it exhibits the exponential concentration with respect to matrix norms, which is possible typically under the sub-Gaussian type assumption (Vershynin, 2012)[Corollary 5.50].

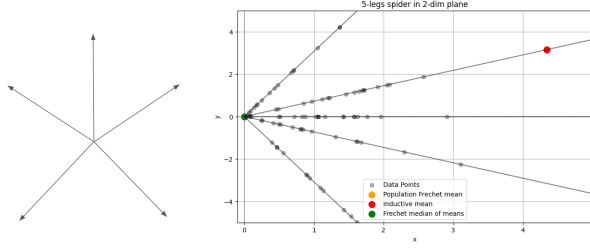


Figure 2: **Left:** A 5-legs spider tree. **Right:** One experiment result on the 5-legs spider tree. The origin denotes the population Fréchet mean, and red and green dot stand for inductive mean and Fréchet median of inductive means respectively.

5 Implementation and experiments

A general procedure to implement our method is displayed in Algorithm 1. To implement the algorithm in practice, the most important part is the choice of the number of blocks k . While there is a precise quantity of k for the given the confidence rate δ as derived in Section 4, in practice k is chosen to guarantee p in Theorem 3.3 to be small enough for $\lfloor n/k \rfloor$ number of data (Lin et al., 2020)[Remark 3.1]. Following this philosophy, the choice of k for each experiment is determined to ensure the original estimator to reasonably concentrate within $\lfloor n/k \rfloor$ data.

We conducted experiments for each example in Section 4: Fréchet mean estimation in a NPC space and boosting the sample covariance estimator with respect to two different metrics d_{AI} and d_{BW} .

As a first experiment, we conducted Fréchet mean estimation problem in a metric tree, a NPC space that is neither a Riemannian manifold nor a Hilbert space. Metric trees are widely used for their theoretical versatility and practical applications (Fakcharoenphol et al., 2003; Abu-Ata and Dragan, 2016). In this regard, estimating Fréchet mean in metric trees is a frequently encountered problem (Romon and Brunel, 2023). Among the various choices of metric trees, we chose a spider tree model. A spider tree is the simplest metric tree that can be embedded in \mathbb{R}^2 . A d -legs spider tree S_d consists of d -copies of the positive real line, called legs, glued together at the origin. Consequently, this space is also an example of a stratified space. Figure 2 illustrates a 5-legs spider tree. Mathematically, S_d can be viewed as a quotient space $\{1, \dots, d\} \times \mathbb{R}_{\geq 0} / \sim$, where the equivalence relation is defined by $(x_1, 0) \sim (y_1, 0)$ for all $x_1, y_1 \in \{1, \dots, d\}$. The metric in this space is defined as follows: for $(x_1, x_2), (y_1, y_2) \in S_d$

$$d(x, y) = \begin{cases} |x_2 - y_2| & \text{if } x_1 = y_1, \\ |x_2| + |y_2| & \text{otherwise.} \end{cases}$$

The equivalence class $(x_1, 0)$ is called the center node. For more explanation on metric trees and spider trees, we refer to Aksoy and Oikhberg (2010).

We chose a 5-legs spider tree for our experiment. For the probability measure, we used $P = \text{Unif}(\{1, \dots, 5\}) \times ((1 - \alpha) |N(1, 1)| + \alpha |N(100, 1)|)$. Here, $|N(\cdot, \cdot)|$ is a distribution of $|X|$ for $X \sim N(\cdot, \cdot)$. α was used to add a small portion of outliers to make the distribution heavy tail. Due to symmetry, the Fréchet mean of P becomes the center node. We used $\alpha = 0.1$, a sample size $n = 100$, and the number of blocks $k = 10$ for this experiment. Figure 2 shows one of the simulation results.

For the covariance estimation problem, we set the population distribution in dimension $d = 10$ as $t_{2.5}(0, \Sigma)$, where Σ was randomly generated while fixing eigenvalues $\lambda_j = j$ for $j = 1, \dots, d$. This distribution has a polynomial tail with a covariance being 5Σ . We used a sample size of $n = 10d^4$ (motivated by the d^4 term in the bound of Proposition 4.5) and the number of blocks $k = 5$ for both d_{AI} and d_{BW} .

Our results from representative experiments are summarized in Table 1 and Figure 3. For all experiments in this section, we conducted 1000 simulations for each task and obtained average squared errors and 95% confidence intervals. Overall, the experiment results indicate that our method achieves higher concentration when the distribution has a heavy tail, provided the block size is chosen appropriately. Additional experiments under different settings (block sizes, population distributions, model spaces) are provided in Appendix C.

Codes for our experiments can be found at https://github.com/wldydd15510/Frechet_median_of_means/, and the implementation details are provided in Appendix C.1.

6 Conclusion

In this work, we extend the Fréchet median of estimators to $\text{CAT}(\kappa)$ spaces, which include almost all spaces of interest in statistics and machine learning. This generalization allows us to obtain exponential concentration on $\text{CAT}(\kappa)$ spaces under mild assumptions. In particular, we leverage this result to famous Fréchet mean and covariance estimation problems. Lastly, supportive numerical evidences are also provided. We conclude the paper with some open questions.

1. Is there a more general space that enables Fréchet median based robust estimation to work? Our analysis on $\text{CAT}(\kappa)$ spaces almost covers the all existing approaches of Fréchet median based robust estimation methods, but not Banach space case proposed in Minsker (2015). This implies

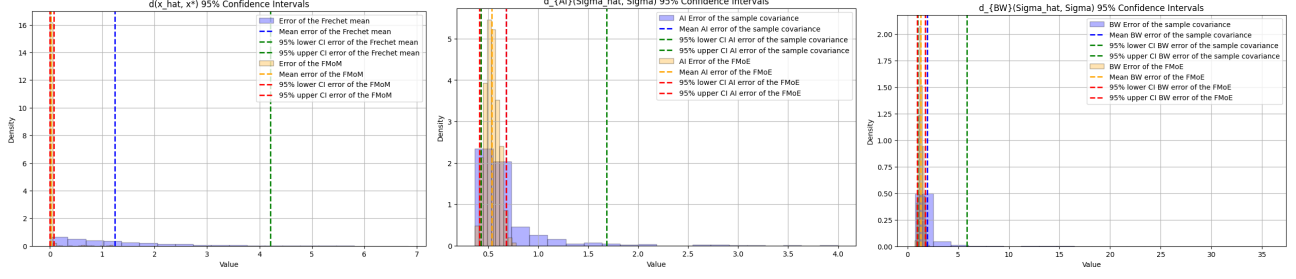


Figure 3: Histogram, mean, and 95% confidence interval for each experiment from 1000 simulations. **Left:** 5-legs spider. **Middle:** (SPD, d_{AI}) . **Right:** (SPD, d_{BW}) . All results indicate our method achieves much stronger concentration as well as much smaller mean squared errors.

Table 1: Mean squared error and 95% confidence interval comparisons from 1000 simulations.

Task	$\mathbb{E}d^2(\hat{\theta}, \theta)$	$\mathbb{E}d^2(\hat{\theta}_{FMoE}, \theta)$	$d(\hat{\theta}, \theta)$ CI	$d(\hat{\theta}_{FMoE}, \theta)$ CI
5-spider tree	3.1244	1.2×10^{-5}	[0.0110, 4.4202]	$[9.6 \times 10^{-5}, 0.0086]$
Covariance (d_{AI})	0.6057	0.2931	[0.4266, 1.6865]	[0.4132, 0.6792]
Covariance (d_{BW})	8.3281	1.7360	[0.9936, 5.8796]	[0.9443, 1.7409]

there is a possibility of more generalizations, for example, to non-Riemannian Finsler manifolds which are not Alexandrov spaces. In NPC spaces, the concept of generalized CAT(0) (Khamisi and Shukri, 2017) incorporates the CAT(0) spaces and Banach spaces. However, we are not aware of the generalization of this concept beyond NPC spaces.

2. Is there a way to overcome the curvature upper bound condition? Our proof crucially relies on the curvature upper bound. However, some spaces without the curvature upper bound are widely studied, e.g., Wasserstein space over \mathbb{R}^d . Whether one can extend the similar methods to such spaces is unknown.
3. Can one obtain bounds independent of population quantities? When one wants to determine the optimal number of blocks or construct the confidence regions, unknown population quantities in the bound hinder us from obtaining the precise value. Some methods were developed to overcome those dependencies, e.g., adaptation methods. Developing such a method for our estimator is an important factor for practical applications.

Acknowledgements

We appreciate Young-Heon Kim, Inyoung Ryu, and Byeongsu Yu for their helpful discussions. In addition, we appreciate anonymous reviewers for their insightful feedback during the review process. JP acknowledges support from NSF DMS 2210689. AB acknowledges NSF DMS 2210689, NSF DMS 1916371 for supporting this project.

References

Bibliography

- Pedro Abdalla and Nikita Zhivotovskiy. Covariance estimation: Optimal dimension-free guarantees for adversarial corruption and heavy tails, 2023. URL <https://arxiv.org/abs/2205.08494>.
- Muad Abu-Ata and Feodor F. Dragan. Metric tree-like structures in real-world networks: an empirical study. *Netw.*, 67(1):49–68, January 2016. ISSN 0028-3045. doi: 10.1002/net.21631. URL <https://doi.org/10.1002/net.21631>.
- Adil Ahidar-Coutrix, Thibaut Le Gouic, and Quentin Paris. Convergence rates for empirical barycenters in metric spaces: curvature, convexity and extendable geodesics. *Probability Theory and Related Fields*, 177:323–368, 2018. URL <https://api.semanticscholar.org/CorpusID:88522339>.
- Asuman Aksoy and Timur Oikhberg. Some results on metric trees. *Banach Center Publications*, 91, 07 2010. doi: 10.4064/bc91-0-1.
- A. D. Aleksandrov. A theorem on triangles in a metric space and some of its applications. *Trudy Mat. Inst. Steklov.*, 38:5–23, 1951. ISSN 0371-9685.
- Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. In *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '96, page 20–29, New York, NY, USA, 1996. Association for Computing Machinery. ISBN 0897917855. doi: 10.1145/237814.237823. URL <https://doi.org/10.1145/237814.237823>.
- J. Anderson. *Hyperbolic Geometry*. Springer Undergraduate Mathematics Series. Springer London, 2005. ISBN 9781852339340. URL https://books.google.com/books?id=NYVnzAX8_qoC.
- T. Ando and J. L. van Hemmen. An inequality for trace ideals. *Communications in Mathematical Physics*, 76(2):143 – 148, 1980.
- Marc Arnaudon, Frédéric Barbaresco, and Le Yang. *Medians and Means in Riemannian Geometry: Existence, Uniqueness and Computation*, pages 169–197. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-30232-9. doi: 10.1007/978-3-642-30232-9_8. URL https://doi.org/10.1007/978-3-642-30232-9_8.
- Jean-Yves Audibert and Olivier Catoni. Robust linear least squares regression. *Annals of Statistics*, 39(5): 2766–2794, 2011. doi: 10.1214/11-AOS918. URL <https://hal.science/hal-00522534>. 29 pages.
- Miroslav Bačák. Computing medians and means in hadamard spaces. *SIAM Journal on Optimization*, 24(3):1542–1566, 2014. doi: 10.1137/140953393. URL <https://doi.org/10.1137/140953393>.
- Rajendra Bhatia and John Holbrook. Riemannian geometry and matrix geometric means. *Linear Algebra and its Applications*, 413(2):594–618, 2006. ISSN 0024-3795. doi: <https://doi.org/10.1016/j.laa.2005.08.025>. URL <https://www.sciencedirect.com/science/article/pii/S0024379505004350>. Special Issue on the 11th Conference of the International Linear Algebra Society, Coimbra, 2004.
- Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the bures–wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019. ISSN 0723-0869. doi: <https://doi.org/10.1016/j.exmath.2018.01.002>. URL <https://www.sciencedirect.com/science/article/pii/S0723086918300021>.
- Louis J. Billera, Susan P. Holmes, and Karen Vogtmann. Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, 27(4):733–767, 2001. ISSN 0196-8858. doi: <https://doi.org/10.1006/aama.2001.0759>. URL <https://www.sciencedirect.com/science/article/pii/S0196885801907596>.
- Nicolas Boria, Sébastien Bogleux, Benoit Gaüzère, and Luc Brun. Generalized median graph via iterative alternate minimizations. In *Graph-Based Representations in Pattern Recognition: 12th IAPR-TC-15 International Workshop, GbRPR 2019, Tours, France, June 19–21, 2019, Proceedings*, page 99–109, Berlin, Heidelberg, 2019. Springer-Verlag. ISBN 978-3-030-20080-0. doi: 10.1007/978-3-030-20081-7_10. URL https://doi.org/10.1007/978-3-030-20081-7_10.
- S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013. ISBN 9780199535255. URL <https://books.google.com/books?id=koNqWRluhPOC>.
- Martin R. Bridson and André Haefliger. *Metric spaces of non-positive curvature*, volume 319 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1999. ISBN 3-540-64324-9. doi: 10.1007/978-3-662-12494-9. URL <https://doi.org/10.1007/978-3-662-12494-9>.
- Samuel Brown. A gluing theorem for negatively curved complexes. *Journal of the London Mathematical Society*, 93(3):741–762, 04 2016. ISSN 0024-6107. doi: 10.1112/jlms/jdw021. URL <https://doi.org/10.1112/jlms/jdw021>.
- Victor-Emmanuel Brunel and Jordan Serres. Concentration of empirical barycenters in metric spaces. In Claire Vernade and Daniel Hsu, editors, *Proceedings*

- of *The 35th International Conference on Algorithmic Learning Theory*, volume 237 of *Proceedings of Machine Learning Research*, pages 337–361. PMLR, 25–28 Feb 2024. URL <https://proceedings.mlr.press/v237/brunel24a.html>.
- D. Burago, I.U.D. Burago, and S. Ivanov. *A Course in Metric Geometry*. Crm Proceedings & Lecture Notes. American Mathematical Society, 2001a. ISBN 9780821821299. URL <https://books.google.com/books?id=dRmIAwAAQBAJ>.
- Dmitri Burago, Yuri Burago, and Sergei Ivanov. *A course in metric geometry*, volume 33 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2001b. ISBN 0-8218-2129-6. doi: 10.1090/gsm/033. URL <https://doi.org/10.1090/gsm/033>.
- Yu. Burago, M. Gromov, and G. Perelman. A. D. Aleksandrov spaces with curvatures bounded below. *Uspekhi Mat. Nauk*, 47(2(284)):3–51, 222, 1992. ISSN 0042-1316,2305-2872. doi: 10.1070/RM1992v047n02ABEH000877. URL <https://doi.org/10.1070/RM1992v047n02ABEH000877>.
- Olivier Catoni. Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148–1185, 2012.
- Xiongtao Dai and Sara López-Pintado. Tukey’s depth for object data. *Journal of the American Statistical Association*, 118:1760 – 1772, 2021.
- Jittat Fakcharoenphol, Satish Rao, and Kunal Talwar. A tight bound on approximating arbitrary metrics by tree metrics. In *Proceedings of the Thirty-Fifth Annual ACM Symposium on Theory of Computing*, STOC ’03, page 448–455, New York, NY, USA, 2003. Association for Computing Machinery. ISBN 1581136749. doi: 10.1145/780542.780608. URL <https://doi.org/10.1145/780542.780608>.
- Charles Fefferman, Sergei Ivanov, Yaroslav Kurylev, Matti Lassas, and Hariharan Narayanan. Reconstruction and interpolation of manifolds i: The geometric whitney problem, 2019. URL <https://arxiv.org/abs/1508.00674>.
- Pierre Fillard, Vincent Arsigny, Xavier Pennec, Paul M. Thompson, and Nicholas Ayache. Extrapolation of sparse tensor fields: Application to the modeling of brain variability. *Lecture Notes in Computer Science*, 3565:27–38, 2005. ISSN 0302-9743. doi: 10.1007/11505730_3. 19th International Conference on Information Processing in Medical Imaging, IPMI 2005 ; Conference date: 10-07-2005 Through 15-07-2005.
- P. Thomas Fletcher, Suresh Venkatasubramanian, and Sarang Joshi. The geometric median on riemannian manifolds with application to robust atlas estimation. *NeuroImage*, 45(1, Supplement 1):S143–S152, 2009. ISSN 1053-8119. doi: <https://doi.org/10.1016/j.neuroimage.2008.10.052>. URL <https://www.sciencedirect.com/science/article/pii/S1053811908012019>. Mathematics in Brain Imaging.
- Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3–5):75–174, February 2010. ISSN 0370-1573. doi: 10.1016/j.physrep.2009.11.002. URL <http://dx.doi.org/10.1016/j.physrep.2009.11.002>.
- Dan Geiger, David Heckerman, Henry King, and Christopher Meek. Stratified exponential families: Graphical models and model selection. *The Annals of Statistics*, 29(2):505 – 529, 2001. doi: 10.1214/aos/1009210550. URL <https://doi.org/10.1214/aos/1009210550>.
- Thibaut Le Gouic, Quentin Paris, Philippe Rigollet, and Austin J. Stromme. Fast convergence of empirical barycenters in alexandrov spaces and the wasserstein space. *Journal of the European Mathematical Society*, 2019. URL <https://api.semanticscholar.org/CorpusID:199405372>.
- Peter Hall, G. S. Watson, and Javier Cabrera. Kernel density estimation with spherical data. *Biometrika*, 74(4):751–762, 1987. ISSN 00063444. URL <http://www.jstor.org/stable/2336469>.
- Susan Holmes. Statistics for phylogenetic trees. *Theoretical Population Biology*, 63(1):17–32, 2003. ISSN 0040-5809. doi: [https://doi.org/10.1016/S0040-5809\(02\)00005-9](https://doi.org/10.1016/S0040-5809(02)00005-9). URL <https://www.sciencedirect.com/science/article/pii/S0040580902000059>.
- Jaehong Jeong, Mikyoung Jun, and Marc G. Genton. Spherical Process Models for Global Spatial Statistics. *Statistical Science*, 32(4):501 – 513, 2017. doi: 10.1214/17-STS620. URL <https://doi.org/10.1214/17-STS620>.
- M. A. Khamsi and S. A. Shukri. Generalized CAT(0) spaces. *Bulletin of the Belgian Mathematical Society - Simon Stevin*, 24(3):417 – 426, 2017. doi: 10.36045/bbms/1506477690. URL <https://doi.org/10.36045/bbms/1506477690>.
- Michael Kunzinger and Roland Steinbauer. Alexandrov spaces: Lecture notes. https://www.mat.univie.ac.at/~mike/teaching/ss18/Alexandrov_spaces.pdf, 2018.
- Michael Kunzinger, Roland Steinbauer, and Milena Stojković. The exponential map of a c1,1-metric. *Differential Geometry and its Applications*, 34:14–24, 2014. ISSN 0926-2245. doi: <https://doi.org/10.1016/j.difgeo.2014.03.005>.

- URL <https://www.sciencedirect.com/science/article/pii/S0926224514000370>.
- Georg Köstenberger and Thomas Stark. Robust signal recovery in hadamard spaces, 2024. URL <https://arxiv.org/abs/2307.06057>.
- M. Lerasle and R. I. Oliveira. Robust empirical mean estimators, 2011. URL <https://arxiv.org/abs/1112.3914>.
- Lizhen Lin, Drew Lazar, Bayan Sarpabayeva, and David B. Dunson. Robust optimization and inference on manifolds. *Statistica Sinica*, 2020. URL <https://api.semanticscholar.org/CorpusID:219636463>.
- Gábor Lugosi and Shahar Mendelson. Risk minimization by median-of-means tournaments. *Journal of the European Mathematical Society*, 2016. URL <https://api.semanticscholar.org/CorpusID:15881119>.
- Gábor Lugosi and Shahar Mendelson. Sub-Gaussian estimators of the mean of a random vector. *The Annals of Statistics*, 47(2):783 – 794, 2019. doi: 10.1214/17-AOS1639. URL <https://doi.org/10.1214/17-AOS1639>.
- Gábor Lugosi and Shahar Mendelson. Robust multivariate mean estimation: The optimality of trimmed mean. *Annals of Statistics*, 49(1):–, 2021. doi: 10.1214/20-AOS1961.
- Estelle Massart, Julien M. Hendrickx, and P.-A. Absil. Curvature of the manifold of fixed-rank positive-semidefinite matrices endowed with the bures–wasserstein metric. In Frank Nielsen and Frédéric Barbaresco, editors, *Geometric Science of Information*, pages 739–748, Cham, 2019. Springer International Publishing. ISBN 978-3-030-26980-7.
- Jonathan C. Mattingly, Ezra Miller, and Do Tran. Central limit theorems for fréchet means on stratified spaces, 2023a. URL <https://arxiv.org/abs/2311.09455>.
- Jonathan C. Mattingly, Ezra Miller, and Do Tran. Geometry of measures on smoothly stratified metric spaces, 2023b. URL <https://arxiv.org/abs/2311.09453>.
- Jonathan C. Mattingly, Ezra Miller, and Do Tran. Shadow geometry at singular points of $\text{cat}(k)$ spaces, 2023c. URL <https://arxiv.org/abs/2311.09451>.
- Stanislav Minsker. Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21(4):2308 – 2335, 2015. doi: 10.3150/14-BEJ645. URL <https://doi.org/10.3150/14-BEJ645>.
- Stanislav Minsker. Uniform bounds for robust mean estimators, 2019. URL <https://arxiv.org/abs/1812.03523>.
- Stanislav Minsker, Sanvesh Srivastava, Lizhen Lin, and David B. Dunson. Robust and scalable bayes via a median of subset posterior measures. *Journal of Machine Learning Research*, 18(124):1–40, 2017. URL <http://jmlr.org/papers/v18/16-655.html>.
- Nina Miolane, Nicolas Guigui, Alice Le Brigant, Johan Mathe, Benjamin Hou, Yann Thanwerdas, Stefan Heyder, Olivier Peltre, Niklas Koep, Hadi Zaatiti, Hatem Hajri, Yann Cabanes, Thomas Gerald, Paul Chauchat, Christian Shewmake, Daniel Brooks, Bernhard Kainz, Claire Donnat, Susan Holmes, and Xavier Pennec. Geomstats: A python package for riemannian geometry in machine learning. *Journal of Machine Learning Research*, 21(223):1–9, 2020. URL <http://jmlr.org/papers/v21/19-027.html>.
- Nina Miolane, Luís F. Pereira, Saiteja Utpala, Nicolas Guigui, Alice Le Brigant, Hzaatiti, Yann Cabanes, Johan Mathe, Niklas Koep, elodiemaignant, ythanwerdas, xpennec, tgeral68, Christian, Tra My Nguyen, Olivier Peltre, John Harvey, pchauchat, julesdeschamps, Quentin Barthélemy, mortenapedersen, Maya95assal, Abdellaoui-Souhail, Adele Myers, Felix Ambellan, Florent-Michel, Stefan Heyder, Shubham Talbar, Yann de Mont-Marin, and Marius. geomstats/geomstats: Geomstats v2.8.0, September 2024. URL <https://doi.org/10.5281/zenodo.13737807>.
- A. S. Nemirovsky and D. B. Yudin. Problem complexity and method efficiency in optimization. 1983.
- Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/59dfa2df42d9e3d41f5b02bfc32229dd-Paper.pdf.
- Frank Nielsen. A simple approximation method for the fisher–rao distance between multivariate normal distributions. *Entropy*, 25(4), 2023. ISSN 1099-4300. doi: 10.3390/e25040654. URL <https://www.mdpi.com/1099-4300/25/4/654>.
- Roberto I. Oliveira and Zoraida F. Rico. Improved covariance estimation: optimal robustness and sub-gaussian guarantees under heavy tails, 2024. URL <https://arxiv.org/abs/2209.13485>.
- Alexander Petersen and Hans-Georg Müller. Functional data analysis for density functions by transformation to a Hilbert space. *The Annals of Statistics*, 44(1): 183 – 218, 2016. doi: 10.1214/15-AOS1363. URL <https://doi.org/10.1214/15-AOS1363>.
- Gabriel Romon and Victor-Emmanuel Brunel. Convex

- generalized fréchet means in a metric tree, 2023. URL <https://arxiv.org/abs/2310.17435>.
- Christof Schötz. Variance inequalities for transformed fréchet means in hadamard spaces, 2024. URL <https://arxiv.org/abs/2310.13668>.
- Karl-Theodor Sturm. Metric spaces of lower bounded curvature. *Expositiones Mathematicae*, 17:035–048, 1999.
- Karl-Theodor Sturm. Probability measures on metric spaces of nonpositive curvature. *Heat Kernels and Analysis on Manifolds*, 2(1):217–240, 2000.
- Asuka Takatsu. On wasserstein geometry of the space of gaussian measures, 2009. URL <https://arxiv.org/abs/0801.2250>.
- Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Yonina C. Eldar and GittaEditors Kutyniok, editors, *Compressed Sensing: Theory and Practice*, pages 210–268. Cambridge University Press, 2012. ISBN 9780511794308. doi: 10.1017/CBO9780511794308.006.
- Takumi Yokota. Convex functions and p -barycenter on CAT(1)-spaces of small radii. *Tsukuba Journal of Mathematics*, 41(1):43 – 80, 2017. doi: 10.21099/tkbjm/1506353559. URL <https://doi.org/10.21099/tkbjm/1506353559>.
- Ho Yun and Byeong U. Park. Exponential concentration for geometric-median-of-means in non-positive curvature spaces. *Bernoulli*, 29(4), November 2023. ISSN 1350-7265. doi: 10.3150/22-bej1569. URL <http://dx.doi.org/10.3150/22-BEJ1569>.
- Yikun Zhang and Yen-Chi Chen. Kernel smoothing, mean shift, and their learning theory with directional data. *Journal of Machine Learning Research*, 22 (154):1–92, 2021. URL <http://jmlr.org/papers/v22/20-1194.html>.

Checklist

1. For all models and algorithms presented, check if you include:
 - (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [✓Yes/No/Not Applicable]
 - We included mathematical detail in Sections 3 and 4.
 - (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [✓Yes/No/Not Applicable]
 - We included the a brief description of the algorithm when the algorithm is feasible in the end of Section 3. However, we did not include the complexity, as it depends on the choice of specific implementation detail, e.g., which Fréchet median algorithm to use or what the parameter space is.
 - (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [✓Yes/No/Not Applicable]
 - We will attach the source code in the supplement.
2. For any theoretical claim, check if you include:
 - (a) Statements of the full set of assumptions of all theoretical results. [✓Yes/No/Not Applicable]
 - Sections 3 and 4 specifically addressed theoretical detail with full explanations.
 - (b) Complete proofs of all theoretical results. [✓Yes/No/Not Applicable]
 - All proofs are provided in Appendix B.
 - (c) Clear explanations of any assumptions. [✓Yes/No/Not Applicable]
 - We made remarks for the most of Theorems discussing the assumptions.
3. For all figures and tables that present empirical results, check if you include:
 - (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [✓Yes/No/Not Applicable]
 - We will attach the source code in the supplement, which contains the exact code one can reproduce.
 - (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [✓Yes/No/Not Applicable]
 - We provided implementation and experiment detail in Appendix C.
 - (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [✓Yes/No/Not Applicable]
 - We explicitly mentioned the measuring criteria and number of simulations in Section 5 and Appendix C.
 - (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [✓Yes/No/Not Applicable]
 - We explicitly mentioned the computing resource we used in Appendix C.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
 - (a) Citations of the creator If your work uses existing assets. [✓Yes/No/Not Applicable]
 - We made a citation of the package we used in Appendix C.
 - (b) The license information of the assets, if applicable. [Yes/No/✓Not Applicable]
 - We only utilized the open-source libraries.
 - (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/✓Not Applicable]
 - (d) Information about consent from data providers/curators. [Yes/No/✓Not Applicable]
 - (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/✓Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
 - (a) The full text of instructions given to participants and screenshots. [Yes/No/✓Not Applicable]
 - (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/ ✓Not Applicable]
 - (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/✓Not Applicable]

Robust Estimation in metric spaces: Achieving Exponential Concentration with a Fréchet Median: Supplementary Materials

A Preliminaries

A.1 $\text{CAT}(\kappa)$ spaces

$\text{CAT}(\kappa)$ spaces, also known as Alexandrov spaces, are a generalization of Riemannian manifolds with a bounded upper sectional curvature. $\text{CAT}(\kappa)$ spaces are locally compact complete length spaces with a uniformly bounded curvature. Due to its generality and rich implications in geometry, they have been extensively studied more than seven decades (Aleksandrov (1951); Burago et al. (1992); Bridson and Haefliger (1999); Sturm (1999); Burago et al. (2001b)).

There are three equivalent ways to define them via the *distance functions*, the *speed of geodesics* and the *comparison triangles*, respectively. For the sake of completeness, we introduce all of them. For more details, see Kunzinger and Steinbauer (2018).

To begin with, we define model spaces M_κ^2 .

Definition A.1 (Model spaces). *For $\kappa \in \mathbb{R}$, we denote by M_κ^2 the following spaces:*

- (i) M_0^2 is the Euclidean plane \mathbb{R}^2 .
- (ii) For $\kappa > 0$, $M_\kappa^2 = \frac{1}{\sqrt{\kappa}}\mathbb{S}^2$ where \mathbb{S}^2 is the 2-dimensional sphere.
- (iii) For $\kappa < 0$, $M_\kappa^2 = \frac{1}{\sqrt{|\kappa|}}\mathbb{H}^2$ where \mathbb{H}^2 is the 2-dimensional hyperbolic space.

We use d_κ to denote the intrinsic metric of M_κ^2 .

Throughout this section, let (\mathcal{X}, d) be a complete, connected metric space.

Definition A.2 (Length space). *(\mathcal{X}, d) is called a length space if the metric d is the intrinsic length metric, i.e. for any $x, y \in \mathcal{X}$, and any path $\alpha : [a, b] \rightarrow \mathcal{X}$ such that $\alpha(a) = x, \alpha(b) = y$,*

$$d(x, y) = \inf_{\alpha} \{\text{length } \alpha\}$$

where the length of path α is defined as

$$\text{length } \alpha := \sup \left\{ \sum_{i=1}^n d(\alpha(t_i), \alpha(t_{i-1})) : a = t_0 \leq t_1 \leq \dots \leq t_n = b \right\}.$$

Definition A.3 (Complete length space). *A path $\alpha : [a, b] \rightarrow \mathcal{X}$ is called a shortest path if for any other paths β connecting $\alpha(a)$ to $\alpha(b)$, $\text{length } \alpha \leq \text{length } \beta$.*

For a length space (\mathcal{X}, d) , α is a shortest path connecting $\alpha(a)$ to $\alpha(b)$ if and only if

$$d(\alpha(a), \alpha(b)) = \text{length } \alpha.$$

If for any $x, y \in \mathcal{X}$, there is a shortest path connecting x to y , then (\mathcal{X}, d) is called strictly intrinsic, and a complete length space.

Definition A.4 (Geodesic space). *We call a complete length space as a geodesic space, and shortest paths as geodesics.*

Definition A.5. *For $x, y \in \mathcal{X}$, write $T = d(x, y)$. α is called a geodesic with unit speed connecting x to y , or a geodesic parametrized by arc length, if $\alpha : [0, T] \rightarrow \mathcal{X}$ satisfies $d(x, \alpha(t)) = t$ for any $t \in [0, T]$.*

We use $[xy] = \alpha$ to denote a geodesic with unit speed connecting x to y .

Now, we provide the first definition of $\text{CAT}(\kappa)$ spaces via distance function. In simple terms, this definition states that curvature distorts the space such that, from any fixed point, any point lies in the middle of the geodesic is further away than it would be in the model space.

Given $x, y \in \mathcal{X}$, fix $p \in \mathcal{X}$. *One-dimensional distance function* (at p) is defined as

$$g(t) := d(p, \alpha(t)).$$

We will compare $g(t)$ to an appropriate one-dimensional distance function in the model space. To this end, given $x, y, p \in \mathcal{X}$, choose $\tilde{x}, \tilde{y}, \tilde{p}$ in the model space M_κ^2 such that

$$d_\kappa(\tilde{x}, \tilde{y}) = d(x, y), d_\kappa(\tilde{p}, \tilde{x}) = d(p, x) \text{ and } d_\kappa(\tilde{p}, \tilde{y}) = d(p, y).$$

For $\kappa > 0$ we further assume

$$d(x, y) + d(p, x) + d(p, y) \leq 2D_\kappa = \frac{2\pi}{\sqrt{\kappa}}. \quad (1)$$

We call $\{\tilde{x}, \tilde{y}, \tilde{p}\}$ as *comparison configuration*, \tilde{p} as a *reference point*, and $[\tilde{x}\tilde{y}] = \tilde{\alpha} : [0, T] \rightarrow M_\kappa^2$, the geodesic connecting \tilde{x} to \tilde{y} with unit speed in M_κ^2 , as a *comparison segment* respectively. Note that comparison configuration is unique up to rigid motions.

Definition A.6 (Comparison distance function). *Define*

$$\tilde{g}_\kappa(t) := d_\kappa(\tilde{p}, \tilde{\alpha}(t))$$

the model space distance from the reference point \tilde{p} to the comparison segment $[\tilde{x}\tilde{y}] = \tilde{\alpha}$. This \tilde{g}_κ is called the comparison distance function of g .

Definition A.7 (Distance condition). *A geodesic space (\mathcal{X}, d) is a $\text{CAT}(\kappa)$ space if every point in \mathcal{X} has a neighbourhood U such that the following holds: for any point $p \in U$ and all $[xy] = \alpha \subseteq U$, the comparison distance function $\tilde{g}_\kappa(t)$ for one-dimensional distance function $g(t) = d(p, \alpha(t))$ satisfies*

$$\tilde{g}_\kappa(t) \geq g(t) \text{ for all } t \in [0, T]. \quad (2)$$

For $\kappa > 0$, we further assume (1).

The next definition is followed by triangle comparison. In fact, it is straightforwardly equivalent to the definition via distance function.

For $x, y, z \in \mathcal{X}$, a *triangle* is a triangle with sides $[xy], [yz], [xz]$, each of which is a geodesic with unit speed of a pair of three points. We write $\triangle xyz$ to denote for the triangle of x, y, z with side lengths $d(x, y), d(y, z)$ and $d(x, z)$. Notice that since there are multiple geodesics, x, y, z cannot determine $\triangle xyz$ uniquely. However, every $\triangle xyz$ has the same side lengths.

A *comparison triangle* $\triangle \tilde{x}\tilde{y}\tilde{z}$ is a triangle of a comparison configuration in the model space; hence $\triangle \tilde{x}\tilde{y}\tilde{z}$ has the same side lengths as $\triangle xyz$. Clearly, a comparison triangle is also unique up to rigid motions. Let $[\tilde{x}\tilde{y}]$ be the geodesic in the model space connecting \tilde{x} to \tilde{y} with the length $d(x, y)$: i.e. the side between \tilde{x} and \tilde{y} of the comparison triangle $\triangle \tilde{x}\tilde{y}\tilde{z}$. If $\alpha = [xy]$, we use $\tilde{\alpha} := [\tilde{x}\tilde{y}]$, the natural comparison geodesic in the model space induced by α .

Definition A.8 (Triangle condition). *A geodesic space (\mathcal{X}, d) is $\text{CAT}(\kappa)$ space if every point in \mathcal{X} has a neighbourhood U such that the following holds: for every triangle $\triangle xyz \subseteq U$ and every point $w \in [xz]$,*

$$d(w, y) \leq d_\kappa(\tilde{w}, \tilde{y}) \quad (3)$$

where $\tilde{w} \in [\tilde{x}\tilde{z}]$ such that $d_\kappa(\tilde{x}, \tilde{w}) = d(x, w)$, or $\tilde{w} = \tilde{\alpha}(d(x, w))$. For $\kappa > 0$, we further assume (1).

The last one to define $\text{CAT}(\kappa)$ spaces is to compare angles. In words, angles in upper bounded curved spaces are smaller than in the model space. This is indeed equivalent to the fact that two geodesics in $\text{CAT}(\kappa)$ spaces starting at the same initial point with a certain angle is indeed further than those with the same angle in the model space.

Definition A.9. Given distinct points $x, y, z \in \mathcal{X}$, the comparison angle $\angle \widetilde{x}\widetilde{y}\widetilde{z}$ at y is the angle at \widetilde{y} of the comparison triangle $\triangle \widetilde{x}\widetilde{y}\widetilde{z}$. Alternatively,

$$\angle \widetilde{x}\widetilde{y}\widetilde{z} := \arccos \frac{d(x, y)^2 + d(y, z)^2 - d(x, z)^2}{2d(x, y)d(y, z)}.$$

Definition A.10. Let $\alpha, \beta : [0, T] \rightarrow \mathcal{X}$ be two paths with the same initial point p . The angle between α and β is defined as

$$\angle(\alpha, \beta) := \lim_{s, t \rightarrow 0} \angle \widetilde{\alpha}(s) \widetilde{p} \widetilde{\beta}(t)$$

if the limit exists. Given three distinct points $x, y, z \in \mathcal{X}$, the angle of $\triangle xyz$ at y is defined as

$$\angle xyz := \angle([xy], [yz]).$$

Remark A.11. It is not trivial that the angle between two paths always exists. However, if $\kappa \leq 0$, it always exists: see Kunzinger and Steinbauer (2018)[3.3.2 Proposition]. If $\kappa > 0$, (1) should be required.

Definition A.12 (Angle condition). A geodesic space (\mathcal{X}, d) is $\text{CAT}(\kappa)$ space if every point in \mathcal{X} has a neighbourhood U such that the following holds: for every triangle $\triangle xyz \subseteq U$ the angles of $\triangle xyz$ satisfy

$$\angle yxz \leq \angle \widetilde{y}\widetilde{x}\widetilde{z}, \angle zyx \leq \angle \widetilde{z}\widetilde{y}\widetilde{x}, \text{ and } \angle xzy \leq \angle \widetilde{x}\widetilde{z}\widetilde{y} \quad (4)$$

where $\triangle \widetilde{x}\widetilde{y}\widetilde{z}$ is a comparison triangle of $\triangle xyz$ in the model space M_κ^2 . For $\kappa > 0$, we further assume (1).

The three conditions, (2), (3) and (4), are equivalent to define $\text{CAT}(\kappa)$ spaces.

Theorem A.13. All the definitions of $\text{CAT}(\kappa)$ spaces, that is the distance condition A.7, the triangle condition A.8, and the angle condition A.12 are equivalent.

Lastly, we note that analogous statements can be made for spaces with bounded lower curvature by reversing the direction of the inequalities in the definitions above. Such spaces arise in the fields of optimal transport and Wasserstein geometry.

A.1.1 Fréchet mean estimation in NPC spaces

In this appendix, we consider two different estimators for population Fréchet mean in NPC spaces: empirical Fréchet mean and inductive mean. Throughout this appendix, we fix (\mathcal{X}, d) to be an NPC space.

First is an empirical Fréchet mean, the most natural way to estimate the Fréchet mean.

Definition A.14 (Empirical Fréchet mean). Given n data points $x_1, \dots, x_n \in \mathcal{X}$, the empirical Fréchet mean is defined by

$$\widehat{x}_{EM} = \operatorname{argmin}_{x \in \mathcal{X}} \frac{1}{n} \sum_{j=1}^n d^2(x, x_j).$$

Inductive mean is another natural way to estimate the Fréchet mean proposed by Sturm (2000), coming from the generalization of the law of large numbers. As the name implies, the inductive mean is defined ‘inductively’.

Definition A.15 (Inductive mean). Set δ and k in a same way. Given n data points $x_1, \dots, x_n \in \mathcal{X}$, define a sequence s_i as follows:

$$s_1 = x_1, \quad s_i = \left(1 - \frac{1}{i}\right) s_{i-1} + \frac{1}{i} x_i \text{ for } i = 2, \dots, n$$

where the summation can be understood as a geodesic interpolation with a given ratio. Then, the resulting $s_n := \widehat{x}_{IM}$ is called inductive mean.

These two estimators coincide to the arithmetic mean in Hilbert space, but not in the general metric space. As pointed out in Sturm (2000), the inductive mean depends on the permutation of x_i 's, unlike the empirical Fréchet mean. However, the inductive mean has certain advantages over the empirical Fréchet mean from both theoretical and practical perspectives. Practically, if you have a geodesic interpolation oracle, computing the inductive mean is straightforward. Theoretically, the following proposition highlights the benefits of using the inductive mean.

Proposition A.16. *Let $x_1, \dots, x_n \stackrel{i.i.d.}{\sim} P \in \mathcal{P}_2(\mathcal{X})$. Let \hat{x} be either \hat{x}_{EM} or \hat{x}_{IM} . If $\hat{x} = \hat{x}_{EM}$, in addition assume $\kappa_{\min}(\mathcal{X}) > -\infty$. Then,*

$$\mathbb{E} [d^2(\hat{x}, x^*)] \leq \frac{\sigma^2}{n}$$

where $\sigma^2 = \mathbb{E}_{y \sim P}[d^2(x^*, y)]$ is the second moment of P .

Proof. For \hat{x}_{EM} case, see Gouic et al. (2019)[Corollary 3.4]. For \hat{x}_{IM} case, see Sturm (2000)[Theorem 4.7]. \square

The fact that \hat{x}_{IM} does not require the additional condition on the curvature lower bound of \mathcal{X} is favorable, as there are cases where we need to deal with spaces of unbounded curvature, such as metric trees or statistical manifolds with Gamma and Dirichlet distributions, to name a few.

B Deferred proofs

This appendix contains detailed proofs of the results that are missing in the main paper.

B.1 Proofs in Section 3

Proof of Lemma 3.1:

Proof. As in Minsker (2015)[Lemma 2.1] suppose the implication is false for the contradiction. Without the loss of generality, it means that $d(x_j, z) \leq r$ for $j = 1, \dots, \lfloor (1 - \alpha)k \rfloor + 1$.

We separately analyze the cases when $\kappa \leq 0$ and $\kappa > 0$.

CASE I: $\kappa \leq 0$.

Recall

$$F(x) := \frac{1}{k} \sum_{j=1}^k d(x, x_j).$$

Note that $F(x)$ always admits a minimizer followed by Bačák (2014)[Lemma 2.3]. Let x^* be a median (a minimizer of $F(\cdot)$) and $z \neq x^*$ be an arbitrary point. Let $\alpha : [0, 1] \rightarrow \mathcal{X}$ be a geodesic curve in \mathcal{X} such that $\alpha(0) = x^*$ and $\alpha(1) = z$. Since $\alpha(0) = x^*$ is a minimizer of F , we have

$$\limsup_{t \rightarrow 0} \frac{F(\alpha(t)) - F(\alpha(0))}{t} \geq 0. \quad (5)$$

Now, notice that

$$\begin{aligned} \limsup_{t \rightarrow 0} \frac{F(\alpha(t)) - F(\alpha(0))}{t} &\leq \sum_{j=1}^k \limsup_{t \rightarrow 0} \frac{d(x_j, \alpha(t)) - d(x_j, x^*)}{t} \mathbb{1}_{\{x_j \neq x^*\}} \\ &\quad + \sum_{j=1}^k \limsup_{t \rightarrow 0} \frac{d(x_j, \alpha(t))}{t} \mathbb{1}_{\{x_j = x^*\}}. \end{aligned} \quad (6)$$

For the first term, the first variation formula in Alexandrov spaces (see Kunzinger and Steinbauer (2018)[Proposition 3.4.2]) with $l(t) = d(x_j, \alpha(t))$ gives

$$\limsup_{t \rightarrow 0} \frac{d(x_j, \alpha(t)) - d(x_j, x^*)}{t} \leq -\cos \gamma_j$$

where $\gamma_j = \angle x_j x^* z$ in Alexandrov sense. Since \mathcal{X} is a NPC space, comparing between $\triangle x_j x^* z$ and its Euclidean comparison triangle $\triangle \tilde{x}_j \tilde{x}^* \tilde{z}$ (see Definitions A.8 and A.12) results in

$$\gamma_j \leq \tilde{\gamma}_j := \angle \tilde{x}_j \tilde{x}^* \tilde{z}, \quad d(x_j, \alpha(t)) \leq \|\tilde{x}_j - \tilde{\alpha}(t)\|. \quad (7)$$

where $\tilde{\alpha} = [\tilde{x}^* \tilde{z}]$. Now, since $\tilde{\gamma}_j$ is the angle inside the triangle, $\tilde{\gamma}_j < \pi$ holds. This implies $\cos(\gamma_j) \geq \cos(\tilde{\gamma}_j)$. Plugging-in Equation (7) to Equation (6) yields

$$\limsup_{t \rightarrow 0} \frac{F(\alpha(t)) - F(\alpha(0))}{t} \leq - \sum_{j=1}^k \cos(\tilde{\gamma}_j) \mathbf{1}_{\{\tilde{x}_j \neq \tilde{x}^*\}} + \sum_{j=1}^k \mathbf{1}_{\{\tilde{x}_j = \tilde{x}^*\}} < -(1-\alpha)k \sqrt{1 - \frac{1}{C_\alpha^2}} + \alpha k \leq 0$$

where the second inequality follows in the same manner as Minsker (2015)[Lemma 2.1]; see Figure 4 as well. Since it contradicts to Equation (5), we prove the claim.

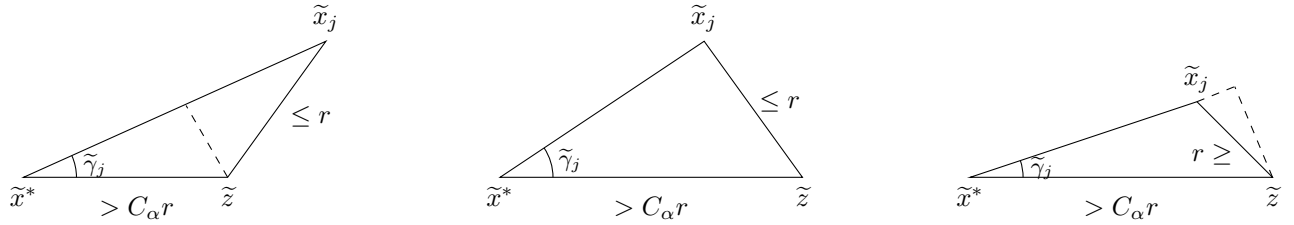


Figure 4: Possible configurations of the triangle $\triangle \tilde{x}^* \tilde{x}_j \tilde{z}$ for $j = 1, \dots, \lfloor (1-\alpha)k \rfloor + 1$. Clearly, the second case with the equality gives the tightest upper bound on $\sin(\tilde{\gamma}_j)$, which is $1/C_\alpha$. This automatically yields the lower bound of the $\cos(\tilde{\gamma}_j)$.

CASE II: $\kappa > 0$.

Since we assumed x^* exists in this case, Equation (5) holds. By the assumption, for any x_j , the triangle inequality implies

$$d(x^*, x_j) + d(x^*, z) + d(x_j, z) \leq 2(d(x^*, x_j) + d(x^*, z)) \leq 2D_\kappa.$$

Hence, $\{x^*, x_j, z\}$ can be embedded to $M_\kappa^2 := \frac{1}{\sqrt{\kappa}} \mathbb{S}^2$. In other words, one can pick points \tilde{x}_j 's, \tilde{x}^* and \tilde{z} on M_κ^2 such that any $\{\tilde{x}^*, \tilde{x}_j, \tilde{z}\}$ forms the comparison triangle of $\{x^*, x_j, z\}$: see Kunzinger and Steinbauer (2018)[Proposition 4.2.20].

Applying the same method as in $\kappa \leq 0$ case, one can check

$$\limsup_{t \rightarrow 0} \frac{F(\alpha(t)) - F(\alpha(0))}{t} \leq - \sum_{j=1}^k \cos(\tilde{\gamma}_j) \mathbf{1}_{\{\tilde{x}_j \neq \tilde{x}^*\}} + \sum_{j=1}^k \mathbf{1}_{\{\tilde{x}_j = \tilde{x}^*\}}$$

where $\tilde{\gamma}_j$ is the angle at \tilde{x}^* of $\triangle \tilde{x}^* \tilde{x}_j \tilde{z}$ on M_κ^2 . Now, we apply Rauch comparison theorem on M_κ^2 (Kunzinger et al., 2014)[Theorem 3.1]. Denoting the intrinsic metric of M_κ^2 as \tilde{d} , we obtain

$$\frac{\sin(\sqrt{\kappa}R)}{\sqrt{\kappa}R} \|u\| \leq \|d_v \exp_{\tilde{x}^*}(u)\|$$

for any $\|u\| \leq R < D_\kappa$. In fact, since M_κ^2 is $\mathbb{S}^2/\sqrt{\kappa}$, we can plug-in $R = D_\kappa/2$, the half of injectivity radius of $\mathbb{S}^2/\sqrt{\kappa}$. Under this choice of R , $B(\tilde{x}^*, D_\kappa/2) \subset M_\kappa^2$ becomes convex. Therefore, the above estimate becomes global in such ball, so that

$$\frac{2}{\pi} \|u - v\|_{\tilde{x}^*} \leq \tilde{d}(\exp_{\tilde{x}^*}(u), \exp_{\tilde{x}^*}(v)).$$

See Fefferman et al. (2019)[Equation (4.1), (4.2)] for more detail of this procedure. This results in the logarithmic map $\log_{\tilde{x}^*} \cdot$ being $\pi/2$ -Lipschitz in the ball $B(\tilde{x}^*, D_\kappa/2)$. Since $\tilde{x}_j, \tilde{z} \in B(\tilde{x}^*, D_\kappa/2)$, we have

$$\|\log_{\tilde{x}^*} \tilde{z} - \log_{\tilde{x}^*} \tilde{x}_j\|_{\tilde{x}^*} \leq \frac{\pi}{2} \tilde{d}(\tilde{z}, \tilde{x}_j) \leq \frac{\pi}{2} r$$

for $j = 1, \dots, \lfloor (1 - \alpha)k \rfloor + 1$. Therefore, by considering the same Figure 4 for the triangle constructed by $0, \log_{\tilde{x}^*} \tilde{x}_j$, and $\log_{\tilde{x}^*} \tilde{z}$ in the tangent space $T_{\tilde{x}^*} M_\kappa^2$, $\sin(\tilde{\gamma}_j)$ is upper bounded by $1/C_\alpha$ for $j = 1, \dots, \lfloor (1 - \alpha)k \rfloor + 1$ again. Therefore, the same procedure as in Case I yields

$$0 \leq \limsup_{t \rightarrow 0} \frac{F(\alpha(t)) - F(\alpha(0))}{t} \leq - \sum_{j=1}^k \cos(\tilde{\gamma}_j) \mathbb{1}_{\{\tilde{x}_j \neq \tilde{x}^*\}} + \sum_{j=1}^k \mathbb{1}_{\{\tilde{x}_j = \tilde{x}^*\}} < -(1 - \alpha)k \sqrt{1 - \frac{1}{C_\alpha^2}} + \alpha k \leq 0$$

whenever $C_\alpha \geq (1 - \alpha) \sqrt{\frac{1}{1 - 2\alpha}}$. This contradicts to Equation (5). \square

Proof of Theorem 3.3:

Proof. For $\kappa \leq 0$, let $\mathcal{E} := \{d(\hat{\theta}_{MoE}, \theta) > C_\alpha \epsilon\}$ the event. Then, under this event, by Lemma 3.1 with $x^* = \hat{\theta}_{MoE}$, $x_j = \hat{\theta}_j$, and $z = \theta$, we have $J \subseteq \{1, \dots, k\}$ such that $|J| > \alpha k$ and $d(\hat{\theta}_j, \theta) > \epsilon$ for $j \in J$. Let $W \sim B(k, p)$ be Binomial random variable. Then,

$$\mathbb{P}(\mathcal{E}) \leq \mathbb{P}\left(\sum_{j=1}^k \mathbb{1}_{\{d(\hat{\theta}_j, \theta) > \epsilon\}} \geq \alpha k\right) \leq \mathbb{P}(W \geq \alpha k) \leq \exp\left(-k(1 - \alpha) \log \frac{1 - \alpha}{1 - p} - k\alpha \log \frac{\alpha}{p}\right)$$

where the second and the last inequalities follow from Lerasle and Oliveira (2011)[Lemma 23], and Chernoff bound respectively.

The case for $\kappa > 0$ goes almost similarly, once one sets the event as $\mathcal{E} := \{\pi C_\alpha \epsilon / 2 < d(\hat{\theta}_{FMoE}, \theta) \leq D_\kappa / 2\}$. The upper bound $d(\hat{\theta}_{FMoE}, \theta) \leq D_\kappa / 2$ vanishes by the assumption. \square

Lastly, for the conditional probability in Remark 3.4, one can check

$$d(\hat{\theta}_{FMoE}, \theta) \leq \min_j \left[d(\hat{\theta}_{FMoE}, \hat{\theta}_j) + d(\hat{\theta}_j, \theta) \right] \leq \frac{D_\kappa}{2} - \epsilon + \min_j d(\hat{\theta}_j, \theta)$$

almost surely. Now, from the assumption,

$$\mathbb{P}\left[\min_j d(\hat{\theta}_j, \theta) \leq \epsilon\right] \geq 1 - p^k.$$

Therefore,

$$\mathbb{P}\left[d(\hat{\theta}_{FMoE}, \theta) \leq \frac{D_\kappa}{2}\right] \geq 1 - p^k.$$

Then, the definition of the conditional probability leads to the claimed bound.

B.2 Proofs in Section 4

Proof of Theorem 4.3:

Proof. From Proposition 4.1, we obtain

$$\mathbb{E}[d^2(\hat{x}, x^*)] \leq \frac{\sigma^2}{\lfloor n/k \rfloor} \leq \frac{2k\sigma^2}{n}.$$

Fixing $\alpha \in (0, 0.5)$ and $p \in (0, \alpha)$, and then define $k := \lfloor \log(1/\delta)/\psi(\alpha, p) \rfloor + 1$ and $\epsilon := \sqrt{(2k\sigma^2)/(np)}$. Applying Markov inequality, one obtains

$$\mathbb{P}[d(\hat{x}, x^*) \geq \epsilon] \leq \frac{\mathbb{E}[d^2(\hat{x}, x^*)]}{\epsilon^2} = \frac{2k\sigma^2}{n\epsilon^2} = p.$$

Then, it follows from Theorem 3.3 that

$$\mathbb{P}[d(\hat{x}_{FMoM}, x^*) \geq C_\alpha \epsilon] \leq \exp(-k\psi(\alpha, p)) \leq \delta$$

by the construction of k . Observe that

$$C_\alpha \epsilon = C_\alpha \sqrt{\frac{2k\sigma^2}{np}} = C_\alpha \sqrt{\frac{2\sigma^2}{np\psi(\alpha, p)}} \sqrt{k\psi(\alpha, p)} \leq C_\alpha \sqrt{\frac{2\sigma^2}{np\psi(\alpha, p)}} \sqrt{\psi(\alpha, p) + \log(1/\delta)}.$$

Since α and p are arbitrary, any choice of α and p will induce the bound. In fact, by minimizing the right-most above term with respect to α and p , we obtain the tighter bound. In this case, we plugged-in $p = 1/10$ and $\alpha = 7/18$, as suggested in Minsker (2015)[Corollary 4.1]. \square

Proof of Proposition 4.5:

Proof. Case I: $d = d_{AI}$.

We write $\Delta := \Sigma^{-1/2} \widehat{\Sigma} \Sigma^{-1/2} - I_d = \Sigma^{-1/2} (\widehat{\Sigma} - \Sigma) \Sigma^{-1/2}$, and let $\tilde{\lambda}, \tilde{\sigma}$ be the eigenvalues and singular values of Δ . Note that the eigenvalues of $\Sigma^{-1/2} \widehat{\Sigma} \Sigma^{-1/2}$ are $\tilde{\lambda} + 1$.

Recall the inequalities between the spectral norm and the Frobenius norm:

$$\|\cdot\|_2 \leq \|\cdot\|_F \leq \sqrt{d} \|\cdot\|_2. \quad (8)$$

Here, $\|\cdot\|_F, \|\cdot\|_2$ stand for Frobenius norm and spectral norm respectively. Equation (8) implies

$$\mathbb{P} \left[d_{AI} \left(\Sigma, \widehat{\Sigma} \right) \geq \epsilon \right] = \mathbb{P} \left[\left\| \log \Sigma^{-1/2} \widehat{\Sigma} \Sigma^{-1/2} \right\|_F \geq \epsilon \right] \leq \mathbb{P} \left[\sqrt{d} \left\| \log \Sigma^{-1/2} \widehat{\Sigma} \Sigma^{-1/2} \right\|_2 \geq \epsilon \right].$$

Note that if λ is the eigenvalue of a positive semi-definite matrix A , then $\log \lambda$ is that of $\log A$. Hence,

$$\begin{aligned} \mathbb{P} \left[\sqrt{d} \left\| \log \Sigma^{-1/2} \widehat{\Sigma} \Sigma^{-1/2} \right\|_2 \geq \epsilon \right] &= \mathbb{P} \left[\max_{j \leq d} \left| \log(\tilde{\lambda}_j + 1) \right| \geq \frac{\epsilon}{\sqrt{d}} \right] \\ &= \mathbb{P} \left[\left\{ \max_{j \leq d} \log(\tilde{\lambda}_j + 1) \geq \frac{\epsilon}{\sqrt{d}} \right\} \cup \left\{ \min_{j \leq d} \log(\tilde{\lambda}_j + 1) \leq -\frac{\epsilon}{\sqrt{d}} \right\} \right] \\ &\leq \mathbb{P} \left[\tilde{\lambda}_{\max} \geq e^{\frac{\epsilon}{\sqrt{d}}} - 1 \right] + \mathbb{P} \left[\tilde{\lambda}_{\min} \leq e^{-\frac{\epsilon}{\sqrt{d}}} - 1 \right] \\ &\leq \mathbb{P} \left[\tilde{\sigma}_{\max} \geq 1 - e^{-\frac{\epsilon}{\sqrt{d}}} \right] \end{aligned}$$

where the last inequality follows from the fact that $\cosh(x) \geq 0$ and singular value bounds the absolute value of eigenvalues. Observe the submultiplicativity of singular values

$$\sigma_{\max}(ABC) \leq \sigma_{\max}(A) \sigma_{\max}(B) \sigma_{\max}(C). \quad (9)$$

Taking $A, C = \Sigma^{-1/2}$ and $B = \widehat{\Sigma} - \Sigma$, Equation (9) leads to

$$\begin{aligned} \mathbb{P} \left[\tilde{\sigma}_{\max} \geq 1 - e^{-\frac{\epsilon}{\sqrt{d}}} \right] &\leq \mathbb{P} \left[\lambda_{\max}(\Sigma^{-1/2})^2 \lambda_{\max} \left(\left| \widehat{\Sigma} - \Sigma \right| \right) \geq 1 - e^{-\frac{\epsilon}{\sqrt{d}}} \right] \\ &= \mathbb{P} \left[\lambda_{\max} \left(\left| \widehat{\Sigma} - \Sigma \right| \right) \geq \lambda_{\min} \left(1 - e^{-\frac{\epsilon}{\sqrt{d}}} \right) \right] \\ &= \mathbb{P} \left[\left\| \widehat{\Sigma} - \Sigma \right\|_2 \geq \lambda_{\min} \left(1 - e^{-\frac{\epsilon}{\sqrt{d}}} \right) \right] \\ &\leq \mathbb{P} \left[\left\| \widehat{\Sigma} - \Sigma \right\|_F \geq \lambda_{\min} \left(1 - e^{-\frac{\epsilon}{\sqrt{d}}} \right) \right] \end{aligned}$$

where, again, the last inequality follows from Equation (8). Combining the above argument,

$$\mathbb{P} \left[d_{AI} \left(\Sigma, \widehat{\Sigma} \right) \geq \epsilon \right] \leq \mathbb{P} \left[\left\| \widehat{\Sigma} - \Sigma \right\|_F \geq \lambda_{\min} \left(1 - e^{-\frac{\epsilon}{\sqrt{d}}} \right) \right]. \quad (10)$$

Now, writing $X_i = (X_i^1, \dots, X_i^d) \in \mathbb{R}^d$, it follows that

$$\begin{aligned}
 \mathbb{P} \left[\left\| \widehat{\Sigma} - \Sigma \right\|_F \geq \lambda_{\min} \left(1 - e^{-\frac{\epsilon}{\sqrt{d}}} \right) \right] &= \mathbb{P} \left[\sum_{k,l=1}^d \left| \frac{1}{n} \sum_{i=1}^n X_i^k X_i^l - \mathbb{E}(X^k X^l) \right|^2 \geq \lambda_{\min}^2 \left(1 - e^{-\frac{\epsilon}{\sqrt{d}}} \right)^2 \right] \\
 &\leq \sum_{k,l=1}^d \mathbb{P} \left[\left| \frac{1}{n} \sum_{i=1}^n X_i^k X_i^l - \mathbb{E}(X^k X^l) \right|^2 \geq \frac{\lambda_{\min}^2}{d^2} \left(1 - e^{-\frac{\epsilon}{\sqrt{d}}} \right)^2 \right] \\
 &\leq \sum_{k,l=1}^d \frac{d^2 \mathbb{E} \left[\left| \frac{1}{n} \sum_{i=1}^n X_i^k X_i^l - \mathbb{E}(X^k X^l) \right|^2 \right]}{\lambda_{\min}^2 \left(1 - \exp \left(-\frac{\epsilon}{\sqrt{d}} \right) \right)^2} \\
 &= \sum_{k,l=1}^d \frac{d^2 \sum_{i,j=1}^n \mathbb{E} (X_i^k X_i^l X_j^k X_j^l) - [\mathbb{E}(X^k X^l)]^2}{n^2 \lambda_{\min}^2 \left(1 - \exp \left(-\frac{\epsilon}{\sqrt{d}} \right) \right)^2}
 \end{aligned}$$

where the second last inequality above follows by Markov inequality. Since X_i and X_j are independent,

$$\sum_{k,l=1}^d \frac{d^2 \sum_{i,j=1}^n \mathbb{E} (X_i^k X_i^l X_j^k X_j^l) - [\mathbb{E}(X^k X^l)]^2}{n^2 \lambda_{\min}^2 \left(1 - \exp \left(-\frac{\epsilon}{\sqrt{d}} \right) \right)^2} = \sum_{k,l=1}^d \frac{d^2 \sum_{i=1}^n \mathbb{E} [(X_i^k X_i^l)^2] - [\mathbb{E}(X^k X^l)]^2}{n^2 \lambda_{\min}^2 \left(1 - \exp \left(-\frac{\epsilon}{\sqrt{d}} \right) \right)^2}.$$

Therefore,

$$\begin{aligned}
 \mathbb{P} \left[\left\| \widehat{\Sigma} - \Sigma \right\|_F \geq \lambda_{\min}(\Sigma) \left(1 - e^{-\frac{\epsilon}{\sqrt{d}}} \right) \right] &\leq \sum_{k,l=1}^d \frac{d^2 \sum_{i=1}^n \mathbb{E} [(X_i^k X_i^l)^2] - [\mathbb{E}(X^k X^l)]^2}{n^2 \lambda_{\min}^2 \left(1 - \exp \left(-\frac{\epsilon}{\sqrt{d}} \right) \right)^2} \\
 &= \frac{d^2}{n \lambda_{\min}^2 \left(1 - \exp \left(-\frac{\epsilon}{\sqrt{d}} \right) \right)^2} \sum_{k,l=1}^d \text{Var}(X^k X^l).
 \end{aligned}$$

Since we assumed $P \in \mathcal{P}_4(\mathbb{R}^d)$, we have $\sum_{k,l=1}^d \text{Var}(X^k X^l) \leq C d^2$ for some constant $C > 0$. The conclusion follows.

Case II: $d = d_{BW}$.

Using the fact that eigenvalues of $\widehat{\Sigma}, \Sigma$ are lower bounded by λ_0 , the following estimate for Bures-Wasserstein distance holds:

$$d_{BW}^2(\widehat{\Sigma}, \Sigma) \stackrel{(i)}{=} \min_{U: \text{Unitary}} \left\| \widehat{\Sigma}^{1/2} - \Sigma^{1/2} U \right\|_F^2 \leq \left\| \widehat{\Sigma}^{1/2} - \Sigma^{1/2} \right\|_F^2 \stackrel{(ii)}{\leq} \frac{1}{4\lambda_0} \left\| \widehat{\Sigma} - \Sigma \right\|_F^2$$

where (i) and (ii) are from Bhatia et al. (2019)[Theorem 1] and Ando and van Hemmen (1980)[Proposition 3.2] respectively.

Hence, we get

$$\mathbb{P} \left[d_{BW}(\widehat{\Sigma}, \Sigma) \geq \epsilon \right] \leq \mathbb{P} \left[\left\| \widehat{\Sigma} - \Sigma \right\|_F \geq 2\sqrt{\lambda_0} \epsilon \right] \leq \frac{C d^4}{4n \lambda_0 \epsilon^2}$$

where the last inequality can be obtained similarly to the last step in Case I. \square

Proof of Theorem 4.7:

Proof. We follow the same technique as in Theorem 4.3. We fix α and p and then set $k = \lfloor \log(1/\delta)/\psi(\alpha, p) \rfloor + 1$.

Case I: $d = d_{AI}$.

From Proposition 4.5:

$$\mathbb{P} \left[d_{AI}(\widehat{\Sigma}, \Sigma) \geq \epsilon \right] \leq \frac{2k C d^4}{n \lambda_{\min}^2 \left(1 - \exp \left(-\frac{\epsilon}{\sqrt{d}} \right) \right)^2} := p$$

by choosing

$$\epsilon = -\sqrt{d} \log \left(1 - \frac{\sqrt{2kCd^2}}{\lambda_{\min} \sqrt{np}} \right)$$

which is always possible from the condition $n > 2kCd^4/\lambda_{\min}^2$. Then, we apply Theorem 3.3 to yield

$$\mathbb{P} \left[d_{AI} \left(\widehat{\Sigma}_{FME}, \Sigma \right) \geq C_\alpha \epsilon \right] \leq \exp(-k\psi(\alpha, p)) \leq \delta$$

by the setting of k . Observe

$$C_\alpha \epsilon = -\sqrt{d} C_\alpha \log \left(1 - \frac{\sqrt{2kCd^2}}{\lambda_{\min} \sqrt{np}} \right) \leq -\sqrt{d} C_\alpha \log \left(1 - \frac{\sqrt{2Cd^2}}{\lambda_{\min} \sqrt{np\psi(\alpha, p)}} \sqrt{\psi(\alpha, p) + \log(1/\delta)} \right).$$

Again, this bound holds for all $\alpha \in (0, 0.5)$ and $p \in (0, \alpha)$. Plugging-in $\alpha = 0.4$ and $p = 0.1$ yields the claimed result.

Case II: $d = d_{BW}$.

For d_{BW} , one can employ the similar technique as in the above with $\epsilon = d^2 \sqrt{(kC)/(2n\lambda_0 p)}$. We used the same $\alpha = 0.4$ and $p = 0.1$. The condition $n > 6\lambda_0 kCd^4$ was imposed to guarantee $\epsilon < D_\kappa/(\pi C_\alpha)$ for such choice of α and p . \square

The precise statement of Remark 4.6:

Notice that

$$|\lambda_{\min}(\widehat{\Sigma}) - \lambda_{\min}(\Sigma)| \leq \|\widehat{\Sigma} - \Sigma\|_2$$

from the Weyl's inequality. Then, again applying Equation (8) and the calculation on the Frobenius norm bound in the proof of Proposition 4.5 lead to

$$\mathbb{P} \left[|\lambda_{\min}(\widehat{\Sigma}) - \lambda_{\min}(\Sigma)| \geq \epsilon \right] \leq \mathbb{P} \left[\|\widehat{\Sigma} - \Sigma\|_2 \geq \epsilon \right] \leq \mathbb{P} \left[\|\widehat{\Sigma} - \Sigma\|_F \geq \epsilon \right] \leq \frac{Cd^4}{n\epsilon^2}.$$

Here, C is the same C in Proposition 4.5. Writing $\delta := Cd^4/(n\epsilon^2)$, one gets

$$\mathbb{P} \left[\lambda_{\min}(\widehat{\Sigma}) \in B \left(\lambda_{\min}(\Sigma), d^2 \sqrt{\frac{C}{n\delta}} \right) \right] \geq 1 - \delta.$$

This implies $\lambda_{\min}(\widehat{\Sigma})$ concentrates in $B(\lambda_{\min}(\Sigma), C/\sqrt{n})$ for some universal $C > 0$ with high probability whenever P allows the finite fourth moment.

C Implementation detail and additional experiments

This section includes implementation detail and more experiments of our algorithm under different settings. First, for all experiments in the main paper, we show how the performance of our method varies by the block size and the population distribution. In addition, for Fréchet mean estimation problem, we conduct the additional experiment on a Poincaré disk model to verify our method works in various domains. All new experiments in this Appendix were conducted 100 times, while we maintained with the results in Section 5 for the same one. We found this number of simulation to be sufficient to obtain the coherent results.

C.1 Implementation detail

For implementations, we used Python package **Geomstats** (Miolane et al., 2020, 2024) to model particular $\text{CAT}(\kappa)$ spaces and compute the geometric quantities like metrics and geodesics. For the Fréchet mean estimation problem in a NPC space, Fréchet mean and median were implemented using inductive mean and Bačák (2014)[Algorithm 4.3]. For the covariance estimation problem, Fréchet mean and median were computed using predefined functions in **Geomstats** (which are based on subgradient methods). All experiments were performed on a free version of Google Colab without any use of GPU. For each task, running 100 simulations did not take more than 5 minutes.

C.2 Effects of the block sizes and population distributions

We first analyze the effect of the block size and the population distributions. For 5-legs spider tree, we used same $n = 100$, and varied the block size from $k = 1$ to 100. Note that $k = 1$ case coincides to the original estimator within n samples. For the population distribution, we maintained to use $P = \text{Unif}(1, \dots, 5) \times ((1 - \alpha) |N(1, 1)| + \alpha |N(100, 1)|)$, while varying α to observe the effect of the tail of the population distribution. The results are summarized in Table 2 and 3.

Table 2: $\mathbb{E}d^2(\hat{x}_{FMoE}, x^*)$ from 100 simulations in 5-legs spider with α -portion outliers.

α	$k = 100$	$k = 50$	$k = 10$	$k = 5$	$k = 1$
0	9.4×10^{-9}	3.4×10^{-8}	1.1×10^{-5}	0.0002	0.0003
0.1	8.6×10^{-9}	3.1×10^{-8}	1.2×10^{-5}	0.0105	3.1244
0.5	1.1×10^{-8}	4.3×10^{-8}	0.0223	0.0313	3.4634
0.9	0.0118	0.3560	0.0127	0.0193	3.8080

Table 3: 95% confidence interval of $d(\hat{x}_{FMoE}, x^*)$ from 100 simulations in 5-legs spider with α -portion outliers.

α	$k = 100$	$k = 50$	$k = 10$	$k = 5$	$k = 1$
0	$[2.6 \times 10^{-5}, 0.0001]$	$[1.5 \times 10^{-5}, 0.0002]$	$[6.9 \times 10^{-5}, 0.0127]$	$[7.1 \times 10^{-5}, 0.0485]$	$[0.0005, 0.0415]$
0.1	$[8.2 \times 10^{-6}, 0.0001]$	$[1.1 \times 10^{-5}, 0.0002]$	$[9.6 \times 10^{-5}, 0.0086]$	$[7.3 \times 10^{-7}, 0.0711]$	$[0.0110, 4.4202]$
0.5	$[4.5 \times 10^{-5}, 0.0001]$	$[0.0001, 0.0002]$	$[0.0004, 0.2928]$	$[5.5 \times 10^{-8}, 0.6888]$	$[0.11584, 5.415]$
0.9	$[8.1 \times 10^{-5}, 0.0004]$	$[3.4 \times 10^{-5}, 2.5411]$	$[0.0009, 0.1423]$	$[7.3 \times 10^{-8}, 0.6190]$	$[0.0566, 4.9039]$

For covariance estimation problem, we again used the same $d = 10$ and $n = 10d^4$, while varying $k = 1$ to 100. We used the same procedure to generate the Σ , and then experimented on $t_\nu(0, \Sigma)$ for different ν to observe the effect of the tail of the distribution. As $\nu \rightarrow 2$, distributions will impose a heavier tail. The results for covariance estimation problem are summarized in Table 4, 5, 6, and 7.

In all experiments, when the tail is relatively light ($\alpha = 0, \nu = 5$), the original estimators ($k = 1$) are comparable to our proposed estimators. When the tail is very heavy ($\alpha = 0.9, \nu = 2.2$), interesting behaviors emerge. In these cases, the proper choices of the block size ($k = 5, 10$) yield higher accuracy as well as stronger concentration, which are in agreement with results in Section 5. However, one can observe that performances worsen if the block sizes are set too large ($k = 50$ for spider, and $k = 100$ for the covariance estimation). This observation aligns with the discussion in the beginning of Section 5; if the block size is too large, the original estimator may not perform well within the subset size of $\lfloor n/k \rfloor$, and our method may not work properly in such cases.

Table 4: $\mathbb{E}d_{AI}^2(\hat{\Sigma}_{FMoE}, \Sigma)$ from 100 simulations for covariance estimation on $t_\nu(0, \Sigma)$.

ν	$k = 100$	$k = 10$	$k = 5$	$k = 1$
2.2	6.5528	3.4604	2.9163	3.3515
2.5	0.9739	0.3454	0.2931	0.6057
3	0.1327	0.0377	0.0387	0.1141
5	0.0055	0.0032	0.0032	0.0034

Table 5: 95% confidence interval of $d_{AI}(\hat{\Sigma}_{FMoE}, \Sigma)$ from 100 simulations for covariance estimation on $t_\nu(0, \Sigma)$.

ν	$k = 100$	$k = 10$	$k = 5$	$k = 1$
2.2	$[1.2948, 2.3295]$	$[1.6055, 2.0854]$	$[1.3301, 1.9865]$	$[1.2816, 3.3372]$
2.5	$[0.9035, 1.0604]$	$[0.4239, 1.7378]$	$[0.4132, 0.6792]$	$[0.4266, 1.6865]$
3	$[0.3027, 0.4070]$	$[0.1422, 0.244011]$	$[0.1485, 0.2562]$	$[0.1488, 0.8103]$
5	$[0.0527, 0.0934]$	$[0.0445, 0.0666]$	$[0.0447, 0.0686]$	$[0.0444, 0.0756]$

Table 6: $\mathbb{E}d_{BW}^2(\widehat{\Sigma}_{FMoE}, \Sigma)$ from 100 simulations for covariance estimation on $t_\nu(0, \Sigma)$.

ν	$k = 100$	$k = 10$	$k = 5$	$k = 1$
2.2	61.0520	36.6296	31.5038	76.8935
2.5	4.6620	2.0343	1.7360	8.3281
3	0.3703	0.1407	0.1484	0.4993
5	0.0082	0.0065	0.0067	0.0074

 Table 7: 95% confidence interval of $d_{BW}(\widehat{\Sigma}_{FMoE}, \Sigma)$ from 100 simulations for covariance estimation on $t_\nu(0, \Sigma)$.

ν	$k = 100$	$k = 10$	$k = 5$	$k = 1$
2.2	[6.5140, 8.8935]	[4.6619, 7.2466]	[4.2740, 6.8831]	[4.0157, 25.6891]
2.5	[1.8019, 2.5497]	[1.0648, 1.8412]	[0.9443, 1.7409]	[0.9936, 5.8796]
3	[0.4492, 0.7285]	[0.2632, 0.4909]	[0.2631, 0.5110]	[0.2777, 1.6146]
5	[0.0680, 0.1222]	[0.0600, 0.1078]	[0.0584, 0.1053]	[0.0578, 0.1126]

C.3 Fréchet mean estimation in Poincaré disk model

Lastly, to verify our method works in various domains, we conduct the Fréchet mean estimation in Poincaré disk model, a widely used space for hierarchical model due to the pioneer work of Nickel and Kiela (2017). Poincaré disk is a 2-dimensional Riemannian manifold with nonpositive sectional curvature (therefore a $\text{CAT}(\kappa)$ space) which can be embedded in the unit ball of \mathbb{R}^2 . A Riemannian metric tensor of Poincaré disk is defined by the following formula:

$$ds^2 = \frac{dx^2 + dy^2}{(1 - x^2 - y^2)^2}.$$

A Poincaré disk can be embedded into an open Euclidean unit ball in \mathbb{R}^2 . Specifically, we can construct a Poincaré disk using a method similar to the stereographic projection of a sphere. Consider the upper hyperboloid described by the equation $t^2 = x^2 + y^2 + 1$ for $t > 1$. We can project this hyperboloid from the point $(t = -1, x = 0, y = 0)$ onto a unit disk at $t = 0$. This projection maps points on the hyperboloid to the unit disk, creating the Poincaré disk model. A distance between two points is given by the Euclidean length of the hyperbolic arc between corresponding points. Figure 5 illustrates the Poincaré disk. Intuitively, points closer to the boundary will have larger distances.

There are extensive theories regarding hyperbolic geometry and the Poincaré disk model. However, since these topics are not our primary interest, we refer interested readers to Anderson (2005)[Chapter 4.1]. Instead, we focus on numerically validating our method for estimating the Fréchet mean in the setting of heavy-tailed distributions in the Poincaré disk.

For the population distribution, we used the following mixture distribution in Poincaré disk:

$$P = (1 - \alpha)N(0, 0.2^2) \Big|_{B(0,1)} + \alpha \text{Unif}((1 - 10^{-7})\mathbb{S}^1).$$

Here, $N(0, 0.2^2) \Big|_{B(0,1)}$ means the distribution of projected Gaussian random variable to the unit ball of Euclidean space. This distribution has Fréchet mean at the origin (due to its symmetricity), but it has a very heavy tail in Poincaré disk due to the effect of the outlier quantities, which are from the uniform distribution around the boundary. α again denotes the portion of outliers. We used the sample size $n = 100$, and experimented by changing the block size k and the portion of outliers α . The results are summarized in Table 8, 9, and Figure 6, 7.

The results are consistent with previous experiments. When the tail is light ($\alpha = 0$), the inductive mean estimator itself achieves high accuracy and a small confidence region around 0. However, as the tail becomes heavy, our proposed estimator with the optimal block size (in this example, $k = 50$ for the most cases) performs significantly better.

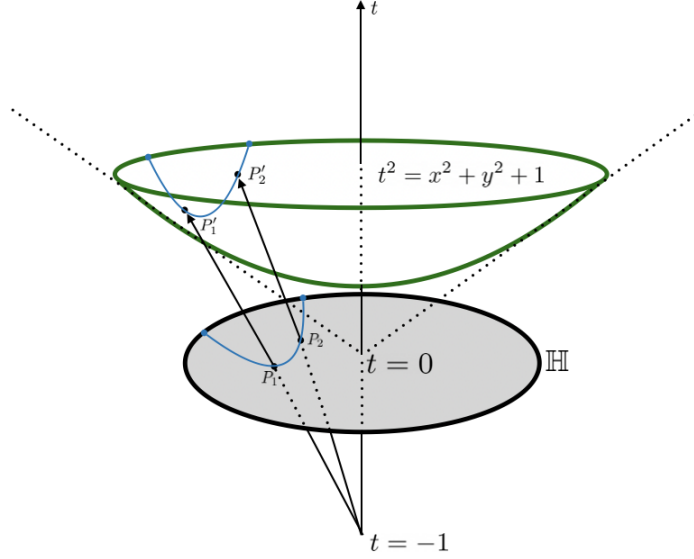


Figure 5: The illustration of the Poincaré disk model. Poincaré disk model is defined by the intersection of the unit disk, e.g., P_1, P_2 , and the projection map from the point $(t = -1, x = 0, y = 0)$ to the upper hyperboloid, e.g., P'_1, P'_2 . The distance between P_1 and P_2 is determined by the Euclidean length of the hyperbolic arc connecting their corresponding points P'_1 and P'_2 .

Table 8: $\mathbb{E}d^2(\hat{x}_{FMoE}, x^*)$ from 100 simulations in Poincaré disk with α -portion outliers.

α	$k = 50$	$k = 10$	$k = 5$	$k = 1$
0	0.0035	0.0047	0.0038	0.0028
0.1	0.0058	0.0323	0.1447	0.1063
0.5	0.0195	0.1553	0.1937	0.1709
0.9	0.0667	0.1776	0.2320	0.1747

Table 9: 95% confidence interval of $d(\hat{x}_{FMoE}, x^*)$ from 100 simulations in Poincaré disk with α -portion outliers.

α	$k = 50$	$k = 10$	$k = 5$	$k = 1$
0	[0.0157, 0.1032]	[0.0081, 0.1358]	[0.0111, 0.1198]	[0.0107, 0.0935]
0.1	[0.0197, 0.1397]	[0.0401, 0.4282]	[0.0504, 0.7766]	[0.0610, 0.6266]
0.5	[0.0216, 0.2845]	[0.0809, 0.7567]	[0.0703, 0.8623]	[0.0852, 0.7389]
0.9	[0.0261, 0.5159]	[0.0565, 0.7992]	[0.1253, 0.8588]	[0.0772, 0.8063]

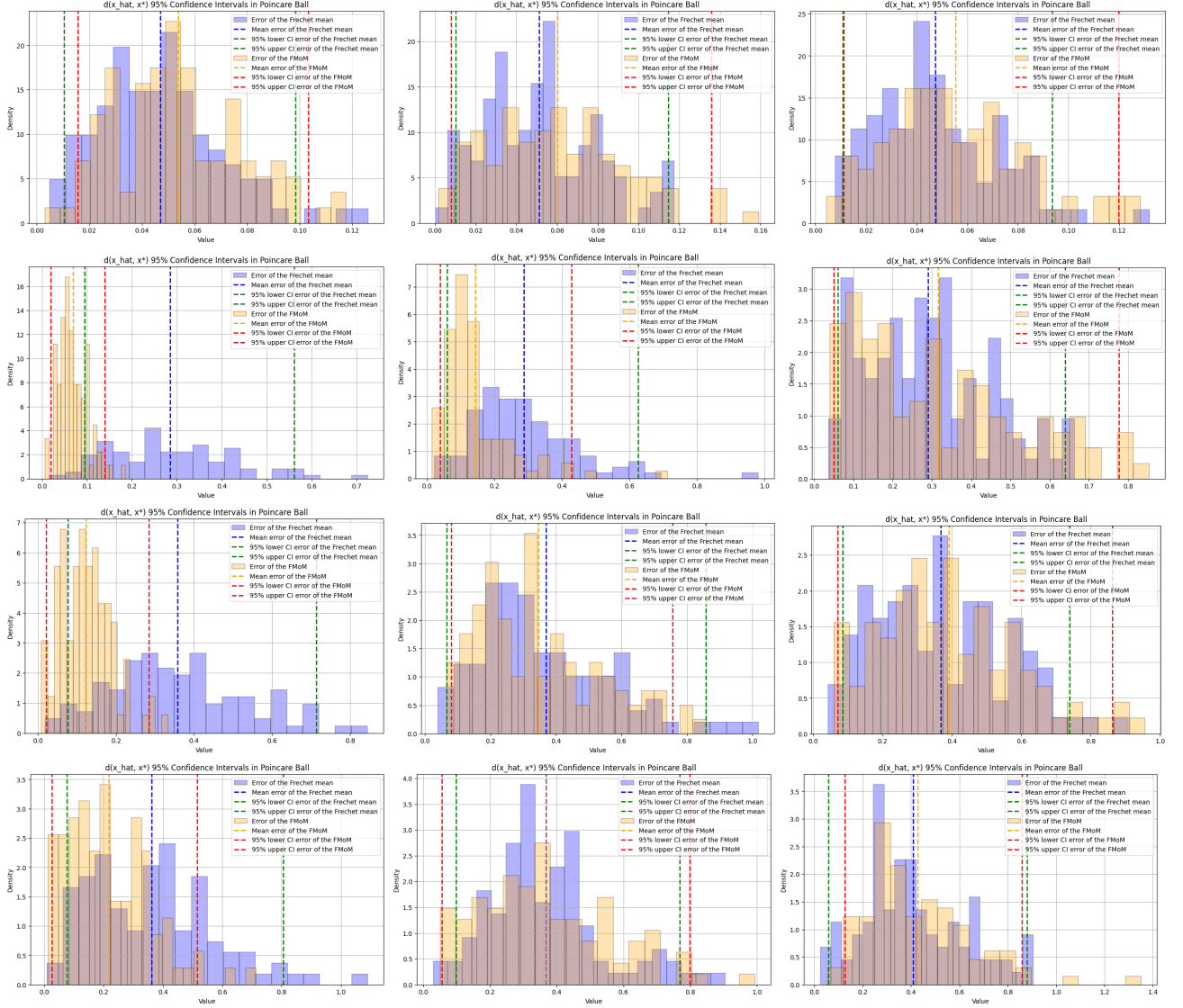


Figure 6: Histogram, mean, and 95% confidence interval for each experiment from 100 simulations. **Rows:** $\alpha = 0, 0.1, 0.5, 0.9$ from top to the bottom. **Columns:** $k = 50, 10, 5$ from left to right. For each experiment, comparison between the inductive mean estimator ($k = 1$) is displayed.

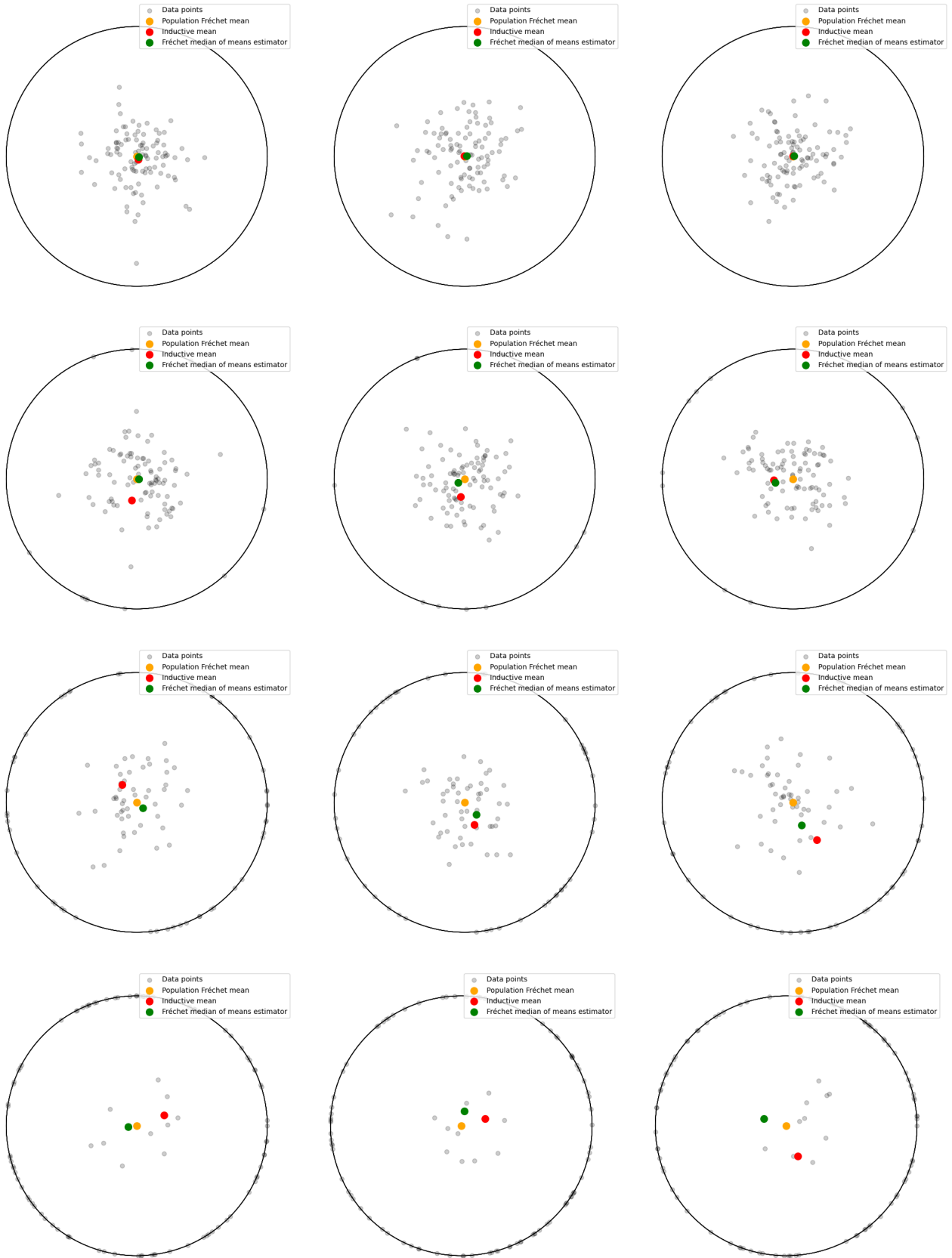


Figure 7: One randomly chosen experiment results for each setting. **Rows:** $\alpha = 0, 0.1, 0.5, 0.9$ from top to the bottom. **Columns:** $k = 50, 10, 5$ from left to right. Grey points denote the samples, and the yellow, red, and green point denote the population Fréchet mean, inductive mean, and FMoM estimator respectively.