Robust Estimation in metric spaces

Achieving Exponential Concentration with a Fréchet Median

Jakwang Kim*[1] Jiyoung Park*[2] Anirban Bhattacharya[2]

[1]PIMS Kantotrovich Initiative
University of British Columbia

[2]Department of Statistics
Texas A&M University

*equal contribution, alphabetically ordered.

March 4, 2025
AISTATS 2025

Outline

Introducion

TEXAS A&M UNIVERSITY

- Main question: Is there a 'robust' estimator that works for general 'parameter spaces'?
- Two aspects:
  1. Robustness: e.g. median is said to be more robust than mean in $\mathbb{R}$ (but in what criterion?).
  2. Parameter space: e.g. means of vectors $\in \mathbb{R}^d$; covariance matrices $\in$ Symmetric Positive Definite (SPD) matrices.
- Robust methods for Euclidean spaces do not extend directly to general parameter spaces.
  - E.g., what does it mean to take 'median' of vectors & covariance matrices?
- Our claim: 'Certain robustification method' (Fréchet median) can be extended to some general metric spaces (CAT(κ) spaces).

- Robust estimations
  - There are many notions of robustness: e.g. breakdown points, influence function, ...
  - In this work: robustness will stand for tail concentrations *i.e.*, the estimator is more robust if the estimator's distribution has a lighter tail.
  - Examples in $\mathbb{R}$:
    - If $X_i \overset{i.i.d}{\sim} N(\mathbb{E}(X), 1)$, then $\mathbb{P}(\left|\overline{X} - \mathbb{E}(X)\right| \geq t) \leq e^{-cnt^2}$ (Exact calculation or Bernstein).
    - If $Y_i \overset{i.i.d}{\sim}$ any distribution with a second moment, then $\mathbb{P}(\left|\overline{Y} - \mathbb{E}(Y)\right| \geq t) \leq \frac{\mathrm{Var}(Y)}{nt^2}$ (Chebyshev).
    - The first example *concentrates exponentially* to the target quantity w.r.t. n; the second example *concentrates polynomially* w.r.t. n. The first estimator is more robust.
- Why robustness?
  - We frequently encounter heavy-tail distributions.
  - Many standard estimators fail to be robust under heavy-tail setting, e.g., sample mean (see the above example).

- We often encounter parameters that are not in $\mathbb{R}^d$.
- Examples:
  - Infinite dimensional spaces: Hilbert spaces, Banach spaces - nonparametric regressions, functional data analysis.
  - Riemannian manifolds: Hypershperes - spatial statistics, SPD matrices (under some metrics) - covariance estimation problems.
  - Discrete spaces: Trees, graphs - Hierarchical models, Graphical models, network data analysis.
- $\Rightarrow$ Whether existing methods can be extended to such domain is an important question.

Preliminaries

- CAT(κ) space: a very general metric space equipped with the minimal essential geometric structure for our purpose.
  - Informal statement: CAT(κ) space is a metric space $(\mathcal{X}, d)$ whose curvatures are upper bounded by κ.
  - What does curvature mean here?
    - How curved the space is determined by comparing a triangle in $\mathcal{X}$ with a triangle in 2-dimensional κ-curved spaces $:= M_\kappa^2$ having same side lengths.
    - M_κ^2 for $\kappa < 0$ is a $1/\sqrt{|\kappa|}$ scaled hyperbolic space $(\mathbb{H}^2/\sqrt{|\kappa|})$, M_0^2 is the Euclidean plane, and M_κ^2 for $\kappa > 0$ is the 2-dimensional sphere of radius $1/\sqrt{\kappa}$ $(\mathbb{S}^2/\sqrt{\kappa})$. Here, the model space is κ-curved in a Gaussian curvature sense.
    - CAT(κ): Any sufficiently small triangle in $\mathcal{X}$ is 'thinner' than its *comparison triangle* in the 2-dimensional model space M_κ^2. See Figure 1.
    - This notion of curvature is called *Alexandrov curvature* (generalization of a sectional curvature in Riemannian geometry).

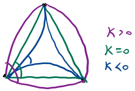

Figure 1: Triangle comparison: A fatter triangle have larger angles than a thinner triangle.

- Examples of CAT(κ) spaces:
  - Riemannian manifolds whose sectional curvatures are upper bounded by κ: Euclidean spaces, hyperspheres, hyperbolic spaces, many statistical manifolds in information geometry, SPD spaces with appropriate metrics.
  - Infinite dimensional Hilbert space.
  - Spaces of phylogenetic trees.
  - Metric graphs and trees: Metric graphs with cycles of length less than 2π is CAT(1). Metric trees are CAT(0).
- CAT(0) spaces are often called Non-Postiviely-Curved (NPC) spaces.

- Frećhet mean (median): Generalizations of Euclidean mean (*resp.* median).
- Analogy from Euclidean:
  - $\mathbb{E}(X) = \text{argmin}_Y \mathbb{E}_X \|X - Y\|^2$.
  - $med(X) \in \text{argmin}_Y \mathbb{E}_X \|X - Y\|$.
- Notations:
  - $(\mathcal{X}, d)$: a metric space.
  - $\mathcal{P}_p(\mathcal{X})$: a set of Borel probability measures in $\mathcal{X}$ with a finite pth moment, *i.e.*, $\int_{\mathcal{X}} d^p(x, y) dP(y) < \infty$ for some $x \in \mathcal{X}$.

Definition

For $P \in \mathcal{P}_p(\mathcal{X})$, suppose there exists an element $x^* \in \mathcal{X}$ such that

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

Such x^* with $p = 2$ ($p = 1$) is called a Fréchet mean (resp. median) of P.

- Some literatures uses the term *barycenter* (*resp. geometric median*) for Fréchet mean (*resp.* median).
- Existence and uniqueness of Fréchet mean and median:
  - In NPC spaces: Fréchet mean (*resp.* median) globally exists for all P with a finite second (*resp.* first) moment (Ba4).
    - Fréchet mean is unique.
    - Fréchet median may not be unique, but they form a single geodesic segment (Analogy: Think of median in $\mathbb{R}$).
  - $\kappa > 0$: Some additional conditions are required (e.g. bounded support, ...) (Yok17).
- Remark: Fréchet mean (median) is not the only generalization of Euclidean mean (*resp.* median) in metric spaces: inductive mean and convex mean (Stu99), tournament-based-median (LM19), Tukey's depth based median (DLP21), ...

- Median-of-means (MoM): an estimator achieving *exponential concentration* under only moment conditions.
  1. Split n data into k disjoint blocks.
  2. Obtain k sample means from each block.
  3. Take a median of k means; median-of-means.
- MoM: Concentrate to mean exponentially (more robust than sample mean).
- Fréchet Median-of-estimators (FMoE): taking a Fréchet median of any weakly (polynomially) concentrating estimators induces a strongly (exponentially) concentrating estimator when a parameter is in Banach spaces (Min15) and some Riemannian manifolds (LLSD20).
- **Our finding**: FMoE can be generalized up to CAT(κ)!
- Remark: FMoE is not the only option to obtain the robust estimator: trimmed mean (LM21), different notions of a median (LM19; YP23), ...

Main Theorems

- The exponential concentration of Fréchet median is a direct consequence of the geometric property of the Fréchet median.

Lemma (Geometric discrepency near the Fréchet median)

Let $(\mathcal{X}, d)$ be a CAT(κ) space, and fix $x_1, \ldots, x_k \in \mathcal{X}$. Denote $x^* := med(x_1, \ldots, x_k)$. 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, $x_j \in B(x^*, D_\kappa/2)$, 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 $J \subseteq \{1, \ldots, k\}$ of the cardinality $|J| > \alpha k$ such that for all $j \in J$, $d(x_j, z) > r$.

- Interpretations of Lemma:
  - If a point z is far away from a Fréchet median x^*, then z also has to be far away from the bulk of the points, x_j's.
  - This property of median is sometimes referred to *majority voting*.
- $\kappa > 0$ case requires some additional conditions, due to the unfavorable behavior of Fréchet median in positively-curved spaces.
- Proof idea:
  - CAT(κ) structure allow us to compare the triangle of $\mathcal{X}$ (precisely, $\triangle x_j x^* z$) to the triangle in the model space M_κ^2 ($\triangle \widetilde{x}_j \widetilde{x}^* \widetilde{z}$).
  - Since M_κ^2 is a 2-dimensional surface, there are direct calculation rules for some geometric quantities (e.g. $\cos(\angle \widetilde{x}_j \widetilde{x}^* \widetilde{z})$).
  - One will assume the claim is false, and show in such case x^* cannot be a Fréchet median (minima of the sum of the distance) using calculation rules in M_κ^2 $\Rightarrow$ contradiction!

TEXAS A&M UNIVERSITY

Theorem

Suppose the parameter space Θ *is CAT*(κ) *space. Let* $\theta \in \Theta$ *be a parameter of interest and* $\widehat{\theta}_j$, $j = 1, \dots, k$ *be independent estimators of* θ. *Let* $\widehat{\theta}_{FMoE} := med\left(\widehat{\theta}_1, \dots, \widehat{\theta}_k\right)$ *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.*

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

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

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

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

- $\kappa > 0$ case requires additional conditions, which are from the condition in previous Lemma.
- The proof is direct once we have Lemma - skecth for $\kappa < 0$:
  - The event $\left\{ d(\widehat{\theta}_{FMoE}, \theta) > C_\alpha \epsilon \right\} \subseteq \left\{ \sum_{j=1}^{k} \mathbb{1}_{\left\{ d(\widehat{\theta}_j, \theta) > \epsilon \right\}} > \alpha k \right\}$ from Lemma.
  - The probability of the latter event can be bounded by $\exp(-k\psi(\alpha, p))$ using Chernoff bound on Binomial random variables.
- This CAT(κ) extension covers the almost all the previous analysis on FMoE, except Banach space (Min15).
- Main takeaway: Whenever an estimator concentrates weakly (polynomially), applying FMoE will make an estimator concentrating strongly (exponentially) $\Rightarrow$ robust estimation!

- One can conduct the same procedure as in MoM to obtain FMoE.
- FMoE is *nearly* fully implementable:
  - Computation of Fréchet median is always possible in NPC spaces (Ba4).
  - If $\kappa > 0$, no universal method for Fréchet median; algorithms tailored to specific domains exist (e.g. Riemannian manifolds).
  - $\therefore$ in many applications, if you can compute the original estimator, you can compute FMoE.
- Time complexity of FMoE $\approx$ the original estimator.
  - If the original estimator has a time complexity $O(n^{\alpha})$ and the Fréchet median has a time complexity $O(n^{\beta})$, then FMoE's time complexity will be $O(n^{\alpha}k^{1-\alpha} + k^{\beta})$.
  - If $k = O(1)$ then it matches to $O(n^{\alpha})$.

Statistical Applications

1. Robust Fréchet mean estimations in NPC spaces

- Empirical Fréchet mean in NPC spaces concentrates weakly under mild moment assumptions.
- Throughout this section, we set $(\mathcal{X}, d)$ to be an NPC space whose Alexandrov curvatures $-\infty < curv(\mathcal{X}) \leq 0$, and $X_i \overset{i.i.d}{\sim} P \in \mathcal{P}_2(\mathcal{X})$ with σ^2 being a second moment of P.

Proposition ((GPRS19))

Let $\widehat{x}$ be an empirical Fréchet mean. Then,

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

Furthermore, for any $\epsilon > 0$

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

- Remark: One *cannot* guarantee such polynomial concentration under positive curvatures without additional assumptions (e.g., extendible geodesic condition (GPRS19)[Theorem 4.2]).

- For exponential concentration of the empirical Fréchet mean, one requires additional sub-Gaussian type assumptions.

Proposition ((BS24))

Suppose P satisfies the following sub-Gaussian type assumption, i.e.,

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

where $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R} \mid f$ is a 1-Lipschitz function$\}$. Then,

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

- Remark: This notion of sub-Gaussian is stronger than usual sub-Gaussian in $\mathbb{R}^d$ $(f(X) = \langle v_f, X\rangle$ for $v_f \in \mathbb{S}^{d-1})$.

- FMoM concentrates exponentially to population Fréchet mean in the absence of sub-Gaussian type assumptions, unlike empirical Fréchet mean.

Theorem

Fix $\delta > 0$. *Let* $\widehat{x}_{FMoM}$ *be a Fréchet median-of-means of* $\widehat{x}_j$*'s where* $\widehat{x}$ *is empirical Fréchet mean. Set* $k = \lfloor \log(1/\delta)/\psi(7/18, 1/10) \rfloor + 1$. *Then*

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

- Proof technique (standard):
  - $\mathbb{P}\left[d(\widehat{x}_j, x^*) \geq \epsilon\right] \leq \frac{2k\sigma^2}{n\epsilon^2} := p.$
  - Apply our main Theorem.
  - Optimize the bound w.r.t. p and α.

Covariance estimation problem

Motivations

- A difficulty in the sample covariance matrix estimation problem: SPD matrix constraint.
- Matrix norm based approaches may potentially violate this constraint.
  - E.g., A Fréchet median of SPD matrices w.r.t. a matrix norm $\Rightarrow$ Is it SPD? Non-trivial...
- Some metrics take into account the SPD constraint:

$$d_{AI}(A,B) := \|\log A^{-1/2}BA^{-1/2}\|_F \qquad \text{(Affine-Invariant metric)},$$
$$d_{BW}^2(A,B) := \operatorname{tr}(A) + \operatorname{tr}(B) - 2\operatorname{tr}(A^{1/2}BA^{1/2})^{1/2} \qquad \text{(Bures-Wasserstein metric)}.$$

- (SPD, d_{AI}): NPC space.
- (SPD, d_{BW}): CAT(κ) space for some $\kappa > 0$ under some additional assumptions.
- $\Rightarrow$ Their Fréchet medians are still SPD!

Robust estimation of covariance

- One can use FMoM proposed earlier for the robust estimator of the covariance matrix.
- However, sample covariance matrix $\widehat{\Sigma} = \sum_{i=1}^{n} X_i X_i^T / n$ is computationally favorable than Fréchet mean w.r.t. d_{AI}, d_{BW}.
- Idea: Use $\widehat{\Sigma}$ as an original estimator and take a Fréchet median.
- As long as $\widehat{\Sigma}$ concentrates polynomially w.r.t. *desired metric*, our method is applicable.

TEXAS A&M UNIVERSITY

Proposition (Polynomial tail bound for covariance matrix estimator)

Let $X_i \overset{i.i.d}{\sim} P \in \mathcal{P}_4(\mathbb{R}^d)$ a distribution with mean 0 and covariance matrix Σ with a fixed dimension d. Let $\widehat{\Sigma} = \sum_i X_i X_i^T / n$ be a sample covariance estimator. Then, writing $\lambda_{\min}$ the smallest eigenvalue of Σ,

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

for some constant $C > 0$ only depends on the moments of P.
For d_{BW}, in addition assume both $\widehat{\Sigma}$ and Σ have the eigenvalue lower bound by $\lambda_0 > 0$. Then

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

with the same C in the above.

- Since we have a polynomial concentration w.r.t. both metrics, the same technique in the previous section yields the exponential concentration of FMoE.

Theorem (Exponential concentration of median of sample covariance matrices)

Under the same setting in Proposition 4.4, we set $\widehat{\Sigma}_{FMoE}$ with the original estimator being a sample covariance matrix and the metric d being either d_{AI} or d_{BW}. Set $k = \lfloor \log(1/\delta)/\psi(0.4, 0.1) \rfloor + 1$.

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

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

(b) *For $d = d_{BW}$, again assume both $\widehat{\Sigma}$ and Σ have the eigenvalue lower bound by $\lambda_0 > 0$. In addition assume conditions in Theorem 3.2(b) holds with $\kappa = 3/(2\lambda_0^2)$, $\theta = \Sigma$, and $\widehat{\theta}_j = \widehat{\Sigma}_j$. Then, whenever $n > 6\lambda_0 kCd^4$, we have*

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

Numerical experiments

- Task: Covariance estimation problem.
- Dimension $d = 10$, number of samples $n = 10000$, number of blocks $k = 5$
- $X_i \sim t_{2.5}(0, \Sigma)$ for randomly generated Σ with fixing $\lambda_j = j$ for $j = 1, \ldots, 10$.
  - Heavy tail distribution with the variance being 5Σ.
- More experiments in our paper!

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

Task	$\mathbb{E}d^2(\widehat{\theta}, \theta)$	$\mathbb{E}d^2(\widehat{\theta}_{FMoE}, \theta)$	$d(\widehat{\theta}, \theta)$ CI	$d(\widehat{\theta}_{FMoE}, \theta)$ CI
Covariance (d_{AI})	0.6057	0.2931	$[0.4266, 1.6865]$	$[0.4132, 0.6792]$
Covariance (d_{BW})	8.3281	1.7360	$[0.9936, 5.8796]$	$[0.9443, 1.7409]$

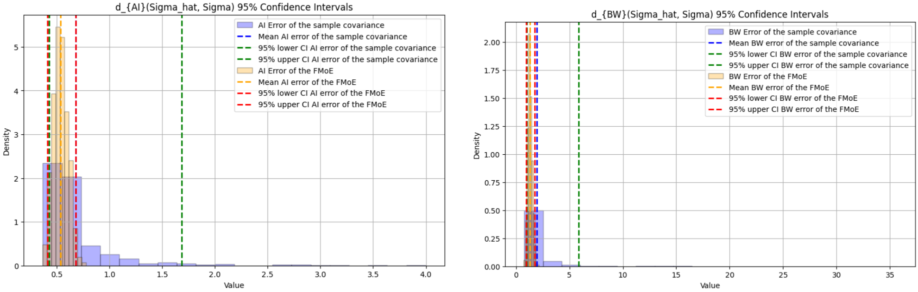

Figure 2: Histogram, mean, and 95% confidence interval for each experiment from 1000 simulations. **Left**: (SPD, d_{AI}). **Right**: (SPD, d_{BW}). All results indicate our method achieves much stronger concentration as well as much smaller mean squared errors.

Conclusions

- Taking Fréchet median of estimators makes more robust estimator.
- This phenomenon can be extended up to general metric spaces as long as it has a (Alexandrov) curvature upper bound.
  - This is due to the geometric property (majority voting) of Fréchet median.
- Our Fréchet median based approach is widely applicable, as well as nearly fully implementable with almost same computational costs.

- Possible further generalizations: Can we make an analysis that even covers Banach space case (Min15)?
- Overcoming the curvature upper bound condition: There are some spaces that are of interest but do not belong to CAT(κ), e.g., Wasserstein space.
- Adaptivity: Our bound relies on population quantities, which is not accessible in practice. This is a problem when we want to construct a sample confidence interval. Can we make the estimator adaptive?

Thank You For Your Attention!

[Ba4] Miroslav Bačák, *Computing medians and means in hadamard spaces*, SIAM Journal on Optimization **24** (2014), no. 3, 1542–1566.

[BS24] Victor-Emmanuel Brunel and Jordan Serres, *Concentration of empirical barycenters in metric spaces*, Proceedings of The 35th International Conference on Algorithmic Learning Theory (Claire Vernade and Daniel Hsu, eds.), Proceedings of Machine Learning Research, vol. 237, PMLR, 25–28 Feb 2024, pp. 337–361.

[DLP21] Xiongtao Dai and Sara López-Pintado, *Tukey's depth for object data*, Journal of the American Statistical Association **118** (2021), 1760 – 1772.

[GPRS19] 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).

[LLSD20] Lizhen Lin, Drew Lazar, Bayan Sarpabayeva, and David B. Dunson, *Robust optimization and inference on manifolds*, Statistica Sinica (2020).

[LM19] Gábor Lugosi and Shahar Mendelson, *Sub-Gaussian estimators of the mean of a random vector*, The Annals of Statistics **47** (2019), no. 2, 783 – 794.

[LM21] Gábor Lugosi and Shahar Mendelson, *Robust multivariate mean estimation: The optimality of trimmed mean*, Annals of Statistics **49** (2021), no. 1, –.

[Min15] Stanislav Minsker, *Geometric median and robust estimation in Banach spaces*, Bernoulli **21** (2015), no. 4, 2308 – 2335.

TEXAS A&M UNIVERSITY

[Stu99] Karl-Theodor Sturm, *Metric spaces of lower bounded curvature*, Expositiones Mathematicae **17** (1999), 035–048.

[Yok17] Takumi Yokota, *Convex functions and p-barycenter on CAT(1)-spaces of small radii*, Tsukuba Journal of Mathematics **41** (2017), no. 1, 43 – 80.

[YP23] Ho Yun and Byeong U. Park, *Exponential concentration for geometric-median-of-means in non-positive curvature spaces*, Bernoulli **29** (2023), no. 4.