Proceedings of Machine Learning Research vol 237:1–25, 2024 35th International Conference on Algorithmic Learning Theory

Concentration of empirical barycenters in metric spaces

Victor-Emmanuel Brunel VICTOR.EMMANUEL.BRUNEL@ENSAE.FR
CREST-ENSAE

Jordan Serres JORDAN.SERRES@ENSAE.FR
CREST-ENSAE

Editors: Claire Vernade and Daniel Hsu

Abstract

Barycenters (aka Fréchet means) were introduced in statistics in the 1940's and popularized in the fields of shape statistics and, later, in optimal transport and matrix analysis. They provide the most natural extension of linear averaging to non-Euclidean geometries, which is perhaps the most basic and widely used tool in data science. In various setups, their asymptotic properties, such as laws of large numbers and central limit theorems, have been established, but their non-asymptotic behaviour is still not well understood. In this work, we prove finite sample concentration inequalities (namely, generalizations of Hoeffding's and Bernstein's inequalities) for barycenters of i.i.d. random variables in metric spaces with non-positive curvature in Alexandrov's sense. As a byproduct, we also obtain PAC guarantees for a stochastic online algorithm that computes the barycenter of a finite collection of points in a non-positively curved space. We also discuss extensions of our results to spaces with possibly positive curvature.

Keywords: Barycenters, Concentration, Curvature, Metric spaces

1. Introduction

Statistics and machine learning are more and more confronted with data that lie in non-linear spaces. For instance, in spatial statistics (e.g., directional data), computational tomography (e.g., data in quotient spaces such as in shape statistics, collected up to rigid transformations), economics (e.g., optimal transport, where data are discrete measures), etc. Moreover, data that are encoded as very high dimensional vectors may have a much smaller intrinsic dimension, for instance, if they are lying on small dimensional submanifolds of the Euclidean space: In that case, leveraging the possibly non-linear geometry of the data can be a powerful tool in order to significantly reduce the dimensionality of the problem at hand. Even though more and more algorithms are developed to work with such data Lim and Pálfia (2014); Ohta and Pálfia (2015); Zhang and Sra (2016, 2018), there are still very little theory for uncertainty quantification, especially in non-asymptotic regimes, which are pervasive in machine learning. In this work, we prove statistical results for barycenters of data points, which are the most natural extension of linear averaging to non-linear geometries. Namely, working with an extension of the notion of sub-Gaussian random variables in the context of metric spaces, and assuming a non-positive curvature condition, we prove analogs of both Hoeffding and Bernstein concentration inequalities. Finally, we discuss extensions of our results to the case of metric spaces with possibly positive curvature.

In this paper, we consider a complete metric space (M, d). It is called non-positively curved (NPC for short) if for all pairs $(x, y) \in M$, there exists $m \in M$ satisfying

$$d(z,m)^2 \le \frac{1}{2}(d(z,x)^2 + d(z,y)^2 - \frac{1}{2}d(x,y)^2), \quad \forall z \in M. \tag{1}$$

NPC spaces are known to enjoy a lot of regularity: For instance, any two points $x, y \in M$ are connected by a unique (constant speed) geodesic $\gamma_{x,y}$, i.e., a continuous mapping $\gamma = \gamma_{x,y} : [0,1] \to M$ satisfying $d(\gamma(s), \gamma(t)) = |s-t|d(x,y)$, for all $s, t \in [0,1]$. Geodesics generalize line segments from Euclidean spaces, as shortest paths from one point x to another point y. Moreover, the point m in (1) is unique, given by the midpoint of x and y, i.e., $z = \gamma_{x,y}(1/2)$. Finally, the distance function d is geodesically convex jointly in both variables and for all $x_0 \in M$, the function $\frac{1}{2}d(x_0, \cdot)^2$ is 1-strongly geodesically convex. A function f is called geodesically convex (resp. α-strongly geodesically convex, for $\alpha > 0$) if it is convex (resp. α-strongly convex) along any geodesic $\gamma_{x,y}$, $x, y \in M$, i.e., $f(\gamma_{x,y}(t)) \leq (1-t)f(x) + tf(y)$ (resp. $f(\gamma_{x,y}(t)) \leq (1-t)f(x) + tf(y) - \frac{\alpha}{2}t(1-t)d(x,y)^2$), for all $x, y \in M$ and $t \in [0,1]$. In fact, a NPC space can also be defined as a geodesic metric space (i.e., a metric space where any two points are connected by at least one geodesic) that is complete and where triangles are thinner than Euclidean triangles with same side lengths. We refer the reader to Sturm (2003); Bridson and Haefliger (2013) for a thourough exposition on NPC spaces and to Bacák (2014) for a detailed account on convexity in metric spaces. Here is a short list of examples of NPC spaces:

- Euclidean and Hilbert spaces;
- Hyperbolic spaces and, more generally, Cartan-Hadamard manifolds, i.e. simply connected, complete, Riemannian manifolds with everywhere non-positive sectional curvature (Sturm, 2003, Proposition 3.1);
- Metric trees: These are combinatorial trees embedded in the Euclidean plane, in which the distance between any two points is the Euclidean length of the (unique) shortest path between them;
- The space $\mathcal{S}_p$ of symmetric positive definite matrices of size p (for any fixed integer $p \geq 1$) equipped with the following distance: For any $A, B \in \mathcal{S}_p$, $d(A,B) = \|\log(A^{-1/2}BA^{-1/2})\|_{\mathsf{F}}$, where $\|\cdot\|_{\mathsf{F}}$ is the Fröbenius norm. This construction is important in the study of matrix geometric means, see Section 3.3 below. For a detailed account on this space and matrix geometric means, we also refer the reader to Bhatia and Holbrook (2006) and the references therein.

Recently, NPC spaces have attracted much attention in the machine learning community: Hyperbolic spaces have been proved handy to embed certain types of data, such as hierarchical data Cho et al. (2022); Montanaro et al. (2022); Yang et al. (2023); Mishne et al. (2023); Desai et al. (2023), non-positive curvature naturally occurs in matrix learning Hosseini and Sra (2015) and optimization techniques taylored to NPC spaces are developing fast Hosseini and Sra (2015); Zhang et al. (2016); Criscitiello and Boumal (2022); Martínez-Rubio and Pokutta (2023); Criscitiello and Boumal (2023).

More generally, one can define spaces with curvature bounded from above by any real number κ: such spaces are called CAT(κ) spaces, coined after Cartan, Alexandrov and Toponogov. NPC spaces are simply CAT(0) spaces. Even though we will mostly work in NPC spaces here, Section 4 will be dedicated to extending our results to CAT(κ) spaces, for $\kappa > 0$. We refer to Section B in the appendix and to Bridson and Haefliger (2013) for a more detailed account on CAT spaces

A natural way to extend the notion of averaging from linear to metric spaces is through the notion of barycenters. Given $x_1, \ldots, x_n \in M$ ($n \geq 1$), a barycenter of $x_1, \ldots, x_n$ is any minimizer $b \in M$ of $n^{-1}\sum_{i=1}^n d(x_i, b)^2$. More generally, given a probability distribution μ on M, a barycenter of μ is any minimizer $b \in M$ of $\int_M d(x,b)^2 \,\mathrm{d}\mu(x)$ (provided μ has two moments, i.e., the latter integral is defined for at least one, and hence all, values of b [1]). Existence and uniqueness of barycenters are, in general, hard problems Afsari (2011); Yokota (2016, 2017). In NPC spaces, though, they are solutions to strongly convex optimization problems and hence, they always exist and are unique (Sturm, 2003, Proposition 4.3).

Barycenters were initially introduced in statistics by Fréchet (1948) in the 1940's, and later by Karcher (1977), where they were better known as Fréchet means, or Karcher means. They were popularized in the fields of shape statistics Kendall et al. (2009) and optimal transport Agueh and Carlier (2011); Cuturi and Doucet (2014); Le Gouic and Loubes (2017); Claici et al. (2018); Kroshnin et al. (2019); Altschuler and Boix-Adsera (2021, 2022). More broadly, barycenters in metric spaces have attracted attention in machine learning applications such as computational biology for phylogenetic trees Billera et al. (2001), shape analysis and computer vision Kendall et al. (2009); Marrinan et al. (2014), directional data analysis Edelman et al. (1998); Absil et al. (2004), modeling networks Lunagómez et al. (2021), matrix estimation Schwartzman (2016); Lodhia et al. (2022) and matrix analysis Bhatia and Holbrook (2006); Bhatia et al. (2019), etc.

An alternative construction of barycenters is given iteratively as follows. Given $x_1, \dots, x_n \in M$, the inductive barycenter is defined as the point $\tilde{b}_n$, where $\tilde{b}_1 = x_1, \tilde{b}_2 = \gamma_{\tilde{b}_1, x_2}(1/2)$ and for $k = 3, \dots, n$, $\tilde{b}_k = \gamma_{\tilde{b}_{k-1}, x_k}(1/k)$ (here, we implicitly assume that any pair of points is connected by a unique geodesic, which is the case when M is NPC). In Euclidean spaces, inductive barycenters coincide with barycenters, which are simply given by linear averages. However, they do not coincide in general, because of the lack of associativity of barycenters in non-linear spaces. Compared to barycenters, inductive barycenters have the advantage that they can be easily updated when the points $x_1, \dots, x_n$ come sequentially, in an online fashion. Moreover, their computation is not bound to an optimization problem and only requires to compute geodesics between pairs in M.

Let μ be a probability measure in (M, d) with two (or, again, one would suffice) moments. Let $X_1, \dots, X_n \sim \mu$ be i.i.d., where $n \geq 1$ is the sample size and is fixed. We are interested in the estimation of the barycenter b^* of μ (referred to as population barycenter) based on $X_1, \dots, X_n$, and we let $\hat{b}_n$ and $\tilde{b}_n$ be their barycenter, referred to as empirical barycenter and their inductive barycenter, respectively.

Asymptotic theory is well understood for empirical barycenters in various setups, particularly laws of large numbers Ziezold (1977) and central limit theorems in Riemannian manifolds (a smooth structure on M is a natural assumption in order to derive central limit theorems) Bhattacharya and Patrangenaru (2003, 2005); Bhattacharya and Lin (2017); Eltzner and Huckemann (2019); Eltzner et al. (2019). Only very few non-asymptotic results have been proven so far, most of which hold under fairly technical conditions. First, (Sturm, 2003, Theorem 4.7) bounded the expected value of $d(\tilde{b}_n, b^*)^2$ in NPC spaces. Namely, $\mathbb{E}[d(\tilde{b}_n, b^*)^2] \leq \frac{\sigma^2}{n}$ where $\sigma^2 = \mathbb{E}[d(X_1, b^*)^2]$ can be interpreted as the variance of X_1. (Le Gouic et al., 2022, Corollary 11) provide the same inequality for $\hat{b}_n$, under the extra constraint that (M, d) has curvature bounded from below. At a high level, this means that the space (M, d) is not branching (i.e., geodesics cannot split, unlike, for instance, in metric trees) and this ensures some regularity of the tangent cones of M, which is used in their analysis. They also extend their result to spaces (M, d) that are not necessarily NPC, but that satisfy a so-called hugging condition. However, except for NPC spaces, there is no explicit metric space that satisfy

1. It is actually sufficient for μ to only have one moment, and barycenters are defined as minimizers $b \in M$ of $\int_M (d(x,b)^2 - d(x,b_0)^2) \,\mathrm{d}\mu(x)$, which do not depend on the choice of a fixed base point $b_0 \in M$

such a condition. Recently, Escande (2023) showed that the above bound in expectation still holds even if the assumption of curvature bounded from below is dropped. Several non-asymptotic, high probability bounds are also known for empirical and inductive barycenters. Le Gouic et al. (2022) propose a definition of sub-Gaussian random variables (eq. (3.10)), closely related to the one we give below, and prove (Theorem 12), under the hugging condition mentioned above, a nearly sub-Gaussian tail bound (with a residual term that decays exponentially fast with n) for $\hat{b}_n$ when the data are sub-Gaussian. Ahidar-Coutrix et al. (2020) obtain concentration inequalities with non-parametric rates for $\hat{b}_n$ when the data are bounded, under some metric entropy conditions on (M, d). Finally, most closely related to our work, Funano (2010) proves a Hoeffding-type inequality for the inductive barycenter $\tilde{b}_n$ of i.i.d., bounded random variables in NPC spaces, with particular focus on metric trees and finite dimensional Hadamard manifolds. Namely, If M is a p-dimensional Hadamard manifold (i.e., a Riemannian manifold that is complete, has non-positive curvature, and has dimension $p \geq 1$) and the data are almost surely contained in some ball of fixed radius $C > 0$, then, for all $\delta \in (0, 1)$, it holds with probability at least $1 - \delta$ that

$$d(\tilde{b}_n, b^*) \leq A_1 C \sqrt{\frac{p \log(A_2/\delta)}{n}},$$

where A_1, A_2 are positive universal constants ($4 \leq A_1 \leq 6$ and A_2 is of order $50,000$).

In Corollary 11 below, we greatly improve this bound in several ways. First, we significantly improve the constants. Second, our bound does not require M to have finite dimension. Moreover, we do not even require M to be a smooth manifold. Third, our upper bound contains two terms, which decouple the probability level δ from the variance σ^2. If M is a Hilbert space, σ^2 is the trace of the covariance operator (see below), and it is well known that unlike the bound proved in Funano (2010), our bound is optimal in that case.

The paper is organized as follows. In Section 2, we introduce tools from the concentration of measure theory in metric spaces, which we use in Section 3 to derive Hoeffding and Bernstein type concentration inequalities for empirical and inductive barycenters in NPC spaces. Extensions of our results from this section to the case of non i.i.d. data are given in Section A in the appendix. Finally, in Section 4, we treat the case of possibly positively curved spaces, and derive an optimal rate concentration bound depending on the radius of the space.

In what follows, for any positive integer n and any metric space (M, d), we denote by d_1 (without mention of the dependence on n, for simplicity) the L^1 product metric on M^n, defined as $d_1((x_1, \ldots, x_n), (y_1, \ldots, y_n)) = d(x_1, y_1) + \ldots + d(x_n, y_n)$.

2. The Laplace transform and Sub-Gaussian random variables in metric spaces

2.1. Laplace transform

In this section, we gather information on the Laplace transform of probability measures on metric spaces. It will allow us to study precisely the concentration phenomenon of barycenters in NPC spaces and in particular to deal with sub-Gaussian random variables in such spaces. Let (M, d) be a metric space (not necessarily NPC). Denote by $\mathcal{F}$ the class of all functions $f : M \to \mathbb{R}$ that are 1-Lipschitz, i.e., such that $|f(x) - f(y)| \leq d(x, y)$ for all $x, y \in M$. Let X be a random variable in M. For all $k \geq 1$, we say that X has k moments if $\mathbb{E}[d(X, x_0)^k]$ is finite, where x_0 is any arbitrary point in M (note that this definition does not depend on the choice of x_0). The Laplace transform of a random variable X that has at least one moment is defined as (see (Ledoux, 2001, Section 1.6))

$$\Lambda_X(\lambda) := \sup_{f \in \mathcal{F}} \mathbb{E}[e^{\lambda(f(X)-\mathbb{E}[f(X)])}], \quad \lambda \geq 0. \tag{2}$$

Note that if X has one moment, then so does $f(X)$, for all $f \in \mathcal{F}$. Let us underline the following property of the Laplace transform, whose proof can be found in (Ledoux, 2001, Proposition 1.15).

Lemma 1 *(Ledoux, 2001, Proposition 1.15) If $X_1, \ldots, X_n$ are independent random variables on (M, d) with at least one moment, then the Laplace transform of the random vector $(X_1, \ldots, X_n)$ in (M^n, d_1) satisfies*

$$\Lambda_{(X_1,\ldots,X_n)} \leq \Lambda_{X_1} \cdots \Lambda_{X_n},$$

where we recall that d_1 is the L^1 product metric on M^n.

Finally, we state the following property of the Laplace transform, when a Lipschitz function is applied to a random variable.

Lemma 2 *Let $(M^{(1)}, d^{(1)})$ and $(M^{(2)}, d^{(2)})$ be metric spaces and $\Phi : M^{(1)} \to M^{(2)}$ be an L-Lipschitz function, where $L > 0$. Then, for all random variables X in $M^{(1)}$ with at least one moment,*

$$\Lambda_{\Phi(X)}(\lambda) \leq \Lambda_X(\lambda L), \quad \forall \lambda \geq 0.$$

Proof Let $f : M^{(2)} \to \mathbb{R}$ be a 1-Lipschitz function. Then, for all $\lambda \geq 0$,

$$\mathbb{E}[e^{\lambda(f(\Phi(X))-\mathbb{E}[f(\Phi(X))])}] = \mathbb{E}[e^{\lambda L \frac{(f(\Phi(X))-\mathbb{E}[f(\Phi(X))])}{L}}] = \mathbb{E}[e^{\lambda L(g(X)-\mathbb{E}[g(X)])}]$$

where $g = (1/L)f \circ \Phi$ is a 1-Lipschitz function. Hence, $\mathbb{E}[e^{\lambda(f(\Phi(X))-\mathbb{E}[f(\Phi(X))])}] \leq \Lambda_X(\lambda L)$ and one concludes by taking the supremum over all 1-Lipschitz functions $f : M^{(2)} \to \mathbb{R}$. ■

2.2. Sub-Gaussian random variables

Here, we adapt the standard definition of sub-Gaussian random variables to the setup of random variables in abstract metric spaces. We refer the reader to (Vershynin, 2018, Section 2.5)

Definition 3 *A random variable X in (M, d) is called K^2-sub-Gaussian ($K \geq 0$) if and only if $\Lambda_X(\lambda) \leq e^{\lambda^2 K^2/2}$, for all $\lambda \in \mathbb{R}$.*

Sub-Gaussian random variables are well understood and play a very important role in the Euclidean setup. In particular, they are known to enjoy good concentration properties (Vershynin, 2018, Section 2.5). The following lemma extends this fact to sub-Gaussian random variables in metric spaces.

Lemma 4 *Let X be a random variable in (M, d) and let $K > 0$. The following statements are equivalent:*

(i) X is K^2-sub-Gaussian

(ii) $f(X)$ *is* K^2*-sub-Gaussian, for all* $f \in \mathcal{F}$

(iii) $\sup_{f\in\mathcal{F}} P(|f(X) - \mathbb{E}[f(X)]| \geq t) \leq 2e^{-t^2/(2K^2)}$*, for all* $t \geq 0$.

Moreover, if $\sup_{f\in\mathcal{F}} \mathbb{E}\left[e^{\frac{(f(X)-\mathbb{E}[f(X)])^2}{2K^2}}\right] \leq 2$*, then* X *is* K^2*-sub-Gaussian.*

Proof This proposition directly follows from (Vershynin, 2018, Proposition 2.5.2). Indeed, it is clear that X is K^2-sub-Gaussian in the sense of Defintiion 3 if and only if for all f$\in \mathcal{F}$, $f(X)$ is K^2-sub-Gaussian in the usual sense (Vershynin, 2018, Section 2.5). ■

Definition 3 is stronger than the standard definition of sub-Gaussian random variables in Euclidean spaces. Indeed, if X is a random variable in $\mathbb{R}^p$ ($p \geq 1$), X is usually said to be K^2-sub-Gaussian if and only if for all unit vectors $u \in \mathbb{R}^d$ and all $\lambda \in \mathbb{R}$, it holds that

$$\mathbb{E}[e^{\lambda u^\top X}] \leq e^{\frac{\lambda^2 K^2}{2}}$$

(in other words, $u^\top X$ must be K^2-sub-Gaussian in the usual sense of (Vershynin, 2018, Definition 2.5.6)). For instance, let Y have the standard Gaussian distribution in $\mathbb{R}^p$ ($p \geq 1$) and let Z be a Bernoulli random variable independent of Y such that $P(Z = 0) = P(Z = 1) = 1/2$. Let $X = YZ$. It is easy to check that for all unit vectors $u \in \mathbb{R}^p$, $u^\top X$ is 1-sub-Gaussian. However, there are 1-Lipschitz functions $f : \mathbb{R}^p \to \mathbb{R}$ for which $f(X)$ is not 1-sub-Gaussian. For instance, simply take $f = \|\cdot\|$ (Euclidean norm in $\mathbb{R}^p$). If $f(X)$ was K^2-sub-Gaussian for some $K > 0$, then it would necessarily hold that

$$P(f(X) < \mathbb{E}[f(X)] - \sqrt{p}/2) \leq e^{-p/(8K^2)}.$$

However, since $\mathbb{E}[f(X)]$ is approximately $\sqrt{p}/2$, when p is large, it holds that the latter probability is at least $1/2$, which yields a contradiction, unless K grows as $\sqrt{p}$.

However, in the context of metric spaces, this definition seems most appropriate because of the lack of linear functions and, as will be seen in Lemma 7 below, allows us to treat bounded random variables. Moreover, as opposed to the definition suggested in (Le Gouic et al., 2022, Section 3.3), it does not depend on any reference point in M.

Finally, let us mention a similar definition for sub-Gaussian random variables that is given in Kontorovich (2014). There, a random variable X is declared K^2-sub-Gaussian ($K \geq 0$) if and only if $\mathbb{E}[e^{\lambda\varepsilon d(X,Y)}](= \mathbb{E}[\cosh(\lambda d(X,Y))]) \leq e^{\lambda^2 K^2/2}$ for all $\lambda \in \mathbb{R}$, where Y is an independent copy of X and ε is a Rademacher random variable independent of (X, Y). Their definition is stronger than ours, by the symmetrization and contraction principle. Indeed, if that condition is satisfied, then for all $\lambda \in \mathbb{R}$ and all $f \in \mathcal{F}$,

$$\mathbb{E}[e^{\lambda(f(X)-\mathbb{E}[f(X)])}] \leq \mathbb{E}[e^{\lambda(f(X)-f(Y))}] = \mathbb{E}[e^{\lambda\varepsilon|f(X)-f(Y)|}] \leq \mathbb{E}[e^{\lambda\varepsilon d(X,Y)}] \leq e^{\lambda^2K^2/2},$$

where the first inequality is Jensen's inequality, the following equality is symmetrization and the next inequality is contraction. Moreover, our definition is more flexible: For instance, as will be clear in the next section, it will allow us to derive a Bernstein-type inequality for barycenters in metric spaces.

The next two propositions show that the sub-Gaussian property is preserved by tensorization and by Lipschitz transformations.

Proposition 5 (Tensorization) *Let $X_1, \ldots, X_n$ be independent random variables in M such that each X_i is K_i^2-sub-Gaussian for some $K_i > 0$. Then, the n-uple $(X_1, \ldots, X_n)$ is $(K_1^2 + \ldots + K_n^2)$-sub-Gaussian on the product metric space (M^n, d_1).*

Proof Let X_i be K_i-sub-Gaussian, for each $i = 1, \ldots, n$. Then, $\Lambda_{X_i}(\lambda) \leq e^{\lambda^2 K_i^2/2}$, for all $i = 1, \ldots, n$ and $\lambda \geq 0$. Therefore, by Lemma 1,

$$\Lambda_{(X_1,\ldots,X_n)}(\lambda) \leq \Lambda_{X_1}(\lambda) \ldots \Lambda_{X_n}(\lambda) \leq \prod_{i=1}^{n} e^{\lambda^2 K_i^2/2} = e^{\lambda^2 (K_1^2 + \ldots + K_n^2)/2},$$

for all $\lambda \geq 0$, which yields the result. ■

Proposition 6 *Let $(M^{(1)}, d^{(1)})$ and $(M^{(2)}, d^{(2)})$ be metric spaces and let X be a random variable in M_1. Let $K, L > 0$. If X is K^2-sub-Gaussian and $\Phi : M^{(1)} \to M^{(2)}$ is L-Lipschitz, then $\Phi(X)$ is $(L^2 K^2)$-sub-Gaussian.*

Proof Let $g : M^{(2)} \to \mathbb{R}$ be a 1-Lipschitz function and let $\tilde{g} := L^{-1} g \circ \Phi$. Then, $\tilde{g} : M^{(1)} \to \mathbb{R}$ is 1-Lipschitz so we have

$$\mathbb{E}\left[e^{\lambda g(\Phi(X)) - \mathbb{E}[g(\Phi(X))]}\right] = \mathbb{E}\left[e^{\lambda L(\tilde{g}(X) - \mathbb{E}[\tilde{g}(X)])}\right] \leq \Lambda_\mu(\lambda L) \leq e^{\lambda^2 L^2 K^2/2}.$$

We conclude by taking the supremum over all such functions g. ■

Let us conclude this section with two lemmas, which provide important examples of sub-Gaussian random variables. The first one is from Ledoux (2001); Similarly to Hoeffding's lemma for real-valued random variables, it indicates that bounded random variables are always sub-Gaussian.

Lemma 7 *(Ledoux, 2001, Proposition 1.16) Let X be a bounded random variable in the metric space (M, d), i.e. $d(x_0, X) \leq C$ a.s. for some $x_0 \in M$ and $C > 0$. Then, X is C^2-sub-Gaussian.*

Note that (Ledoux, 2001, Proposition 1.16) actually indicates that X is $4C^2$-sub-Gaussian, because the proof uses a simple, yet slightly loose, symmetrization argument. In fact, if X is almost surely bounded in a domain of diameter at most $2C$, then for all 1-Lipschitz functions $f : M \to \mathbb{R}$, $f(X)$ is almost surely bounded in an interval of length at most $2C$, hence, $f(X)$ is C^2-sub-Gaussian by standard arguments, see, e.g., (Vershynin, 2018, Example 2.5.8 (c)).

The second lemma holds under some extra assumptions on the metric space (M, d). Here, we assume that (M, d) is a Riemannian manifold, i.e., that M is a differentiable manifold and that d is inherited from a Riemannian metric on M. Moreover, we assume that M has Ricci curvature bounded from below and that X has a density with respect to the Riemannian volume. Since Riemannian geometry is not at the heart of this work, we refer to Lee (2012) and Lee (2018) for details on smooth manifolds, Riemannian metrics, different notions of curvature (including Ricci) and Riemannian volume. For the intuition, it is enough to assume that the sectional curvature (which is the most intuitive notion of curvature, and which agrees with the definition of curvature bounds in CAT spaces, see Section B) is bounded from below by a constant, in order to ensure that the Ricci curvature is bounded by the same constant.

Lemma 8 *Let M be a Riemannian manifold with Ricci curvature bounded from below by $\kappa \in \mathbb{R}$. Assume that X has a density ϕ with respect to the Riemannian volume, and that*

$$\phi(x) \leq Ce^{-\beta d(x,x_0)^2}, \forall x \in M,$$

where $C, \beta > 0$ and $x_0 \in M$ are given. Then, X is K^2-sub-Gaussian, for some $K > 0$ that depends on C, β and κ.

A closed form for K can be deduced from the proof, but we do not make it explicit here, for the sake of the simplicity of our presentation. The proof of this lemma can be found in the Appendix, Section D.1.

3. Concentration of empirical barycenters in NPC spaces

In this section, we assume that (M, d) is an NPC space.

3.1. Lipschitz property of barycenters

The last ingredient in order to prove concentration of empirical barycenters and inductive barycenters in NPC spaces is their Lipschitz property, which we state in the next proposition. Denote by $\hat{B}_n : M^n \to M$ the function that maps any n-uple to its (uniquely defined) barycenter and by $\tilde{B}_n$ the function that maps any n-uple to its inductive barycenter. Recall that given $x_1, \ldots, x_n \in M$, their empirical barycenter $\hat{B}_n(x_1, \ldots, x_n)$ is the unique minimizer of $n^{-1} \sum_{i=1}^n d(b, x_i)^2, b \in M$ and their inductive barycenter is defined inductively by setting $\tilde{B}_1(x_1) = x_1$ and, for all $k = 2, \ldots, n$, $\tilde{B}_k(x_1, \ldots, x_k) = \gamma_{\tilde{B}_{k-1}(x_1,\ldots,x_{k-1}),x_k}(1/k)$.

Proposition 9 *(Funano, 2010, Lemma 3.1),(Lim and Pálfia, 2014, Theorem 3.4),(Sturm, 2003, Theorem 6.3) Both functions $\hat{B}_n$ and $\tilde{B}_n$ are $(1/n)$-Lipschitz, M^n being equipped with the L^1 product metric d_1.*

Proposition 9 is well known in the literature, but we give three proofs that we believe are instructive, in Section D.2 in the appendix.

3.2. A concentration inequality for barycenters of sub-Gaussian random variables

We are now in position to state our first main result, which implies concentration of the empirical barycenter and the inductive barycenter of i.i.d. sub-Gaussian random variables in an NPC metric space.

Theorem 10 *Let $X_1, ..., X_n$ be independent random variables in (M, d) such that for all $i = 1, \ldots, n$, X_i is K_i^2-sub-Gaussian, for some $K_i > 0$. Then both the empirical and inductive barycenters of $X_1, ..., X_n$ are $\frac{K_1^2+\ldots+K_n^2}{n^2}$-sub-Gaussian.*

Proof Let B_n be either the inductive or the empirical barycenter of $X_1, \ldots, X_n$. Then, B_n can be written as $\Phi(X_1, \ldots, X_n)$, where Φ is either $\hat{B}_n$ or $\tilde{B}_n$, which, by Proposition 9 are both $(1/n)$-Lipschitz functions on M^n, equipped with the L^1 product metric d_1. By Proposition 5, the n-uple $(X_1, \ldots, X_n)$ is $(K_1^2 + \ldots + K_n^2)$-sub-Gaussian and by Proposition 6, $\Phi(X_1, \ldots, X_n)$ is therefore $n^{-2}(K_1^2 + \ldots + K_n^2)$-sub-Gaussian. ∎

As a consequence, thanks to Lemma 4, for all 1-Lipschitz functions $f : M \to \mathbb{R}$ and for all $\delta \in (0,1)$, it holds with probability at least $1-\delta$ that $f(B_n) \le \mathbb{E}[f(B_n)] + \bar{K}_n\sqrt{\frac{\log(1/\delta)}{n}}$ where B_n is either the empirical or the inductive barycenter of $X_1, \dots, X_n$ and $\bar{K}_n = \sqrt{K_1^2 + \dots + K_n^2}$.

Assume that $X_1, \dots, X_n$ are i.i.d. and let b^* be their (population) barycenter. Denote by $\hat{b}_n = \hat{B}_n(X_1, \dots, X_n)$ their empirical barycenter and by $\tilde{b}_n = \tilde{B}_n(X_1, \dots, X_n)$ their inductive barycenter. Then, of particular interest is taking $f = d(\cdot, b^*)$. Let $\sigma^2 = \mathbb{E}[d(X_1, b^*)^2]$ be the variance of X_1. When B_n is the inductive barycenter of $X_1, \dots, X_n$, it follows from Sturm's law of large numbers (Sturm, 2003, Theorem 4.7) that $\mathbb{E}[d(B_n, b^*)] \le \frac{\sigma}{\sqrt{n}}$. When B_n is the empirical barycenter of $X_1, \dots, X_n$, the same inequality holds thanks to (Le Gouic et al., 2022, Corollary 11) under the extra assumption that the space (M,d) has its curvature bounded from below (see Le Gouic et al. (2022) for more details). Recently, Escande (2023) proved that the same inequality actually holds without any lower bound assumption on the curvature of the space. Therefore, we get the following corollary.

Corollary 11 *Let $X_1, \dots, X_n$ be i.i.d. K^2-sub-Gaussian, for some $K > 0$. Let B_n be either $\hat{b}_n$ or $\tilde{b}_n$. Then, for all $\delta \in (0,1)$, it holds with probability at least $1-\delta$ that*

$$d(B_n, b^*) \le \frac{\sigma}{\sqrt{n}} + K\sqrt{\frac{\log(1/\delta)}{n}}$$

We can now state Hoeffding's inequality in NPC spaces.

Corollary 12 (Hoeffding's inequality in NPC spaces) *Let $X_1, ..., X_n$ be i.i.d random variables in M. Assume that $d(X_1, x_0) \le C$ almost surely, for some $x_0 \in M$ and $C > 0$. Let B_n be either $\hat{b}_n$ or $\tilde{b}_n$. Then, for all $\delta \in (0,1)$, it holds with probability at least $1-\delta$ that*

$$d(B_n, b^*) \le \frac{\sigma}{\sqrt{n}} + C\sqrt{\frac{\log(1/\delta)}{n}}.$$

Proof In Corollary 11, K can be replaced with $2C$ thanks to Lemma 7, yielding the result. ■

The right hand side of this last concentration inequality contains two terms: A bias term, which simply controls the expected distance from B_n to b^*, and a stochastic term, which depends on the probability level δ. If (M,d) is a Hilbert space (which is a special instance of NPC spaces), this inequality actually reads as

$$\|\bar{X}_n - \mathbb{E}[X_1]\| \le \sqrt{\frac{\mathrm{tr}\Sigma}{n}} + C\sqrt{\frac{\log(1/\delta)}{n}},$$

where $\|\cdot\|$ is the Euclidean or Hilbert norm, $\bar{X}_n$ is the empirical mean of $X_1, \dots, X_n$ and Σ is the covariance operator of X_1. For instance, if $M = \mathbb{R}^p$ equipped with the Euclidean structure and Σ is the identity matrix, then $\mathrm{tr}\Sigma = p$ and the dimension of M only appears in the bias term, not the stochastic one. This is mainly why Corollary 11 is a significant improvement over Funano's result Funano (2010). Also note that in the Euclidean or Hilbert case, our inequality is optimal in terms of n and δ, in general. Only when the variance parameter σ^2 is very small, the dependence of this bound in σ and C can be improved, just as in Bernstein's inequality: This is what we show in Section 3.4 below.

Note that Corollary 12 was obtained independently in Escande (2023), using a completely different technique than ours (namely, a quadruple inequality, which holds in NPC spaces), which only applies to the bounded case.

3.3. Application: fast stochastic approximation of barycenters

Corollary 11 yields an algorithmic PAC guarantee for the stochastic approximation of barycenters of finitely many points in NPC spaces. Let $x_1, \ldots, x_n$ be given (deterministic) points in M. Here, the goal is to approximate their barycenter $b_n = \hat{B}_n(x_1, \ldots, x_n)$. Recall that b_n is the solution of an optimization problem, which may be hard to solve numerically. Fix some positive integer m and follow the following steps:

- Sample m integers $I_1, \ldots, I_m$ independently, uniformly at random between 1 and n;
- Set $X_1 = x_{I_1}, \ldots, X_m = x_{I_m}$;
- Compute $\tilde{b}_m = \tilde{B}_m(X_1, \ldots, X_m)$, the inductive barycenter of $X_1, \ldots, X_m$.

The random variables $X_1, \ldots, X_m$ obtained in the second step are i.i.d. with distribution $\mu = n^{-1} \sum_{i=1}^n \delta_{x_i}$. In other words, they are obtained by a bootstrap procedure based on the collection $\{x_1, \ldots, x_n\}$. In particular, their population barycenter is given by b_n.

In general, if m is not too large, computing $\tilde{b}_m$ is simpler than computing b_n directly, as long as one has access to an oracle that gives geodesics between any two points of M. The following result provides a PAC guarantee for $\tilde{b}_m$, as a stochastic approximation of b_n.

Corollary 13 *Let $\varepsilon > 0$ and $\delta \in (0, 1)$. Let D be the diameter of the set $\{x_1, \ldots, x_n\}$. Then, if $m \geq \frac{4D^2}{\varepsilon^2} \max(1, \log(1/\delta))$, it holds that $d(\tilde{b}_m, b_n) \leq \varepsilon$ with probability at least $1 - \delta$.*

Proof Let σ^2 be the variance of X_1, i.e., $\sigma^2 = \mathbb{E}[d(X_1, b_n)^2]$. Since for all $x \in M$, $d(x, \cdot)^2$ is convex, Jensen's inequality (see, e.g., (Sturm, 2003, Theorem 6.2)) yields that $d(X_1, b_n)^2 \leq \frac{1}{n} \sum_{j=1}^m d(X_1, x_j)^2$ almost surely, and each term in the sum is bounded by D^2. Hence, $\sigma^2 \leq D^2$. Therefore, Corollary 12 yields that with probability at least $1 - \delta$, $d(\tilde{b}_m, b_n) \leq \frac{D}{\sqrt{n}}(1 + \sqrt{\log(1/\delta)})$, which implies the desired result. ■

Perhaps surprisingly, the algorithm complexity given by Corollary 13 is dimension free. Moreover, the dependence in n of the algorithm complexity only comes from the uniform sampling of integers between 1 and n.

Note that in some cases, σ^2 might be much smaller than D, and the bound given in Corollary 13 can actually be further improved, using Theorem 14 below, see Section 3.4.

In comparison with this guarantee, (Lim and Pálfia, 2014, Theorem 3.4) gives a deterministic guarantee for finding an ε-approximation of the barycenter of $x_1, \ldots, x_n$, after $\frac{n(D^2+\sigma^2)}{\varepsilon^2}$ steps: The complexity of their algorithm is n times worse than ours, where n is the number of input points.

An important example where this guarantee is useful is that of metric trees, where the computation of inductive barycenters simply requires to identify the shortest paths between any two points, which can be done efficiently. Another important example, in matrix analysis, is that of computing matrix geometric means. Recall that the geometric mean of positive definite matrices $A_1, \dots, A_n \in \mathcal{S}_p$ $(n, p \geq 1)$ is their barycenter, associated with the metric $d(A, B) = \| \log(A^{-1/2}BA^{-1/2})\|_{\mathsf{F}}$, which makes $\mathcal{S}_p$ an NPC space (Bhatia and Holbrook, 2006, Proposition 5). The geometric mean of two matrices $A, B \in \mathcal{S}_p$ is the matrix $A\#B = A^{1/2}(A^{-1/2}BA^{-1/2})^{1/2}A^{1/2}$ and more generally, the geodesic segment between A and B is given by $\gamma_{A,B}(s) = A^{1/2}(A^{-1/2}BA^{-1/2})^{s}A^{1/2}$, also denoted by $A\#_s B$, for all $s \in [0, 1]$. Hence, computing the sequence of inductive barycenters of positive definite matrices boils down to computing expressions such as $A^{1/2}(A^{-1/2}BA^{-1/2})^{s}A^{1/2}$ for $s = 1/2, 1/3, \dots$ which can be done exactly with matrix products and eigendecompositions, whose complexities depend on the size p of the matrices. In fact, there are faster ways to compute good approximations of $A\#_s B$, for $A, B \in \mathcal{S}_p$ and $s \in [0, 1]$, e.g., by using integral representations and Gaussian quadrature: We refer, for instance, to Bhatia (2009); Simon (2019) for more details.

3.4. Bernstein inequality: A refinement in the case of small variance

In this section, we derive a Bernstein inequality refining Hoeffding's lemma in the case where the X_i's have a small variance, i.e., much smaller than the diameter of their support.

Theorem 14 (Bernstein's inequality in NPC spaces) *Recall the notation and assumption of Corollary 12. Then, with probability at least $1 - \delta$, it holds that*

$$d(B_n, b^*) \leq \frac{\sigma}{\sqrt{n}} + \max\left(2\sigma\sqrt{\frac{\log(1/\delta)}{n}}, \frac{8C\log(1/\delta)}{3n}\right).$$

When σ is much smaller than the size of the support of the X_1, this inequality greatly improves Corollary 11. Indeed, it is clear that $\sigma^2 \leq C^2$, since $\sigma^2 = \min_{x\in M} \mathbb{E}[d^2(X_1, x)] \leq \mathbb{E}[d^2(X_1, x_0)] \leq C^2$. However, note that in general, if X_1 is K^2-sub-Gaussian, we do not know whether it holds, like in the Euclidean case, that $\sigma^2 \leq K^2$.

Proof

Let $\lambda \geq 0$ and $f \in \mathcal{F}$ and denote by $\psi(\lambda, f) = \log \mathbb{E}[e^{\lambda(f(X_1)-\mathbb{E}[f(X_1)])}]$. Using the inequality $\log(u) \leq 1 - u$, for all $u > 0$, it holds that

$$\psi(\lambda, f) \leq \mathbb{E}\left[e^{\lambda(f(X_1)-\mathbb{E}[f(X_1)])} - 1\right] = \mathbb{E}\left[e^{\lambda(f(X_1)-\mathbb{E}[f(X_1)])} - 1 - \lambda(f(X_1) - \mathbb{E}f(X_1))\right].$$

Since $d(X_1, x_0) \leq C$ almost surely, we have that $f(X_1) - \mathbb{E}f(X_1) \leq 2C$ almost surely. Therefore, since the map $u \mapsto \frac{e^u-1-u}{u^2}$ is increasing, we obtain that

$$\psi(\lambda, f) \leq \mathbb{E}\left[\frac{e^{2\lambda C} - 1 - 2\lambda C}{4C^2}(f(X_1) - \mathbb{E}f(X_1))^2\right] = \frac{e^{2\lambda C} - 1 - 2\lambda C}{4C^2}\mathrm{Var}(f(X_1)).$$

Moreover, $\mathrm{Var}(f(X_1)) \leq \sigma^2$. Indeed,

$$\mathrm{Var}(f(X_1)) = \mathbb{E}\left[(f(X_1) - \mathbb{E}[f(X_1)])^2\right] \leq \mathbb{E}\left[(f(X_1) - f(b^*))^2\right] \leq \mathbb{E}\left[d(X_1, b^*)^2\right] = \sigma^2,$$

where the first inequality follows from the fact that $\mathbb{E}\left[(f(X_1) - \mathbb{E}[f(X_1)])^2\right] \leq \mathbb{E}\left[(f(X_1) - a)^2\right]$ for all $a \in \mathbb{R}$, and in particular for $a = f(b^*)$.

Finally, by Lemma 1, we obtain that $\Lambda_{X_1,\dots,X_n}(\lambda) \leq e^{\frac{n\sigma^2}{4C^2}(e^{2\lambda C}-1-2\lambda C)}$, for all $\lambda \geq 0$. Now, let B_n be either the empirical or the inductive barycenter of $X_1, \dots, X_n$. It follows from Proposition 9 and Lemma 2 that $\Lambda_{B_n}(\lambda) \leq e^{\frac{n\sigma^2}{4C^2}(e^{2\lambda C/n}-1-2\lambda C/n)}$, for all $\lambda \geq 0$.

By Chernoff's bound, we obtain, for all $f \in \mathcal{F}$ and all $t \geq 0$,

$$\begin{aligned} P\left(f(B_n) - \mathbb{E}f(B_n) \geq t\right) &\leq e^{-\lambda t}\mathbb{E}e^{\lambda(f(B_n)-\mathbb{E}f(B_n))} \leq e^{-\lambda t + \log \Lambda_{B_n}(\lambda)} \\ &\leq \exp\left[-\lambda t + \left(e^{\frac{2\lambda C}{n}} - 1 - \frac{2\lambda C}{n}\right)\frac{n\sigma^2}{4C^2}\right]. \end{aligned}$$

By optimizing in $\lambda \geq 0$, we find the best one to be $\lambda = \frac{n}{2C}\log\left(1 + \frac{2Ct}{\sigma^2}\right)$ and therefore,

$$P\left(f(B_n) - \mathbb{E}f(B_n) \geq t\right) \leq \exp\left(-\frac{n\sigma^2}{4C^2}h\left(\frac{2Ct}{\sigma^2}\right)\right)$$

where $h(u) = (1+u)\log(1+u) - u$, for all $u \geq 0$. Since $h(u) \geq \frac{u^2}{2(1+\frac{u}{3})}$ for all $u \geq 0$, we obtain

$$P\left(f(B_n) - \mathbb{E}f(B_n) \geq t\right) \leq \exp\left(-\frac{nt^2}{2(\sigma^2 + \frac{2Ct}{3})}\right),$$

which yields the desired result by taking $f = d(\cdot, b^*) \in \mathcal{F}$ and setting t so the right hand side is bounded by δ. ■

Theorem 14 yields the following algorithmic corollary, which is a refinement of Corollary 13.

Corollary 15 *Recall the notation of Corollary 13. Let D be the diameter of the set $\{x_1, \ldots, x_n\}$ and $\sigma^2 = \frac{1}{n}\sum_{i=1}^n d(x_i, b_n)^2$. Then, if*

$$m \geq \frac{16}{3}\max\left(\frac{\sigma^2}{\varepsilon^2}, \frac{D}{\varepsilon}\right)\max(1, \log(1/\delta)), \tag{3}$$

it holds that $d(\tilde{b}_m, b_n) \leq \varepsilon$ with probability at least $1 - \delta$.

In practice, of course, computing σ^2 can be costly. However, one can replace it in (3) by any upper bound, such as $\frac{1}{n}\sum_{i-1}^n d(x_i, b)^2$ for any fixed $b \in M$, or $\frac{1}{n^2}\sum_{1\leq i,j\leq n} d(x_i, x_j)^2$. Computing $\frac{1}{n}\sum_{i=1}^n d(x_i, b)^2$ only requires n computations, but the choice of b can be suboptimal. On the other hand, computing $\frac{1}{n^2}\sum_{1\leq i,j\leq n} d(x_i, x_j)^2$ requires a quadratic (in n) number of computations, but yields a tight bound on σ^2, up to a factor 2. Indeed, one has the following lemma, whose proof is deferred to Section D.3 in the appendix.

Lemma 16 *Let X be a random variable in M with two moments and let Y be an independent copy of M. Let $\sigma^2 = \min_{b\in M}\mathbb{E}[d(X,b)^2]$. Then, $\sigma^2 \leq \mathbb{E}[d(X,Y)^2] \leq 2\sigma^2$.*

Remark 17 *The concentration inequalities given in Theorems 11 and 14 can be extended to the case when the X_i's are independent but not identically distributed: In order to keep our presentation as simple as possible, we state and prove these more general results in Section A in the appendix.*

4. Beyond NPC spaces

In this section, we tackle the question of concentration of barycenters in metric spaces that are not NPC, but that still have a curvature upper bound: CAT(κ) spaces, for $\kappa > 0$. Recall that, intuitively, a metric space is said to be CAT(κ), $\kappa \in \mathbb{R}$, when its triangles are thinner than they would be in the model space of curvature κ (hyperbolic plane if $\kappa < 0$, Euclidean plane if $\kappa = 0$ and 2-dimensional sphere of radius $1/\sqrt{\kappa}$ if $\kappa > 0$): See Bridson and Haefliger (2013); Burago et al. (2022) for an introduction to CAT spaces (see also Sections B and C in the appendix). In this terminology, NPC spaces are CAT(0) spaces. An important example of a CAT(κ) space, which is closely related to optimal transport theory, for $\kappa > 0$, is the class of all p-variate ($p \geq 1$) Gaussian distributions whose covariance matrices have eigenvalues that are at least $\sqrt{3/(2\kappa)}$ (Massart et al., 2019, Proposition 2), equipped with the 2-Wasserstein distance. An open question, which seems important to us, is which more general families of distributions, equipped with the 2-Wasserstein distance, are CAT(κ) for some fixed $\kappa > 0$.

4.1. Barycenters in CAT spaces

It is well known that barycenters may no longer be unique in CAT(κ) spaces when $\kappa > 0$. For instance, on a sphere, any point on the equator is a barycenter of the North and South poles. However, it is known that if a probability measure on a CAT(κ) space, with $\kappa > 0$, is supported within a small enough ball, then it does have a unique barycenter (Yokota, 2016, Theorem B). We provide more useful facts on barycenters in CAT(κ) spaces, for $\kappa > 0$, in Section C in the appendix.

In order to derive a non asymptotic bounds on the convergence of the empirical barycenter in CAT(κ) spaces for $\kappa > 0$, we require the following assumption, which controls the complexity of the metric space (M, d).

Assumption 1 *The metric space (M, d) is CAT(κ) for some $\kappa > 0$ and there are positive constants $A, p > 0$ such that for all $x \in M$ and for all $\alpha, r > 0$ with $\alpha \leq r$, $N(B(x, r), \alpha) \leq \left(\frac{Ar}{\alpha}\right)^p$, where $N(B(x, r), \alpha)$ denotes the smallest integer $N \geq 1$ such that the metric ball $B(x, r)$ can be covered by N balls of radius α.*

This assumption is satisfied, for instance, if M is a Riemannian manifold of dimension p (Chavel, 2006, Section III). More generally, in some way, the parameter p in this assumption plays the role of a dimension of the metric space M.

We can now state the following concentration inequality in the framework of CAT(κ) spaces ($\kappa > 0$).

Theorem 18 *Let Assumption 1 hold. Let $x_0 \in M$ and $\varepsilon \in (0, \frac{\pi}{2\sqrt{\kappa}}]$. Let $X_1, ..., X_n$ be i.i.d random variables such that $d(X_1, x_0) \leq \frac{1}{2}\left(\frac{\pi}{2\sqrt{\kappa}} - \varepsilon\right)$ almost surely. Denote by b^* the population barycenter of X_1 and let $\hat{b}_n$ be the empirical barycenter of $X_1, \ldots, X_n$. Then, for all $\delta \in (0, 1)$, it holds with probability at least $1 - \delta$ that*

$$d(\hat{b}_n, b^*) \leq \frac{c}{\varepsilon\kappa}\left(A\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(2/\delta)}{\varepsilon n}}\right),$$

where $c > 0$ is a universal constant and A is the constant appearing in Assumption 1.

The proof of this theorem is deferred to Section D.4 in the appendix. Note that unlike the case of NPC spaces, this bound depends on the parameter p, which plays the role of a dimension. We do not know whether a dimension free bound can be achieved in CAT(κ) spaces for $\kappa > 0$. However, note that as desired, see discussion above, the dimension parameter p is decoupled from the confidence level δ. A similar bound for the inductive barycenter $\tilde{b}_n$ is unknown, but would be interesting in practice, since $\tilde{b}_n$ is generally easier to compute. We leave this question for future work.

5. Conclusion

To summarize, non-positive curvature of the space has allowed to derive a sensitivity analysis of barycenter functionals in order to obtain sharp versions of Hoeffding's and Bernstein's inequalities, without requiring the space to have finite dimension in any sense. Unlike previous existing finite sample bounds for empirical and inductive barycenters, out setup makes use of the right definition of Laplace transform in metric spaces, due to Ledoux, in order to obtain, using simple arguments, sharp concentration bounds. When the curvature is bounded from above, but can be positive, similar bounds can be obtained, by means of empirical process theory, imposing a finite Hausdorff dimension of the ambient space. However, we expect that in a CAT(κ)-space, for $\kappa > 0$, a careful sensitivity analysis can be made possible for barycenter functionals, allowing to relax the finite dimension assumption. This question is left for future work.

Acknowledgments

This work was funded by the French Agence Nationale de la Recherche, as part of the project ANR-22-ERCS-0015-01.

References

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. Riemannian geometry of grassmann manifolds with a view on algorithmic computation. *Acta Applicandae Mathematica*, 80:199–220, 2004.

Bijan Afsari. Riemannian l^p-center of mass: existence, uniqueness, and convexity. *Proceedings of the American Mathematical Society*, 139(2):655–673, 2011.

Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

A. Ahidar-Coutrix, T. Le Gouic, and Q. Paris. Convergence rates for empirical barycenters in metric spaces: curvature, convexity and extendable geodesics. *Probab. Theory Related Fields*, 177(1-2):323–368, 2020. ISSN 0178-8051. doi: 10.1007/s00440-019-00950-0. URL https://doi.org/10.1007/s00440-019-00950-0.

Stephanie Alexander, Vitali Kapovitch, and Anton Petrunin. *An invitation to Alexandrov geometry*. SpringerBriefs in Mathematics. Springer, Cham, 2019. ISBN 978-3-030-05311-6; 978-3-030-05312-3. doi: 10.1007/978-3-030-05312-3. URL https://doi.org/10.1007/978-3-030-05312-3. CAT(0) spaces.

Jason M Altschuler and Enric Boix-Adsera. Wasserstein barycenters can be computed in polynomial time in fixed dimension. *J. Mach. Learn. Res.*, 22:44–1, 2021.

Jason M Altschuler and Enric Boix-Adsera. Wasserstein barycenters are np-hard to compute. *SIAM Journal on Mathematics of Data Science*, 4(1):179–203, 2022.

Miroslav Bacák. Convex analysis and optimization in hadamard spaces. In *Convex Analysis and Optimization in Hadamard Spaces*. de Gruyter, 2014.

Rajendra Bhatia. Positive definite matrices. In *Positive Definite Matrices*. Princeton university press, 2009.

Rajendra Bhatia and John Holbrook. Riemannian geometry and matrix geometric means. *Linear algebra and its applications*, 413(2-3):594–618, 2006.

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the bures–wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019.

Rabi Bhattacharya and Lizhen Lin. Omnibus clts for Fréchet means and nonparametric inference on non-euclidean spaces. *Proceedings of the American Mathematical Society*, 145(1):413–428, 2017.

Rabi Bhattacharya and Vic Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifolds. *Annals of statistics*, 31(1):1–29, 2003.

Rabi Bhattacharya and Vic Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifolds: II. *Annals of statistics*, pages 1225–1259, 2005.

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.

Martin R Bridson and André Haefliger. *Metric spaces of non-positive curvature*, volume 319. Springer Science & Business Media, 2013.

Dmitri Burago, Yuri Burago, and Sergei Ivanov. *A course in metric geometry*, volume 33. American Mathematical Society, 2022.

Isaac Chavel. *Riemannian geometry: a modern introduction*, volume 98. Cambridge university press, 2006.

Seunghyuk Cho, Juyong Lee, Jaesik Park, and Dongwoo Kim. A rotated hyperbolic wrapped normal distribution for hierarchical representation learning. *Advances in Neural Information Processing Systems*, 35:17831–17843, 2022.

Sebastian Claici, Edward Chien, and Justin Solomon. Stochastic wasserstein barycenters. In *International Conference on Machine Learning*, pages 999–1008. PMLR, 2018.

Christopher Criscitiello and Nicolas Boumal. Negative curvature obstructs acceleration for strongly geodesically convex optimization, even with exact first-order oracles. In *Conference on Learning Theory*, pages 496–542. PMLR, 2022.

Christopher Criscitiello and Nicolas Boumal. Curvature and complexity: Better lower bounds for geodesically convex optimization. In *Conference on Learning Theory*, pages 2969–3013. PMLR, 2023.

Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *International conference on machine learning*, pages 685–693. PMLR, 2014.

Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *International Conference on Machine Learning*, pages 7694–7731. PMLR, 2023.

Alan Edelman, Tomás A Arias, and Steven T Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

Benjamin Eltzner and Stephan F Huckemann. A smeary central limit theorem for manifolds with application to high-dimensional spheres. *The Annals of Statistics*, 47(6):3360–3381, 2019.

Benjamin Eltzner, Fernando Galaz-Garcia, Septhan F Huckemann, and Wilderich Tuschmann. Stability of the cut locus and a central limit theorem for Fréchet means of Riemannian manifolds. *arXiv preprint arXiv:1909.00410*, 2019.

Paul Escande. On the concentration of the minimizers of empirical risks. *arXiv preprint arXiv:2304.00809*, 2023.

Maurice Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. H. Poincaré*, 10:215–310, 1948. ISSN 0365-320X. URL http://www.numdam.org/item?id=AIHP_1948__10_4_215_0.

Kei Funano. Rate of convergence of stochastic processes with values in $\mathbb{R}$-trees and Hadamard manifolds. *Osaka J. Math.*, 47(4):911–920, 2010. ISSN 0030-6126. URL http://projecteuclid.org/euclid.ojm/1292854310.

Reshad Hosseini and Suvrit Sra. Matrix manifold optimization for gaussian mixtures. *Advances in neural information processing systems*, 28, 2015.

Hermann Karcher. Riemannian center of mass and mollifier smoothing. *Communications on pure and applied mathematics*, 30(5):509–541, 1977.

David George Kendall, Dennis Barden, Thomas K Carne, and Huiling Le. *Shape and shape theory*, volume 500. John Wiley & Sons, 2009.

Aryeh Kontorovich. Concentration in unbounded metric spaces and algorithmic stability. In *International Conference on Machine Learning*, pages 28–36. PMLR, 2014.

Alexey Kroshnin, Nazarii Tupitsa, Darina Dvinskikh, Pavel Dvurechensky, Alexander Gasnikov, and Cesar Uribe. On the complexity of approximating wasserstein barycenters. In *International conference on machine learning*, pages 3530–3540. PMLR, 2019.

T. Le Gouic, Q. Paris, P. Rigollet, and A.J. Stromme. Fast convergence of empirical barycenters in Alexandrov spaces and the Wasserstein space. *J. Eur. Math. Soc.*, 2022. doi: 10.4171/jems/1234. URL https://doi.org/10.4171/jems/1234.

Thibaut Le Gouic and Jean-Michel Loubes. Existence and consistency of wasserstein barycenters. *Probability Theory and Related Fields*, 168(3):901–917, 2017.

Michel Ledoux. *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2001. ISBN 0-8218-2864-9. doi: 10.1090/surv/089. URL https://doi.org/10.1090/surv/089.

John M. Lee. *Introduction to smooth manifolds*. Springer, 2012.

John M. Lee. *Introduction to Riemannian manifolds*, volume 2. Springer, 2018.

Yongdo Lim and Miklós Pálfia. Weighted deterministic walks for the least squares mean on hadamard spaces. *Bulletin of the London Mathematical Society*, 46, 05 2014. doi: 10.1112/blms/bdu008.

Asad Lodhia, Keith Levin, and Elizaveta Levina. Matrix means and a novel high-dimensional shrinkage phenomenon. *Bernoulli*, 28(4):2578–2605, 2022.

Simón Lunagómez, Sofia C Olhede, and Patrick J Wolfe. Modeling network populations via graph distances. *Journal of the American Statistical Association*, 116(536):2023–2040, 2021.

Tim Marrinan, J Ross Beveridge, Bruce Draper, Michael Kirby, and Chris Peterson. Finding the subspace mean or median to fit your need. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1082–1089, 2014.

David Martínez-Rubio and Sebastian Pokutta. Accelerated riemannian optimization: Handling constraints with a prox to bound geometric penalties. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 359–393. PMLR, 2023.

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 *Geometric Science of Information: 4th International Conference, GSI 2019, Toulouse, France, August 27–29, 2019, Proceedings*, pages 739–748. Springer, 2019.

Gal Mishne, Zhengchao Wan, Yusu Wang, and Sheng Yang. The numerical stability of hyperbolic representation learning. In *International Conference on Machine Learning*, pages 24925–24949. PMLR, 2023.

Antonio Montanaro, Diego Valsesia, and Enrico Magli. Rethinking the compositionality of point clouds through regularization in the hyperbolic space. *Advances in Neural Information Processing Systems*, 35:33741–33753, 2022.

Shin-ichi Ohta. Convexities of metric spaces. *Geom. Dedicata*, 125:225–250, 2007. ISSN 0046-5755. doi: 10.1007/s10711-007-9159-3. URL https://doi.org/10.1007/s10711-007-9159-3.

Shin-ichi Ohta and Miklós Pálfia. Discrete-time gradient flows and law of large numbers in Alexandrov spaces. *Calc. Var. Partial Differential Equations*, 54(2):1591–1610, 2015. ISSN 0944-2669. doi: 10.1007/s00526-015-0837-y. URL https://doi.org/10.1007/s00526-015-0837-y.

Armin Schwartzman. Lognormal distributions and geometric averages of symmetric positive definite matrices. *International statistical review*, 84(3):456–486, 2016.

Barry Simon. *Loewner's Theorem on Monotone Matrix Functions*. Springer, 2019.

Karl-Theodor Sturm. Probability measures on metric spaces of nonpositive curvature. In *Heat kernels and analysis on manifolds, graphs, and metric spaces (Paris, 2002)*, volume 338 of *Contemp. Math.*, pages 357–390. Amer. Math. Soc., Providence, RI, 2003. doi: 10.1090/conm/338/06080. URL https://doi.org/10.1090/conm/338/06080.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Menglin Yang, Min Zhou, Rex Ying, Yankai Chen, and Irwin King. Hyperbolic representation learning: Revisiting and advancing. In *International Conference on Machine Learning*. PMLR, 2023.

Takumi Yokota. Convex functions and barycenter on CAT(1)-spaces of small radii. *J. Math. Soc. Japan*, 68(3):1297–1323, 2016. ISSN 0025-5645. doi: 10.2969/jmsj/06831297. URL https://doi.org/10.2969/jmsj/06831297.

Takumi Yokota. Convex functions and p-barycenter on $\mathrm{CAT}(1)$-spaces of small radii. *Tsukuba J. Math.*, 41(1):43–80, 2017. ISSN 0387-4982. doi: 10.21099/tkbjm/1506353559. URL https://doi.org/10.21099/tkbjm/1506353559.

Hongyi Zhang and Suvrit Sra. First-order methods for geodesically convex optimization. In *Conference on Learning Theory*, pages 1617–1638. PMLR, 2016.

Hongyi Zhang and Suvrit Sra. Towards riemannian accelerated gradient methods. *arXiv preprint arXiv:1806.02812*, 2018.

Hongyi Zhang, Sashank J Reddi, and Suvrit Sra. Riemannian svrg: Fast stochastic optimization on riemannian manifolds. *Advances in Neural Information Processing Systems*, 29, 2016.

Herbert Ziezold. On expected figures and a strong law of large numbers for random elements in quasi-metric spaces. In *Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes and of the 1974 European Meeting of Statisticians*, pages 591–602. Springer, 1977.

Appendix A. Concentration inequalities for non-identically distributed random variables

In this section, we extend Corollary 11 and Theorem 14 to the case where $X_1, \ldots, X_n$ are independent, but not identically distributed, and share the same barycenter b^*. For instance, if (M, d) is a hyperbolic space and each X_i has a distribution with density (with respect to the Riemannian volume) proportional to $g(\beta_i d(\cdot, b^*))$, where $g : [0, \infty) \to [0, \infty)$ is a fixed function and $\beta_1, \ldots, \beta_n > 0$ are scale parameters, then all X_i's have the same barycenter, namely, b^*.

Theorem 19 *Let (M, d) be an NPC space, and let $X_1, ..., X_n$ be independent random variables with same barycenter $b^* \in M$. Further assume that for all $i = 1, \ldots, n$, $d(X_i, x_i) \le C_i$ almost surely, for some $x_i \in M$ and $C_i > 0$. Let B_n be either the empirical or the inductive barycenter* *of $X_1, \ldots, X_n$. If $B_n = \hat{b}_n$, further assume that M has a curvature lower bound. Then, for all $\delta \in (0,1)$, it holds with probability at least $1-\delta$ that*

$$d(B_n, b^*) \le \frac{\overline{\sigma}_n}{\sqrt{n}} + \overline{C}_n \sqrt{\frac{\log(1/\delta)}{n}},$$

where $\overline{\sigma}_n = \sqrt{\frac{\sigma_1^2+\ldots+\sigma_n^2}{n}}$ and $\overline{C}_n = \sqrt{\frac{C_1^2+\ldots+C_n^2}{n}}$.

Proof The proof is similar to that of Proposition 11, and the main difference is in bounding the bias term $\mathbb{E}[d(B_n, b^*)]$. When B_n is the empirical barycenter, a close inspection of the proof of (Le Gouic et al., 2022, Corollary 11) indicates that the X_i's need not be identically distributed and one readily obtains $\mathbb{E}[d(\hat{b}_n, b^*)] \le \frac{\overline{\sigma}_n}{\sqrt{n}}$. When B_n is the inductive barycenter, we adapt the proof of (Sturm, 2003, Theorem 4.7) and obtain the following lemma. ■

Lemma 20 *Let $X_1, \ldots, X_n$ be independent random points with two moments in an NPC space (M, d) and with same barycenter b^*. Then,*

$$\mathbb{E}[d(S_n, b^*)^2] \le \frac{\sigma_1^2 + \ldots + \sigma_n^2}{n^2},$$

where $\sigma^2 = \mathbb{E}[d(X_i, b^*)^2]$ is the variance of X_i, for $i = 1, \ldots, n$.

Proof The proof of this lemma follows the same lines as the proof of (Sturm, 2003, Theorem 4.7) and proceeds by induction on n. Denote by $V_n = \mathbb{E}[d(S_n, b^*)^2]$ and by $F_n(x) = \mathbb{E}[d(X_n, x)^2]$, for all $x \in M$. Then, by the 2-geodesic strong convexity of the squared distance to any given point, one obtains

$$\begin{aligned} V_n &\le \mathbb{E}\left[\left(1 - \frac{1}{n}\right) d(S_{n-1}, b^*)^2 + \frac{1}{n} d(X_n, b^*)^2 - \frac{n-1}{n^2} d(S_{n-1}, X_n)^2\right] \\ &= \frac{n-1}{n} V_{n-1} + \frac{\sigma_n^2}{n} - \frac{n-1}{n^2} \mathbb{E}[F_n(S_{n-1})], \end{aligned}$$

where we use the fact that X_n and S_{n-1} are independent, by construction of S_{n-1} (which only depends on $X_1, \ldots, X_{n-1}$). Now, again by the 2-geodesic strong convexity of the squared distance to any given point, we obtain that F_n is also 2-geodescailly strongly convex, yielding $F(S_n) \ge F(b^*) + d(S_n, b^*)^2$ almost surely, since b^* is the minimizer of F_n by assumption. Therefore, it follows that

$$\begin{aligned} V_n &\le \frac{n-1}{n} V_{n-1} + \frac{\sigma_n^2}{n} - \frac{n-1}{n^2} \sigma_n^2 - \frac{n-1}{n^2} V_{n-1} \\ &= \frac{(n-1)^2}{n^2} V_{n-1} + \frac{\sigma_n^2}{n^2}. \end{aligned}$$

In other words, $n^2 V_n \le (n-1)^2 V_{n-1} + \sigma_n^2$. Since, by definition of S_1, $V_1 = \sigma_1^2$, the result follows by induction. ■

Finally, we also have the following generalization of Theorem 14 when the X_i's are not identically distributed.

Theorem 21 *Let (M, d) be an NPC space and $X_1, \ldots, X_n$ be independent random variables in (M, d). Assume that $d(X_i, x_0) \le C$ almost surely, for all $i = 1, \ldots, n$, for some fixed $x_0 \in M$ and $C > 0$. let B_n be either the empirical or the inductive barycenter of $X_1, \ldots, X_n$. Then, for all $\delta \in (0, 1)$, it holds with probability at least $1 - \delta$ that*

$$d(B_n, b^*) \le \frac{\overline{\sigma}_n}{\sqrt{n}} + \min\left(2\overline{\sigma}_n\sqrt{\frac{\log(1/\delta)}{n}}, \frac{8C\log(1/\delta)}{3n}\right),$$

where $\overline{\sigma}_n$ is as in Theorem 19.

In Theorem 19, when $X_1, \ldots, X_n$ do not even share the same barycenter, we still have the following fact, for any fixed $b \in M$. For any $\delta \in (0, 1)$, it holds with probability at least $1 - \delta$ that

$$d(B_n, b) \le \mathbb{E}[d(B_n, b)] + \overline{C}_n\sqrt{\frac{\log(1/\delta)}{n}}.$$

However, it is not clear what b to choose in order to make the first term as small as possible. If $b_1, \ldots, b_n$ are the respective population barycenters of $X_1, \ldots, X_n$ (i.e., each b_i minimizes $\mathbb{E}[d(X_i, b)^2]$ over $b \in M$), a natural candidate for b would be the barycenter of $b_1, \ldots, b_n$.

Open question *Let $X_1, \ldots, X_n$ be independent random variables in M, with two moments. For each $i = 1, \ldots, n$, let b_i be the barycenter of X_i and $\sigma_i^2 = \mathbb{E}[d(X_i, b_i)^2]$ its variance. Is it true that*

$$\mathbb{E}[d(\hat{b}_n, b_n^*)^2] \le \frac{\overline{\sigma}_n^2}{n},$$

where b_n^* is the barycenter of $b_1, \ldots, b_n$ and $\overline{\sigma}_n^2 = \frac{\sigma_1^2 + \ldots + \sigma_n^2}{n}$, as in Theorem 19?

Appendix B. Background on CAT spaces

In this section, we give the precise definition of CAT spaces. We refer the reader to the book Alexander et al. (2019) for a complete view on the topic. First, we introduce a family of model spaces that will allow us to define local and global curvature bounds in the sequel. Let $\kappa \in \mathbb{R}$.

$\kappa = 0$: Euclidean plane Set $M_0 = \mathbb{R}^2$, equipped with its Euclidean metric. This model space corresponds to zero curvature, is a geodesic space where geodesics are unique and given by line segments.

$\kappa > 0$: Sphere Set $M_\kappa = \frac{1}{\sqrt{\kappa}}\mathbb{S}^2$: This is the 2-dimensional Euclidean sphere, embedded in $\mathbb{R}^3$, with center 0 and radius $1/\sqrt{\kappa}$, equipped with the arc length metric: $d_\kappa(x, y) = \frac{1}{\sqrt{\kappa}}\arccos(\kappa x^\top y)$, for all $x, y \in M_\kappa$. This is a geodesic space where the geodesics are unique except for antipodal points, and given by arcs of great circles. Here, a great circle is the intersection of the sphere with any plane going through the origin.

$\kappa < 0$**: Hyperbolic space** Set $M_\kappa = \frac{1}{\sqrt{\kappa}}\mathbb{H}^2$, where $\mathbb{H}^2 = \{(x_1, x_2, x_3) \in \mathbb{R}^3 : x_3 > 0, x_1^2 + x_2^2 - x_3^2 = -1\}$. The metric is given by $d_\kappa = \frac{1}{\sqrt{-\kappa}}\text{arccosh}(-\kappa\langle x, y\rangle)$, for all $x, y \in M_\kappa$, where $\langle x, y\rangle = x_1y_1 + x_2y_2 - x_3y_3$. This is a geodesic space where geodesics are always unique and are given by arcs of the intersections of M_κ with planes going through the origin.

For $\kappa \in \mathbb{R}$, let D_κ be the diameter of the model space M_κ, i.e., $D_\kappa = \begin{cases} \infty \text{ if } \kappa \le 0 \\ \frac{\pi}{\sqrt{\kappa}} \text{ if } \kappa > 0 \end{cases}$.

Let (M, d) be a geodesic space, i.e., a metric space where any two points have at least one geodesic between them. The notion of curvature (lower or upper) bounds for (M, d) is defined by comparing the triangles in M with triangles with the same side lengths in model spaces.

Definition 22 *A (geodesic) triangle in M is a set of three points in M (the vertices) together with three geodesic segments connecting them (the sides).*

Given three points $x, y, z \in S$, we abusively denote by $\Delta(x, y, z)$ a triangle with vertices x, y, z, with no mention to which geodesic segments are chosen for the sides (geodesics between points are not necessarily unique, as seen for example on a sphere, between any two antipodal points). The perimeter of a triangle $\Delta = \Delta(x, y, z)$ is defined as $\text{per}(\Delta) = d(x, y) + d(y, z) + d(x, z)$. It does not depend on the choice of the sides.

Definition 23 *Let $\kappa \in \mathbb{R}$ and Δ be a triangle in M with $\text{per}(\Delta) < 2D_\kappa$. A comparison triangle for Δ in the model space M_κ is a triangle $\bar{\Delta} \subseteq M_\kappa$ with same side lengths as Δ, i.e., if $\Delta = \Delta(x, y, z)$, then $\bar{\Delta} = \Delta(\bar{x}, \bar{y}, \bar{z})$ where $\bar{x}, \bar{y}, \bar{z}$ are any points in M_κ satisfying*

$$\begin{cases} d(x, y) = d_\kappa(\bar{x}, \bar{y}) \\ d(y, z) = d_\kappa(\bar{y}, \bar{z}) \\ d(x, z) = d_\kappa(\bar{x}, \bar{z}). \end{cases}$$

Note that a comparison triangle in M_κ is always unique up to rigid motions. We are now ready to define curvature bounds. Intuitively, we say that (M, d) has global curvature bounded from above (resp. below) by κ if all its triangles (with perimeter smaller than $2D_\kappa$) are thinner (resp. fatter) than their comparison triangles in the model space M_κ.

Definition 24 *Let $\kappa \in \mathbb{R}$. We say that (M, d) has global curvature bounded from above (resp. below) by κ if and only if for all triangles $\Delta \subseteq M$ with $\text{per}(\Delta) < 2D_\kappa$ and for all $x, y \in \Delta$, $d(x, y) \le d_\kappa(\bar{x}, \bar{y})$ (resp. $d(x, y) \ge d_\kappa(\bar{x}, \bar{y})$), where $\bar{x}$ and $\bar{y}$ are the points on a comparison triangle $\bar{\Delta}$ in M_κ that correspond to x and y. We then call (M, d) a CAT(κ) (resp. CAT$^+(\kappa)$) space.*

Appendix C. Strong convexity and barycenters in CAT spaces

The following strong convexity property of the squared distance to any fixed point can be found in (Ohta, 2007, Proposition 3.1).

Lemma 25 *Let (M, d) be a CAT(κ) space, with $\kappa > 0$ and let $x_0 \in M$. Then, for all $\varepsilon \in (0, \frac{\pi}{2\sqrt{\kappa}}]$, $\frac{1}{2}d(x_0, \cdot)^2$ is k_ε-geodesically strongly convex on the ball $B\left(x_0, \frac{1}{2}\left(\frac{\pi}{2\sqrt{\kappa}} - \varepsilon\right)\right)$, with $k_\varepsilon = (\pi - 2\sqrt{\kappa}\varepsilon)\tan(\varepsilon\sqrt{\kappa})$.*

This strong convexity property implies a variance inequality in CAT(κ) spaces of small diameter for $\kappa > 0$.

Lemma 26 *Let (M, d) be a CAT(κ) space with $\kappa > 0$, let $x_0 \in M$ and let $\varepsilon \in (0, \frac{\pi}{2\sqrt{\kappa}}]$. Let X be a random variable in M with values in the ball $B\left(x_0, \frac{1}{2}\left(\frac{\pi}{2\sqrt{\kappa}} - \varepsilon\right)\right)$ almost surely. Then it satisfies the following variance inequality*

$$d^2(z, b^*) \leq \frac{2}{k_\varepsilon}\mathbb{E}\left[d^2(z, X) - d^2(b^*, X)\right], \quad \forall z \in M, \tag{4}$$

where $k_\varepsilon = (\pi - 2\sqrt{\kappa}\varepsilon)\tan(\varepsilon\sqrt{\kappa})$, and b^* denotes the population barycenter of X.

Proof

By (Ohta, 2007, Proposition 3.1), we know that for all $p \in M$, the functions $z \mapsto d^2(z, p)$ are k_ε-convex on the ball $B\left(x_0, \frac{\pi}{2\sqrt{\kappa}} - \varepsilon\right)$. Moreover, b^* is in the ball by (Yokota, 2016, Theorem B). It follows that the function $z \mapsto \mathbb{E}\left[d^2(z, X) - d^2(b^*, X)\right]$ is also k_ε-convex on the ball. Therefore, by taking z_t the joining geodesic between b^* and z (it exists and is unique because the ball is small enough), and for $p = b^*$, by definition of the barycenter, we get that for $t \in [0, 1]$,

$$\begin{aligned} 0 &\leq \mathbb{E}\left[d^2(z_t, X) - d^2(b^*, X)\right] \\ &\leq t\,\mathbb{E}\left[d^2(z, X) - d^2(b^*, X)\right] - \frac{k_\varepsilon}{2}t(1-t)d^2(z, b^*) \end{aligned}$$

Therefore,

$$t\,\mathbb{E}\left[d^2(z, X) - d^2(b^*, X)\right] \geq \frac{k_\varepsilon}{2}t(1-t)d^2(z, b^*),$$

yielding the result by dividing by t and letting t goes to zero.

■

Note that $k_\varepsilon \in (0, 2)$, for all $\varepsilon \in (0, \frac{\pi}{2\sqrt{\kappa}})$. Moreover k_ε tends to zero when ε tends to zero, and it tends to 2 when ε tends to $\frac{\pi}{2\sqrt{\kappa}}$. So when the diameter tends to zero, the convexity tends to be the same as in the Euclidean case, whereas when the diameter tends to be maximal (i.e. $\frac{\pi}{2\sqrt{\kappa}}$), the geodesic strong convexity may no longer hold.

Appendix D. Proofs

D.1. Proof of Lemma 8

In order to prove this lemma, we need the following result. The key argument in its proof is that the lower bound on the Ricci curvature allows to control the volume of arbitrary large balls, by comparison theorems.

Lemma 27 *Let M be a p-dimensional Riemannian manifold with Ricci curvature bounded from below by $(p-1)\kappa \in \mathbb{R}$, where $\kappa \leq 0$. Then, for all $x_0 \in M$ and for all $\alpha > 0$,*

$$\int_M e^{-\alpha d(x, x_0)^2}\,\mathrm{dVol}(x) \leq \begin{cases} \frac{c_{p-1}}{(p-1)\alpha^{(p-1)/2}} J_p \text{ if } \kappa = 0 \\ \frac{5c_{p-1}e^{(p-1)\sqrt{-\kappa/\alpha} - \frac{\kappa(p-1)^2}{\alpha}}}{(p-1)(-\kappa)^{p/2}} \text{ otherwise} \end{cases}$$

where $c_{p-1} = \frac{2\pi^{p/2}}{\Gamma(p/2)}$ and $J_p = \sum_{r=0}^{\infty}(r+1)^p e^{-r^2}$.

Here, Vol stands for the Riemannian volume.

Proof By Bishop-Gromov volume comparison theorem (Lee, 2018, Theorem 11.19), it holds that for all $r \geq 0$,

$$\mathrm{V}ol(B(x_0, r)) \leq V_{p,\kappa}(r),$$

where $V_{p,\kappa}(r)$ is the volume of any ball of radius r in the p-dimensional hyperbolic space of constant curvature κ (which we identify with $\mathbb{R}^p$ is $\kappa = 0$). It is known (Chavel, 2006, Section III) that

$$V_{p,\kappa}(r) = c_{p-1} \int_0^r \left(\frac{\sinh(\sqrt{-\kappa}t)}{\sqrt{-\kappa}} \right)^{p-1} \mathrm{d}t$$

where $c_{p-1} = \frac{2\pi^{p/2}}{\Gamma(p/2)}$ and where the integral should be understood as r^p/p if $\kappa = 0$. If $\kappa < 0$, we readily obtain the inequality

$$V_{p,\kappa}(r) = \frac{c_{p-1} e^{(p-1)\sqrt{-\kappa}r}}{(p-1)(-\kappa)^{p/2}}.$$

Now, we write that, for any choice of $c > 0$,

$$\begin{aligned} I(\alpha) := \int_M e^{-\alpha d(x,x_0)^2} \, \mathrm{dVol}(x) &= \sum_{r=0}^{\infty} \int_{B(x_0,c(r+1)) \setminus B(x_0,cr)} e^{-\alpha d(x,x_0)^2} \, \mathrm{dVol}(x) \\ &\leq \sum_{r=0}^{\infty} e^{-\alpha c^2 r^2} V_{p,\kappa}(c(r+1)) \end{aligned} \tag{5}$$

For simplicity, let us distinguish the two cases when $\kappa = 0$ or $\kappa < 0$. First, assume $\kappa = 0$. Then, (5) with $c = 1/\sqrt{\alpha}$ yields

$$I(\alpha) \leq \frac{c_{p-1}}{(p-1)\alpha^{(p-1)/2}} \sum_{r=0}^{\infty} e^{-r^2} (r+1)^p.$$

Now, let us assume that $\kappa < 0$. Then, (5) with $c = 1/\sqrt{\alpha}$ again yields

$$\begin{aligned} I(\alpha) &\leq \frac{c_{p-1}}{(p-1)(-\kappa)^{p/2}} \sum_{r=0}^{\infty} e^{-r^2} e^{(p-1)\sqrt{-\kappa}(r+1)/\sqrt{\alpha}} \\ &= \frac{c_{p-1} e^{(p-1)\sqrt{-\kappa}+\sqrt{\alpha}}}{(p-1)(-\kappa)^{p/2}} \sum_{r=0}^{\infty} e^{-r^2} e^{(p-1)\sqrt{-\kappa}r/\sqrt{\alpha}} \\ &= \frac{c_{p-1} e^{(p-1)\sqrt{-\kappa/\alpha} - \frac{\kappa(p-1)^2}{\alpha}}}{(p-1)(-\kappa)^{p/2}} \sum_{r=0}^{\infty} e^{-\left(r - \frac{(p-1)\sqrt{-\kappa}}{2\sqrt{\alpha}}\right)^2}. \end{aligned}$$

Now, using the inequality $\sum_{r=0}^{\infty} e^{-(r-m)^2} \leq 5$, for any $m > 0$, we obtain that

$$I(\alpha) \leq \frac{5c_{p-1} e^{(p-1)\sqrt{-\kappa/\alpha} - \frac{\kappa(p-1)^2}{\alpha}}}{(p-1)(-\kappa)^{p/2}}.$$

■

Proof of Lemma 8 If $\kappa > 0$, then by (Lee, 2018, Corollary 11.18), M has finite diameter, bounded from above by $D = \pi/\sqrt{\kappa}$. Hence, X is bounded and, by Lemma 7, it is K^2-sub-Gaussian, with $K^2 = 4\pi^2/\kappa$.

Now, assume that $\kappa \leq 0$ and let $f \in \mathcal{F}$. By Jensen's inequality, for all $K > 0$, $\mathbb{E}[e^{\frac{(f(X)-\mathbb{E}[f(X)])^2}{2K^2}}] \leq \mathbb{E}[e^{\frac{(f(X)-f(Y))^2}{2K^2}}]$, where Y is independent of X and has the same distribution. Therefore,

$$\mathbb{E}[e^{\frac{(f(X)-\mathbb{E}[f(X)])^2}{2K^2}}] \leq \mathbb{E}[e^{\frac{d(X,Y)^2}{2K^2}}] \leq \mathbb{E}\left[e^{\frac{d(X,x_0)^2+d(Y,x_0)^2}{K^2}}\right] = \mathbb{E}\left[e^{\frac{d(X,x_0)^2}{K^2}}\right]^2$$
$$\leq \left(C\int_M e^{-\left(\beta-\frac{1}{K^2}\right)d(x,x_0)^2}\,\mathrm{dVol}(x)\right)^2.$$

Lemma 27 then yields the result, by taking K large enough. $\square$

D.2. Proof of Proposition 9

The proof for the inductive barycenter function $\tilde{B}_n$ follows from a simple induction and it can be found in (Funano, 2010, Lemma 3.1).

The 1-Lipschitz property of $\hat{B}_n$ follows from two different arguments, which we both give here, because they are both instructive. Let $\mathcal{P}^1(M)$ be the set of all probability measures on (M,d) with finite first moment. For $\mu \in \mathcal{P}^1(M)$, let $B(\mu)$ be its barycenter, i.e., the (unique) minimizer $b \in M$ of $\mathbb{E}[d(X,b)^2 - d(X,b_0)^2]$, where $X \sim \mu$ and $b_0 \in M$ is arbitrarily fixed.

The first argument follows from (Lim and Pálfia, 2014, Theorem 3.4), which provides a deterministic connection between barycenters and their inductive versions. Let $(x_1,\ldots,x_n)$ and $(y_1,\ldots,y_n)$ be two n-uples in M. We extend these n-uples into periodic infinite sequences by setting, for all positive integers k, $x_k = x_{r_k}$ and $y_k = y_{r_k}$, where r_k is the unique integer between 1 and n such that $k - r_k$ is a multiple of n. Then, (Lim and Pálfia, 2014, Theorem 3.4) indicates that $\tilde{B}_k(x_1,\ldots,x_k)$ and $\tilde{B}_k(y_1,\ldots,y_k)$ converge to $\hat{B}_n(x_1,\ldots,x_n)$ and $\hat{B}_n(y_1,\ldots,y_n)$ respectively, as $k \to \infty$. Moreover, thanks to the $(1/(qn))$-Lipschitz feature of $\tilde{B}_{qn}$ proved above, one has, for all positive integers q,

$$d(\tilde{B}_{qn}(x_1,\ldots,x_{qn}),\tilde{B}_{qn}(y_1,\ldots,y_{qn}))$$
$$\leq \frac{1}{qn}\sum_{k=1}^{qn} d(x_k,y_k) = \frac{1}{n}\sum_{k=1}^{n} d(x_k,y_k).$$

Taking the limit as $q \to \infty$ yields the $(1/n)$-Lipschitz property of $\hat{B}_n$.

The second argument uses Jensen's inequality, which implies that the barycenter functional B is contractive over $\mathcal{P}^1(M)$, equipped with the Wasserstein distance W_1 (Sturm, 2003, Theorem 6.3). More precisely, for all probability measures $\mu,\nu \in \mathcal{P}^1(M)$, it holds that $d(B(\mu),B(\nu)) \leq W_1(\mu,\nu)$, where $W_1(\mu,\nu) = \inf_{X\sim\mu,Y\sim\nu}\mathbb{E}[d(X,Y)]$. Now, fix two n-uples $(x_1,\ldots,x_n)$ and $(y_1,\ldots,y_n)$ in M^n and set $\mu = n^{-1}\sum_{i=1}^n \delta_{x_i}$ and $\nu = n^{-1}\sum_{i=1}^n \delta_{y_i}$. It is clear that $B(\mu) = \hat{B}_n(x_1,\ldots,x_n)$ and $B(\nu) = \hat{B}_n(y_1,\ldots,y_n)$. Moreover, $W_1(\mu,\nu) \leq \frac{1}{n}(d(x_1,y_1)+\ldots+d(x_n,y_n))$, which can be seen by taking the coupling (X,Y) of μ and ν such that $P(X=a_i,Y=b_i) = \frac{1}{n}, i=1,\ldots,n$.

D.3. Proof of Lemma 16

The right hand side is obvious thanks to the triangle inequality (and it does not require independence of X and Y). For the left hand side, denote by $F(b) = \mathbb{E}[d(X,b)^2]$, for all $b \in M$. Then, $\sigma^2 = \min_{b\in M} F(b)$ and independence of X and Y yields that $\sigma^2 \leq F(Y) = \mathbb{E}[d(X,Y)^2|Y]$ almost surely. The result follows by taking the expectation on both sides.

D.4. Proof of Theorem 18

The proof is based on an application of (Ahidar-Coutrix et al., 2020, Theorem 2.1). It is possible to use this theorem thanks to Assumption 1 and Lemma 26. By a careful analysis of its proof with the constants $D = p$, $C = A$, $K_1 = \sqrt{\frac{\pi}{2\sqrt{\kappa}} - \varepsilon}$, $K_2 = 2\left(\frac{\pi}{2\sqrt{\kappa}} - \varepsilon\right)$, $K_3 = \frac{2}{(\pi - 2\sqrt{\kappa}\varepsilon)\tan(\varepsilon\sqrt{\kappa})}$, and $\alpha = \beta = 1$, we obtain that for all $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\sqrt{\frac{k_\varepsilon}{2}}d(\hat{b}_n, b^*) \leq 3c_1\sqrt{\frac{p}{n}} + 3c_2\sqrt{\frac{\log(2/\delta)}{n}},$$

where

$$k_\varepsilon = (\pi - 2\sqrt{\kappa}\varepsilon)\tan(\varepsilon\sqrt{\kappa}),$$

$$c_1 = \frac{96\sqrt{2A}\left(\frac{\pi}{2\sqrt{\kappa}} - \varepsilon\right)}{\sqrt{(\pi - 2\sqrt{\kappa}\varepsilon)\tan(\varepsilon\sqrt{\kappa})}} = \frac{96\sqrt{A}\sqrt{\frac{\pi}{2\sqrt{\kappa}} - \varepsilon}}{\sqrt{2\kappa\tan(\varepsilon\sqrt{\kappa})}}$$

and

$$c_2 = \frac{4\left(\frac{\pi}{2\sqrt{\kappa}} - \varepsilon\right)}{\sqrt{(\pi - 2\sqrt{\kappa}\varepsilon)\tan(\varepsilon\sqrt{\kappa})}} + \frac{16}{3}\sqrt{\left(\frac{\pi}{2\sqrt{\kappa}} - \varepsilon\right)} = \sqrt{\frac{\pi - 2\varepsilon\sqrt{\kappa}}{\kappa}}\left(\frac{2}{\sqrt{\tan(\varepsilon\sqrt{\kappa})}} + \frac{16}{3\sqrt{2}}\right).$$

Hence, with probability at least $1-\delta$, one has

$$d(\hat{b}_n, b^*) \leq \frac{288\sqrt{A}}{\sqrt{\kappa}\tan(\varepsilon\sqrt{\kappa})}\sqrt{\frac{p}{n}} + \frac{1}{\sqrt{\kappa\tan(\varepsilon\sqrt{\kappa})}}\left(\frac{6\sqrt{2}}{\sqrt{\tan(\varepsilon\sqrt{\kappa})}} + 16\right)\sqrt{\frac{\log(2/\delta)}{n}},$$

which proves the result by noting that $\tan x \geq x$ for all $x \in [0, \pi/2)$.