On the rate of convergence of empirical barycentres in metric spaces: curvature, convexity and extendible geodesics

Adil Ahidar-Coutrix, Thibaut Le Gouic, Quentin Paris

Convergence rates for empirical barycenters in metric spaces: curvature, convexity and extendable geodesics

A. Ahidar-Coutrix[*], T. Le Gouic[†] and Q. Paris[‡]

May 30, 2019

Abstract

This paper provides rates of convergence for empirical (generalised) barycenters on compact geodesic metric spaces under general conditions using empirical processes techniques. Our main assumption is termed a variance inequality and provides a strong connection between usual assumptions in the field of empirical processes and central concepts of metric geometry. We study the validity of variance inequalities in spaces of non-positive and non-negative Aleksandrov curvature. In this last scenario, we show that variance inequalities hold provided geodesics, emanating from a barycenter, can be extended by a constant factor. We also relate variance inequalities to strong geodesic convexity. While not restricted to this setting, our results are largely discussed in the context of the 2-Wasserstein space.

[*]Aix Marseille Univ., CNRS, Centrale Marseille, I2M, Marseille, France. Email:adil.ahidar@outlook.com

[†]Aix Marseille Univ., CNRS, Centrale Marseille, I2M, Marseille, France & National Research University Higher School of Economics, Moscow, Russia. This work has been funded by the Russian Academic Excellence Project '5-100'. Email:thibaut.le_gouic@math.cnrs.fr

[‡]National Research University Higher School of Economics, Moscow, Russia. This work has been funded by the Russian Academic Excellence Project '5-100'. Email:qparis@hse.ru

1 Introduction

Given a separable and complete metric space (M, d), define $\mathcal{P}_2(M)$ as the set of Borel probability measures P on M such that

$$\int_M d(x,y)^2 \,\mathrm{d}P(y) < +\infty,$$

for all $x \in M$. A barycenter of $P \in \mathcal{P}_2(M)$, also called a Fréchet mean [27], is any element $x^* \in M$ such that

$$x^* \in \operatorname*{arg\,min}_{x \in M} \int_M d(x,y)^2 \,\mathrm{d}P(y). \tag{1.1}$$

When it exists, a barycenter stands as a natural analog of the mean of a (square integrable) probability measure on $\mathbb{R}^d$. Alternative notions of mean value include local minimisers [31], p-means [64], exponential barycenters [25] or convex means [25]. Extending the notion of mean value to the case of probability measures on spaces M with no Euclidean (or Hilbert) structure has a number of applications ranging from geometry [51] and optimal transport [59, 60, 49, 23] to statistics and data science [46, 17, 14, 36], and the context of abstract metric spaces provides a unifying framework encompassing many non-standard settings.

Properties of barycenters, such as existence and uniqueness, happen to be closely related to geometric characteristics of the space M. These properties are addressed in the context of Riemannian manifolds in [1]. Many interesting examples of metric spaces, however, cannot be described as smooth manifolds because of their singularities or infinite dimensional nature. More general geometrical structures are geodesic metric spaces which include many more examples of interest (precise definitions and necessary background on metric geometry are reported in Appendix A). The barycenter problem has been addressed in this general setting. The scenario where M has non-positive curvature (from here on, curvature bounds are understood in the sense of Aleksandrov) is considered in [51]. More generally, the case of metric spaces with upper bounded curvature is studied in [63] and [64]. The context of spaces M with lower bounded curvature is discussed in [62] and [45].

Focus on the case of metric spaces with non-negative curvature may be motivated by the increasing interest for the theory of optimal transport and its applications. Indeed, a space of central importance in this context is the Wasserstein space $M = \mathcal{P}_2(\mathbb{R}^d)$, equipped with the Wasserstein metric W_2, known to be geodesic and with non-negative curvature (see Section 7.3 in [6]). In this framework, the barycenter problem was first studied by [2] and has since gained considerable momentum. Existence and uniqueness of barycenters in $\mathcal{P}_2(\mathbb{R}^d)$ has further been studied in [38].

A number of objects of interest, including barycenters as a special case, may be described as minimisers of the form

$$x^* \in \operatorname*{arg\,min}_{x \in M} \int_M F(x,y) \,\mathrm{d}P(y), \tag{1.2}$$

for some probability measure P on metric space M and some functional $F : M \times M \to \mathbb{R}$. While we obviously recover the definition of barycenters whenever $F(x,y) = d(x,y)^2$, many functionals of interest are not of this specific form. With a slight abuse of language, minimisers such as x^* will be called generalised barycenters in the sequel. A first example we have in mind, in the context where $M = \mathcal{P}_2(\mathbb{R}^d)$, is the case where functional F is an f-divergence, i.e.

$$F(\mu,\nu) := \begin{cases} \int f\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\nu & \text{if } \mu \ll \nu, \\ +\infty & \text{otherwise,} \end{cases}$$

for some convex function $f : \mathbb{R}_+ \to \mathbb{R}$. Known for their importance in statistics [37, 55], and information theory [56], f-divergences have become a crucial tool in a number of other fields such as geometry and optimal transport [52, 53, 39] or machine learning [29]. Other examples arise when the squared distance $d(x,y)^2$ in (1.1) is replaced by a regularised version $F(x,y)$ aiming at enforcing computationally friendly properties, such as convexity, while providing at the same time a sound approximation of

$d(x,y)^2$. A significant example in this spirit is the case where functional F is the entropy-regularised Wasserstein distance (also known as the Sinkhorn divergence) largely used as a proxy for W_2 in applications [22, 23, 4, 24].

In the paper, our main concern is to provide rates of convergence for empirical generalised barycenters, defined as follows. Given a collection $Y_1, \dots, Y_n$ of independent and M-valued random variables with same distribution P, we call empirical generalised barycenter any

$$x_n \in \arg\min_{x \in M} \frac{1}{n} \sum_{i=1}^{n} F(x, Y_i). \tag{1.3}$$

Any such x_n provides a natural empirical counterpart of a generalised barycenter x^* defined in (1.2). The statistical properties of x_n have been studied in a few specific scenarios. In the case where $F(x,y) = d(x,y)^2$ and M is a Riemannian manifold, significant contributions, establishing in particular consistency and limit distribution under general conditions, are [12, 13] and [32]. Asymptotic properties of empirical barycenters in the Wasserstein space are studied in [38]. We are only aware of a few contributions providing finite sample bounds on the statistical performance of x_n. Paper [14] provides upper and lower bounds on convergence rates for empirical barycenters in the context of the Wasserstein space over the real line. Independently of the present contribution, [50] studies a similar problem and provides results complementary to ours.

In addition to more transparent conditions, our results are based on the fundamental assumption that there exists constants $K > 0$ and $\beta \in (0, 1]$ such that, for all $x \in M$,

$$d(x, x^*)^2 \leq K \left(\int_M (F(x,y) - F(x^*, y)) \, \mathrm{d}P(y) \right)^{\beta}. \tag{1.4}$$

We show that condition (1.4) provides a connection between usual assumptions in the field of empirical processes and geometric characteristics of the metric space M. First, the reader familiar with the theory of empirical processes will identify in the proof of Theorems 2.1 and 2.5 that condition (1.4) implies a Bernstein condition on the class of functions indexing our empirical process, that is an equivalence between their L_2 and L_1 norms. Many authors have emphasised the role of this condition for obtaining fast rates of convergence of empirical minimisers. Major contributions in that direction are for instance [41, 42, 15, 9, 10, 34, 11] and [43]. In particular, this assumption may be understood in our context as an analog of the Mammen-Tsybakov low-noise assumption [41] used in binary classification. Second, we show that condition (1.4) carries a strong geometrical meaning. In the context where $F(x,y) = d(x,y)^2$, [51] established a tight connection between (1.4), with $K = \beta = 1$, and the fact that M has non-positive curvature. When $F(x,y) = d(x,y)^2$, we show that (1.4) actually holds with $K > 0$ and $\beta = 1$ in geodesic spaces of non-negative curvature under flexible conditions related to the possibility of extending geodesics emanating from a barycenter. Finally, for a general functional F, we connect (1.4) to its strong convexity properties. Using terminology introduced in [51] in a slightly more specific context, we will call by extention (1.4) a variance inequality.

The paper is organised has follows. Section 2 provides convergence rates for generalised empirical barycenters under several assumptions of functional F and two possible complexity assumptions on metric space M. Section 3 investigates in details the validity of the variance inequality (1.4) in different scenarios. In particular, we focus on studying (1.4) under curvature bounds of the metric M whenever $F(x,y) = d(x,y)^2$. Additional examples where our results apply are discussed in Section 4. Proofs are postponed to Section 5. Finally, Appendix A presents an overview of basic concepts and results in metric geometry for convenience.

2 Rates of convergence

In this section, we provide convergence rates for generalised empirical barycenters. Paragraph 2.1 defines our general setup and mentions our main assumptions on functional F. Paragraphs 2.2 and

2.3 present rates under different assumptions on the complexity of metric space M. Subsection 2.4 discusses the optimality of our results.

2.1 Setup

Let (M, d) be a separable and complete metric space and $F : M \times M \to \mathbb{R}$ a measurable function. Let P be a Borel probability measure on M. Suppose that, for all $x \in M$, the function $y \in M \mapsto F(x, y)$ is integrable with respect to P, and let

$$x^* \in \operatorname*{arg\,min}_{x \in M} \int_M F(x, y) \, \mathrm{d}P(y), \tag{2.1}$$

which we suppose exists. Given a collection $Y_1, \dots, Y_n$ of independent and M-valued random variables with same distribution P, we consider an empirical minimiser

$$x_n \in \operatorname*{arg\,min}_{x \in M} \frac{1}{n} \sum_{i=1}^{n} F(x, Y_i). \tag{2.2}$$

The present section studies the statistical performance of x_n under the following assumptions on F.

(A1) There exists a constant $K_1 > 0$ such that, for all $x, y \in M$,

$$|F(x, y)| \leq K_1.$$

(A2) There exist constants $K_2 > 0$ and $\alpha \in (0, 1]$ such that, for all $x, x', y \in M$,

$$|F(x, y) - F(x', y)| \leq K_2 d(x, x')^\alpha.$$

(A3) (Variance inequality) There exist constants $K_3 > 0$ and $\beta \in (0, 1]$ such that, for all $x \in M$,

$$d(x, x^*)^2 \leq K_3 \left(\int_M (F(x, y) - F(x^*, y)) \, \mathrm{d}P(y) \right)^\beta.$$

Assumptions (A1) and (A2) are transparent boundedness and regularity conditions. For instance if $F(x, y) = d(x, y)^2$, these assumptions are satisfied whenever M is bounded with $K_1 = \mathrm{diam}(M)^2$, $K_2 = 2\mathrm{diam}(M)$ and $\alpha = 1$, by the triangular inequality. The meaning of condition (A3) is less obvious at first sight. A detailed discussion of (A3) is postponed to section 3. For now, we mention three straightforward implications of (A3). First, note that imposing both (A2) and (A3) requires M to be bounded. Indeed, plugging (A2) into (A3) yields

$$\mathrm{diam}(M) \leq 2(K_2^\beta K_3)^{\frac{1}{2-\alpha\beta}}.$$

More importantly, (A3) implies that minimiser x^* is unique. Finally, condition (A3) applied to minimiser x_n reads

$$d(x_n, x^*)^2 \leq K_3 \left(\int_M (F(x_n, y) - F(x^*, y)) \, \mathrm{d}P(y) \right)^\beta. \tag{2.3}$$

The left hand side of this inequality is the estimation performance of x_n. The integral, under power β, may be called the learning performance of x_n, or its excess risk. Having this comparison in mind, we will focus on controlling the learning performance of x_n knowing that an upper bound on $d(x_n, x^*)^2$ may be readily deduced from our results. The remainder of the section therefore presents upper bounds for the right hand side of (2.3) under specific complexity assumptions on M. For that purpose, we recall the definition of covering numbers. For $A \subset M$ and $\varepsilon > 0$, an ε-net for A is a finite subset $\{x_1, \dots, x_N\} \subset M$ such that

$$A \subset \bigcup_{j=1}^{N} B(x_j, \varepsilon),$$

where $B(x, \varepsilon) := \{u \in M : d(x, u) < \varepsilon\}$. The ε-covering number $N(A, d, \varepsilon) \in (1, +\infty]$ is the smallest integer $N \geq 1$ such that there exists an ε-net of size N for A in M. The function $\varepsilon \mapsto \log N(A, d, \varepsilon)$ will be referred to as the metric entropy of A.

2.2 *Doubling condition*

Our first complexity assumption is the following.

(B1) (Doubling condition) There exist constants $C, D > 0$ such that, for all $0 < \varepsilon \leq r$,

$$N(B(x^*, r), d, \varepsilon) \leq \left(\frac{Cr}{\varepsilon}\right)^D.$$

Condition (B1) essentially characterises M as a D-dimensional space and implies the following result.

Theorem 2.1. *Assume that (A1), (A2), (A3) and (B1) hold. Then, for all $n \geq 1$ and all $t > 0$,*

$$kd(x^*, x_n)^{\frac{2}{\beta}} \leq \int_M (F(x_n, y) - F(x^*, y))\, \mathrm{d}P(y) \leq A \cdot \max\left\{\left(\frac{D}{n}\right)^{\frac{1}{2-\alpha\beta}}, \left(\frac{t}{n}\right)^{\frac{1}{2-\alpha\beta}}\right\},$$

with probability at least $1 - 2e^{-t}$, where $k = K_3^{-\frac{1}{\beta}}$ and A is an explicit constant independent of n.

Note that bounds in expectation may be derived from this result, using classical arguments. As described in section 3, (A2) and (A3) hold in several interesting cases for $\alpha = \beta = 1$. In this case, Theorem 2.1 exhibits an upper bound of order D/n. A discussion on the optimality of this result is postponed to paragraph 2.4 below. Next, we shortly comment condition (B1).

Remark 2.2. *With a slight abuse of terminology, condition (B1) is termed doubling condition. In the literature, the doubling condition usually refers to the situation where inequality*

$$\sup_{x \in M} \sup_{\varepsilon > 0} \log_2 N(B(x, 2\varepsilon), d, \varepsilon) \leq D$$

holds for some $D \in (0, +\infty)$. It may be seen that this inequality implies (B1) with $C = 2$. Note however that (B1) is slightly less restrictive as it requires only the control of the covering numbers of balls centered at x^*. This fact is sometimes useful as described in example 2.4 below.

We now give two examples where assumption (B1) holds.

Example 2.3. *Suppose there exists a positive Borel measure μ on (M, d) such that, for some $D > 0$, we have*

$$\forall (x, r) \in M \times (0, +\infty), \quad \alpha_- r^D \leq \mu(B(x, r)) \leq \alpha_+ r^D, \tag{2.4}$$

for some constants $0 < \alpha_- \leq \alpha_+ < +\infty$. Then, for all $x \in M$ and all $0 < \varepsilon \leq r$,

$$\frac{\alpha_-}{\alpha_+}\left(\frac{r}{\varepsilon}\right)^D \leq N(B(x, r), d, \varepsilon) \leq \frac{\alpha_+}{\alpha_-}\left(\frac{3r}{\varepsilon}\right)^D, \tag{2.5}$$

and thus (B1) is satisfied. The proof is given in Section 5. Measures μ satisfying condition (2.4) are called D-regular or Ahlfors-David regular. Many examples of such spaces are discussed in section 12 in [30] or section 2.2 in [7]. Note that the present example includes the case where M is a D-dimensional compact Riemannian manifold equipped with the volume measure μ.

A direct and simple consequence of Example 2.3 is that (B1) holds in any D-dimensional vector space equipped with any norm since the Lebesgue measure satisfies (2.4) with $\alpha_- = \alpha_+$. While simple in essence, this observation allows to exhibit more general parametric families satisfying (B1) as in the next example.

Example 2.4 (Location-scatter family)**.** *Here, we detail an example of a subset M of the Wasserstein space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ for which assumption (B1) holds. We say that $M \subset \mathcal{P}_2(\mathbb{R}^d)$ is a location-scatter family if the following two requirements hold:*

(1) *All elements of M have a non-singular covariance matrix.*

(2) *For every two measures $\mu_0, \mu_1 \in M$, with expectations m_0 and m_1 and with covariance matrices Σ_0 and Σ_1 respectively, the map*

$$T_{01} : x \mapsto (m_1 - m_0) + \Sigma_0^{-1/2} \left(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2}\right)^{1/2} \Sigma_0^{-1/2} x$$

pushes forward μ_0 to μ_1, i.e. $\mu_0(T_{01}^{-1}(A)) = \mu_1(A)$ for any Borel set $A \subset \mathbb{R}^d$, which we denote $(T_{01})_{\#}\mu_0 = \mu_1$.

Such sets have been studied for instance in [5]. The map T_{01} being the gradient of a convex function, the theory of optimal transport guarantees that the coupling $(\mathrm{id}, T_{01})_{\#}\mu_0$ is optimal, so that

$$\begin{aligned} W_2^2(\mu_0, \mu_1) &= \int \|x - T_{01}(x)\|^2 \mathrm{d}\mu_0(x) \\ &= \|m_0 - m_1\|^2 + \mathrm{tr}\left(\Sigma_0 + \Sigma_1 - 2\left(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2}\right)^{1/2}\right), \end{aligned} \tag{2.6}$$

where $\mathrm{tr}(A)$ denotes the trace of matrix A and $\|.\|$ refers to the standard euclidean norm. Next, we show that such a family satisfies (B1). Let P be a probability measure on M and denote $\mu^* \in M$ a barycenter of P with mean m_* and covariance matrix Σ_*. For any two measures $\mu_0, \mu_1 \in M$, set

$$T_{*i}(x) = m_i - m_* + \Sigma_*^{-1/2} \left(\Sigma_*^{1/2}\Sigma_i\Sigma_*^{1/2}\right)^{1/2} \Sigma_*^{-1/2} x,$$

where m_i and Σ_i denote the mean and covariance matrix of μ_i, $i = 0, 1$. The pushforward $(T_{*0}, T_{*1})_{\#}\mu^*$ is a (possibly suboptimal) coupling between μ_0 and μ_1. Therefore,

$$\begin{aligned} W_2^2(\mu_0, \mu_1) &\leq \int \|T_{*0}(x) - T_{*1}(x)\|^2 \mathrm{d}\mu^*(x) \\ &= \|m_0 - m_1\|^2 + \|\Sigma_*^{-1/2}\left(\Sigma_*^{1/2}\Sigma_0\Sigma_*^{1/2}\right)^{1/2} - \Sigma_*^{-1/2}\left(\Sigma_*^{1/2}\Sigma_1\Sigma_*^{1/2}\right)^{1/2}\|_F^2, \end{aligned} \tag{2.7}$$

where $\|.\|_F$ stands for the Frobenius norm. Note that $\|(m, A)\|_*^2 = \|m\|^2 + \|A\|_F^2$ defines a norm $\|.\|_*$ on the vector space $\mathbb{R}^d \times S_d$ where S_d denotes the space of symmetric matrices of size $d \times d$. Then, define the function ϕ that maps each $\mu_{m,\Sigma}$ in the location-scatter family, with mean m and covariance Σ, to

$$\phi(\mu_{m,\Sigma}) = \left(m, \Sigma_*^{-1/2}\left(\Sigma_*^{1/2}\Sigma\Sigma_*^{1/2}\right)^{1/2}\right).$$

Then, combining (2.6) and (2.7), it follows that

$$W_2(\mu_0, \mu_1) \leq \|\phi(\mu_0) - \phi(\mu_1)\|_*, \tag{2.8}$$

with equality if $\mu_0 = \mu^*$ or $\mu_1 = \mu^*$. Therefore, since $\phi(M)$ is a subset of a vector space of dimension $D = d + d(d+1)/2$, there exists $C > 0$ such that for all $\varepsilon > 0$,

$$N(B(\mu^*, r), W_2, \varepsilon) \leq N(B((m_*, \Sigma_*^{1/2}), r) \cap \phi(M), \|.\|_*, \varepsilon) \leq \left(\frac{Cr}{\varepsilon}\right)^D.$$

Hence, (B1) holds.

The result, derived in example 2.4, may be generalised to other parametric subsets of the Wasserstein space (or more generally parametric subsets of geodesic Polish spaces with non-negative curvature). Indeed, since the Wasserstein space over $\mathbb{R}^d$ has non-negative curvature, the support of P pushed forward to the tangent cone at a barycenter μ^* is isometric to a Hilbert space (this result follows by combining Theorem 5.5 and Lemma 5.8) and its norm satisfies (2.8) (see Proposition A.13). Therefore it is enough, for (B1) to hold, to require that the image of the support of P by the $\log_{\mu^*}$ map (see paragraph A.5 for a definition) is included in a finite dimensional vector space.

2.3 Polynomial metric entropy

Condition (B1) is essentially a finite dimensional behaviour and does not apply in some scenarios of interest. This paragraph addresses the situation where the complexity of set M, measured by its metric entropy, is polynomial.

(B2) (Polynomial metric entropy) There exists constants $C, D > 0$ such that, for all $\varepsilon > 0$,

$$\log N(M, d, \varepsilon) \leq \left(\frac{C}{\varepsilon}\right)^D.$$

Theorem 2.5. *Assume that (A1), (A2), (A3) and (B2) hold. Then, for all $n \geq 1$ and all $t > 0$,*

$$kd(x^*, x_n)^{\frac{2}{\beta}} \leq \int_M (F(x_n, y) - F(x^*, y))\, \mathrm{d}P(y) \leq A \cdot \max\left\{v_n, \left(\frac{t}{n}\right)^{\frac{1}{2-\alpha\beta}}\right\},$$

with probability at least $1 - 2e^{-t}$, where

$$v_n = \begin{cases} n^{-\frac{2}{4-(2\alpha-D)\beta}} & \text{if } D < 2\alpha, \\ (\log n)/\sqrt{n} & \text{if } D = 2\alpha, \\ n^{-\frac{\alpha}{D}} & \text{if } D > 2\alpha, \end{cases}$$

where $k = K_3^{-\frac{1}{\beta}}$ and A is an explicit constant independent on n.

As for Theorem 2.1, bounds in expectation may be easily derived from this result. The optimality of Theorem 2.5 is addressed in paragraph 2.4. Next is an example where assumption (B2) applies.

Example 2.6 (Wasserstein space)**.** *Let $B = \{x \in \mathbb{R}^d : \|x - x_0\|_2 \leq \rho\}$ be a closed Euclidean ball in $\mathbb{R}^d$ and let $M = \mathcal{P}_2(B)$ be the set of square-integrable probability measures supported on B equipped with the 2-Wassertein metric W_2. Combining the result of Appendix A in [18] with a classical bound on the covering number of euclidean balls, it follows that for all $0 < \varepsilon \leq \rho$,*

$$\log N(M, W_2, \varepsilon) \leq 2\left(\frac{6\rho}{\varepsilon}\right)^d \log\left(\frac{8e\rho}{\varepsilon}\right).$$

In particular, for any $D > d$, there exists $C > 0$ depending on D and ρ such that, for all $0 < \varepsilon \leq \rho$,

$$\log N(M, W_2, \varepsilon) \leq \left(\frac{C}{\varepsilon}\right)^D,$$

so that (B2) is satisfied for all $D > d$.

We finally point towards Theorem 3 in [61] which may be used to derive upper bounds on the covering number of subsets of the 2-Wasserstein space composed of measures, absolutely continuous with respect to the Lebesgue measure, and with density belonging to some Besov class.

2.4 On optimality

At the level of generality considered by Theorems 2.1 and 2.5, we have not been able to assess the optimality of the given rates for all choices of functional F satisfying the required assumptions and all values of constants D, α, β. In particular, it is likely that the rates displayed in Theorem 2.5 are artefacts of our proof techniques and that results may be improved in some specific scenarios using additional information on the problem at hand. However, we discuss below some regimes where our results appear sharp and, on the contrary, settings where our results should allow for improvements.

To start our discussion, consider the barycenter problem, i.e. the case where $F(x, y) = d(x, y)^2$. In the context where (M, d) is a Hilbert space equipped with its usual metric and P is square integrable,

explicit computations reveal that $x_n = \sum_{i=1}^n Y_i/n$ is an empirical barycenter of P in the sense of (2.2) and that $x^* = \mathbb{E}[Y_1]$ (in the sense of the Pettis or Bochner integral) is the unique barycenter of P. In addition, we check that, for all $n \geq 1$,

$$\mathbb{E}d(x_n, x^*)^2 = \mathbb{E}\int_M (d(x_n, x)^2 - d(x^*, x)^2)\, \mathrm{d}P(x) = \frac{1}{n}\int_M d(x, x^*)^2\, \mathrm{d}P(x). \tag{2.9}$$

We notice that, under assumptions much more general than those considered in the present paper, the rate of convergence (in expectation) of empirical barycenters in a Hilbert space is of order $1/n$. While this observation concerns the very special case of Hilbert spaces, we conjecture that the rates of convergence of empirical barycenters is of order $1/n$ in a wide family of metric spaces including Hilbert spaces as a special case. Identifying precisely this wider family remains an open question but it appears from this discussion that boundedness and complexity restrictions, such as (A1), (B1) and (B2), may be unnecessary for the barycenter problem. Whenever (M, d) is $\mathbb{R}^D$ equipped with a general norm, a very interesting recent contribution, connected to that question, is [40]. On a more positive note, we point towards two encouraging aspects encoded in our results in the context of the barycenter problem. First, consider the case where (M, d) is $\mathbb{R}^D$ equipped with the euclidean metric and suppose that the Y_i's are independent with gaussian distribution $\mathcal{N}(x^*, \sigma^2)$. Then, identity (2.9) reads in this case

$$\mathbb{E}d(x_n, x^*)^2 = \mathbb{E}\int_M (d(x_n, x)^2 - d(x^*, x)^2)\, \mathrm{d}P(x) = \frac{\sigma^2 D}{n}.$$

It is known, furthermore, that $\sigma^2 D/n$ corresponds (up to universal constants) to the minimax rate of estimation of x^* in the context where the Y_i's are i.i.d. subgaussian random variables with mean x^* and variance proxy σ^2 (see Chapter 4 in [48]). Therefore, provided (M, d) is a bounded metric space (which guarantees (A1) and (A2) with $\alpha = 1$) and provided assumption (A3) holds for $\beta = 1$ (which is often the case as discussed in paragraphs 3.1 and 3.2 below) Theorem 2.1 recovers the optimal rate of convergence D/n, up to constants, in a fairly wide context. Finally, note that while possibly suboptimal in some cases, the rates provided by Theorems 2.1 and 2.5, combined with examples 2.4, 2.6 and discussions of paragraph 3.2, provide up to our knowledge the first rates for the Wasserstein barycenter problem at this level of generality. An exception is the Wasserstein space over the real line (studied, for instance, in [14]) which happens to be isometric to a convex subset of a Hilbert space as can be deduced for instance from combining statement (iii) of Proposition 3.5 in [51] and Proposition 4.1 in [33].

Outside from the setting of the barycenter problem, not much is known on optimal rates of estimation or learning (in the sense described at the end of paragraph 2.1) of x^* defined in (2.1). We believe this question remains mainly open. It is our impression that the $1/n$ rate, conjectured to hold for empirical barycenters in a wide setup, is a behavior very specific to the case $F(x, y) = d(x, y)^2$. For more general functionals, we suspect that the complexity of M should have an impact as it is classically the case in nonparametric statistics or learning theory. Note in particular that whenever parameters $\alpha = \beta = 1$ in (A2) and (A3), the rate v_n in Theorem 2.5 becomes

$$v_n = \begin{cases} n^{-\frac{2}{2+D}} & \text{if } D < 2, \\ (\log n)/\sqrt{n} & \text{if } D = 2, \\ n^{-\frac{1}{D}} & \text{if } D > 2, \end{cases}$$

which corresponds to known state of the art learning rates, under complexity assumptions in the same flavor as (B2), as displayed for instance by Theorem 2 in [47]. However, exact situations under which Theorem 2.5 provides optimal rates of convergence remains unclear to us.

Finally, note that the second inequality in both Theorems 2.1 and 2.5 hold for the limiting case $\beta = 0$ (with A remaining finite), which correspond to dropping assumption (A3). In the context of Theorem 2.1 (or that of Theorem 2.5 with $D < 2\alpha$) the case $\beta = 0$ gives rise to a bound of order

$$\int_M (F(x_n, y) - F(x^*, y))\, \mathrm{d}P(y) \leq \frac{C}{\sqrt{n}},$$

with high probability. Note however that this limiting case does not allow to provide any bound for $d(x_n, x^*)$.

3 Variance inequalities

This section studies conditions implying the validity of (A3). The first three paragraphs below focus on the barycenter problem, i.e. the case where $F(x, y) = d(x, y)^2$, and investigate (A3) in the light of curvature bounds. Aleksandrov curvature bounds of a geodesic space (see paragraph A.3) is a key concept of comparison geometry and many geometric phenomena are known to depend on whether the space M has a curvature bounded from below or above. In paragraphs 3.1 and 3.2, devoted respectively to non-positively and non-negatively curved spaces, we show that curvature bounds also affect statistical procedures through their relation with (A3). Finally, paragraph 3.3 addresses the case of a general F and connects (A3) to its convexity properties. The material presented in this section relies heavily on background in metric geometry gathered in appendix A for convenience.

3.1 Non positive curvature

This first paragraph introduces a fundamental insight due to K.T. Sturm, in the context of geodesic spaces of non-positive curvature, that has strongly influenced our study. To put the following result in perspective, we recall that a geodesic space (M, d) is said to have non-positive curvature ($\mathrm{curv}(M) \leq 0$ for short) if, for any $p, x, y \in M$ and any geodesic $\gamma : [0, 1] \to M$ such that $\gamma(0) = x$ and $\gamma(1) = y$,

$$d(p, \gamma(t))^2 \leq (1 - t)d(p, x)^2 + td(p, y)^2 - t(1 - t)d(x, y)^2,$$

for all $t \in [0, 1]$. Non-positive curvature is given a probabilistic description in the next result.

Theorem 3.1 (Theorem 4.9 in [51]). *Let (M, d) be a separable and complete metric space. Then, the following properties are equivalent.*

(1) *(M, d) is geodesic and $\mathrm{curv}(M) \leq 0$.*

(2) *Any probability measure $Q \in \mathcal{P}_2(M)$ has a unique barycenter $x^* \in M$ and, for all $x \in M$,*

$$d(x, x^*)^2 \leq \int_M (d(x, y)^2 - d(x^*, y)^2)\, \mathrm{d}Q(y). \tag{3.1}$$

In words, Theorem 3.1 states in particular that (A3) holds for any possible probability measure P on M, with $K_3 = 1$ and $\beta = 1$, provided $\mathrm{curv}(M) \leq 0$. It is worth mentioning again that (A1) and (A2) also hold, provided in addition $\mathrm{diam}(M) < +\infty$, so that the case of bounded metric spaces with non-positive curvature fits very well our basic assumptions. Condition $\mathrm{curv}(M) \leq 0$ is satisfied in a number of interesting examples. Such examples include (convex subsets of) Hilbert spaces or the case where (M, d) is a simply connected Riemannian manifold with non-positive sectional curvature. Other examples are metric trees and other metric constructions such as products or gluings of spaces of non-positive curvature (see [20], [21] or [3] for more details).

3.2 Non negative curvature and extendable geodesics

The present paragraph investigates the case of spaces of non-negative curvature. Contrary to the case of spaces of non-positive curvature, condition (A3) may not hold for every probability measure P on M if $\mathrm{curv}(M) \geq 0$. Indeed, note that unlike in the case when $\mathrm{curv}(M) \leq 0$, there might exist probability measures $P \in \mathcal{P}_2(M)$ with more than one barycenter whenever $\mathrm{curv}(M) \geq 0$. A simple example when $M = S^{d-1}$, the unit euclidean sphere in $\mathbb{R}^d$ with angular metric, is the uniform measure on the equator having the north and south poles as barycenters. Since (A3) implies uniqueness of barycenter x^*, this condition disqualifies such probability measures. Hence, establishing conditions under which (A3) holds is more delicate whenever $\mathrm{curv}(M) \geq 0$. The next result provides an important first step in this direction.

Theorem 3.2. *Let (M,d) be a separable and complete geodesic space such that* $\mathrm{curv}(M) \geq 0$. *Let $P \in \mathcal{P}_2(M)$ and x^* be a barycenter of P. Then, for all $x \in M$,*

$$d(x,x^*)^2 \int_M k^x_{x^*}(y)\,\mathrm{d}P(y) = \int_M (d(x,y)^2 - d(x^*,y)^2)\,\mathrm{d}P(y), \tag{3.2}$$

where, for all $x \neq x^*$ and all y,

$$k^x_{x^*}(y) = 1 - \frac{\|\log_{x^*}(x) - \log_{x^*}(y)\|^2_{x^*} - d(x,y)^2}{d(x,x^*)^2}. \tag{3.3}$$

Therefore, P satisfies (A3) with $K_3 = 1/k$ and $\beta = 1$ if and only if, for all $x \in M$,

$$k \leq \int_M k^x_{x^*}(y)\,\mathrm{d}P(y). \tag{3.4}$$

By definition of a barycenter, the right hand side of (3.2) is non negative. In addition $\mathrm{curv}(M) \geq 0$ implies that $d(x,y) \leq \|\log_p(x) - \log_p(y)\|_p$ for all $x,y,p \in M$ (see Proposition A.13 in appendix A). Combining these two observations with the definition of $k^x_{x^*}(y)$ implies that

$$0 \leq \int_M k^x_{x^*}(y)\,\mathrm{d}P(y) \leq 1.$$

The next result identifies a condition under which a variance inequality holds.

Theorem 3.3. *Let (M,d) be a separable and complete geodesic space such that* $\mathrm{curv}(M) \geq 0$. *Let $P \in \mathcal{P}_2(M)$ and x^* be a barycenter of P. Fix $\lambda > 0$ and suppose that the following properties hold.*

(1) *For P-almost all $y \in M$, there exists a geodesic $\gamma_y : [0,1] \to M$ connecting x^* to y that can be extended to a function $\gamma^+_y : [0, 1+\lambda] \to M$ that remains a shortest path between its endpoints.*

(2) *The point x^* remains a barycenter of the measure $P_\lambda = (e_\lambda)_\# P$ where $e_\lambda : M \to M$ is defined by $e_\lambda(y) = \gamma^+_y(1+\lambda)$.*

Then, for all $x \in M$,

$$d(x,x^*)^2 \leq \frac{1+\lambda}{\lambda} \int_M (d(x,y)^2 - d(x^*,y)^2)\,\mathrm{d}P(y),$$

and thus (A3) holds with $K_3 = (1+\lambda)/\lambda$ and $\beta = 1$.

Examples of geodesic spaces of non-negative curvature include (convex subsets of) Hilbert spaces or the case where (M,d) is a simply connected Riemannian manifold with non-negative sectional curvature. Next is a simple example where the condition (1) of Theorem 3.3, i.e. the ability to extend geodesics, takes a simple form.

Example 3.4 (Unit sphere)**.** *Let $M = S^{d-1}$ be the unit Euclidean sphere in $\mathbb{R}^d$ equipped with the angle metric. Let $P \in \mathcal{P}_2(M)$ be such that it has a unique barycenter $x^* \in M$. In M, a shortest path between two points is a part of a great circle and a part of a great circle is a shortest path between its endpoints if, and only if, it has length less than π. Therefore, if a neighborhood V of $C(x^*) = \{-x^*\}$, the cut locus of x^*, satisfies $P(V) = 0$, then condition* (1) *of Theorem 3.3 is satisfied for some $\lambda > 0$. Note however that condition* (1) *is not enough to give a variance inequality in general. Indeed, consider the uniform measure P on the equator with the north and south poles for barycenters. Then, for P-almost all $y \in M$, the geodesic connecting the south pole x^* to y can be extended by a factor $1+\lambda = 2$ in the sense of* (1) *in the above theorem. However, since there is no unique barycenter, no variance inequality can hold in this case. Therefore, requirement* (2) *cannot be dropped in Theorem 3.3.*

In the rest of this paragraph, we provide a sufficient condition for the extendable geodesics condition (1) of Theorem 3.3 to hold in the context where $(M,d) = (\mathcal{P}_2(H), W_2)$ is the Wasserstein space over a Hilbert space H. In this case, (M,d) is known to have non-negative curvature (see Section 7.3 in [6]). We recall the following definition. For a convex function $\phi : H \to \mathbb{R}$, its subdifferential $\partial\phi \subset H^2$ is defined by

$$(x,y) \in \partial\phi \Leftrightarrow \forall z \in H, \quad \phi(z) \geq \phi(x) + \langle y, z - x\rangle.$$

Then we can prove the following result.

Theorem 3.5. *Let $(M,d) = (\mathcal{P}_2(H), W_2)$ be the Wasserstein space over a Hilbert space H. Let μ and ν be two elements of M and let $\gamma : [0,1] \to S$ be a geodesic connecting μ to ν in M. Then, γ can be extended by a factor $1+\lambda$ (in the sense of (1) in Theorem 3.3) if, and only if, the support of the optimal transport plan π of (μ,ν) lies in the subdifferential $\partial\phi$ of a $\frac{\lambda}{1+\lambda}$-strongly convex map $\phi : H \to \mathbb{R}$.*

Again, like in example 3.4, this gives a condition that ensures the validity of (1) in Theorem 3.5, but it can be shown that it is not enough to obtain a variance inequality.

3.3 Convexity

Here, we connect (A3) to convexity properties of functional F along paths in M.

Definition 3.6 ((k,β)-convexity)**.** *Given $k > 0$, $\beta \in (0,1]$ and a path $\gamma : [0,1] \to M$, a function $f : M \to \mathbb{R}$ is called (k,β)-convex along γ if the function*

$$t \in [0,1] \mapsto f(\gamma(t)) - kd(\gamma(0),\gamma(1))^{\frac{2}{\beta}} t^2$$

is convex. If (M,d) is geodesic, a function $f : M \to \mathbb{R}$ is called (k,β)-geodesically convex if, for all $x, x' \in M$, f is (k,β)-convex along at least one geodesic connecting x to x'.

In the sequel, we abbreviate $(k,1)$-convexity by k-convexity. When M is geodesic, a (k,β)-convex function $f : M \to \mathbb{R}$ refers to a (k,β)-geodesically convex function unless stated otherwise. Note that (k,β)-convexity is a special case of uniform convexity (see Definition 1.6. in [51]). We start by a general result.

Theorem 3.7. *Let $k > 0$ and $\beta \in (0,1]$. Suppose that, for all $x \in M$, there exists a path connecting x^* to x along which the function*

$$x \in M \mapsto \int_M F(x,y)\,\mathrm{d}P(y)$$

is (k,β)-convex. Then, for all $x \in M$,

$$d(x,x^*)^2 \leq \left(\frac{1}{k}\int_M (F(x,y) - F(x^*,y))\,\mathrm{d}P(y)\right)^{\beta}, \tag{3.5}$$

and hence (A3) holds for $K_3 = k^{-\beta}$. In particular, the assumption of the theorem holds whenever, for all $x \in M$ and (P-almost) all $y \in M$, there exists a path connecting x^* to x along which the function $F(.,y)$ is itself (k,β)-convex. A special case in when (M,d) is geodesic and, for (P-almost) all $y \in M$, the functional $x \in M \mapsto F(x,y)$ is itself (k,β)-geodesically convex.

The previous result is deliberately stated in a general form. This allows to investigate several notions of convexity that may coexists as in the space of probability measures.

Remark 3.8. *For some spaces, such as the Wasserstein space over $\mathbb{R}^d$, there exist two canonical paths between two points μ and ν. One is the geodesic γ_t defined by the fact that $W_2(\gamma_t,\gamma_s) = |s-t|W_2(\mu,\nu)$ for $s,t \in [0,1]$ (which needs not be unique). A second one is the linear interpolation between the two measures $\ell_t = (1-t)\mu + t\nu$. It may be that $x \mapsto \int F(x,y)\mathrm{d}P(y)$ is strongly convex along only one of these two paths. This case is further discussed in section 4.*

We end by discussing the notion of a k-convex metric space. In the remainder of the paragraph, we consider $F(x,y) = d(x,y)^2$.

Definition 3.9. *A geodesic space (M,d) is said to be k-convex if, for all $y \in M$, the function $x \in M \mapsto d(x,y)^2$ is k-geodesically convex.*

Using Theorem 3.7, it follows that if M is k-convex, then (A3) holds with $\beta = 1$ and $K_3 = 1/k$. When $k = 1$, note that a geodesic space is k-convex if and only if $\mathrm{curv}(M) \leq 0$ (see Proposition A.6). When $k \neq 1$, the connection between curvature bounds and k-convexity of a metric space is not straightforward. In particular, k-convex metric spaces include many interesting spaces, for which condition $\mathrm{curv}(M) \leq 0$ does not hold. We give two examples.

Example 3.10 (Proposition 3.1. in [44]). *Let M be a geodesic space with $\mathrm{curv}(M) \leq 1$ and $\mathrm{diam}(M) < \pi/2$. Then M is k-convex for*

$$k = 4\,\mathrm{diam}(M)\tan\left(\frac{\pi}{2} - \mathrm{diam}(M)\right).$$

Implication of this result can be compared to Theorem 3.3. In the context of $M \subset S^2$, considered in example 3.4, the above result states that M is k-convex provided it is included in the interior of an $1/8$th of a sphere. In comparison, Theorem 3.3 expresses that a variance inequality (A3) may hold for a measure supported on the whole sphere minus a neighbourhood of the cut locus of the barycenter.

Example 3.11 (Theorem 1 in [8]). *If $(S, \mathcal{S}, \mu)$ is a measured space and $p \in (1, 2]$, then $M = L^p(\mu)$ is $(p-1)$-convex.*

4 Further examples

We describe additional examples, different from the usual barycenter problem, where our results apply. In these examples, we focus mainly on functionals over the Wasserstein space. Paragraphs 4.1 addresses the case of f-divergences. Subsection 4.2 discusses interaction energy. In 4.3 we consider several examples related to the approximation of barycenters.

4.1 f-divergences

We call f-divergence a functional of the form

$$D_f(\mu, \nu) := \begin{cases} \int f\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\nu & \text{if } \mu \ll \nu, \\ +\infty & \text{otherwise,} \end{cases} \tag{4.1}$$

for some convex function $f : \mathbb{R}_+ \to \mathbb{R}$. Such functionals are also known as internal energy, relative functionals or Csiszár divergences. Specific choices for function f give rise to well known examples. For instance, $f(x) = x \log x$ gives rise to the Kullback-Leibler divergence (or relative entropy). Minimisers of the average Kullback-Leibler divergence, or its symmetrised version, have been considered for instance in [58] for speech synthesis. Other functions like $f(x) = (x-1)^2$ or $f(x) = |x-1|/2$ lead respectively to the chi-squared divergence or the total variation. The next results present sufficient conditions under which (A1), (A2) and (A3) hold in this case. First, suppose $M \subset \mathcal{P}_2(E)$ where $E \subset \mathbb{R}^d$ is a convex set. Note that (A1) holds if there exists $0 < c_- < c_+ < +\infty$ and a reference measure such that all $\mu \in M$ have a density g_μ with respect to this measure with values in $[c_-, c_+]$ on their support. Next, we show that (A2) holds under related conditions.

Theorem 4.1. *Suppose that all measures $\mu \in M$ have a density g_μ with respect to some reference measure. Suppose there exist $0 < c_- < c_+ < +\infty$ such that all g_μ take values in $[c_-, c_+]$. Suppose in addition that there exists $\Lambda > 0$ such that all g_μ are Λ-Lipschitz on E. Assume finally that f is differentiable and that f' is L-Lipschitz on E. Then, for all $\nu, \mu, \mu' \in M$,*

$$|D_f(\mu, \nu) - D_f(\mu', \nu)| \leq \frac{2L\Lambda c_+}{c_-^2} W_2(\mu, \mu'),$$

so that (A2) holds for $K_2 = 2L\Lambda c_+/c_-^2$ and $\alpha = 1$.

The next result is devoted to condition (A3) and is a slight modification of Theorem 9.4.12 in [6]. Below we say that $\nu \in M$ is λ-log-concave if $\mathrm{d}\nu(x) = e^{-U(x)}\mathrm{d}x$ for some λ-convex $U : E \to \mathbb{R}$.

Theorem 4.2. *Consider $h(s) = e^s f(e^{-s})$. Suppose that h is convex and that there exists $c > 0$ such that, for all $u \in \mathbb{R}$ and all $\varepsilon > 0$, $h(u) - h(u+\varepsilon) \geq c\varepsilon$. Suppose finally that M is a geodesically convex subset of $\mathcal{P}_2(E)$. Then the following holds.*

(1) *For any λ-log-concave $\nu \in M$, the functional $\mu \in M \mapsto D_f(\mu, \nu)$ is $c\lambda$-geodesically convex.*

(2) *Let $\lambda > 0$ and suppose $P \in \mathcal{P}(M)$ is supported on λ-log-concave measures in M. Then for any minimizer*

$$\mu^* \in \operatorname*{arg\,min}_{\mu \in M} \int_M D_f(\mu, \nu) \, \mathrm{d}P(\nu),$$

and any $\mu \in M$,

$$W_2(\mu, \mu^*)^2 \leq \frac{1}{c\lambda} \int_M (D_f(\mu, \nu) - D_f(\mu^*, \nu)) \, \mathrm{d}P(\nu),$$

and thus (A3) holds with $K_3 = 1/c\lambda$ and $\beta = 1$.

Note that the requirements made on f in Theorem 4.2 are compatible with $f(x) = x \log x$ for $c = 1$. However, these requirements exclude examples such as $f(x) = (x-1)^2$. At the price of a smaller exponent $\beta = 1/2$, the next result shows that (A3) holds for any choice of a strongly convex f.

Theorem 4.3. *Let $f : \mathbb{R}_+ \to \mathbb{R}$ be k-convex. Then the following holds.*

(1) *Let $\nu, \mu_0, \mu_1 \in \mathcal{P}_2(E)$ be such that $\mu_0 \ll \nu$ and $\mu_1 \ll \nu$. Suppose in addition that*

$$m_4(\nu) := \int \|x\|^4 \, \mathrm{d}\nu(x) < +\infty.$$

Then, $D_f(., \nu)$, is $(\frac{k}{4m_4(\nu)}, \frac{1}{2})$-convex along the linear interpolation $\ell_t = (1-t)\mu_0 + t\mu_1$.

(2) *Let $m > 0$. Suppose that M is a linearly convex subset of $\mathcal{P}_2(E)$ and suppose $P \in \mathcal{P}(M)$ is supported on measures $\nu \in M$ such that $m_4(\nu) \leq m$. Then, for any minimizer*

$$\mu^* \in \operatorname*{arg\,min}_{\mu \in M} \int_M D_f(\mu, \nu) \, \mathrm{d}P(\nu),$$

and any $\mu \in M$,

$$W_2(\mu, \mu^*)^2 \leq 2 \left(\frac{m}{k} \int_M (D_f(\mu, \nu) - D_f(\mu^*, \nu)) \, \mathrm{d}P(\nu) \right)^{1/2},$$

and thus (A3) holds with $K_3 = 2(m/k)^{1/2}$ and $\beta = 1/2$.

We end this paragraph with an important remark. Using Theorem 3.7, and as in Theorem 4.2, condition (A3) can be deduced from k-convexity of the f-divergence. Strong convexity of the f-divergence was used by Sturm, Lott and Villani (see [52, 39] and references therein) to define the celebrated synthetic notion of Ricci curvature bounds. In particular, for an underlying Riemannian D-dimensional manifold E equipped with a measure with density e^{-V} with respect to the Riemannian measure, k-convexity of such f-divergence is equivalent to

$$\mathrm{Ric} + \mathrm{Hess}(V) \geq k$$

where Ric stands for the Ricci curvature tensor and $\mathrm{Hess}(V)$ is the Hessian of V, given $s \mapsto s^D f(s^{-D})$ is geodesically convex.

4.2 Interaction energy

Interaction energy of a measure μ, for a given function g on $M \times M$ called interaction potential, is usually defined as the integral of g w.r.t. the product measure $\mu \otimes \mu$. Here, we keep the same terminology for a larger class of functionals on the Wasserstein space. Let (E, ρ) be a separable, complete, locally compact and geodesic space. Let $M = \mathcal{P}_2(E)$ be equipped with the Wasserstein metric W_2. Consider functional $I_g : M \times M \to \mathbb{R}_+$ defined by

$$I_g(\mu, \nu) = \int_{E \times E} g(x, y) \, \mathrm{d}\mu(x) \mathrm{d}\nu(y),$$

for a measurable $g : E \times E \to \mathbb{R}_+$. Note first that (A1) is clearly satisfied provided g is upper bounded or g is continuous and E is compact. The next two results present sufficient conditions for (A2) and (A3) to hold.

Theorem 4.4. *Fix $L > 0$ and suppose that g is L-Lipschitz in the first variable when the second is fixed. Then, for all $\mu, \mu', \nu \in \mathcal{P}_2(E)$,*

$$|I_g(\mu, \nu) - I_g(\mu', \nu)| \leq L W_2(\mu, \mu'),$$

and (A2) holds with $K_2 = L$ and $\alpha = 1$.

Theorem 4.5. *Fix $k \geq 0$ and $\beta \in (0, 1]$. Suppose that g is (k, β)-geodesically convex in the first variable when the second is fixed. Then, the following holds.*

(1) *For any fixed $\nu \in \mathcal{P}_2(E)$, the functional $I_g(., \nu)$ is (k, β)-geodesically convex on $\mathcal{P}_2(E)$.*

(2) *Let $P \in \mathcal{P}(\mathcal{P}_2(E))$. Then for any minimizer*

$$\mu^* \in \operatorname*{arg\,min}_{\mu \in \mathcal{P}_2(E)} \int I_g(\mu, \nu) \, \mathrm{d}P(\nu),$$

and any $\mu \in \mathcal{P}_2(E)$,

$$W_2(\mu, \mu^*)^2 \leq \left(\frac{1}{k} \int (I_g(\mu, \nu) - I_g(\mu^*, \nu)) \, \mathrm{d}P(\nu) \right)^{\beta},$$

and (A3) holds with $K_3 = 1/k^{\beta}$.

4.3 Regularised Wasserstein distance

In this final paragraph, we present a few functionals of interest whose study, in the light of the previous results, could provide interesting research perspectives. In certain cases, like for the Wasserstein space, the distance is difficult to compute. In the field of computational optimal transport, a lot of work has been devoted to construct computationally friendly approximations of the Wasserstein distance. Such approximations include the sliced Wasserstein distance, penalised Wasserstein distances or regularised Wasserstein distances (see [23] for more details). In some of these approximations, one hope is that one may enforce some form of convexity allowing for condition (A3) to be valid in a wide setting. The Sinkhorn divergence, defined below, is an approximation of the Wasserstein distance that has been widely used in practice due to its attractive computationally properties. Denote by D the relative entropy, i.e. $D = D_f$ for $f(x) = x \log x$ where D_f is as in (4.1), and $\Gamma(\mu, \nu)$ the set of probability measures π on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ an ν respectively. Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, and $\gamma > 0$. The entropy-regularised Wasserstein distance or Sinkhorn divergence between μ and ν is defined by

$$W_\gamma(\mu, \nu)^2 := \inf_{\pi \in \Gamma(\mu, \nu)} \int d(x, y)^2 \, \mathrm{d}\pi(x, y) + \gamma D(\pi, \mu \otimes \nu). \tag{4.2}$$

Identifying scenarios in which W_γ satisfies conditions (A1)-(A3) remains for us an open question. Note however that, using Lagrange multipliers, the Sinkhorn divergence can be alternatively expressed as

$$\inf \left\{ \int d(x, y)^2 \, \mathrm{d}\pi(x, y) : \pi \in \Gamma(\mu, \nu), D(\pi, \mu \otimes \nu) \leq \delta \right\},$$

for some $\delta \geq 0$. Letting $\delta = 0$, the above expression reduces to the interaction energy I_g studied in paragraph 4.2 for $g(x, y) = d(x, y)^2$. This observation suggests that, for large values of γ, functional $W_\gamma(\mu, \nu)$ should satisfy (A1)-(A3) under reasonable conditions. An interesting alternative notion of regularised Wasserstein distance, via factored couplings, was introduced in [26]. In order to define this notion, we introduce some notation. For a partition $\mathcal{C} = (C_1, \ldots, C_n)$, and a measure μ, denote by $\mu_i^{\mathcal{C}}$ the measure μ restricted to C_i. For $n \in \mathbb{N}$, denote by $\Gamma_n(\mu, \nu)$ the set of measures γ such that there exists two partition $\mathcal{C}^0$ and $\mathcal{C}^1$ such that $\gamma = \sum_{i=1}^n \lambda_i \mu_i^{\mathcal{C}^0} \otimes \nu_i^{\mathcal{C}^1}$, with $\lambda_i := \mu(C_i^0) = \nu(C_i^1)$. For two measures μ and ν, the regularised Wasserstein distance between μ and ν is defined by

$$RW_n^2(\mu, \nu) := \inf_{\gamma \in \Gamma_n(\mu, \nu)} \int d^2 \mathrm{d}\gamma = \inf \left\{ \sum_{i=1}^n \lambda_i \int d^2 \mathrm{d}(\mu_i \otimes \nu_i) : \sum_{i=1}^n \lambda_i \mu_i \otimes \nu_i = \gamma \in \Gamma_n(\mu, \nu) \right\}.$$

While the trivial case $n = 1$ reduces to the interaction energy studied above, it remains unclear to us in which setting our assumptions apply to this functional in general.

5 Proofs

5.1 Proof of Theorem 2.1

The proof is based on three auxiliary lemmas. The first lemma provides an upper bound on the largest fixed point of a random nonnegative function. The proof follows from a combination of arguments presented in Theorem 4.1, Corollary 4.1 and Theorem 4.3 in [35].

Lemma 5.1. *Let $\{\phi(\delta) : \delta \geq 0\}$ be non-negative random variables (indexed by all deterministic $\delta \geq 0$) such that, almost surely, $\phi(\delta) \leq \phi(\delta')$ if $\delta \leq \delta'$. Let $\{b(\delta, t) : \delta \geq 0, t \geq 0\}$, be (deterministic) real numbers such that $b(\delta, t) \leq b(\delta, t')$, as soon as $t \leq t'$, and such that*

$$\mathbb{P}(\phi(\delta) \geq b(\delta, t)) \leq e^{-t}.$$

Finally, let $\hat{\delta}$ be a nonnegative random variable, a priori upper bounded by a constant $\bar{\delta} > 0$, and such that, almost surely,

$$\hat{\delta} \leq \phi(\hat{\delta}).$$

Then defining, for all $t \geq 0$,

$$b(t) := \inf\left\{\alpha > 0 : \sup_{\delta \geq \alpha} \frac{b\left(\delta, \frac{t\delta}{\alpha}\right)}{\delta} \leq 1\right\},$$

we obtain, for all $t \geq 0$,

$$\mathbb{P}(\hat{\delta} \geq b(t)) \leq 2e^{-t}.$$

The second lemma is due to [19] and improves upon the work of [54] by providing explicit constants. Given a family $\mathcal{F}$ of functions $f : M \to \mathbb{R}$, denote

$$|P - P_n|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} (P - P_n) f \quad \text{and} \quad \sigma_{\mathcal{F}}^2 := \sup_{f \in \mathcal{F}} P(f - Pf)^2,$$

where M and P are as defined in paragraph 2.1 and $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$.

Lemma 5.2. *Suppose that all functions in $\mathcal{F}$ are $[a, b]$-valued, for some $a < b$. Then, for all $n \geq 1$ and all $t > 0$,*

$$|P - P_n|_{\mathcal{F}} \leq \mathbb{E}|P - P_n|_{\mathcal{F}} + \sqrt{\frac{2t}{n}\left(\sigma_{\mathcal{F}}^2 + 2(b-a)\,\mathbb{E}|P - P_n|_{\mathcal{F}}\right)} + \frac{(b-a)t}{3n},$$

with probability larger than $1 - e^{-t}$.

For background on empirical processes, including the proof of Lemma 5.2, we refer the reader to [28]. Finally, the third result we need is the following generalized version of Dudley's entropy bound (see, for instance, Theorem 5.31 in [57]).

Lemma 5.3. *Let $(X_t)_{t \in E}$ be a real valued process indexed by a pseudo metric space (E, d). Suppose that the three following conditions hold.*

(1) (Separability) *There exists a countable subset $E' \subset E$ such that, for any $t \in E$,*

$$X_t = \lim_{t' \to t, t' \in E'} X_{t'}, \quad a.s.$$

(2) (Subgaussianity) *For all $s, t \in E$, $X_s - X_t$ is subgaussian in the sense that*

$$\forall \theta \in \mathbb{R}, \quad \log \mathbb{E} e^{\theta(X_s - X_t)} \leq \frac{\theta^2 d(s,t)^2}{2}.$$

(3) (Lipschitz property) *There exists a random variable L such that, for all $s, t \in E$,*

$$|X_s - X_t| \leq L d(s,t), \quad a.s.$$

Then, for any $S \subset E$ and any $\varepsilon \geq 0$, we have

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

We are now in position to prove Theorem 2.1.

Proof of Theorem 2.1. (1) For any $\delta \geq 0$, denote

$$M(\delta) := \{x \in M : P(F(x,.) - F(x^*,.)) \leq \delta\},$$

and

$$\phi_n(\delta) := \sup\{(P - P_n)(F(x,.) - F(x^*,.)) : x \in M(\delta)\}.$$

As a consequence of Assumption (B1), the set M is separable. Hence, the quantity $\phi_n(\delta)$ is measurable, as well as all suprema involved in the rest of the proof. Define

$$\delta_n := P(F(x_n,.) - F(x^*,.)).$$

By definition of x_n, $P_n(F(x_n,.) - F(x^*,.)) \leq 0$ so that

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

As a result, in order to upper bound δ_n with high probability, it is enough to upper bound $\phi_n(\delta)$, for fixed $\delta \geq 0$, and apply Lemma 5.1. Denoting

$$\sigma^2(\delta) := \sup\{P(F(x,.) - F(x^*,.))^2 : x \in M(\delta)\},$$

and observing that $-2K_1 \leq F(x,y) - F(x^*,y) \leq 2K_1$, for all $x, y \in M$ due to (A1), it follows from Lemma 5.2 that inequality

$$\phi_n(\delta) \leq \mathbb{E}\phi_n(\delta) + \sqrt{\frac{2t}{n}\left(\sigma^2(\delta) + 8K_1\,\mathbb{E}\phi_n(\delta)\right)} + \frac{4K_1 t}{3n},$$

holds with probability at least $1 - e^{-t}$. Using basic inequalities $\sqrt{u+v} \leq \sqrt{u} + \sqrt{v}$ and $2\sqrt{uv} \leq u + v$ for positive numbers, we further deduce that

$$\phi_n(\delta) \leq 2\mathbb{E}\phi_n(\delta) + \sigma(\delta)\sqrt{\frac{2t}{n}} + \frac{16K_1 t}{3n}, \tag{5.1}$$

with probability at least $1 - e^{-t}$. Combining (A2) and (A3), we deduce that for all $x \in M$,

$$P(F(x,.) - F(x^*,.))^2 \leq K_2^2 K_3^{\alpha}(P(F(x,.) - F(x^*,.)))^{\alpha\beta}, \tag{5.2}$$

and therefore,

$$\sigma^2(\delta) \leq K_2^2 K_3^{\alpha} \delta^{\alpha\beta}. \tag{5.3}$$

Next, we provide an upper bound for $\mathbb{E}\phi_n(\delta)$. Let $\sigma_1, \ldots, \sigma_n$ be a sequence of i.i.d. random signs, *i.e.* such that $\mathbb{P}(\sigma_i = -1) = \mathbb{P}(\sigma_i = 1) = 1/2$, independent from the Y_i's. The symmetrization principle (see, e.g., Lemma 7.4 in [57]) indicates that

$$\begin{aligned}\mathbb{E}\phi_n(\delta) &\leq 2\mathbb{E}\sup\left\{\frac{1}{n}\sum_{i=1}^n \sigma_i(F(x,Y_i) - F(x^*,Y_i)) : x \in M(\delta)\right\} \\ &= 2\mathbb{E}\sup\left\{\frac{1}{n}\sum_{i=1}^n \sigma_i F(x,Y_i) : x \in M(\delta)\right\},\end{aligned}$$

where the last line follows from the fact that the σ_i's are centered and independent of the Y_i's. Now, introduce the set $\mathcal{F} = \{F(x,.) : x \in M\}$, and observe that, conditionally on the Y_i's, the process

$$X_f := \frac{1}{\sqrt{n}}\sum_{i=1}^n \sigma_i f(Y_i), \quad f \in \mathcal{F},$$

satisfies the separability condition of Lemma 5.3 due to the separability of M. In addition, this process is subgaussian since

$$\forall f, g \in \mathcal{F}, \forall \theta \in \mathbb{R}, \quad \log \mathbb{E}[e^{\theta(X_f - X_g)} | Y_1, \ldots, Y_n] \leq \frac{\theta^2 d_n(f,g)^2}{2},$$

where

$$d_n(f,g)^2 = \frac{1}{n}\sum_{i=1}^{n}(f(Y_i) - g(Y_i))^2$$

is the natural metric in $L^2(P_n)$. Also, it satisfies the Lipchitz condition

$$|X_f - X_g| \leq \sqrt{n} d_n(f,g).$$

Hence, denoting $\mathcal{F}(\delta) = \{F(x,.) : x \in M(\delta)\}$ and applying Lemma 5.3, we obtain

$$\mathbb{E}\phi_n(\delta) \leq 2\mathbb{E} \inf_{\varepsilon \geq 0} \left\{ 2\varepsilon + \frac{12}{\sqrt{n}} \int_{\varepsilon}^{+\infty} \sqrt{\log N(\mathcal{F}(\delta), d_n, u)}\, \mathrm{d}u \right\}.$$

But combining (A2) and (A3), we deduce that, almost surely,

$$\begin{aligned} N(\mathcal{F}(\delta), d_n, u) &\leq N\left(M(\delta), d, \left(\frac{u}{K_2}\right)^{\frac{1}{\alpha}}\right) \\ &\leq N\left(B(x^*, \sqrt{K_3\delta^\beta}), d, \left(\frac{u}{K_2}\right)^{\frac{1}{\alpha}}\right). \end{aligned}$$

As a result,

$$\begin{aligned} \mathbb{E}\phi_n(\delta) &\leq 2 \inf_{\varepsilon \geq 0} \left\{ 2\varepsilon + \frac{12}{\sqrt{n}} \int_{\varepsilon}^{+\infty} \sqrt{\log N\left(B(x^*, \sqrt{K_3\delta^\beta}), d, \left(\frac{u}{K_2}\right)^{\frac{1}{\alpha}}\right)}\, \mathrm{d}u \right\} \\ &= 2 \inf_{\varepsilon \geq 0} \left\{ 2\varepsilon + \frac{12}{\sqrt{n}} \int_{\varepsilon}^{K_2 K_3^{\alpha/2}\delta^{\alpha\beta/2}} \sqrt{\log N\left(B(x^*, \sqrt{K_3\delta^\beta}), d, \left(\frac{u}{K_2}\right)^{\frac{1}{\alpha}}\right)}\, \mathrm{d}u \right\} \\ &\leq 2 \inf_{\varepsilon \geq 0} \left\{ 2\varepsilon + \frac{12}{\sqrt{n}} \int_{\varepsilon}^{K_2 K_3^{\alpha/2}\delta^{\alpha\beta/2}} \sqrt{\frac{D}{\alpha} \log\left(\frac{C^\alpha K_2 K_3^{\alpha/2}\delta^{\alpha\beta/2}}{u}\right)}\, \mathrm{d}u \right\}, \end{aligned}$$

where the last inequality follows from (B1). Assuming without loss of generality that $C \geq 1$ and using the simple upper bound $\log(x) \leq x - 1 \leq x$, for all $x > 0$, it follows from straightforward computations that

$$\mathbb{E}\phi_n(\delta) \leq \frac{48 C^{\frac{\alpha}{2}} D^{\frac{1}{2}} K_2 K_3^{\frac{\alpha}{2}}}{\alpha^{\frac{1}{2}}} \cdot \delta^{\frac{\alpha\beta}{2}} n^{-\frac{1}{2}}. \tag{5.4}$$

Combining (5.1),(5.3) and (5.4) implies therefore that we have

$$\phi_n(\delta) \leq b_n(\delta, t) := c_1 \delta^{\frac{\alpha\beta}{2}} \sqrt{\frac{D}{n}} + c_2 \delta^{\frac{\alpha\beta}{2}} \sqrt{\frac{t}{n}} + \frac{c_3 t}{n},$$

with probability at least $1 - e^{-t}$ where

$$c_1 = \frac{96 C^{\frac{\alpha}{2}} K_2 K_3^{\frac{\alpha}{2}}}{\alpha^{\frac{1}{2}}}, \quad c_2 = \sqrt{2} K_2 K_3^{\frac{\alpha}{2}} \quad \text{and} \quad c_3 = \frac{16 K_1}{3}. \tag{5.5}$$

Using the fact that $\delta_n \leq \phi_n(\delta_n)$ and Lemma 5.1, it follows that

$$\delta_n \leq b_n(t) := \inf\left\{ \tau > 0 : \sup_{\delta \geq \tau} \delta^{-1} b_n\left(\delta, \tfrac{t\delta}{\tau}\right) \leq 1 \right\},$$

with probability larger that $1 - 2e^{-t}$. It therefore remains to provide an upper bound for $b_n(t)$. Observing that $\alpha\beta \leq 1$, and that for nonincreasing functions $h_j : [0, +\infty) \to [0, +\infty)$ we have

$$\inf\{\tau > 0 : h_1(\tau) + \cdots + h_m(\tau) \leq 1\} \leq \max_{1 \leq j \leq m} \inf\{\tau > 0 : h_j(\tau) \leq 1/m\},$$

it follows that

$$b_n(t) \leq \max\left\{(3c_1)^{\frac{2}{2-\alpha\beta}} \left(\frac{D}{n}\right)^{\frac{1}{2-\alpha\beta}}, (3c_2)^{\frac{2}{2-\alpha\beta}} \left(\frac{t}{n}\right)^{\frac{1}{2-\alpha\beta}}, \frac{3c_3 t}{n}\right\},$$

where c_1, c_2, c_3 are as in (5.5). To sum up, we have shown that, for all $n \geq 1$ and all $t > 0$,

$$\begin{aligned}
&\int_M (F(x_n, y) - F(x^*, y))\, \mathrm{d}P(y) \\
&\leq \max\left\{(3c_1)^{\frac{2}{2-\alpha\beta}} \left(\frac{D}{n}\right)^{\frac{1}{2-\alpha\beta}}, (3c_2)^{\frac{2}{2-\alpha\beta}} \left(\frac{t}{n}\right)^{\frac{1}{2-\alpha\beta}}, \frac{3c_3 t}{n}\right\},
\end{aligned}$$

with probability at least $1 - 2e^{-t}$ where c_1, c_2, c_3 are as in (5.5). Finally note that, at the price of a slightly worst dependence on the constants, the last term in the maximum can be removed. Indeed, if $t < n$, we have

$$\frac{t}{n} \leq \left(\frac{t}{n}\right)^{\frac{1}{2-\alpha\beta}},$$

while, for $t \geq n$, assumption (A3) implies that inequality

$$\int_M (F(x_n, y) - F(x^*, y))\, \mathrm{d}P(y) \leq 2K_3 \left(\frac{t}{n}\right)^{\frac{1}{2-\alpha\beta}}$$

trivially holds. This completes the proof. □

5.2 Proof of Example 2.3

We prove inequalities (2.5). The proof uses a classical strategy but is reproduced for completeness. Define $\alpha_D : M \times (0, +\infty) \to [0, +\infty)$ by

$$\alpha_D(x, r) = \frac{\mu(B(x, r))}{r^D}, \tag{5.6}$$

so that $\alpha_- \leq \alpha_D(x, r) \leq \alpha_+$ by assumption. Let $x \in M$ and $0 < \varepsilon \leq r$ be fixed. Let $x_1, \ldots, x_N$ be a minimal ε-net for $B(x, r)$ of size $N = N(B(x, r), d, \varepsilon)$. Then, by monotonicity and subadditivity of μ, it follows that

$$\alpha_D(x, r) r^D = \mu(B(x, r)) \leq \sum_{i=1}^N \mu(B(x_i, \varepsilon)) = \varepsilon^D \sum_{i=1}^N \alpha_D(x_i, \varepsilon).$$

By definition of α_- and α_+, we deduce that $\alpha_- r^D \leq \alpha_+ N \varepsilon^D$ which proves the first inequality. To prove the second inequality, define the ε-packing number $N_{\mathrm{pack}}(B(x, r), d, \varepsilon)$ of $B(x, r)$ as the maximal number m of points $x_1, \ldots, x_m \in B(x, r)$ such that $d(x_i, x_j) > \varepsilon$ for all $i \neq j$. A collection of such points is called an ε-packing of $B(x, r)$. It is a classical fact that the covering and packing numbers satisfy the duality property

$$N(B(x, r), d, \varepsilon) \leq N_{\mathrm{pack}}(B(x, r), d, \varepsilon) \leq N(B(x, r), d, \varepsilon/2).$$

In particular, to upper bound the ε-covering number of $B(x, r)$ it suffices to upper bound its ε-packing number. Hence, let $x_1, \ldots, x_m$ be a maximal ε-packing of $B(x, r)$ of size $m = N_{\mathrm{pack}}(B(x, r), d, \varepsilon)$.

Notice that the balls $B(x_i, \varepsilon/2)$, $i = 1, \dots, m$ are disjoint by definition and included in $B(x, r + \varepsilon/2)$. Hence, by monotonicity and additivity of μ, it follows that

$$\begin{aligned}\alpha_D\left(x, r+\frac{\varepsilon}{2}\right)\left(r+\frac{\varepsilon}{2}\right)^D = \mu\left(B\left(x, r+\frac{\varepsilon}{2}\right)\right) &\geq \mu\left(\bigcup_{i=1}^m B\left(x_i, \frac{\varepsilon}{2}\right)\right) \\ &= \sum_{i=1}^m \mu\left(B\left(x_i, \frac{\varepsilon}{2}\right)\right) \\ &= \left(\frac{\varepsilon}{2}\right)^D \sum_{i=1}^m \alpha_D\left(x_i, \frac{\varepsilon}{2}\right).\end{aligned}$$

The definition of α_- and α_+ implies once again that

$$m\alpha_-\left(\frac{\varepsilon}{2}\right)^D \leq \alpha_+\left(r+\frac{\varepsilon}{2}\right)^D,$$

and therefore

$$m \leq \frac{\alpha_+}{\alpha_-}\left(\frac{2r}{\varepsilon}+1\right)^D \leq \frac{\alpha_+}{\alpha_-}\left(\frac{3r}{\varepsilon}\right)^D,$$

which concludes the proof.

5.3 *Proof of Theorem 2.5*

The proof is identical to that of Theorem 2.1 up to a minor modification. Here, the control on the complexity of set M is only global. Hence, the inequality

$$\mathbb{E}\phi_n(\delta) \leq 2\inf_{\varepsilon \geq 0}\left\{2\varepsilon + \frac{12}{\sqrt{n}}\int_\varepsilon^{K_2 K_3^{\alpha/2}\delta^{\alpha\beta/2}} \sqrt{\log N\left(B(x^*, \sqrt{K_3\delta^\beta}), d, \left(\frac{u}{K_2}\right)^{\frac{1}{\alpha}}\right)}\,\mathrm{d}u\right\},$$

used in the proof of Theorem 2.1, while still valid, cannot be exploited as such. We simply replace it by the upper bound

$$\mathbb{E}\phi_n(\delta) \leq 2\inf_{\varepsilon \geq 0}\left\{2\varepsilon + \frac{12}{\sqrt{n}}\int_\varepsilon^{K_2 K_3^{\alpha/2}\delta^{\alpha\beta/2}} \sqrt{\log N\left(M, d, \left(\frac{u}{K_2}\right)^{\frac{1}{\alpha}}\right)}\,\mathrm{d}u\right\}.$$

From then on, the proof is similar to that of Theorem 2.1.

5.4 *Proof of Theorem 3.2*

The proof of Theorem 3.2 follows by combining Lemmas 5.7 and 5.8 below.

Definition 5.4 (Exponential barycenter)**.** *For any $P \in \mathcal{P}_2(M)$, a point $x^* \in M$ is said to be an exponential barycenter of P if*

$$\int_M \int_M \langle \log_{x^*}(x), \log_{x^*}(y)\rangle_{x^*}\,\mathrm{d}P(x)\mathrm{d}P(y) = 0.$$

For the definition of $\log_x$ and $\langle ., . \rangle_x$ we refer the reader to Appendix A. Exponential barycenters where introduced in [25]. The definition of an exponential barycenter mimics that of the Pettis integral of a Hilbert valued function and stands as an alternative way to define the analog of the mean value of an element of $\mathcal{P}_2(M)$. Next is an important property of exponential barycenters.

Theorem 5.5 (Theorem 45 in [62])**.** *Suppose that* $\mathrm{curv}(M) \geq 0$*. Let x^* be an exponential barycenter of $P \in \mathcal{P}_2(M)$. Then the linear hull of the support of $P \circ \log_{x^*}^{-1}$ is isometric to a Hilbert space.*

Remark 5.6. *Note that Theorem 45 of [62] requires the tangent cone to be separable. It is not clear to us whether the tangent cone of a separable geodesic space with curvature bounded below is always separable. However, separability is used only to approximate (a countable number) of integrals w.r.t a measure by the same integral w.r.t. a finitely supported measure. This can be derived from the law of large number and thus separability is not necessary.*

Using Theorem 5.5 we prove next that the identity of Theorem 3.2 holds for exponential barycenters.

Lemma 5.7. *Suppose that* $\mathrm{curv}(M) \geq 0$. *Let* $P \in \mathcal{P}_2(M)$ *and* x^* *be an exponential barycenter of* P. *Then, for all* $x \in M$,

$$d(x,x^*)^2 \int_M k^x_{x^*}(y)\,\mathrm{d}P(y) = \int_M (d(x,y)^2 - d(x^*,y)^2)\,\mathrm{d}P(y),$$

where, for all $x \neq x^*$ *and all* y, $k^x_{x^*}(y)$ *is as in* (3.3).

Proof of Lemma 5.7. Fix an exponential barycenter x^* of P. For brevity, denote $\log = \log_{x^*}$, $\|.\| = \|.\|_{x^*}$ and $\langle .,. \rangle = \langle .,. \rangle_{x^*}$. Given any $x \in M$, let x_t be a geodesic connecting x^* to x in M. It is an easy exercise to check that $t \mapsto \log(x_t)$ is a geodesic connecting $\log(x^*)$ to $\log(x)$ in $T_{x^*}M$. According to Theorem 5.5, and the geometry of a Hilbert space, it follows that, for all $x, y \in M$ and all $t \in [0,1]$,

$$\|y - x_t\|^2 = (1-t)\|y-x^*\|^2 + t\|y-x\|^2 - t(1-t)\|x-x^*\|^2,$$

where here and throughout, we denote both a point $x \in M$ and its image $\log(x)$ in the tangent cone $T_{x^*}M$ by the same symbol x when there is no risk of confusion. Now since $\mathrm{curv}(M) \geq 0$, we have $d(x,y) \leq \|x-y\|$ for all $x, y \in M$, with equality if $x = x^*$ or $y = x^*$. Hence, it follows from the previous identity that, for all $x, y \in M$ and all $t \in (0,1)$,

$$\begin{aligned} t(1-t)d(x,x^*)^2 &= (1-t)d(x^*,y)^2 + t\|y-x\|^2 - \|y-x_t\|^2 \\ &= t(d(x,y)^2 - d(x^*,y)^2) + (\|y-x^*\|^2 - \|y-x_t\|^2) \\ &\quad + t\left(1 - k^x_{x^*}(y)\right) d(x,x^*)^2. \end{aligned}$$

Reordering terms and dividing by t, we obtain

$$(k^x_{x^*}(y) - t)d(x,x^*)^2 = (d(x,y)^2 - d(x^*,y)^2) + \frac{1}{t}(\|y-x^*\|^2 - \|y-x_t\|^2). \tag{5.7}$$

Integrating with respect to $P(\mathrm{d}y)$, we obtain

$$(Pk^x_{x^*}(.) - t)d(x,x^*)^2 = P(d(x,.)^2 - d(x^*,.)^2) + \frac{P(\|x^*-.\|^2 - \|x_t-.\|^2)}{t}.$$

Finally, using Theorem 5.5, we get that

$$P(\|x^*-.\|^2 - \|x_t-.\|^2) = -t^2\|x^*-x\|^2.$$

Hence, letting $t \to 0$ in the previous identity leads to the desired result. □

Lemma 5.8. *Suppose that* $\mathrm{curv}(M) \geq 0$ *and let* $P \in \mathcal{P}_2(M)$. *Then a barycenter of* P *is an exponential barycenter of* P.

Proof of Lemma 5.8. Fix a barycenter x^* of P. As in the proof of the previous Lemma, denote $\log = \log_{x^*}$, $\|.\| = \|.\|_{x^*}$ and $\langle .,. \rangle = \langle .,. \rangle_{x^*}$ for brevity. Also, we denote both a point $x \in M$ and its image $\log(x)$ in the tangent cone $T_{x^*}M$ by the same symbol x when there is no risk of confusion. Since $\mathrm{curv}(M) \geq 0$, the first statement of Theorem 45 in [62] implies that

$$\int_M \int_M \langle x, y \rangle\,\mathrm{d}P(x)\mathrm{d}P(y) \geq 0.$$

Moreover, we know that $d(x,y) \leq \|x-y\|$ for all $x, y \in M$ with equality if $x = x^*$ or $y = x^*$. In particular, x^* also minimises

$$y \mapsto \int_M \|x-y\|^2 \, \mathrm{d}P(x).$$

Thus, for all $y \in M$, letting y_t be a geodesic connecting x^* to y we get for all $t \in (0,1]$,

$$\begin{aligned}
\int_M \|x\|^2 \, \mathrm{d}P(x) &= \int_M \|x-x^*\|^2 \, \mathrm{d}P(x) \\
&\leq \int_M \|x-y_t\|^2 \, \mathrm{d}P(x) \\
&= \int_M \|x - t\cdot y\|^2 \, \mathrm{d}P(x) \\
&= \int_M (\|x\|^2 - 2t\langle x, y\rangle + t^2\|y\|^2) \, \mathrm{d}P(x),
\end{aligned}$$

where we have used the properties log and $\|.\|$ stated in section A.5. Simplifying the above expression, we obtain, for all $t \in (0,1]$,

$$2\int_M \langle x, y\rangle \, \mathrm{d}P(x) \leq t\|y\|^2.$$

Letting $t \to 0$, we get

$$\int_M \langle x, y\rangle \mathrm{d}P(x) \leq 0.$$

Integrating with respect to y, we obtain

$$\int_M \int_M \langle x, y\rangle \, \mathrm{d}P(x)\mathrm{d}P(y) \leq 0.$$

Combining this observation with the first inequality of the proof shows that x^* is an exponential barycenter. □

5.5 *Proof of Theorem 3.3*

Consider y in the support of P, denote $y_\lambda = \gamma_y^+(1+\lambda) = e_\lambda(y)$ and consider the map $\sigma_y : [0,1] \to M$ defined by $\sigma_y(t) = \gamma_y^+(t(1+\lambda))$. By assumption, σ_y is a geodesic connecting x^* to y_λ. In addition, we have by construction that

$$\sigma_y(\tau) = y \quad \text{where} \quad \tau = \frac{1}{1+\lambda}.$$

It follows from the properties of the map $\log_y$ listed in appendix A that

$$y = (1-\tau)x^* + \tau y_\lambda,$$

where we identify a point u and its image $\log_y(u)$ in T_yM. Now since $\mathrm{curv}(M) \geq 0$ (see Proposition A.8), we know that $\mathrm{curv}(T_yM) \geq 0$ so that, for all $x \in T_yM$,

$$\begin{aligned}
\|x-y\|_y^2 &\geq (1-\tau)\|x-x^*\|_y^2 + \tau\|x-y_\lambda\|_y^2 - \tau(1-\tau)\|y_\lambda - x^*\|_y^2 \\
&= \frac{\lambda}{1+\lambda}\|x-x^*\|_y^2 + \frac{1}{1+\lambda}\|x-y_\lambda\|_y^2 - \frac{\lambda}{(1+\lambda)^2}\|y_\lambda - x^*\|_y^2.
\end{aligned}$$

Using the fact that $\|y_\lambda - x^*\|_y^2 = (1+\lambda)^2\|y-x^*\|_y^2$ and the fact that $d(u,v) \leq \|u-v\|_y$, with equality if $u = y$ or $v = y$, we deduce from the inequality above that

$$\begin{aligned}
\frac{\lambda}{1+\lambda} d(x,x^*)^2 &\leq \frac{\lambda}{1+\lambda}\|x-x^*\|_y^2 \\
&\leq d(x,y)^2 - \frac{1}{1+\lambda}\|x-y_\lambda\|_y^2 + \lambda d(x^*,y)^2 \\
&= (d(x,y)^2 - d(x^*,y)^2) - \frac{1}{1+\lambda}\|x-y_\lambda\|_y^2 + (1+\lambda)d(x^*,y)^2.
\end{aligned}$$

Integrating this inequality with respect to $\mathrm{d}P(y)$, it follows that

$$\frac{\lambda}{1+\lambda}d(x,x^*)^2 \leq \int_M (d(x,y)^2 - d(x^*,y)^2)\mathrm{d}P(y) + \int_M \left((1+\lambda)d(x^*,y)^2 - \frac{1}{1+\lambda}\|x-y_\lambda\|_y^2\right)\mathrm{d}P(y).$$

To conclude the proof, it remains to show that, for all $x \in M$,

$$\rho(x) := \int_M \left((1+\lambda)d(x^*,y)^2 - \frac{1}{1+\lambda}\|x-y_\lambda\|_y^2\right)\mathrm{d}P(y) \leq 0.$$

Observing that

$$(1+\lambda)^2 d(x^*,y)^2 = d(x^*,y_\lambda)^2 \quad \text{and that} \quad d(x,y_\lambda) \leq \|x-y_\lambda\|_y,$$

we deduce that

$$\begin{aligned}\rho(x) &\leq \frac{1}{1+\lambda}\int_M (d(x^*,y_\lambda)^2 - d(x,y_\lambda)^2)\mathrm{d}P(y) \\ &= \frac{1}{1+\lambda}\int_M (d(x^*,y)^2 - d(x,y)^2)\mathrm{d}P_\lambda(y),\end{aligned}$$

where we have used the fact that $y_\lambda = e_\lambda(y)$ and the fact that $P_\lambda = (e_\lambda)_* P$. Hence, for all $x \in M$, inequality $\rho(x) \leq 0$ follows from the fact that x^* is a barycenter of P_λ by assumption.

5.6 Proof of Theorem 3.5

We start by a technical lemma. Below H is a Hilbert space with scalar product $\langle .,. \rangle$ and associated norm $\|.\|$.

Lemma 5.9. *Let $\phi : H \to \mathbb{R}$ be a convex function and $\partial\phi \subset H^2$ its subdifferential defined by*

$$(x,y) \in \partial\phi \Leftrightarrow \forall z \in H, \quad \phi(z) \geq \phi(x) + \langle y, z-x\rangle.$$

Then, for all $c > 0$,

$$(x,y) \in \partial\phi \Leftrightarrow \forall z \in H, \quad \phi(z) \geq \phi(x) + \langle y, z-x\rangle - c\|x-z\|^2.$$

Proof of the Lemma. One implication is obvious. For the second implication, suppose $(x,y) \in H^2$ is such that

$$\forall z \in H, \quad \phi(z) \geq \phi(x) + \langle y, z-x\rangle - c\|z-x\|^2. \tag{5.8}$$

Then, on the one hand, we get by convexity of φ that, for all $z \in H$ and all $t \in (0,1)$,

$$\phi(z) - \phi(x) \geq \frac{1}{t}(\phi(tz + (1-t)x) - \phi(x)). \tag{5.9}$$

On the other hand, applying (5.8), we obtain for all $z \in H$ and all $t \in (0,1)$,

$$\phi(tz + (1-t)x) - \phi(x) \geq t\langle y, z-x\rangle - ct^2\|z-x\|^2. \tag{5.10}$$

Hence, combining (5.9) and (5.10), we deduce that for all $z \in H$ and all $t \in (0,1)$,

$$\phi(z) - \phi(x) \geq \langle y, z-x\rangle - ct\|z-x\|^2.$$

Letting $t \to 0$ proves the other implication. □

We are now in position to prove Theorem 3.5. Fix $\mu, \nu \in S = \mathcal{P}_2(H)$ and denote $\gamma : [0,1] \to S$ a constant speed shortest path between μ and ν. By the Knott-Smith optimality criterion (see Theorem 2.12 in [59]), π is an optimal transport plan of (μ, ν) if and only if its support lies in the graph of the subdifferential of a convex function ϕ, i.e.

$$(x,y) \in \mathrm{supp}\pi \Rightarrow \forall z \in H, \quad \phi(z) \geq \phi(x) + \langle y, z - x \rangle. \tag{5.11}$$

Suppose first that, for some $\lambda > 0$, $\gamma : [0,1] \to S$ can be extended to a function $\gamma^+ : [0, 1+\lambda] \to S$ that remains a shortest path between its endpoints $\mu = \gamma^+(0) = \gamma(0)$ and $\nu^\lambda := \gamma^+(1+\lambda)$. Then, by Theorem 7.2.2 in [6], there exists an optimal transport plan π^λ of (μ, ν^λ) such that $\{\pi\} = \Gamma_o(\mu, \nu)$ is that the law of

$$(X,Y) := \left(X, \frac{\lambda}{1+\lambda} X + \frac{1}{1+\lambda} Y^\lambda \right)$$

where $(X, Y^\lambda) \sim \pi^\lambda$. In particular $Y^\lambda = (1+\lambda Y) - \lambda X$. Therefore, there exists a convex function ϕ^λ such that denoting $y^\lambda = (1+\lambda)y - \lambda x$

$$\begin{aligned}
(x,y) \in \mathrm{supp}\pi &\Leftrightarrow (x, y^\lambda) \in \mathrm{supp}\pi^\lambda \\
&\Rightarrow \forall z \in H, \phi^\lambda(z) \geq \phi^\lambda(x) + \langle y^\lambda, z - x \rangle \\
&\Leftrightarrow \forall z \in H, \phi^\lambda(z) \geq \phi^\lambda(x) + (1+\lambda)\langle y, z - x \rangle - \lambda \langle x, z - x \rangle \\
&\Leftrightarrow \forall z \in H, \phi^\lambda(z) + \frac{\lambda}{2}\|z\|^2 \geq \phi^\lambda(x) + \frac{\lambda}{2}\|x\|^2 + (1+\lambda)\langle y, z - x \rangle + \lambda \frac{1}{2}\|x - z\|^2 \\
&\Rightarrow \forall z \in H, \frac{\phi^\lambda(z)}{1+\lambda} + \frac{\lambda \|z\|^2}{2(1+\lambda)} \geq \frac{\phi^\lambda(x)}{1+\lambda} + \frac{\lambda \|x\|^2}{2(1+\lambda)} + \langle y, z - x \rangle \\
&\Leftrightarrow \forall z \in H, \phi(z) \geq \phi(x) + \langle y, z - x \rangle,
\end{aligned} \tag{5.12}$$

where we denote

$$\phi = \frac{\phi^\lambda}{1+\lambda} + \frac{1}{2}\frac{\lambda}{1+\lambda}\|.\|^2.$$

Thus, $\mathrm{supp}\pi$ lies in the subdifferential of ϕ that is $\frac{\lambda}{1+\lambda}$-convex, since ϕ^λ is convex.

Conversely, suppose that there exists a $\frac{\lambda}{1+\lambda}$-convex function ϕ such that $\mathrm{supp}\pi$ for $\pi \in \Gamma_o(\mu, \nu)$ lies in the subdifferential of ϕ. Denote $(X,Y) \sim \pi$ and set $Y^\lambda = Y + \lambda(Y - X) \sim \nu^\lambda$ and $(X, Y^\lambda) \sim \pi^\lambda$. Then π^λ is an optimal transport plan between μ and ν^λ if and only if there exists a convex function ϕ^λ such that the $\mathrm{supp}\pi^\lambda$ lies in $\partial\phi^\lambda$. In that case,

$$W_2^2(\mu, \nu^\lambda) = E\|Y^\lambda - X\|^2 = (1+\lambda)^2 W_2^2(\mu, \nu) = \frac{(1+\lambda)^2}{\lambda^2} W_2^2(\nu, \nu^\lambda),$$

so that by Lemma 7.2.1 of [6], ν is in the shortest path joining μ to ν^λ, which is the desired result. It thus just remains to prove that there exists a convex function ϕ^λ such that $\mathrm{supp}\pi^\lambda$ lies in $\partial\phi^\lambda$.

Set

$$\phi^\lambda = (1+\lambda)\phi - \frac{1}{2}\lambda\|.\|^2.$$

ϕ^λ is convex since ϕ is $\frac{\lambda}{1+\lambda}$-convex. Then, denoting again $y^\lambda = (1+\lambda)y - \lambda x$,

$$\begin{aligned}
(x, y^\lambda) \in \mathrm{supp}\pi^\lambda &\Leftrightarrow (x,y) \in \mathrm{supp}\pi \\
&\Rightarrow \forall z \in H, \phi(z) \geq \phi(x) + \langle y, z - x \rangle \\
&\Leftrightarrow \forall z \in H, \frac{\phi^\lambda(z)}{1+\lambda} + \frac{\lambda\|z\|^2}{2(1+\lambda)} \geq \frac{\phi^\lambda(x)}{1+\lambda} + \frac{\lambda\|x\|^2}{2(1+\lambda)} + \langle y, z - x \rangle \\
&\Leftrightarrow \forall z \in H, \phi^\lambda(z) \geq \phi^\lambda(x) + \langle y^\lambda, z - x \rangle - \lambda\frac{1}{2}\|x - z\|^2.
\end{aligned} \tag{5.13}$$

Since ϕ^λ is convex, by Lemma 5.9, (5.13) is equivalent to

$$\forall z \in H, \phi^\lambda(z) \geq \phi^\lambda(x) + \langle y^\lambda, z - x \rangle,$$

that is, $\mathrm{supp}\pi^\lambda$ lies in the subdifferential of the convex function ϕ^λ.

5.7 *Proof of Theorem 3.7*

For all $x \in M$, denote

$$V(x) = \int_M F(x,y)\,\mathrm{d}P(y).$$

Now fix $x \in M$ and let $\gamma : [0,1] \to M$ be a path connecting x^* to x and along which V is (k,β)-convex. Then, for all $t \in [0,1]$,

$$V(\gamma_t) \leq (1-t)V(x^*) + tV(x) - kt(1-t)d(x,x^*)^{\frac{2}{\beta}}.$$

Reordering these terms we obtain, for all $t \in (0,1)$,

$$\begin{aligned} kd(x,x^*)^{\frac{2}{\beta}} &\leq \frac{V(x)-V(x^*)}{1-t} + \frac{V(x^*)-V(\gamma_t)}{t(1-t)} \\ &\leq \frac{V(x)-V(x^*)}{1-t}, \end{aligned}$$

since $V(x^*) \leq V(\gamma_t)$ by definition of x^*. Letting t tend to 0 concludes the proof.

5.8 *Proof of Theorem 4.1*

For any $\mu \in M \subset \mathcal{P}_2(E)$, let g_μ be the density of μ with respect to the reference measure m. Fix $\nu, \mu, \mu' \in M$ and denote for brevity

$$a_\mu(x) = \frac{g_\mu}{g_\nu}(x), \quad a_{\mu'}(x) = \frac{g_{\mu'}}{g_\nu}(x) \quad \text{and} \quad \delta_{\mu,\mu'}(x) = \frac{f(a_\mu(x)) - f(a_{\mu'}(x))}{a_\mu(x) - a_{\mu'}(x)}.$$

Then, we obtain

$$\begin{aligned} D_f(\mu,\nu) - D_f(\mu',\nu) &= \int (f(a_\mu(x)) - f(a_{\mu'}(x)))g_\nu(x)\mathrm{d}m(x) \\ &= \int \frac{f(a_\mu(x)) - f(a_{\mu'}(x))}{a_\mu(x) - a_{\mu'}(x)}(\mathrm{d}\mu(x) - \mathrm{d}\mu'(x)) \\ &= \int \delta_{\mu,\mu'}(x)(\mathrm{d}\mu(x) - \mathrm{d}\mu'(x)). \end{aligned} \tag{5.14}$$

Let us prove that $\delta_{\mu,\mu'}$ is Lipschitz. To that aim, observe that since f' is L-Lipschitz, for all $x, y \in E$,

$$\begin{aligned} |\delta_{\mu,\mu'}(x) - \delta_{\mu,\mu'}(y)| &= |\int_0^1 \left(f'((1-t)a_\mu(x) + ta_{\mu'}(x)) - f'((1-t)a_\mu(y) + ta_{\mu'}(y))\right)\mathrm{d}t| \\ &\leq \int_0^1 |f'((1-t)a_\mu(x) + ta_{\mu'}(x)) - f'((1-t)a_\mu(y) + ta_{\mu'}(y))|\mathrm{d}t \\ &\leq L\int_0^1 |(1-t)(a_\mu(x) - a_\mu(y)) + t(a_{\mu'}(x) - a_{\mu'}(y))|\mathrm{d}t \\ &\leq L\max\{|a_\mu(x) - a_\mu(y)|, |a_{\mu'}(x) - a_{\mu'}(y)|\}. \end{aligned} \tag{5.15}$$

Then, we see that

$$\begin{aligned} a_\mu(x) - a_\mu(y) &= \frac{g_\mu(x)g_\nu(y) - g_\mu(y)g_\nu(x)}{g_\nu(x)g_\nu(y)} \\ &= \frac{g_\mu(x)(g_\nu(y) - g_\nu(x)) + g_\nu(x)(g_\mu(x) - g_\mu(y))}{g_\nu(x)g_\nu(y)}, \end{aligned}$$

which implies, under the conditions of the theorem, that

$$|a_\mu(x) - a_\mu(y)| \leq \frac{2\Lambda c_+}{c_-^2}\|x - y\|. \tag{5.16}$$

Combining (5.15) and (5.16) we deduce that $\delta_{\mu,\mu'}$ is Lipschitz with constant at most $2L\Lambda c_+/c_-^2$. Using (5.14) and the Kantorovich-Rubinstein formula (see remark 6.5 in [60]) we therefore obtain that

$$|D_f(\mu,\nu) - D_f(\mu',\nu)| \leq \frac{2L\Lambda c_+}{c_-^2} W_1(\mu,\mu').$$

The result follows by observing that $W_1(\mu,\mu') \leq W_2(\mu,\mu')$ (see remark 6.6 in [60]).

5.9 *Proof of Theorem 4.2*

First note that the second statement of the theorem follows by Theorem 3.7. Hence, we need only to prove that, for all $\mu_0, \mu_1 \in M$, there exists a geodesic μ_t connecting μ_0 to μ_1 such that, for all $0 \leq t \leq 1$,

$$D_f(\mu_t,\nu) \leq (1-t)D_f(\mu_0,\nu) + tD_f(\mu_1,\nu) - c\lambda t(1-t)W_2(\mu_0,\mu_1)^2.$$

If either μ_0 or μ_1 is not absolutely continuous with respect to ν, then the right hand side is $+\infty$ and the inequality trivially holds. Suppose now that $\mu_0 \ll \nu$ and $\mu_1 \ll \nu$. Since ν, and therefore μ_0, has a density with respect to the Lebesgue measure, there exists an optimal transport map $T : E \to E$ pushing μ_0 to μ_1. Letting $T_t(x) = (1-t)x + tT(x)$, the curve $\mu_t = (T_t)_{\#}\mu_0$ defines a geodesic connecting μ_0 to μ_1 in M. For all $t \in [0,1]$, we denote

$$\rho_t(x) = \frac{\mathrm{d}\mu_t}{\mathrm{d}x}(x),$$

so that, letting $\mathrm{d}\nu(x) = e^{-U(x)}\mathrm{d}x$,

$$\mathrm{d}\mu_t(x) = \rho_t(x)e^{U(x)}\mathrm{d}\nu(x).$$

Letting $DT_t(x) = (1-t)I + tDT(x)$ denote the differential of T_t at x, the change of variables formula

$$\rho_0(x) = \rho_t(T_t(x))\det(DT_t(x)),$$

implies that

$$\begin{aligned} D_f(\mu_t,\nu) &= \int f(\rho_t(x)e^{U(x)})e^{-U(x)}\,\mathrm{d}x \\ &= \int f\left(\frac{\rho_0(x)e^{U(T_t(x))}}{\det(DT_t(x))}\right)\frac{\det(DT_t(x))}{\rho_0(x)e^{U(T_t(x))}}\rho_0(x)\,\mathrm{d}x \\ &= \int h(s(t,x))\rho_0(x)\,\mathrm{d}x, \end{aligned}$$

where

$$s(t,x) = -U(T_t(x)) + \log\det(DT_t(x)) - \log\rho_0(x). \tag{5.17}$$

The transport map T being the gradient of a convex function, the map $t \in [0,1] \mapsto \log\det(DT_t(x))$ is concave for any fixed $x \in E$. The λ-convexity of U therefore implies that, for all $t \in [0,1]$,

$$s(t,x) \geq (1-t)s(0,x) + ts(1,x) + \lambda t(1-t)\|x - \mathrm{T}(x)\|^2.$$

Using the assumptions on h, which imply in particular that it is decreasing, we deduce that

$$\begin{aligned} h(s(t,x)) &\leq h((1-t)s(0,x) + ts(1,x) + \lambda t(1-t)\|x-\mathrm{T}(x)\|^2) \\ &\leq h((1-t)s(0,x) + ts(1,x)) - c\lambda t(1-t)\|x-\mathrm{T}(x)\|^2 \\ &\leq (1-t)h(s(0,x)) + th(s(1,x)) - c\lambda t(1-t)\|x-\mathrm{T}(x)\|^2. \end{aligned}$$

Integrating the last inequality with respect to $\rho_0(x)\mathrm{d}x$ yields

$$D_f(\mu_t,\nu) \leq (1-t)D_f(\mu_0,\nu) + tD_f(\mu_1,\nu) - c\lambda t(1-t)W_2^2(\mu_0,\mu_1),$$

which is the desired result.

5.10 *Proof of Theorem 4.3*

Denote $\mathrm{d}\mu_i = g_i \mathrm{d}\nu$. The linear interpolation $\ell_t = (1-t)\mu_0 + t\mu_1$ therefore satisfies

$$\frac{\mathrm{d}\ell_t}{\mathrm{d}\nu} = (1-t)g_0 + tg_1.$$

It follows from the k-convexity of f that

$$\begin{aligned} D_f(\ell_t, \nu) &= \int f\left(\frac{\mathrm{d}\ell_t}{\mathrm{d}\nu}\right) \mathrm{d}\nu \\ &= \int f((1-t)g_0 + tg_1)\,\mathrm{d}\nu \\ &\leq (1-t)\int f(g_0)\,\mathrm{d}\nu + t\int f(g_1)\,\mathrm{d}\nu - kt(1-t)\int |g_1 - g_0|^2\,\mathrm{d}\nu \\ &= (1-t)D_f(\mu_0,\nu) + tD_f(\mu_1,\nu) - kt(1-t)\int |g_1 - g_0|^2\,\mathrm{d}\nu. \end{aligned}$$

According to Theorem 6.15 in [60], we know that

$$\begin{aligned} W_2(\mu_0,\mu_1)^2 &\leq 2\int \|x\|^2 |g_1(x) - g_0(x)|\,\mathrm{d}\nu(x) \\ &\leq 2m_4(\nu)^{1/2}\left(\int |g_1(x) - g_0(x)|^2\,\mathrm{d}\nu(x)\right)^{1/2}. \end{aligned}$$

Hence, we deduce that

$$D_f(\ell_t,\nu) \leq (1-t)D_f(\mu_0,\nu) + tD_f(\mu_1,\nu) - \frac{k}{4m_4(\nu)}t(1-t)W_2(\mu_0,\mu_1)^4,$$

which proves the first statement. The second statement follows directly from Theorem 3.7.

5.11 *Proof of Theorem 4.4*

For fixed $y \in E$, the fact that $x \in E \mapsto g(x,y)$ is L-Lipschitz, the Kantorovich-Rubinstein formula (see remark 6.5 in [60]) and the fact that $W_1 \leq W_2$ (see remark 6.6 in [60]) implies that, for all $\mu, \mu' \in \mathcal{P}_2(E)$,

$$\left|\int g(x,y)(\mathrm{d}\mu(x) - \mathrm{d}\mu'(x))\right| \leq LW_1(\mu,\mu') \leq LW_2(\mu,\mu').$$

Integrating with respect to $\nu \in \mathcal{P}_2(E)$ implies that

$$\begin{aligned} |I_g(\mu,\nu) - I_g(\mu',\nu)| &\leq \int \left|\int g(x,y)(\mathrm{d}\mu(x) - \mathrm{d}\mu'(x))\right| \mathrm{d}\nu(y) \\ &\leq LW_2(\mu,\mu'). \end{aligned}$$

5.12 *Proof of Theorem 4.5*

Given that (E,ρ) is separable, complete, locally compact and geodesic, the space $(\mathcal{P}_2(E), W_2)$ is also geodesic according to Corollary 7.22 in [60]. In addition, given $\mu_0, \mu_1 \in \mathcal{P}_2(E)$, a curve $(\mu_t)_{0\leq t\leq 1}$ in $\mathcal{P}_2(E)$ is a geodesic connecting μ_0 to μ_1 if and only if there exists a probability measure Π on the set $\mathcal{G}(E)$ of all geodesics $\gamma : [0,1] \to E$ in E such that

$$\mu_t = (\mathrm{eval}_t)_*(\Pi),$$

where $\mathrm{eval}_t(\gamma) = \gamma_t$, for all $\gamma \in \mathcal{G}(E)$, and where $(\mathrm{eval}_0, \mathrm{eval}_1)_*(\Pi)$ is an optimal coupling of μ_0 and μ_1. For such a geodesic,

$$
\begin{aligned}
I_g(\mu_t,\nu) &= \int\int g(x,y)\,\mathrm{d}\mu_t(x)\mathrm{d}\nu(y)\\
&= \int\int g(\gamma_t,y)\,\mathrm{d}\Pi(\gamma)\mathrm{d}\nu(y)\\
&\le \int\int \left((1-t)g(\gamma_0,y)+tg(\gamma_1,y)-kt(1-t)\rho(\gamma_0,\gamma_1)^{\frac{2}{\beta}}\right)\mathrm{d}\Pi(\gamma)\mathrm{d}\nu(y)\\
&= (1-t)I_g(\mu_0,\nu)+tI_g(\mu_1,\nu)-kt(1-t)\int \rho(\gamma_0,\gamma_1)^{\frac{2}{\beta}}\,\mathrm{d}\Pi(\gamma)\\
&\le (1-t)I_g(\mu_0,\nu)+tI_g(\mu_1,\nu)-kt(1-t)\left(\int \rho(\gamma_0,\gamma_1)^2\,\mathrm{d}\Pi(\gamma)\right)^{\frac{1}{\beta}}\\
&= (1-t)I_g(\mu_0,\nu)+tI_g(\mu_1,\nu)-kt(1-t)W_2(\mu_0,\mu_1)^{\frac{2}{\beta}}.
\end{aligned}
$$

This completes the proof of (1). Statement (2) follows directly by combining the first statement and Theorem 3.7.

A Metric geometry

A.1 Geodesic spaces

Let (M,d) be a metric space. We call path in M a continuous map $\gamma : I \to M$ defined on an interval $I \subset \mathbb{R}$. The length $L(\gamma) \in [0,+\infty]$ of a path $\gamma : I \to M$ is defined by

$$L(\gamma) := \sup \sum_{i=0}^{n-1} d(\gamma(t_i),\gamma(t_{i+1})),$$

where the supremum is taken over all $n \ge 1$ and all $t_0 \le \cdots \le t_n$ in I. A path is called rectifiable if it has finite length. Two paths γ_1 and γ_2 are said to be equivalent if $\gamma_1 \circ \varphi_1 = \gamma_2 \circ \varphi_2$ for non-decreasing and continuous functions φ_1 and φ_2. In this case, γ_1 is said to be a reparametrisation of γ_2 and we check that $L(\gamma_1) = L(\gamma_2)$. A path $\gamma : [a,b] \to M$ is said to have constant speed if for all $a \le s \le t \le b$,

$$L(\gamma_{[s,t]}) = \frac{t-s}{b-a}L(\gamma), \tag{A.1}$$

where $\gamma_{[s,t]}$ denotes the restriction of γ to $[s,t]$.

Proposition A.1. *Any rectifiable path has a constant speed reparametrisation $\gamma : [0,1] \to M$.*

Given $x, y \in M$, a path $\gamma : [a,b] \to M$ is said to connect x to y if $\gamma(a) = x$ and $\gamma(b) = y$. By construction of the length function L, $d(x,y) \le L(\gamma)$ for any path γ connecting x to y. The space M is called a length space if, for all $x, y \in M$,

$$d(x,y) = \inf_{\gamma} L(\gamma), \tag{A.2}$$

where the infimum is taken over all paths γ connecting x to y. A length space is said to be a geodesic space if, for all $x, y \in M$, the infimum on the right hand side of (A.2) is attained.

Definition A.2. *In a geodesic space, we call geodesic between x and y any constant speed reparametrisation $\gamma : [0,1] \to M$ of a path attaining the infimum in (A.2).*

For a geodesic γ, it follows from its minimising properties that

$$d(\gamma(s),\gamma(t)) = L(\gamma_{[s,t]}),$$

for all $0 \le s \le t \le 1$. In particular, (A.1) translates in this case as

$$d(\gamma(s),\gamma(t)) = (t-s)d(\gamma(0),\gamma(1)),$$

for all $0 \le s \le t \le 1$. We end by a general characterization of geodesic spaces.

Proposition A.3. *Let (M, d) be a metric space.*

(1) *If M is a geodesic space, then any two points $x, y \in M$ admit a midpoint, i.e. a point $z \in M$ such that*
$$d(x, z) = d(y, z) = \frac{1}{2} d(x, y).$$

(2) *Conversely, if M is complete and if any two points in M admit a midpoint, then M is a geodesic space.*

A.2 Model spaces

Given a real number $\kappa \in \mathbb{R}$, a geodesic space of special interest is the (complete and simply connected) 2-dimensional Riemannian manifold with constant sectional curvature κ. For given $\kappa \in \mathbb{R}$, this metric space (M^2_κ, d_κ) is unique up to an isometry, and modelled as follows.

- If $\kappa < 0$, (M^2_κ, d_κ) is the hyperbolic plane with metric multiplied by $1/\sqrt{-\kappa}$.
- If $\kappa = 0$, (M^2_0, d_0) is the Euclidean plane equipped with its Euclidean metric.
- If $\kappa > 0$, (M^2_κ, d_κ) is the Euclidean sphere in $\mathbb{R}^3$ of radius $1/\sqrt{\kappa}$ with the angular metric.

The diameter ϖ_κ of M^2_κ is
$$\varpi_\kappa := \begin{cases} +\infty & \text{if} \quad \kappa \leq 0, \\ \pi/\sqrt{\kappa} & \text{if} \quad \kappa > 0. \end{cases}$$

For $\kappa \in \mathbb{R}$, there is a unique geodesic connecting x to y in (M^2_κ, d_κ) provided $d_\kappa(x, y) < \varpi_\kappa$. By convention, we call triangle in M^2_κ any set of three distinct points $\{p, x, y\} \subset M^2_\kappa$, with perimeter

$$\mathrm{peri}\{p, x, y\} := d_\kappa(p, x) + d_\kappa(p, y) + d_\kappa(x, y) < 2\varpi_\kappa.$$

Side lengths of triangle $\{p, x, y\}$ are the numbers $d_\kappa(p, x)$, $d_\kappa(p, y)$ and $d_\kappa(x, y)$. Given $a, b, c > 0$ satisfying the triangle inequality and such that $a + b + c < 2\varpi_\kappa$, there exists a unique (up to an isometry) triangle $\{p, x, y\}$ in M^2_κ such that $d_\kappa(p, x) = a$, $d_\kappa(p, y) = b$ and $d_\kappa(x, y) = c$. The angle $\sphericalangle^\kappa_p(x, y)$ at p in $\{p, x, y\} \subset M^2_\kappa$ is defined by

$$\cos \sphericalangle^\kappa_p(x, y) := \begin{cases} \dfrac{a^2 + b^2 - c^2}{2ab} & \text{if } \kappa = 0, \\ \dfrac{c_\kappa(c) - c_\kappa(a) \cdot c_\kappa(b)}{\kappa \cdot s_\kappa(a) s_\kappa(b)} & \text{if } \kappa \neq 0, \end{cases}$$

where $a = d_\kappa(p, x)$, $b = d_\kappa(p, y)$, $c = d_\kappa(x, y)$ and $c_\kappa := s'_\kappa$ with

$$s_\kappa(r) := \begin{cases} \sin(r\sqrt{\kappa})/\sqrt{\kappa} & \text{if } \kappa > 0, \\ \sinh(r\sqrt{-\kappa})/\sqrt{-\kappa} & \text{if } \kappa < 0. \end{cases} \tag{A.3}$$

We end by observing that the angle is constant along geodesics in the model space (M^2_κ, d_κ).

Proposition A.4. *Let $\kappa \in \mathbb{R}$ and $\{p, x, y\} \subset (M^2_\kappa, d_\kappa)$ be a triangle. If γ_x and γ_y are geodesics from p to x and from p to y respectively, then for all $(s, t) \in (0, 1]^2$,*

$$\sphericalangle^\kappa_p(\gamma_x(s), \gamma_y(t)) = \sphericalangle^\kappa_p(x, y).$$

A.3 Curvature

In this section, we describe the notion of curvature bounds of metric spaces. Curvature bounds in general metric spaces are defined by comparison arguments involving the model surfaces (M^2_κ, d_κ) discussed in the previous section. The fundamental device allowing for this comparison is that of a comparison triangle. Given a metric space (M, d), we define a triangle in M as any set of three points $\{p, x, y\} \subset M$. For $\kappa \in \mathbb{R}$, a comparison triangle for $\{p, x, y\}$ in M^2_κ is an isometric embedding of $\{p, x, y\}$ in M^2_κ, i.e. a set $\{p_\kappa, x_\kappa, y_\kappa\} \subset M^2_\kappa$ such that

$$d_\kappa(p_\kappa, x_\kappa) = d(p, x), \quad d_\kappa(p_\kappa, y_\kappa) = d(p, y) \quad \text{and} \quad d_\kappa(x_\kappa, y_\kappa) = d(x, y).$$

Such a comparison triangle always exists (and is unique up to an isometry) provided

$$\mathrm{peri}\{p, x, y\} := d(p, x) + d(p, y) + d(x, y) < 2\varpi_\kappa.$$

We are now in position to define curvature bounds for geodesic spaces.

Definition A.5. *Let $\kappa \in \mathbb{R}$ and (M, d) be a geodesic space.*

(1) *We say that* $\mathrm{curv}(M) \geq \kappa$ *if for any triangle $\{p, x, y\} \subset M$ satisfying* $\mathrm{peri}\{p, x, y\} < 2\varpi_\kappa$*, any comparison triangle $\{p_\kappa, x_\kappa, y_\kappa\} \subset M^2_\kappa$, any geodesic γ joining x to y in M and any geodesic γ_κ joining x_κ to y_κ in M^2_κ , we have for all $t \in [0, 1]$,*

$$d(p, \gamma(t)) \geq d_\kappa(p_\kappa, \gamma_\kappa(t)). \tag{A.4}$$

(2) *We say that* $\mathrm{curv}(M) \leq \kappa$ *if the above definition holds with opposite inequality in* (A.4).

The previous definition has a natural geometric interpretation: if $\mathrm{curv}(M) \geq \kappa$ (resp. $\mathrm{curv}(M) \leq \kappa$) a triangle $\{p, x, y\}$ looks thicker (resp. thiner) than a corresponding comparison triangle $\{p_\kappa, x_\kappa, y_\kappa\}$ in M^2_κ. In the context of $\kappa = 0$, the above definition may be given an alternative form of practical interest.

Proposition A.6. *Let (M, d) be a geodesic space. Then* $\mathrm{curv}(M) \geq 0$ *if, and only if, for any points $p, x, y \in M$ and any geodesic γ joining x to y, we have*

$$\forall t \in [0, 1], \quad d(p, \gamma(t))^2 \geq (1 - t)d(p, x)^2 + td(p, y)^2 - t(1 - t)d(x, y)^2.$$

We have $\mathrm{curv}(M) \leq 0$ *if, and only if, the same statement holds with opposite inequality.*

The proof follows immediately from Definition A.5 by exploiting the geometry of the Euclidean plane. Note indeed that, whenever $\{p, x, y\} \subset \mathbb{R}^2$ and $\mathbb{R}^2$ is equipped with the Euclidean metric $\|. - .\|$, the unique geodesic from x to y is $\gamma(t) = (1 - t)x + ty$ and, for all $t \in [0, 1]$,

$$\|p - \gamma(t)\|^2 = (1 - t)\|p - x\|^2 + t\|p - y\|^2 - t(1 - t)\|x - y\|^2.$$

For $\kappa \neq 0$, an equivalent formulation of Definition A.5, given only in terms of the ambient metric d, is given in the next subsection using the notion of angle. A (complete) geodesic space (M, d) with $\mathrm{curv}(M) \leq \kappa$ for some $\kappa \geq 0$ is sometimes called a CAT(κ) space in reference to contributions of E. Cartan, A.D. Alexandrov and V.A. Toponogov. A CAT(0) space is also referred to as an NPC (non positively curved) space or an Hadamard space. Similarly, M is also called an PC (positively curved) space if $\mathrm{curv}(M) \geq 0$. If (M, d) is a Riemannian manifold (complete for instance) with sectional curvature lower (resp. upper) bounded by κ at every point, then $\mathrm{curv}(M) \geq \kappa$ (resp $\leq \kappa$) in the sense of Definition A.5. It is worth noting that the previous definitions are of global nature as they require comparison inequalities to be valid for all triangles (that admit a comparison triangle in the relevant model space). Some definitions of curvature require the previous comparison inequalities to hold only locally. The local validity of these comparison inequalities is known, under suitable conditions depending on the value of κ, to imply their global validity. Results in this direction are known as globalisation theorems.

A.4 Angles and space of directions

Angles, as defined below, allow to provide alternative characterisations of curvature bounds. Let (M,d) be a metric space and let $\kappa \in \mathbb{R}$. Given a triangle $\{p,x,y\}$ in M with $\mathrm{peri}\{p,x,y\} < 2\varpi_\kappa$, we define the comparison angle $\sphericalangle_p^\kappa(x,y) \in [0,\pi]$ at p by

$$\cos \sphericalangle_p^\kappa(x,y) := \begin{cases} \dfrac{d(p,x)^2 + d(p,y)^2 - d(x,y)^2}{2d(p,x)d(p,y)} & \text{if } \kappa = 0, \\ \dfrac{c_\kappa(d(x,y)) - c_\kappa(d(p,x)) \cdot c_\kappa(d(p,y))}{\kappa \cdot s_\kappa(d(p,x)) s_\kappa(d(p,y))} & \text{if } \kappa \neq 0, \end{cases}$$

where c_κ and s_κ are as in (A.3). In other words, given any comparison triangle $\{p_\kappa, x_\kappa, y_\kappa\}$ of $\{p,x,y\}$ in M^2_κ,

$$\sphericalangle_p^\kappa(x,y) = \sphericalangle_{p_\kappa}^\kappa(x_\kappa, y_\kappa).$$

This allows to give an equivalent definition of curvature lower bounds that has the advantage of making sense on arbitrary metric spaces, not necessarily geodesic.

Definition A.7 (Quadruple comparison)**.** *Let (M,d) be a metric space. Let $\kappa \in \mathbb{R}$. We say that* $\mathrm{curv}(M) \geq \kappa$*, if for any four disctinct points $p,x,y,z \in M$ such that every three points have a perimeter less than ϖ_κ,*

$$\sphericalangle_p^\kappa(x,y) + \sphericalangle_p^\kappa(y,z) + \sphericalangle_p^\kappa(z,x) \leq 2\pi.$$

Proposition A.8. *If (M,d) is a geodesic space, then curvature lower bounds as defined in Definition A.5 and Definition A.7 are equivalent. Moreover, even if (M,d) is not geodesic, and satisfies Definition A.7, then, equation (A.4) is satisfied for all $t \in [0,1]$ and $\gamma(t) \in M$ such that*

$$d(\gamma(0),\gamma(t))/t = d(\gamma(t),\gamma(1))/(1-t) = d(\gamma(1),\gamma(0)).$$

The next result presents a characterization of curvature bounds in terms of the monotonicity of the comparison angle.

Proposition A.9 (Angle monotonicity)**.** *Let (M,d) be a geodesic space and let $\kappa \in \mathbb{R}$. Then* $\mathrm{curv}(M) \geq \kappa$ *(resp.* $\mathrm{curv}(M) \leq \kappa$*), in the sense of Definition A.5, if and only if, for any triangle $\{p,x,y\}$ in M and any geodesics γ_x and γ_y from p to x and from p to y respectively, the function*

$$(s,t) \in [0,1]^2 \mapsto \sphericalangle_p^\kappa(\gamma_x(s), \gamma_y(t)),$$

is non-increasing (resp. non-decreasing) in each variable when the other is fixed.

In a geodesic space (M,d), if $p \in M$ and γ_x and γ_y are two geodesics connecting p to x and y respectively, we define

$$\sphericalangle^\kappa(\gamma_x,\gamma_y) := \lim_{s,t\to 0} \sphericalangle_p^\kappa(\gamma_x(s),\gamma_y(t)),$$

when it exists. It follows from Proposition A.9 that this limit exists provided $\mathrm{curv}(M) \geq \kappa$ or $\mathrm{curv}(M) \leq \kappa$. It may be shown furthermore that this limit is independent of κ. Hence, whenever M has upper or lower curvature bound, we denote $\sphericalangle(\gamma_x,\gamma_y)$ the angle between these two geodesics. Given a third geodesic $\gamma_z : [0,1] \to S$ such that $\gamma_z(0) = p$ and $\gamma_z(1) = z$, we have the triangular inequality

$$\sphericalangle(\gamma_x,\gamma_y) \leq \sphericalangle(\gamma_x,\gamma_z) + \sphericalangle(\gamma_z,\gamma_y), \tag{A.5}$$

so that $\sphericalangle$ defines a pseudo metric on the set $G(p)$ of all geodesics emanating from p. Defining the equivalence relation $\sim$ on $G(p)$ by $\alpha \sim \beta \Leftrightarrow \sphericalangle(\alpha,\beta) = 0$, the angle $\sphericalangle$ induces a metric (still denoted $\sphericalangle$) on the quotient set $G(p)/\sim$ and we call space of directions the completion $(\Sigma_p, \sphericalangle)$ of $(G(p)/\sim, \sphericalangle)$. An element of Σ_p is called a direction.

A.5 Tangent cones

Metric spaces considered so far have a priori no differentiable structure. In this context, an analog of a tangent space is provided by the notion of a tangent cone. This section shortly reviews this notion. Below, M denotes a geodesic space with lower or upper bounded curvature in the sense of Definition A.5.

Definition A.10 (Tangent cone)**.** *Let $p \in M$. The tangent cone T_pM at p is the Euclidean cone over the space of directions $(\Sigma_p, \sphericalangle)$. In other words, T_pM is the metric space:*

- *Whose underlying set consists in equivalent classes in $\Sigma_p \times [0, +\infty)$ for the equivalence relation $\sim$ defined by*
$$(\alpha, s) \sim (\beta, t) \Leftrightarrow ((s = 0 \text{ and } t = 0) \text{ or } (s = t \text{ and } \alpha = \beta)).$$
A point in T_pM is either the tip of the cone o_p, i.e. the class $\Sigma_p \times \{0\}$, or a couple $(\alpha, s) \in \Sigma_p \times (0, +\infty)$ (identified to the class $\{(\alpha, s)\}$).
- *Whose metric d_p is defined (without ambiguity) by*
$$d_p((\alpha, s), (\beta, t)) := \sqrt{s^2 + t^2 - 2st \cos \sphericalangle(\alpha, \beta)}.$$

For $u = (\alpha, s)$ and $v = (\beta, t) \in T_pM$, we often denote $\|u - v\|_p := d_p(u, v)$, $\|u\|_p := d_p(o_p, u) = s$ and
$$\langle u, v \rangle_p := st \cos \sphericalangle(\alpha, \beta) = (\|u\|_p^2 + \|v\|_p^2 - \|u - v\|_p^2)/2.$$

Proposition A.11. *Let (M, d) be a geodesic space and $p \in M$ be fixed. If $\mathrm{curv}(M) \geq \kappa$ for some $\kappa \in \mathbb{R}$, then T_pM is a metric space with $\mathrm{curv}(T_pM) \geq 0$ (in the sense of Definition A.7).*

Note that the tangent space is not always a geodesic space, see discussion before Proposition 28 in [62] and the Proposition itself for the proof. Notation $\|.\|_p$ and $\langle ., . \rangle_p$ introduced above is justified by the fact that the cone T_pM possesses a Hilbert-like structure described as follows. For a point $u = (\alpha, t)$ and $\lambda \geq 0$, we define $\lambda \cdot u := (\alpha, \lambda t)$. Then, the sum of points $u, v \in T_pM$ is defined as the mid-point of $2 \cdot u$ and $2 \cdot v$ as defined in Definition A.3. Finally, it may be checked using the previous definitions that, for any $u, v \in T_pM$ and any $\lambda \geq 0$, we get
$$\|\lambda \cdot u\|_p = \lambda \|u\|_p \quad \text{and} \quad \langle \lambda \cdot u, v \rangle_p = \langle u, \lambda \cdot v \rangle_p = \lambda \langle u, v \rangle_p.$$

Next we define logarithmic maps. Fix $p \in M$. Since M has upper or lower bounded curvature, the angle monotonicity imposes the following observation. If there are two geodesics γ_x^1 and γ_x^2 connecting p to x such that $\sphericalangle(\gamma_x^1, \gamma_x^2) = 0$, then $\gamma_x^1 = \gamma_x^2$. In other words, the set of points $x \in M$ for which there is not only one equivalence class of geodesics connecting p to x is exactly the set of points x connected to p by at least two distinct geodesics. This set of points is denoted $C(p)$ and called the cut-locus of p. For all $x \in M \setminus C(p)$, we denote $\uparrow_p^x$ the direction of the unique geodesic $\gamma_x : [0, 1] \to M$ connecting p to x.

Definition A.12 (Logarithmic map)**.** *Let (M, d) be a geodesic space with curvature bounded from above or below in the sense of Definition A.5 and fix $p \in M$. We denote*
$$\log_p : x \in M \setminus C(p) \mapsto (\uparrow_p^x, d(p, x)) \in T_pM.$$

For all $t \in [0, 1]$ and all $x \in M \setminus C(p)$, one checks in particular that
$$\log_p(\gamma_x(t)) = t \cdot \log_p(x).$$

More generally, if $\gamma : [0, 1] \to M$ is a geodesic in M and if we denote $p = \gamma(t)$ for some $t \in [0, 1]$, then provided $\gamma(0) = x$ and $\gamma(1) = y$ both belong to $M \setminus C(p)$, we check that, for all $s \in [0, 1]$,
$$\log_p(\gamma(s)) = (1 - s) \cdot \log_p(x) + s \cdot \log_p(y).$$

Note finally that $C(p)$ is empty for spaces of curvature bounded from above by 0. It is often practical to extend the definition of $\log_p$ to the cut locus $C(p)$. To that aim, it suffices to select, for any $x \in C(p)$, a geodesic connecting p to x. Denoting as above $\uparrow_p^x$ the direction associated to this chosen geodesic, we simply extend $\log_p$ to M by setting $\log_p(x) = (\uparrow_p^x, d(p,x))$. This extension depends a priori on the choice of geodesics connecting p to $x \in C(p)$. However, note that when (M,d) is a Polish space, this extension $\log_p$ can be chosen to be measurable on M, by application of the measurable selection theorem of Kuratowski and Ryll-Nardzewski (see Theorem 6.9.3 in [16]). Next is a fundamental result.

Proposition A.13. *Let (M,d) be a geodesic space and $p \in M$ be fixed. If $\mathrm{curv}(M) \geq 0$, then for all $x, y \in M$,*

$$d(x,y) \leq \| \log_p(x) - \log_p(y) \|_p,$$

with equality if $x = p$ or $y = p$.

Acknowledgements

We would like to thank the associate editor, as well as two anonymous referees, for valuable comments that helped improve significantly the original version of the manuscript. We also express our gratitude to Philippe Rigollet for stimulating conversations and useful remarks on a preliminary version of the paper.

References

[1] B. Afsari. Riemannian L^p center of mass: existence, uniqueness and convexity. *Proceedings of the American Mathematical Society*, 139(2):655–673, 2011.

[2] M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[3] S. Alexander, V. Kapovitch, and A. Petrunin. *Alexandrov geometry.* Book in preparation, 2017.

[4] J. Altschuler, J. Weed, and P. Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. *Advances in Neural Information Processing*, 2017.

[5] P. C. Álvarez-Esteban, E. del Barrio, J. Cuesta-Albertos, and C. Matrán. A fixed-point approach to barycenters in wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016.

[6] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures.* Springer, 2008.

[7] L. Ambrosio and P. Tilli. *Topics on analysis on metric spaces.* Oxford University Press, 2004.

[8] K. Ball, E. Carlen, and E. Lieb. Sharp uniform convexity and smoothness inequalities for trace norms. *Inventiones Mathematicae*, 115:463–482, 1995.

[9] P. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33:1497–1537, 2005.

[10] P. Bartlett, M. Jordan, and J. McAuliffe. Convexity, classification and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 2006.

[11] P. Bartlett and S. Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135:311–334, 2006.

[12] R. Bhattacharya and V. Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifold - i. *The Annals of Statistics*, 31(1):1–29, 2003.

[13] R. Bhattacharya and V. Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifold - ii. *The Annals of Statistics*, 33(3):1225–1259, 2005.

[14] J. Bigot, R. Gouet, T. Klein, and A. López. Upper and lower risk bounds for estimating the wasserstein barycenter of random measures on the real line. *Electronic Journal Statistics*, 12(2):2253–2289, 2018.

[15] G. Blanchard, G. Lugosi, and N. Vayatis. On the rate of convergence of regularized boosting classifiers. *Journal of Machine Learning Research*, 4:861–894, 2003.

[16] V. Bogachev. *Measure Theory*. Springer-Verlag, Berlin Heidelberg, 2007.

[17] E. Boissard, T. Le Gouic, and J.-M. Loubes. Distribution's template estimate with wasserstein metrics. *Bernoulli*, 21(2):740–759, 2015.

[18] F. Bolley, A. Guillin, and C. Villani. Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probability Theory and Related Fields*, 137:541–593, 2007.

[19] O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus de l'Académie des Sciences de Paris*, 334:495–500, 2002.

[20] M. Bridson and A. Häfliger. *Metric spaces of nonpositive curvature*. Springer, 1999.

[21] D. Burago, Y. Burago, and S. Ivanov. *A course in metric geometry*. American Mathematical Society, 2001.

[22] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

[23] M. Cuturi and G. Peyré. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

[24] P. Dvurechensky, A. Gasnikov, and A. Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. *arxiv:1802.04367*, 2018.

[25] M. Émery and G. Mokobodzki. Sur le barycentre d'une probabilité dans une variété. In *Séminaire de probabilités de Strasbourg*, volume 25, pages 220–233. Springer, 1991.

[26] A. Forrow, J.-C. Hütter, M. Nitzan, P. Rigollet, G. Schiebinger, and J. Weed. Statistical optimal transport via factored couplings. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2454–2465, 2019.

[27] M. Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l'Institut Henry Poincaré, Section B, Probabilités et Statistiques*, 10:235–310, 1948.

[28] E. Giné and R. Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge University Press, 2015.

[29] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 2672–2680. MIT Press, 2014.

[30] S. Graf and H. Luschgy. *Foundations of quantization for probability distributions*. Springer-Verlag, New-York, 2000.

[31] H. Karcher. Riemannian center of mass and so called karcher mean. *arXiv*, 1407.2087, 2014.

[32] W. Kendall and H. Le. Limit theorems for empirical Fréchet means of independent and non-identically distributed manifold-valued random variables. *Brazilian Journal of Probability and Statistics*, 25(3):323–352, 2011.

[33] B. Kloeckner. A geometric study of Wasserstein spaces: Euclidean spaces. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze-Serie V*, 9(2):297–323, 2010.

[34] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34:2593–2656, 2006.

[35] V. Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems.* Springer, 2011.

[36] A. Kroshnin, V. Spokoiny, and A. Suvorikova. Statistical inference for Bures-Wasserstein barycenters. *arXiv:1901.00226*, 2019.

[37] L. M. Le Cam. *Asymptotic Methods in Statistical Decision Theory.* Springer, 1986.

[38] T. Le Gouic and J.-M. Loubes. Existence and consistency of Wasserstein barycenters. *Probability Theory and Related Fields*, 168(3-4):901–917, 2017.

[39] J. Lott and C. Villani. Ricci curvature for metric-measure spaces via optimal transport. *Annals of Mathematics*, 169:903–991, 2009.

[40] G. Lugosi and S. Mendelson. Near-optimal mean estimators with respect to general norms. *Probability Theory and Related Fields*, To appear, 2019.

[41] E. Mammen and A. Tsybakov. Smooth discriminant analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

[42] P. Massart. Some applications of concentration inequalities to statistics. *Annales de la Faculté des Sciences de Toulouse*, 9:245–303, 2000.

[43] S. Mendelson. Learning without concentration. *Journal of the ACM*, 62(3):1–25, 2015.

[44] S.-I. Ohta. Convexities of metric spaces. *Geometriae Dedicata*, 125(1):225–250, 2007.

[45] S.-I. Ohta. Barycenters in Alexandrov spaces of curvature bounded below. *Advances in geometry*, 14:571–587, 2012.

[46] B. Pelletier. Informative barycentres in statistics. *Annals of the Institute of Statistical Mathematics*, 57(4):767–780, 2005.

[47] A. Rakhlin, K. Sridharan, and A. Tsybakov. Empirical entropy, minimax regret and minimax risk. *Bernoulli*, 23(2):789–824, 2017.

[48] P. Rigollet and J.-C. Hütter. High dimensional statistics. *MIT lecture notes (unpublished)*, 2017.

[49] F. Santambrogio. *Optimal transport for applied mathematicians.* Birkhauser, 2015.

[50] C. Schötz. Convergence rates for the generalized Fréchet mean via the quadruple inequality. *arXiv:1812.08037*, 2018.

[51] K.-T. Sturm. Probability measures on metric spaces of nonpositive curvature. *Heat Kernels and Analysis on Manifolds, Graphs, and Metric Spaces: Lecture Notes from a Quarter Program on Heat Kernels, Random Walks, and Analysis on Manifolds and Graphs: April 16-July 13, 2002, Emile Borel Centre of the Henri Poincaré Institute, Paris, France*, 338:357, 2003.

[52] K. T. Sturm. On the geometry of metric measure spaces - i. *Acta Mathematica*, 196(1):65–131, 2006.

[53] K. T. Sturm. On the geometry of metric measure spaces - ii. *Acta Mathematica*, 196(1):133–177, 2006.

[54] M. Talagrand. New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126:505–563, 1996.

[55] A. B. Tsybakov. *Introduction to Nonparametric estimation.* Springer, 2009.

[56] I. Vajda. *Theory of Statistical Inference and Information.* Kluwer Academic Press, 1989.

[57] R. van Handel. *Probability in high dimension.* Princeton Lecture Notes, 2016.

[58] R. Veldhuis. The centroid of the symmetrical kullback-leibler distance. *IEEE signal processing letters*, 9(3):96–99, 2002.

[59] C. Villani. *Topics in Optimal Transportation.* Graduate Studies in Mathematics, American Mathematical Society, 2003.

[60] C. Villani. *Optimal transport: Old and new.* Springer, 2008.

[61] J. Weed and Q. Berthet. Estimation of smooth densities in Wasserstein distance. *arXiv:1902.01778*, 2019.

[62] T. Yokota. A rigidity theorem in Alexandrov spaces with lower curvature bound. *Mathematische Annalen*, 353(2):305–331, 2012.

[63] T. Yokota. Convex functions and barycenter on CAT(1)-spaces of small radii. *Journal of the Mathematical Society of Japan*, 68(3):1297–1323, 2016.

[64] T. Yokota. Convex functions and p-barycenter on CAT(1)-spaces of small radii. *Tsukuba Journal of Mathematics*, 41(1):43–80, 2017.