Wasserstein Barycenters: Statistics and Optimization

by

Austin J. Stromme

B.S., Mathematics, University of Washington (2018)
B.S., Computer Science, University of Washington (2018)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2020

© Massachusetts Institute of Technology 2020. All rights reserved.

Author . . . .
Department of Electrical Engineering and Computer Science
May 15th, 2020

Certified by . . . .
Philippe Rigollet
Associate Professor of Mathematics
Thesis Supervisor

Accepted by . . . .
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee for Graduate Students

Wasserstein Barycenters: Statistics and Optimization

by

Austin J. Stromme

Submitted to the Department of Electrical Engineering and Computer Science
on May 15th, 2020, in partial fulfillment of the
requirements for the degree of
Master of Science

Abstract

We study a geometric notion of average, the barycenter, over 2-Wasserstein space. We significantly advance the state of the art by introducing extendible geodesics, a simple synthetic geometric condition which implies non-asymptotic convergence of the empirical barycenter in non-negatively curved spaces such as Wasserstein space. We further establish convergence of first-order methods in the Gaussian case, overcoming the non-convexity of the barycenter functional. These results are accomplished by various novel geometrically inspired estimates for the barycenter functional including a variance inequality, new so-called quantitative stability estimates, and a Polyak-Łojasiewicz (PL) inequality. These inequalities may be of independent interest.

Thesis Supervisor: Philippe Rigollet
Title: Associate Professor of Mathematics

Acknowledgments

Foremost thanks goes to my collaborators: Sinho Chewi, Thibaut Le Gouic, Quentin Paris, Tyler Maunu, and Philippe Rigollet. Without them none of this would have been possible. I am especially grateful to Philippe for his talented and well-rounded support as an advisor. Extra thanks as well to Sinho for always patiently teaching me mathematics. At this point in my PhD, I have learned far more from Sinho than from any class or book. Thanks again to Matt and Dheeraj for showing Enric and me how it's done.

Thank you to my friends in the chemistry department and in Washington state. Finally, highest gratitude and love for Steve, Peggy, Emily, Kamran, Lang, Sarah, and Stephanie.

Contents

Chapter 1

Introduction

This thesis is concerned with *averaging probability measures*. It combines the theory of optimal transport, termed Wasserstein space, with geometric averages, termed barycenters, to provide novel quantitative understanding of *barycenters on Wasserstein space*.

1.1 Background

Optimal transport is a geometrically meaningful way of measuring distances between probability distributions. Essentially, optimal transport distances measure the most efficient means of moving one distribution to another. Stated at the level of random variables, the optimal transport distance between two random variables X, Y is the infimum, over all of couplings π of X, Y, of $\mathbb{E}_\pi[\|X - Y\|_2^2]$. Note the contrast between this distance measure and classical information divergences, such as the Kullback-Liebler divergence, which measure purely pointwise differences of densities. This distinction is often summed up by saying that optimal transport is “horizontal" whereas divergences are “vertical". We shall call the space of measures with the optimal transport metric Wasserstein space.

Optimal transport was first studied in 1781 by Gaspard Monge who wondered about the optimal way to fill holes with sand [42]. The theory developed in fits and starts until the 1980's when Knott and Smith [34] and Brenier [16] independently discovered a simple and beautiful form for the optimal coupling. Since these major advances, the theory has deepened and expanded at an accelerating pace.

Although this rapid theoretical progress inspired a few practical forays into optimal transport, it was Cuturi's 2013 introduction of an efficient approximation algorithm via entropic regularization that heralded the current wave of interest amongst applied scientists [23]. His efficient algorithm, combined with optimal transport's appealing geometric semantics for comparing probability distributions, led to an explosion of applications, see e.g. [49] and the references therein. The aspect of optimal transport at stake in this present thesis is a means of using the geometry to construct meaningful summaries, i.e., "averages," of collections of distributions. To incorporate the geometry of optimal transport into this average, we use the concept of barycenters.

The barycenter, also dubbed center of mass, center of gravity, or (mistakenly) the Karcher mean, is a natural generalization of averages to curved spaces that has been studied since at least the early 20th century [58, 31]. Ignoring issues of existence and uniqueness, the **barycenter** b^* of a distribution P on a metric space (M, d_M), is defined as

$$b^* := \operatorname*{arg\,min}_{b \in M} F(b) := \frac{1}{2}\mathbb{E}_{p\sim P}\left[d_M^2(b,p)\right].$$

As can be easily verified, this coincides with the usual Euclidean average in the case where $(M, d_M) = (\mathbb{R}^d, \|\cdot\|_2)$.

Wasserstein barycenters are thus barycenters in the case of probability measures on $\mathbb{R}^d$ metrized by the optimal transport distance. Wasserstein barycenters offer practitioners a highly non-linear yet meaningful average for applications where the pointwise average of measures is inappropriate. Whenever data can be embedded as a probability distribution over $\mathbb{R}^d$, Wasserstein barycenters may offer a significant advantage over more traditional techniques of summarization [24]. They have thus been applied in a broad variety of areas, including graphics, neuroimaging, Bayestian statistics, dimensionality reduction, and economics [51, 50, 55, 27, 15, 56, 21].

The theoretical study of Wasserstein barycenters dates back to the 1990's [45, 35, 41]. Several special cases were considered, the most important being the case where P is supported on two points and the corresponding theory of Wasserstein geodesics by McCann in [41]. Although there was significant theoretical understanding of barycenters in some general contexts around that time [58], the existing theory did not apply in a significant way due to the *positive curvature* of Wasserstein space [7, 44]. Because of this, insight into the Wasserstein space case was not provided until the work of Agueh and Carlier in 2011 [1], where the authors used the structure of Wasserstein space to develop a primaldual proof establishing existence, uniqueness, and optimality conditions under general assumptions. One beautiful consequence of the work of Agueh and Carlier is a surprising connection between the barycenter problem and the original optimal transport problem via the concept of multi-marginal transport. This connection, combined with the fundamental relationship between barycenters and geometry, serve to theoretically motivate the study of Wasserstein barycenters.

1.2 Main Results

Theoretical work subsequent to the Agueh and Carlier paper has focused on both statistical and computational aspects of Wasserstein barycenters. In this section, we present the two primary results contained in this thesis. The first result is on the convergence of empirical Wasserstein barycenters to their population counterparts, from [26]. The second result is on the convergence of popular first order methods for computing Wasserstein barycenters, from [22].

The failure of standard techniques for both of these problems essentially stems from the (sometimes infinitely) *positive curvature* of Wasserstein space. As such, the proofs rely on building a powerful quantitative understanding of the positive curvature of Wasserstein space and carefully applying it to the problems of interest.

1.2.1 Empirical barycenters

Empirical averages and their convergence to population averages are at the heart of statistics. We thus study the corresponding problem on Wasserstein space. Consider a distribution P on Wasserstein space, and let $\hat{b}_n$ be the barycenter of the empirical distribution $P_n := \frac{1}{n}\sum_{i=1}^n \delta_{p_i}$ for independent samples $p_i \sim P$. We call $\hat{b}_n$ the *empirical barycenter*. It has been shown that $\hat{b}_n$ is a consistent estimator of b^* [38]. It is then natural to ask about the finite sample variance for this estimator, which we can define by $\mathbb{E}[W_2^2(b^*, \hat{b}_n)]$ where the expectation is over all samples of size n. Note that in the Euclidean case, the variance of the empirical mean is always exactly

$$\mathbb{E}\left[\|b^* - \hat{b}_n\|_2^2\right] = \frac{\sigma^2}{n}$$

where $\sigma^2 := \mathbb{E}[\|x - b^*\|_2^2]$ is the variance of the distribution. We are thus motivated to understand when a *parametric rate*[1] holds for the empirical barycenter over Wasserstein space. Some special cases have been considered [47, 12, 14, 2]. In [36], parametric rates are obtained for the Gaussian case. Prior to the following theorem, the best rates in a general context were of the form $O(n^{-1/d})$ and were established using techniques from empirical process theory [3].

Before we can state our first main result, we review the discovery of Brenier and Agueh and Carlier: consider the 2-Wasserstein problem between two measures μ, ν. If μ is absolutely continuous w.r.t. the Lebesgue measure then the infimium over couplings in the W_2 problem is uniquely attained by a deterministic mapping which we will often write as $T_{\mu\to\nu}$. Moreover, this mapping is the gradient of a convex function; we write this as $T_{\mu\to\nu} = \nabla\varphi_{\mu\to\nu}$. The first main result in this thesis is the following:

Theorem 1.2.1 (Main Theorem 1)**.** *Suppose P is a distribution supported on absolutely continuous measures over $\mathbb{R}^d$ with finite second moment. Suppose further that P has a barycenter b^* such that for each $\mu \in \operatorname{supp}(P)$, the optimal potential $\varphi_{b^*\to\mu}$ is α-strongly convex and β-smooth, where $\beta - \alpha < 1$. Then b^* is the unique barycenter of P and moreover*

$$\mathbb{E}\left[W_2^2(\hat{b}_n, b^*)\right] \leqslant \frac{4\sigma^2}{k^2 n}.$$

where $k := 1 - (\beta - \alpha) > 0$ and $\sigma^2 := \mathbb{E}[W_2^2(p, b^*)]$ is the variance of P.

A couple of remarks are now in order. First, the condition on the optimal potential is geometrically natural. In fact, it is merely the statement that the W_2-geodesic between b^* and μ can be extended past b^* and past μ by an amount depending on β and α, respectively. The second remark is that this result is actually a special case of a much more general theorem proved by the author and collaborators in [26].

1.2.2 First-order methods for barycenters on Gaussians

The above theorem provides finite-sample convergence rates for empirical Wasserstein barycenters in a broad variety of situations. However, given that there is no explicit form for the Wasserstein barycenter, the question becomes how to compute one. A common approach considered in a number of applications is to use a first order optimization method

[1] precisely: a rate of the form c/n for c independent of dimension

such as gradient descent or stochastic gradient descent to approximate the barycenter. This can be made formally rigorous (see [7]), but for our purposes we will simply define the W_2-gradient of the barycenter functional to be, for all b absolutely continuous and with finite second moment,

$$\nabla_{W_2} F(b) := -\mathbb{E}_{p \sim P}[(T_{b \to p} - \mathrm{id})]$$

where $T_{b\to p}$ is the optimal map from $b \to p$ and $\mathrm{id} \colon \mathbb{R}^d \to \mathbb{R}^d$ is the identity. Panaretos and Zemel were able to show that under mild assumptions on the distribution P, a simple gradient descent algorithm converges to the population barycenter asymptotically [46]. The case where P is supported on Gaussians was also studied in [6, 9, 62]. Fast convergence of first order methods was observed in each of these three works, and proving such convergence was left open in [6]. The following theorem, from our work [22], resolves this open problem.

Theorem 1.2.2 (Main Theorem 2). *Let P be a distribution supported on mean-zero Gaussians whose covariance matrices have eigenvalues uniformly bounded between $\lambda_{\min}$ and $\lambda_{\max}$. Let $\kappa := \lambda_{\max}/\lambda_{\min}$. Then P has a unique barycenter b^*. Moreover, W_2 gradient descent for the barycenter functional initialized at $b_0 \in \operatorname{supp}(P)$ obeys*

$$W_2^2(b_T, b^*) \leqslant 2\kappa \left(1 - \frac{1}{4\kappa^2}\right)^T (F(b_0) - F(b^*)).$$

Performing stochastic gradient descent for the barycenter functional initialized at $b_0 \in \operatorname{supp}(P)$, we obtain the bound

$$\mathbb{E}\left[W_2^2(b_n, b^*)\right] \leqslant \frac{96\sigma^2\kappa^5}{n}.$$

We remark that this theorem is an example of a non-convex optimization problem where convex methods provably work. Specifically, the positive curvature of Wasserstein space means the barycenter functional is not geodesically convex - in fact, over Gaussians, it can actually be concave! The proof overcomes this non-convexity by establishing a Polyak-Łojasiewicz inequality using novel theory for the Wasserstein barycenter functional. In fact, a significant portion of the technical work is valid at the level of general first-order optimization for Wasserstein barycenters.

1.3 Guide to the thesis

Chapter 2 begins by summarizing facts from optimal transport and then dives into a significant exploration of curvature in $W_2(\mathbb{R}^d)$ and its connection with the barycenter functional. We believe that this chapter is the most independently interesting, so we attempted to develop a cohesive discussion of the relevant ideas. Sections marked with an asterisk are not used in the remainder of the work, though we feel they are of independent interest. Chapter 3 takes the machinery developed in the previous chapter and applies it to provide a quick solution for the empirical barycenters problem. Chapter 4 studies first-order methods for the barycenter problem.

The most generally useful and novel inequalities are in Theorem 2.3.7, Theorem 2.7.11 (see also Theorem 2.6.1 and Theorem 2.9.2 for specializations), and Lemma 4.4.1.

The appendices contain some empirical work as well as omitted proofs.

Chapter 2

Geometry of optimal transport

In this chapter we develop some of the most general and conceptually important results of the thesis. We start by summarizing background on optimal transport, discussing some simple notions of curvature in metric spaces, and then explaining their close relationship with the barycenter functional. We then study the curvature of Wasserstein space. We conclude with a series of related inequalities that, roughly, allow us to control the positive curvature under the assumption of extendible geodesics.

2.1 Background on optimal transport

In this section, we define frequently used notation and state the most important theorems about optimal transport. We refer the reader to the books [60, 61, 7, 52]. We let $\mathcal{P}_2(\mathbb{R}^d)$ be the set of probability measures on $\mathbb{R}^d$ with finite second moment, namely μ such that $\mathbb{E}_{x\sim\mu}[\|x\|_2^2] < \infty$. We write the collection of probability measures that are absolutely continuous w.r.t. the Lebesgue measure with finite second moment as $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$. We shall often write the identity map $\mathrm{id}\colon \mathbb{R}^d \to \mathbb{R}^d$.

Given two measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, the set $\Pi(\mu, \nu)$ is the set of all couplings of μ and ν, i.e. the set of measures π on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and ν. The 2**-Wasserstein distance** between μ and ν is defined as

$$W_2^2(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \int \|x-y\|_2^2 \mathrm{d}\pi(x,y).$$

The infimum can be shown to be attained by compactness of $\Pi(\mu, \nu)$ w.r.t. the weak topology. A minimizer is referred to as an **optimal transport plan**. Moreover, $W_2(\mu, \nu)$ thus defined is a metric and, when restricted to probability measures over a compact set, the W_2 distance metrizes weak convergence of measures. In the non-compact case, W_2 metrizes weak convergence and convergence of second moments [60, Thm 7.12]. The 2-Wasserstein problem admits a dual formulation, called the **dual Kantorovich problem**, given by

$$\sup_{(f,g)\in S_{\mu,\nu}} \left(\int f d\mu + \int g d\nu \right),$$

where

$$S_{\mu,\nu} := \{(f,g) \in L^1(\mu) \times L^1(\nu) \colon f(x) + g(y) \leqslant \|x-y\|_2^2\}.$$

Given a map $T \colon \mathbb{R}^d \to \mathbb{R}^d$, we let $T_{\#}\mu$ be the **push-forward** of μ under T, namely the law of $T(x)$ when $x \sim \mu$. We shall call a convex function φ **proper** if $\varphi(x) < +\infty$ for some x and if $\varphi(x) > -\infty$ for all x. The **domain** of a convex function $\mathrm{dom}(\varphi)$ shall be the set of points at which it is finite. The **convex conjugate** of a convex function φ is denoted φ^* and defined as $\varphi^*(y) := \sup_{x\in\mathbb{R}^d} \langle x, y\rangle - \varphi(x)$. We shall often use that for a proper lower semicontinous function φ, φ convex if and only if $\varphi = \varphi^{**}$, and as well that $\nabla\varphi^*$ and $\nabla\varphi$, when defined uniquely, are inverses for one another.

We can now state the fundamental theorem of optimal transport.

Theorem 2.1.1. *Suppose $\mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ and $\nu \in \mathcal{P}_2(\mathbb{R}^d)$. Then the following are equivalent:*

1. *$\pi \in \Pi(\mu, \nu)$ is an optimal transport plan*

2. *$\pi = (\mathrm{id}, \nabla\varphi)_{\#}\mu$ for a proper convex function φ*

3. *Strong duality holds between the W_2 problem and the dual Kantorovich problem:*

$$\int \|x-y\|_2^2 d\pi(x,y) = \sup_{(f,g)\in S_{\mu,\nu}} \left(\int f d\mu + \int g d\nu \right).$$

Moreover, the supremum is attained for $f = \|x\|_2^2 - 2\varphi(x)$ and $g = \|y\|_2^2 - 2\varphi^*(y)$.

Finally, there is a unique $\nabla\varphi$ such that the above holds, in the sense that if $\nabla\psi$ is also optimal then $\nabla\varphi(x) = \nabla\psi(x)$ μ-a.e.

The potentials φ, φ^* are referred to as **Kantorovich potentials** for the pair (μ, ν). We shall often write the optimal transport map between $\mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ and $\nu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ as

$T_{\mu\to\nu}$ and an optimal Kantorovich potential as $\varphi_{\mu\to\nu}$ so that $T_{\mu\to\nu} = \nabla\varphi_{\mu\to\nu}$.

We shall always implicitly assume metric spaces (X, d_X) are complete and separable. A **constant speed geodesic** in (X, d_X) is a curve $\gamma\colon [a,b] \to X$ such that for all $t, s \in [a,b]$,

$$d_X(\gamma(t), \gamma(s)) = \frac{|t-s|}{b-a} d_X(\gamma(a), \gamma(b)).$$

We then say that (X, d_X) is **geodesically complete** if each pair of points can be connected by a constant speed geodesic.

Theorem 2.1.2. *$W_2(\mathbb{R}^d)$ is a geodesically complete metric space. For any $\mu, \nu \in W_2$, let $\pi_t := (1-t)x + ty$ for $t \in [0,1]$. Let $\pi^* \in \Pi(\mu,\nu)$ be an optimal transport plan for μ to ν. Then the path $\omega(t) = (\pi_t)_{\#}\pi^*$ is a constant-speed geodesic in W_2 connecting $\omega(0) = \mu$ to $\omega(1) = \nu$. Moreover, all constant-speed geodesics are of this form. Hence, if μ is absolutely continuous with respect to Lebesgue, there is in fact a* ***unique*** *geodesic joining μ to ν.*

Remark 2.1.3. *We remark that geodesics between non-absolutely continuous measures need not be unique. Consider as an example the measures $\mu_0 := \frac{1}{2}\delta_{(0,1)} + \frac{1}{2}\delta_{(0,-1)}$ and $\mu_1 := \frac{1}{2}\delta_{(1,0)} + \frac{1}{2}\delta_{(-1,0)}$. Then both*

$$\mu_t := \frac{1}{2}\delta_{(t,(1-t))} + \frac{1}{2}\delta_{(-t,-(1-t))}$$

and

$$\tilde{\mu}_t := \frac{1}{2}\delta_{(-t,1-t)} + \frac{1}{2}\delta_{(t,-(1-t))}$$

are distance-minimizing geodesics between μ_0 and μ_1.

Lastly, we shall need to define strong convexity and smoothness: for $\alpha \geqslant 0$ and $\beta > 0$, we will say that a convex function φ is **α-strongly convex** if for all $x, y \in \mathrm{dom}(\varphi)$

$$\varphi((1-t)x + ty) \leqslant (1-t)\varphi(x) + t\varphi(y) - \frac{\alpha}{2} t(1-t)\|x-y\|_2^2.$$

It is **β-smooth** if the reverse inequality holds with α replaced by β.

2.2 A simple notion of curvature in geodesic metric spaces

It is not within scope for this thesis to give an introduction to curvature in Riemannian and metric geometry. Suffice it to say that curvature is a word with many meanings, all generally related to controlling the deviation of an object from it's flat Euclidean counterpart. As Gromov says in [28]: "The curvature tensor of a Riemannian manifold is a little monster of (multi)linear algebra whose full geometric meaning remains obscure." Luckily, there are intuitive simplifications of the full curvature tensor which generalize well beyond Riemannian manifolds. In this section we shall define one such notion that will be fundamental for the rest of the thesis.

Definition 2.2.1. *We say a geodesically complete metric space (X, d_X) is* **non-positively curved (NPC)** *if for every $y, x_0, x_1 \in X$, and every constant speed geodesic $\gamma\colon [0,1] \to (X, d_X)$ joining x_0 to x_1,*

$$d_X^2(\gamma(t), y) \leqslant (1-t)d_X^2(x_0, y) + td_X^2(x_1, y) - t(1-t)d_X^2(x_0, x_1). \tag{2.2.1.1}$$

If the reverse inquality holds for all triples y, x_0, x_1 and constant speed geodesics $\gamma\colon [0,1] \to (X, d_X)$, namely

$$d_X^2(\gamma(t), y) \geqslant (1-t)d_X^2(x_0, y) + td_X^2(x_1, y) - t(1-t)d_X^2(x_0, x_1). \tag{2.2.1.2}$$

then we say that (X, d_X) is **non-negatively curved (NNC)***.*

Remark 2.2.2. *We remark that the parallelogram identity in $\mathbb{R}^d$ is the equality case of the above inequalities, and so we can say that $\mathbb{R}^d$ has zero curvature, or is* **flat***. The standard examples of NPC and NNC spaces are the constant curvature surfaces: hyperbolic manifolds and spheres, respectively. In fact, the sectional curvature of these surfaces, and indeed any Riemannian manifold, is reflected at least locally by a constant factor times the third term [43]. For our purposes these simple definitions are sufficient.*

Remark 2.2.3. *The theory of curvature in metric spaces goes back to work of Alexandrov in 1951 [4]. There is a huge literature on the subject, see e.g. [58, 57, 19] as well as the books [18, 17, 5]. In the case where (X, d_X) is a Riemannian manifold, these definitions coincide with complete Riemannian manifolds of non-positive sectional curvature and complete Riemannian manifolds of non-negative sectional curvature, respectively. They arise not merely as an idle generalization of Riemannian manifolds, but as limits of sequences of Riemannian manifolds cropping up in geometric flows. As evidenced in this thesis, they also provide a useful way of thinking about the structure of infinite-dimensional, non-Riemannian spaces such as $W_2(\mathbb{R}^d)$.*

2.3 Curvature and convexity of the barycenter functional

In this section we explore the close relationship between curvature of the metric space and convexity properties of barycenter functionals.

Definition 2.3.1 (Barycenters)**.** *Given a distribution P on (X, d_X) we define the* **barycenter functional** $F_P \colon (X, d_X) \to \mathbb{R}$ *as*

$$F_P(b) := \frac{1}{2}\mathbb{E}_{x\sim P}[d_X^2(x, b)].$$

If $F_P(b)$ is finite for any $b \in X$, then we say that P has **finite second moment***. The collection of all distributions supported on X with finite second moment is denoted $\mathcal{P}_2(X)$. The* **variance** *of P is defined as*

$$\sigma^2(P) := 2 \inf_b F_P(b).$$

Lastly, a **barycenter** *of P is a minimizer of F_P.*

Remark 2.3.2. *We shall typically suppress the distribution P and write simply $F = F_P$ and $\sigma^2 = \sigma^2(P)$. Note as well that by the triangle inequality $F_P(b)$ is finite for any fixed $b \in X$ if and only if it is finite for all $b \in X$.*

Definition 2.3.3. *Given a geodesically complete metric space (X, d_X), we say that a function $f \colon (X, d_X) \to \mathbb{R}$ is* α**-geodesically convex** *for $\alpha \geqslant 0$ if along all constant speed geodesics $\gamma \colon [0, 1] \to (X, d_X)$ we have the analog of the Euclidean inequality:*

$$f(\gamma(t)) \leqslant (1-t)f(\gamma(0)) + tf(\gamma(1)) - \frac{\alpha}{2}t(1-t)d_X^2(\gamma(0), \gamma(1)).$$

We say it is β**-geodesically smooth** *for $\beta \geqslant 0$ if along all constant speed geodesics $\gamma \colon [0, 1] \to (X, d_X)$ we have the opposite inequality:*

$$f(\gamma(t)) \geqslant (1-t)f(\gamma(0)) + tf(\gamma(1)) - \frac{\beta}{2}t(1-t)d_X^2(\gamma(0), \gamma(1)).$$

Using these definitions we can derive the following simple but important observation:

Proposition 2.3.4. *(X, d_X) is non-positively curved (resp. non-negatively curved) if and only if for all distributions $P \in \mathcal{P}_2(X)$ the barycenter functional F_P is 1-geodesically convex (resp. smooth).*

Proof. For the forward direction apply our definition of curvature pointwise in the expectation, and for the reverse direction take the distribution $P = \delta_x$ for each $x \in X$. $\square$

We thus see that the convexity of the barycenter functional is a direct reflection of the underlying curvature of (X, d_X). In this thesis the central issue shall be that we want the barycenter functional to be convex, but it is not because of the positive curvature of $W_2(\mathbb{R}^d)$ (see Section 2.4 below). To circumvent this issue we then marshal various weak and quantitative forms of convexity that allow us to achieve our ends.

One important weak form of convexity that shall play a major role is termed a quadratic growth condition. This is a standard notion in optimization which says that if a function f has a minimizer x^* then $f(x) - f(x^*) \geqslant c\|x - x^*\|_2^2$ for some constant $c > 0$. In particular, it is a consequence of strong convexity. In the context of barycenters, this is known as a variance inequality.

Definition 2.3.5 (Variance Inequality)**.** *Suppose (X, d_X) is a geodesically complete metric space and $P \in \mathcal{P}_2(X)$. Then we say that F_P satisfies a* **variance inequality** *with constant $C_{\mathsf{var}} > 0$ if there exists a point $b^* \in X$ such that for all $b \in X$,*

$$\frac{C_{\mathsf{var}}}{2} d^2(b, b^*) \leqslant F(b) - F(b^*).$$

Let's look at this inequality when (X, d_X) is NPC. In that case, for any $P \in \mathcal{P}_2(X)$ with a barycenter b^*, fix a point $b \in X$. Let γ be the constant speed geodesic joining b^* to b. Then by Prop. 2.3.4, for each $t \in [0, 1]$, we have

$$F(\gamma(t)) \leqslant (1-t)F(b^*) + tF(b) - \frac{1}{2}t(1-t)d_X^2(b, b^*).$$

Observe that by definition $F(b^*) \leqslant F(\gamma(t))$ for all $t \in [0, 1]$. Whence

$$0 \leqslant F(\gamma(t)) - F(b^*) \leqslant t(F(b) - F(b^*)) - \frac{1}{2}t(1-t)d_X^2(b, b^*).$$

Dividing by t and re-arranging yields

$$\frac{1}{2}(1-t)d_X^2(b, b^*) \leqslant F(b) - F(b^*).$$

Since t is arbitrary we conclude that in NPC spaces, barycenter functionals (with minimizers) always satisfy a variance inequality with $C_{\mathsf{var}} = 1$. In fact, more is true.

Theorem 2.3.6 ([58])**.** *Suppose (X, d_X) is a geodesically complete metric space. Then (X, d_X) is non-positively curved if and only if F_P obeys a variance inequality with $C_{\mathsf{var}} = 1$ for every $P \in \mathcal{P}_2(X)$.*

We omit the proof for brevity. As this theorem evidences, even *relaxations* of convexity for the barycenter functional are equivalent to non-positive curvature. Variance inequalities for barycenter functionals are fundamental to the two central problems of this thesis. For the empirical barycenters problem considered in Chapter 3, it has been shown that a variance inequality leads to dimension-dependent rates [3]. Moreover, our approach to the statistical barycenters problem crucially uses a variance inequality. And a general variance inequality shall be central to our approach to analyzing first-order methods for barycenters in Chapter 4.

We conclude this section with the first significant result of this thesis, which provides general conditions under which distributions on Wasserstein space satisfy a variance inequality.

Theorem 2.3.7 (Variance inequality in $W_2(\mathbb{R}^d)$ [22])**.** *Let $P \in \mathcal{P}_2(\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d))$ be a distribution with barycenter $b^* \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$. Assume there exists a mapping $\varphi \colon \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$ such that*

(1) φ is measurable

(2) for P-a.e. μ, φ_μ is an optimal Kantorovich potential for b^* to μ

(3) for almost all $x \in \mathbb{R}^d$

$$\mathbb{E}_{\mu \sim P}[\varphi_\mu(x)] = \frac{1}{2}\|x\|_2^2.$$

(4) for P-a.e. $\mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, the mapping φ_μ is $\alpha(\mu)$-strongly convex for some measurable function $\alpha \colon \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) \to \mathbb{R}_+$.

Then, P satisfies a variance inequality for all $b \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ with constant

$$C_{\mathsf{var}} = \int \alpha(\mu)\, \mathrm{d}P(\mu)\,.$$

Remark 2.3.8. *Assumption (3) is roughly a first-order optimality statement for b^*, and so should be thought of as essentially following from the assumption that b^* is the barycenter of P. We shall have* *more to say about assumption (3) in Chapter 4. Assumption (4) is related to extendible geodesics, see Section 2.7. Finally, we remark that a similar result appears in the unpublished note [37].*

Proof. Fix a $\mu \in \mathrm{supp}(P)$ and $b \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$. By lemma A.1.1, for arbitrary $x \in \mathbb{R}^d$ and almost every $y \in \mathbb{R}^d$ we have

$$\varphi_\mu(x) + \varphi_\mu^*(y) \geqslant \langle x, y\rangle + \frac{\alpha(\mu)}{2}\|x - \nabla\varphi_\mu^*(y)\|_2^2.$$

Rearranging this is

$$\frac{1}{2}\|x\|_2^2 - \varphi_\mu(x) + \frac{1}{2}\|y\|_2^2 - \varphi_\mu^*(y) \leqslant \frac{1}{2}\|x-y\|_2^2 - \frac{\alpha(\mu)}{2}\|x - \nabla\varphi_\mu^*(y)\|_2^2.$$

Hence we can integrate over the optimal coupling of μ to b to yield

$$\begin{aligned}\frac{1}{2}W_2^2(b,\mu) \geqslant &\int (\frac{1}{2}\|x\|_2^2 - \varphi_\mu(x))\mathrm{d}b(x) + \int (\frac{1}{2}\|y\|_2^2 - \varphi_\mu^*(y))\mathrm{d}\mu(y) \\ &+ \frac{\alpha(\mu)}{2}\|T_{\mu\to b} - T_{\mu\to b^*}\|_{L^2(\mu)}^2 \\ \geqslant &\int (\frac{1}{2}\|x\|_2^2 - \varphi_\mu(x))\mathrm{d}b(x) + \int (\frac{1}{2}\|y\|_2^2 - \varphi_\mu^*(y))\mathrm{d}\mu(y) \\ &+ \frac{\alpha(\mu)}{2}W_2^2(b,b^*).\end{aligned}$$

where in the last inequality we used the definition of W_2 distance. We now integrate over $\mu \sim P$ and argue the first integral is 0 irrespective of b. To do this, we first observe that

$$\begin{aligned}\iint |\frac{1}{2}\|x\|_2^2 - \varphi_\mu(x)|\mathrm{d}P(\mu)\mathrm{d}b(x) &\leqslant \frac{1}{2}\sigma^2(b) + \iint |\varphi_\mu(x)|\mathrm{d}P(\mu)\mathrm{d}b(x) \\ &= \frac{1}{2}\sigma^2(b) + \iint \varphi_\mu(x)\mathrm{d}P(\mu)\mathrm{d}b(x) \\ &= \frac{1}{2}\sigma^2(b) < \infty.\end{aligned}$$

where we applied triangle inequality, added and subtracted a linear function lower bounding $\varphi_\mu(x)$ (the linear function is integrable by the assumption that $P \in \mathcal{P}_2(\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d))$), and assumption (3), respectively. It follows that $(\|x\|_2^2/2 - \varphi_\mu(x)) \in L^1(b \otimes P)$ so that we can swap integrals and apply assumption (3) again to conclude

$$\iint (\frac{1}{2}\|x\|_2^2 - \varphi_\mu(x))\mathrm{d}b(x)\mathrm{d}P(\mu) = \iint (\frac{1}{2}\|x\|_2^2 - \varphi_\mu(x))\mathrm{d}P(\mu)\mathrm{d}b(x) = 0.$$

In fact, the same reasoning applied to $b = b^*$ shows

$$\iint (\frac{1}{2}\|x\|_2^2 - \varphi_\mu(x))\mathrm{d}b(x)\mathrm{d}P(\mu) = 0 = \iint (\frac{1}{2}\|x\|_2^2 - \varphi_\mu(x))\mathrm{d}b^*(x)\mathrm{d}P(\mu).$$

Hence

$$\begin{aligned}
F(b) \geqslant& \iint (\frac{1}{2}\|x\|_2^2 - \varphi_\mu(x))\mathrm{d}b(x)\mathrm{d}P(\mu) + \iint (\frac{1}{2}\|y\|_2^2 - \varphi_\mu^*(y))\mathrm{d}\mu(y)\mathrm{d}P(\mu) \\
&+ \frac{\alpha(\mu)}{2} W_2^2(b, b^*) \\
&= \iint (\frac{1}{2}\|x\|_2^2 - \varphi_\mu(x))\mathrm{d}b^*(x)\mathrm{d}P(\mu) + \iint (\frac{1}{2}\|y\|_2^2 - \varphi_\mu^*(y))\mathrm{d}\mu(y)\mathrm{d}P(\mu) \\
&+ \frac{\alpha(\mu)}{2} W_2^2(b, b^*) \\
&= F(b^*) + \frac{\alpha(\mu)}{2} W_2^2(b, b^*).
\end{aligned}$$

This proves the result. □

2.4 Wasserstein space is non-negatively curved

Theorem 2.4.1. *The space $W_2(\mathbb{R}^d)$ is non-negatively curved.*

Proof. We know that it is geodesically complete by 2.1.2. Fix $\mu, \nu_0, \nu_1 \in \mathcal{P}_2(\mathbb{R}^d)$. Choose a W_2 geodesic ν_t from ν_0 to ν_1. Then $\nu_t = ((1-t)\pi_1 + t\pi_2)_\# \gamma$ where γ is an optimal coupling of ν_0 to ν_1 and the π_i are the projection operators. Fix $t \in [0, 1]$, let $y_t := (1-t)y_0 + ty_1$, and let α_t be an optimal coupling from ν_t to μ. We describe a coupling (x, y_0, y_1) where $x \sim \mu$, $y_0 \sim \nu_0$, and $y_1 \sim \nu_1$. We take $y_0 \sim \nu_0$, y_1 from γ conditional on y_0, and x from α_t conditional on y_t. Then we can calculate:

$$\begin{aligned}
W_2^2(\mu, \nu_t) &= \mathbb{E}[\|x - y_t\|_2^2] \\
&= (1-t)\mathbb{E}[\|x - y_0\|_2^2] + t\mathbb{E}[\|x - y_1\|_2^2] - t(1-t)\mathbb{E}[\|y_0 - y_1\|_2^2] \\
&= (1-t)\mathbb{E}[\|x - y_0\|_2^2] + t\mathbb{E}[\|x - y_1\|_2^2] - t(1-t)W_2^2(\nu_0, \nu_1).
\end{aligned}$$

We observe that the joint couplings of (x, y_0) and (x, y_1) have marginals μ, ν_0 and μ, ν_1

respectively, but may not in fact be optimal. Hence we get the lower bound

$$W_2^2(\mu,\nu_t) \geqslant (1-t)W_2^2(\mu,\nu_0) + tW_2^2(\mu,\nu_1) - t(1-t)W_2^2(\nu_0,\nu_1).$$

Which proves the result. □

We remark that this theorem can be significantly extended, though we omit the proof for brevity.

Theorem 2.4.2 (Thm. A.2 [39])**.** *A smooth compact connected Riemannian manifold M has non-negative curvature if and only if $W_2(M)$ has non-negative curvature.*

2.5 Wasserstein space has infinite curvature

In this section, we describe an example which shows, in a certain sense, that "Wasserstein space has infinite curvature at every scale." The idea of the construction is that we can embed the square counterexample in remark 2.1.3 at an arbitrarily small scale within a given measure. We thereby show that any W_2 ball around any measure has elements with non-unique midpoints, which means there cannot be a uniform upper curvature bound on any open subset of Wasserstein space.

Fix a measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\varepsilon > 0$. More specifically, we shall show that there exist measures $\tilde{\mu_0}$ and $\tilde{\mu_1}$, such that $W_2(\mu,\tilde{\mu_0}), W_2(\mu,\tilde{\mu_1}) < \varepsilon$ and $\tilde{\mu_0}$ and $\tilde{\mu_1}$ have two midpoints contained in the W_2 ε-ball around μ.

Begin by letting $\tilde{\mu}$ be the (normalized) restriction of μ to the complement of a ball B. We claim that it is possible to choose the ball B such that $W_2(\mu,\tilde{\mu}) < \varepsilon/2$. If μ has a continuous density this ball can be chosen within the support, and otherwise it can be chosen at infinity. Suppose the ball $B = B(m,R)$. For a small parameter $\tau > 0$, let

$$\nu_0 := \frac{1}{2}\delta_{m+\tau e_1} + \frac{1}{2}\delta_{m-\tau e_1},$$
$$\nu_1 := \frac{1}{2}\delta_{m+\tau e_2} + \frac{1}{2}\delta_{m-\tau e_2}.$$

Let the union of these four points be S. Finally, for a small parameter $\eta > 0$, set

$$\tilde{\mu}_0 := (1-\eta)\tilde{\mu} + \eta\nu_0$$
$$\tilde{\mu}_1 := (1-\eta)\tilde{\mu} + \eta\nu_1.$$

We claim that by taking τ sufficiently small, the optimal coupling π_{01}^* between $\tilde{\mu}_0$ and $\tilde{\mu}_1$ can be decomposed as

$$\pi_{01}^* = (1-\eta)(\mathrm{id},\mathrm{id})_{\#}\tilde{\mu} + \eta q_{01},$$

where q_{01} is an optimal coupling between ν_0 and ν_1. To see this, consider any coupling $\pi \in \Pi(\tilde{\mu}_0, \tilde{\mu}_1)$. Let $p_{cc} = \mathbb{P}_\pi[(x_0, x_1) \in B^c \times B^c]$, $p_{cs} = \mathbb{P}_\pi[(x_0, x_1) \in B^c \times S]$ and similarly for p_{sc} and p_{ss}. Then

$$\begin{aligned}\mathbb{E}_\pi[\|x-y\|_2^2] &\geqslant (R-\tau)^2 p_{cs} + \tau^2 p_{ss} \\ &\geqslant \tau^2(p_{cs} + p_{ss}) \\ &= \eta\tau^2 = \mathbb{E}_{\pi_{01}^*}[\|x_0 - x_1\|_2^2].\end{aligned}$$

Where the first inequality follows by ignoring the contribution of p_{cc} and p_{sc}, the second by choosing $\tau < R/2$, the third equality by the fact that $\pi \in \Pi(\tilde{\mu}_0, \tilde{\mu}_1)$, and the last equality by our choice of π_{01}^*. We observe that if $p_{cs} > 0$ then by our choice of $\tau < R/2$, the second inequality is in fact strict. Hence it follows that the optimal couplings must indeed be of the form π_{01}^*.

But then, since there are two optimal couplings of ν_0 to ν_1, there are two optimal couplings of $\tilde{\mu}_0$ and $\tilde{\mu}_1$. Hence they don't have a unique midpoint. Lastly, we verify that $\eta > 0$ can be chosen small enough that both W_2-geodesics between $\tilde{\mu}_0$ and $\tilde{\mu}_1$ are always within ε of μ in W_2 distance. Fix one of the optimal couplings from ν_0 to ν_1 as q_{01}, and let ν_t be the point along this geodesic from ν_0 to ν_1. Let

$$\tilde{\mu}_t := (1-\eta)\tilde{\mu} + \eta\nu_t.$$

Observe that

$$((1-\eta)(\mathrm{id},\mathrm{id})_{\#}\tilde{\mu} + \eta\tilde{\mu} \otimes \nu_t) \in \Pi(\tilde{\mu}, \tilde{\mu}_t)$$

and using this coupling we have the upper bound

$$W_2(\tilde{\mu}, \tilde{\mu}_t) \leqslant \eta(2\mathbb{E}_{x\sim\tilde{\mu}}[\|x\|_2^2] + 2\|m\|_2^2 + \tau^2)^{1/2}.$$

Take η small enough so that this is smaller than $\varepsilon/2$. Using these settings we get

$$\begin{aligned} W_2(\mu, \tilde{\mu}_t) &\leqslant W_2(\mu, \tilde{\mu}) + W_2(\tilde{\mu}, \tilde{\mu}_t) \\ &< \varepsilon/2 + \varepsilon/2 = \varepsilon. \end{aligned}$$

We summarize this construction in the next proposition.

Proposition 2.5.1. *Suppose $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\varepsilon > 0$. Then there exist distinct distance minimizing W_2-geodesics $\gamma_0, \gamma_1 \colon [0,1] \to B_{W_2}(\mu, \varepsilon)$ such that $\gamma_0(0) = \gamma_1(0)$ and $\gamma_0(1) = \gamma_1(1)$.*

Combining this with the fact that geodesic spaces with bounded curvature (the reader can regard "bounded curvature" as strong geodesic convexity of the squared distance for the moment) have locally unique geodesics we get the following theorem.

Theorem 2.5.2. *No open subset of $W_2(\mathbb{R}^d)$ has bounded curvature.*

2.6 An ounce of convexity

Not discouraged by the previous section, we will look for some weak form of geodesic convexity of the squared 2-Wasserstein distance.

Specifically, fix measures $\mu_0, \mu_1, \mu_2 \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, and consider the W_2-geodesic $\mu_t \colon [0,1] \to W_{2,\mathrm{ac}}(\mathbb{R}^d)$. Consider the following approach to establishing a form of convexity for $W_2^2(\mu_2, \mu_t)$ by applying the definition of Wasserstein space and the flatness of Hilbert spaces:

$$\begin{aligned} W_2^2(\mu_2, \mu_t) &\leqslant \|T_{\mu_0\to\mu_t} - T_{\mu_0\to\mu_2}\|_{L^2(\mu_0)}^2 \\ &= (1-t)W_2^2(\mu_2, \mu_0) + t\|T_{\mu_0\to\mu_1} - T_{\mu_0\to\mu_2}\|_{L^2(\mu_0)}^2 - t(1-t)W_2^2(\mu_0, \mu_1). \end{aligned}$$

Compare to the K-strong convexity inequality for $W_2^2(\mu_2, \mu_t)$,

$$W_2^2(\mu_2, \mu_t) = (1-t)W_2^2(\mu_2, \mu_0) + tW_2^2(\mu_2, \mu_1) - \frac{K}{2}t(1-t)W_2^2(\mu_0, \mu_1) \qquad (2.6.0.1)$$

Hence we see that if we could show an inequality of the form

$$\|T_{\mu_0\to\mu_1} - T_{\mu_0\to\mu_2}\|_{L^2(\mu_0)}^2 \leqslant W_2^2(\mu_2,\mu_1) + (1-K/2)(1-t)W_2^2(\mu_0,\mu_1),$$

we'd be able to get the K-strong convexity inequality (2.6.0.1). We remark that when each μ_i is simply a translate of some base measure μ, we in fact expect this inequality to hold with $K = 1$, since in that case the 2-Wasserstein geometry is totally flat. Hence, this inequality can actually hold. On the other hand, by the construction in Section 2.5, we know that it cannot hold uniformly over open subsets of $W_2(\mathbb{R}^d)$.

In this section, we show that under a simple regularity condition on the Kantorovich potentials generating the optimal coupling, we can get inequalities of this form.

The crucial estimate is as follows.

Theorem 2.6.1. *Suppose $\mu \in W_{2,\text{ac}}(\mathbb{R}^d)$, $\nu_0,\nu_1 \in W_2(\mathbb{R}^d)$, and the optimal Kantorovich potential $\varphi_{\mu\to\nu_0}$ is α-strongly convex and β-smooth.*

$$\|T_{\mu\to\nu_0} - T_{\mu\to\nu_1}\|_{L^2(\mu)}^2 \leqslant W_2^2(\nu_0,\nu_1) + (\beta-\alpha)W_2^2(\mu,\nu_1).$$

Before we give the proof, we note that combining the preceding discussion, this estimate, and a tedious calculation, we obtain the corollary.

Corollary 2.6.2. *Fix $\mu \in W_{2,\text{ac}}(\mathbb{R}^d)$ and $\nu_0,\nu_1 \in W_2(\mathbb{R}^d)$. Suppose that the optimal Kantorovich potentials $\varphi_{\mu\to\nu_i}$ are α_i-strongly convex and β_i-smooth, $i = 0,1$. Let $\kappa_i := 1/\alpha_i - 1/\beta_i$ and $\kappa_i < 1$, $i = 0,1$. Then the K-strong convexity inequality* (2.6.0.1) *holds with $K = 2(\kappa_0 + \kappa_1)$.*

We remark as well that the most useful applications of this estimate will generally be when only one of the optimal potentials is known to be strongly convex and smooth, which corresponds to a K-strong convexity inequality but only for small t.

Proof of Theorem 2.6.1. We begin by writing

$$\begin{aligned}&\|T_{\mu\to\nu_0} - T_{\mu\to\nu_1}\|_{L^2(\mu)}^2\\ &\qquad = W_2^2(\nu_0,\mu) + W_2^2(\mu,\nu_1) - 2\mathbb{E}_\mu[\langle T_{\mu\to\nu_0}(x) - x, T_{\mu\to\nu_1}(x) - x\rangle].\end{aligned}$$

We will focus on bounding this last term. Write $T_{\mu\to\nu_0} = \nabla f_0$. We can integrate the β-

smoothness condition to obtain

$$\begin{aligned}
&\mathbb{E}_\mu[\langle T_{\mu\to\nu_0}(x), T_{\mu\to\nu_1}(x) - x\rangle] \\
&\qquad \geqslant \mathbb{E}_\mu[f_0(T_{\mu\to\nu_1}(x)) - f_0(x)] - \frac{\beta}{2}\mathbb{E}_\mu[\|T_{\mu\to\nu_1}(x) - x\|_2^2] \\
&\qquad = \mathbb{E}_{\nu_1}[f_0] - \mathbb{E}_\mu[f_0] - \frac{\beta}{2}W_2^2(\mu,\nu_1).
\end{aligned}$$

Let z be such that $(T_{\mu\to\nu_0}(x), z)$ is an optimal coupling of ν_0 to ν_1 when $x \sim \mu$. In this case, we can apply the strong convexity assumption to continue the lower bound:

$$\begin{aligned}
&\mathbb{E}_\mu[\langle T_{\mu\to\nu_0}(x), T_{\mu\to\nu_1}(x) - x\rangle] \\
&\qquad \geqslant \mathbb{E}[\langle \nabla f_0(x), z - x\rangle] + \frac{\alpha}{2}\mathbb{E}\|z - x\|_2^2 - \frac{\beta}{2}W_2^2(\mu,\nu_1) \\
&\qquad \geqslant \mathbb{E}[\langle \nabla f_0(x), z - x\rangle] - \frac{\beta-\alpha}{2}W_2^2(\mu,\nu_1),
\end{aligned}$$

where in the last line we used that the induced joint on x and z is a valid coupling of μ to ν_1. We also calculate

$$W_2^2(\nu_0,\mu) = \mathbb{E}_{x\sim\mu}[\|x\|_2^2] + \mathbb{E}_{x\sim\nu_0}[\|x\|_2^2] - 2\mathbb{E}_\mu[\langle T_{\mu\to\nu_0}(x), x\rangle],$$

and similarly

$$W_2^2(\nu_1,\mu) = \mathbb{E}_{x\sim\mu}[\|x\|_2^2] + \mathbb{E}_{x\sim\nu_1}[\|x\|_2^2] - 2\mathbb{E}_\mu[\langle T_{\mu\to\nu_1}(x), x\rangle].$$

Putting this together we find

$$\begin{aligned}
\|T_{\mu\to\nu_0} - T_{\mu\to\nu_1}\|_{L^2(\mu)}^2 &\leqslant \mathbb{E}[\|\nabla f_0(x) - z\|_2^2] + (\beta-\alpha)W_2^2(\mu,\nu_1) \\
&= W_2^2(\nu_0,\nu_1) + (\beta-\alpha)W_2^2(\mu,\nu_1),
\end{aligned}$$

where the last equality is by our assumption on z. Hence the result. □

2.7 Extendible geodesics and curvature bounds

In this section, we shall generalize quite extensively theorem 2.6.1. The starting point for this work is the observation that the hypotheses in theorem 2.6.1 are related to simple geometric conditions on the Wasserstein space.

Definition 2.7.1. *A constant-speed geodesic $\gamma\colon [0,1] \to (X, d_X)$ is said to be $(\lambda_{\text{in}}, \lambda_{\text{out}})$* **extendible** *for $\lambda_{\text{in}}, \lambda_{\text{out}} > 0$ if there exists a constant-speed geodesic $\gamma^+\colon [-\lambda_{\text{in}}, 1+\lambda_{\text{out}}] \to (X, d_X)$ such that γ is the restriction of γ^+ to $[0,1]$.*

Theorem 2.7.2 (Thm. 3.5 [3])**.** *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\gamma\colon [0,1] \to W_2(\mathbb{R}^d)$ be a geodesic connecting μ to ν. Then, γ is $(0, \lambda)$ extendible if, and only if, the support of the optimal transport plan of μ to ν lies in the subdifferential $\partial\varphi_{\mu\to\nu}$ of a $\lambda/(1+\lambda)$-strongly convex map $\varphi_{\mu\to\nu}$.*

Hence, we can re-interpret theorem 2.6.1 as follows.

Corollary 2.7.3. *Fix $\mu \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ and $\nu_0, \nu_1 \in \in(\mathbb{R}^d)$. Suppose the W_2-geodesic connecting μ to ν_0 is $(\lambda_{\text{in}}, \lambda_{\text{out}})$-extendible. Then*

$$\|T_{\mu\to\nu_0} - T_{\mu\to\nu_1}\|^2_{L^2(\mu)} \leqslant W_2^2(\nu_0, \nu_1) + \left(1 + \frac{1}{\lambda_{\text{in}}} - \frac{\lambda_{\text{out}}}{1+\lambda_{\text{out}}}\right) W_2^2(\mu, \nu_1). \tag{2.7.3.1}$$

In this section, we shall show that an analogous statement holds in fact over *all* non-negatively curved metric spaces. The power and utility of this result will be made clear subsequently. To complete our work, we will need a certain amount of technical machinery from the theory of Alexandrov spaces with (lower) curvature bounds. Specifically, we will need to find the proper generalization of the left-hand side of (2.7.3.1). We begin with this, and refer to [18] for a comprehensive treatment.

Definition 2.7.4. *Suppose (X, d_X) is a non-negatively curved metric space. For $p, x, y \in X$, the* **comparison angle** *$\sphericalangle_p^0(x,y) \in [0,\pi]$ at p is defined by*

$$\cos \sphericalangle_p^0(x,y) := \frac{d_X^2(p,x) + d_X^2(p,y) - d_X^2(x,y)}{2 d_X(p,x) d_X(p,y)}.$$

Applying triangle inequality shows that this is well-defined. Now, fix $p \in X$, and let Γ_p be the set of all geodesics $\gamma\colon [0,1] \to (X, d_X)$ with $\gamma(0) = p$. For $\gamma, \sigma \in \Gamma_p$ the **Alexandrov angle** *$\sphericalangle_p(\gamma,\sigma)$ is defined as*

$$\sphericalangle_p(\gamma,\sigma) := \lim_{s,t\to 0} \sphericalangle_p^0(\gamma(s), \sigma(t)).$$

It can be shown that when (X, d_X) is non-negatively curved, this is well defined and is in fact a (pseudo-)metric on Γ_p. Quotient by this equivalence relationship and take the completion of the *result to yield the* **space of directions** $(\Sigma_p, \sphericalangle_p)$*. Then the* **tangent cone** T_pX **of** (X, d_X) **at** p *is the set* $\Sigma_p \times \mathbb{R}_{\geqslant 0}$ *modulo the equivalence* $(\gamma, s) \approx (\sigma, t)$ *when* $s = t = 0$*. We denote the unique element* $(\gamma, 0)$ *by* o_p*, and call it the* **tip** *of the cone. For* $u, v \in T_pX$ *with* $u = (\gamma, s)$ *and* $v = (\sigma, t)$ *the metric is*

$$\|u - v\|_p^2 := s^2 + t^2 - 2st \cos \sphericalangle_p(\gamma, \sigma).$$

We use the notation $\|u\|_p := \|u - o_p\|_p$ *and* $\langle u, v\rangle_p := st \cos \sphericalangle_p(\gamma, \sigma)$*, which means*

$$\|u - v\|_p^2 = \|u\|_p^2 + \|v\|_p^2 - 2\langle u, v\rangle_p.$$

Let $C_p \subset S$ *be the cut-locus of* p *nad for all* $x \in S \setminus C_p$*, let* $\gamma_{p \to x} \in \Sigma_p$ *denote the direction of the unique geodesic connecting* p *to* x*. Then the* **log map** *at* p *is the map* $\log_p \colon S \setminus C_p \to T_pS$ *which sends*

$$x \mapsto \log_p(x) := (\gamma_{p \to x}, d_X(p, x)).$$

We will extend this to $x \in C_p$ *by selecting an arbitrary direction from* p *to* x*.*

Using these definitions, we now collect the relevant facts we will need about non-negatively curved spaces, tangent cones, and barycenters. We'll omit proofs as they would take us quite far afield; full details are given in our work [26].

Theorem 2.7.5. *Suppose* (X, d_X) *is non-negatively curved and* $P \in \mathcal{P}_2(X)$ *with barycenter* b^**.*

1. *For any* $p, x, y \in X$
$$d_X^2(x, y) \leqslant \| \log_p(x) - \log_p(y)\|_p^2. \tag{2.7.5.1}$$

2. *We have*
$$\iint \langle \log_{b^*}(x), \log_{b^*}(y)\rangle_{b^*} \mathrm{d}P(x)\mathrm{d}P(y) = 0. \tag{2.7.5.2}$$

3. *There exists a subset* $\mathcal{L}_{b^*}X \subset T_{b^*}X$ *which is a Hilbert space when equipped with the restricted metric and such that* $\log_{b^*}(\mathrm{supp}(P)) \subset \mathcal{L}_{b^*}X$*.*

4. *For any* $Q \in \mathcal{P}_2(X)$ *with* $\log_{b^*}(\mathrm{supp}(Q)) \subset \mathcal{L}_{b^*}X$ *and* $b \in X$*, we have*
$$\int \langle \log_{b^*}(x), \log_{b^*}(b)\rangle_{b^*} \mathrm{d}Q(x) = \left\langle \int \log_{b^*}(x)\mathrm{d}Q(x), \log_{b^*}(b) \right\rangle_{b^*}. \tag{2.7.5.3}$$

The following property is sometimes called being an **exponential barycenter** in the literature [3].

Corollary 2.7.6. *Suppose (X, d_X) is non-negatively curved and $P \in \mathcal{P}_2(X)$ with barycenter b^*. Then*

$$\int \log_{b^*}(x) \mathrm{d}P(x) = 0.$$

Definition 2.7.7. *Suppose (X, d_X) is non-negatively curved, $P \in \mathcal{P}_2(X)$, and b^* is a barycenter of P. Then for all $b, x \in X$ we define* **hugging function at** *b^* as*

$$k_{b^*}^b(x) := 1 - \frac{\| \log_{b^*}(x) - \log_{b^*}(b)\|_{b^*}^2 - d_X^2(x, b)}{d_X^2(b, b^*)}.$$

Remark 2.7.8. *In Section 2.8, we shall show that in case $(X, d_X) = (\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d), W_2)$, the tangent cone metric is the "$L^2(b^*)$ norm on transport maps." This will mean that*

$$\|T_{\mu \to \nu_0} - T_{\mu \to \nu_1}\|_{L^2(\mu)}^2 = W_2^2(\nu_0, \nu_1) + (1 - k_\mu^{\nu_1}(\nu_0)) W_2^2(\mu, \nu_1).$$

Hence, contingent upon Section 2.8, we are indeed studying a generalization of the previous estimates.

We start with the following fact.

Proposition 2.7.9. *Suppose (X, d_X) is non-negatively curved, $P \in \mathcal{P}_2(X)$, and b^* is a barycenter of P. Then, for all $b \in X$,*

$$\frac{1}{2} d_X^2(b, b^*) \int k_{b^*}^b(x) \mathrm{d}P(x) = F(b) - F(b^*).$$

Proof. Expanding and using the definition of the cone metric we calculate:

$$\begin{aligned} d_X^2(b, b^*) k_{b^*}^b(x) &= d_X^2(x, b) + d_X^2(b, b^*) - \| \log_{b^*}(x) - \log_{b^*}(b)\|_{b^*}^2 \\ &= d_X^2(x, b) - d_X^2(x, b^*) + 2\langle \log_{b^*}(x), \log_{b^*}(b) \rangle_{b^*}. \end{aligned}$$

Taking the expectation over $x \sim P$, the second term is 0 by Corollary 2.7.6 and (2.7.5.3). This yields the result. □

Theorem 2.7.10 (Theorem 3.3 [3])**.** *Suppose (X, d_X) is non-negatively curved. Let $P \in \mathcal{P}_2(X)$ with barycenter b^*. Suppose that, for each $x \in \mathrm{supp}(P)$, there exists a geodesic $\gamma_x : [0, 1] \to X$*

connecting b^* to x which is $(0, \lambda)$-extendible. Suppose in addition that b^* remains a barycenter of distribution $P_\lambda = (e_\lambda)_\# P$ where $e_\lambda(x) = \gamma_x^+(1+\lambda)$. Then for all $b \in X$, $F(b)$ satisfies a variance inequality with $C_{\mathsf{VI}} = \lambda/(1+\lambda)$.

We give a simpler proof.

Proof. Fix any $b \in X$. Write the NNC inequality:

$$d_X^2(b,x) \geqslant \frac{\lambda}{1+\lambda} d_X^2(b,b^*) + \frac{1}{1+\lambda} d_X^2(b, e_\lambda(x)) - \frac{\lambda}{(1+\lambda)^2} d_X^2(b^*, x^\lambda).$$

Re-arrange to yield

$$\frac{\lambda}{1+\lambda} d_X^2(b^*, b) \leqslant d_X^2(b,x) - \frac{1}{1+\lambda} d_X^2(b, e_\lambda(x)) + \frac{\lambda}{(1+\lambda)^2} d_X^2(b^*, e_\lambda(x)).$$

We calculate that

$$\frac{\lambda}{(1+\lambda)^2} d_X^2(b^*, x^\lambda) = \frac{1}{1+\lambda} d_X^2(b^*, e_\lambda(x)) - d_X^2(b^*, x).$$

Hence

$$\frac{\lambda}{1+\lambda} d_X^2(b^*, b) \leqslant d_X^2(b,x) - d_X^2(b^*, x) + \frac{1}{1+\lambda}\left(d_X^2(b^*, e_\lambda(x)) - d_X^2(b, e_\lambda(x))\right).$$

We assumed that b^* is a barycenter of the extended distribution, so by taking expectation over P the second term is non-positive and we get the result. □

According to proposition 2.7.9, a variance inequality with $C_{\mathsf{var}} = \lambda/(1+\lambda)$ is equivalent to the statement that, for all $b \in X$,

$$\int k_{b^*}^b(x) \mathrm{d}P(x) \geqslant \frac{\lambda}{1+\lambda}.$$

We'll use this observation next.

Theorem 2.7.11. *Suppose that (X, d_X) is non-negatively curved and let $x, b, b^* \in X$. Assume that for some $\lambda_{\text{in}}, \lambda_{\text{out}} > 0$, there is a geodesic connecting b^* to x which is $(\lambda_{\text{in}}, \lambda_{\text{out}})$-extendible. Then*

$$k_{b^*}^b(x) \geqslant \frac{\lambda_{\text{out}}}{1+\lambda_{\text{out}}} - \frac{1}{\lambda_{\text{in}}}.$$

Remark 2.7.12. *We note that we can combine this bound with the analogous argument at the beginning of Section 2.6 (also using (2.7.5.1)) to get similar strong convexity statements for the squared distance function as in Corollary 2.7.3.*

Proof. Let $\gamma : [0,1] \to X$ be a $(\lambda_{\text{in}}, \lambda_{\text{out}})$-extendible geodesic connecting b^* to x and let $\gamma^+ : [-\lambda_{\text{in}}, 1+\lambda_{\text{out}}] \to S$ be its extension. Let $z = \gamma^+(-\xi)$ where $\xi = \lambda_{\text{in}}/(1+\lambda_{\text{out}})$. Then, it may be easily checked that b^* is a barycenter of the probability measure

$$P := \frac{\xi}{1+\xi}\delta_x + \frac{1}{1+\xi}\delta_z.$$

Now, we wish to apply theorem 2.7.10 to P. To this aim, note that the geodesic γ from b^* to x is $(0, 1+\lambda_{\text{out}})$-extendible by assumption with $e_{\lambda_{\text{out}}}(x) = \gamma^+(1+\lambda_{\text{out}})$. Similarly, we check that the geodesic $\sigma : [0,1] \to S$ connecting b^* to z and defined by $\sigma(t) = \gamma^+(-t\xi)$ is $(0, 1+\lambda_{\text{out}})$-extendible by construction with $e_{\lambda_{\text{out}}}(z) = \gamma^+(-\lambda_{\text{in}})$. Finally, one checks that b^* remains a barycenter of the probability measure $P_{\lambda_{\text{out}}} = (e_{\lambda_{\text{out}}})_\# P$. As a result, theorem 2.7.10 implies that

$$\begin{aligned}\frac{\lambda_{\text{out}}}{1+\lambda_{\text{out}}} &\leqslant \mathbb{E}_{x\sim P}[k^b_{b^*}(x)]\\ &= \frac{\xi}{1+\xi}k^b_{b^*}(x) + \frac{1}{1+\xi}k^b_{b^*}(z).\end{aligned}$$

Finally, the fact (X, d_X) is non-negatively curved implies that $d(x,y) \leqslant \|\log_{b^*}(x) - \log_{b^*}(y)\|_{b^*}$, for all $x, y \in X$. Thus, $k^b_{b^*}(z) \leqslant 1$ for all $b, z \in S$. Hence we obtain

$$\begin{aligned}k^b_{b^*}(x) &\geqslant \frac{1+\xi}{\xi}\left(\frac{\lambda_{\text{out}}}{1+\lambda_{\text{out}}} - \frac{1}{1+\xi}\right)\\ &= \frac{\lambda_{\text{out}}}{1+\lambda_{\text{out}}} - \frac{1}{\lambda_{\text{in}}},\end{aligned}$$

which completes the proof. □

2.8 The tangent cone in $W_2(\mathbb{R}^d)$*

In this section we give a characterization of the tangent cone of Definition 2.7.4, in case $(X, d_X) = (\mathcal{P}_{2,\text{ac}}(\mathbb{R}^d), W_2)$. This allows us to prove that the result of theorem 2.7.11 is a total generalization of theorem 2.6.1. We remark that our proof uses theorem 2.7.11 to provide a geometrically natural analysis of the tangent cone.

Start by fixing a measure $\mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$. At stated in Definition 2.7.4, since $W_2(\mathbb{R}^d)$ is non-negatively curved the Alexandrov angles always exist. Fix two elements of the tangent cone, $u, v \in T_\mu W_2(\mathbb{R}^d)$, and assume $u = [\gamma_0, s_0]$ and $v = [\gamma_1, s_1]$ with representatives such that $W_2(\gamma_i(1), \mu) = s_i$, $i = 1, 2$. Since the angle $\sphericalangle_\mu(\gamma_0, \gamma_1)$ exists we can take the limit with $t = s$. Applying continuity of $\cos$ and re-arranging we find that in this case

$$\|u - v\|_\mu^2 = \lim_{t \to 0} \frac{W_2^2(\gamma_0(t), \gamma_1(t))}{t^2}.$$

The key point of this section is to show another form for the right side. Since our assumptions will always be satisfied in particular by u and v of the form $\log_\mu(\nu)$, this will prove that the squared norm in the definition of the hugging function (Definition 2.7.7) is precisely the L^2 distance between transport maps, as claimed in 2.7.8.

Theorem 2.8.1. *Suppose $\mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ and $\nu_0, \nu_1 \in W_2(\mathbb{R}^d)$. Let γ_i be the constant-speed geodesic $\gamma_i \colon [0,1] \to W_2(\mathbb{R}^d)$ joining μ and ν_i, $i = 1, 2$. Then*

$$\lim_{t \to 0} \frac{W_2^2(\gamma_0(t), \gamma_1(t))}{t^2} = \|T_{\mu \to \nu_0} - T_{\mu \to \nu_1}\|_{L^2(\mu)}^2.$$

Proof. Using the definition of $W_2(\mathbb{R}^d)$ we get that

$$\begin{aligned} W_2^2(\gamma_0(t), \gamma_1(t)) &\leqslant \|T_{\mu \to \gamma_0(t)} - T_{\mu \to \gamma_1(t)}\|_{L_2(\mu)}^2 \\ &= \|((1-t)\operatorname{id} + tT_{\mu \to \nu_0} - ((1-t)\operatorname{id} + tT_{\mu \to \nu_1})\|_{L_2(\mu)}^2 \\ &= t^2 \|T_{\mu \to \nu_0} - T_{\mu \to \nu_1}\|_{L_2(\mu)}^2. \end{aligned}$$

Dividing by t^2 and taking $t \to 0^+$ we get one side of the equality. The other side is a bit more subtle. We start by assuming that the transport map $T_{\mu \to \nu_0} = \nabla \varphi$ and φ is β-smooth for some $\beta \in \mathbb{R}_{>0}$. Then $T_{\mu \to \gamma_0(t)}$ is the gradient of a $(1-t) + t\beta$ smooth and at least $1 - t$ convex function. Apply theorem 2.6.1 to yield

$$\begin{aligned} t^2 \|T_{\mu \to \nu_0} - T_{\mu \to \nu_1}\|_{L_2(\mu)}^2 &= \|T_{\mu \to \gamma_0(t)} - T_{\mu \to \gamma_1(t)}\|_{L_2(\mu)}^2 \\ &\leqslant ((1 - t + t\beta) - (1-t)) W_2^2(\mu, \gamma_0(t)) + W_2^2(\gamma_0(t), \gamma_1(t)) \\ &= \beta t^3 W_2^2(\mu, \nu_0) + W_2^2(\gamma_0(t), \gamma_1(t)). \end{aligned}$$

Dividing by t^2 and letting $t \to 0^+$ shows that

$$\lim_{t\to 0} \frac{W_2^2(\gamma_0(t), \gamma_1(t))}{t^2} = \|T_{\mu\to\nu_0} - T_{\mu\to\nu_1}\|^2_{L_2(\mu)},$$

at least when $T_{\mu\to\nu_0}$ is β-smooth for some β. As we remarked above, this implies that, for smooth $T_{\mu\to\nu_0}$,

$$\|[\gamma_0, W_2(\mu, \gamma_0(1))] - [\gamma_1, W_2(\mu, \gamma_1(1))]\|^2_\mu = \|T_{\mu\to\nu_0} - T_{\mu\to\nu_1}\|^2_{L_2(\mu)}. \tag{2.8.1.1}$$

To finish, we can take a sequence of smooth transport maps approximating a given one (say by Moreau-Yosida regularization): each side of (2.8.1.1) will converge, and so we will have (2.8.1.1) for all $\gamma_0, \gamma_1 \in \Sigma_p$. Thence we can apply the discussion above the theorem to conclude the result. □

Lastly, we mention that this result can be strengthened into a full isomorphism.

Theorem 2.8.2 (Thm. 12.4.4 [7]). *Suppose $\mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$. Then*

$$T_\mu(W_2(\mathbb{R}^d)) \cong \overline{\{\lambda(\nabla\varphi - \mathrm{id})\colon \lambda > 0,\ \nabla\varphi(x) = T_{\mu\to(\nabla\varphi)_\#(\mu)}(x)\ \mu\text{-a.e.}\}}$$

where the overline indicates a completion with respect to $L^2(\mu)$.

2.9 New quantitative stability bounds*

In this section we demonstrate the power of our bound on the hugging function k through the following application. Fix a compact smooth manifold without boundary and with non-negative sectional curvatures M and consider the Wasserstein geometry on $\mathcal{P}_2(M)$. Then $\mathcal{P}_2(M)$ is non-negatively curved and moreover theorem 2.8.1 can be generalized [39, Prop. A.8, A.33], so that we can write, for any $\nu, \mu_1, \mu_2 \in \mathcal{P}_2(M)$,

$$\int_M d(T_{\nu\to\mu_1}, T_{\nu\to\mu_2})^2 d\nu = (1 - k(\mu_1, \mu_2, \nu))W_2^2(\mu_2, \nu) + W_2^2(\mu_1, \mu_2). \tag{2.9.0.1}$$

Hence, analogous to the Euclidean case, lower bounds on the quantity $k(\mu_1, \mu_2, \nu)$ are directly related to upper bounds on $\mathbb{E}[d(T_{\nu\to\mu_1}, T_{\nu\to\mu_2})^2]$. This was stated explicitly as a problem in [8]. Desire for upper bounds on this quantity arises naturally in applications beyond that setting, such as when obtaining fast rates for estimating optimal transport maps [30]. We can apply theorem 2.7.11 and optimal transport theory to get different quantitative stability estimates in terms of analytic conditions on the potentials generating the transport maps.

We make an example explicit. Unfortunately, the manifold case does not exhibit the same simple relationship between the Brenier theorem optimality condition, namely $d^2/2$-convexity, and analytic conditions [61, Chapter 13]. However, under fairly restrictive conditions on the potential φ, we may still obtain sufficient conditions for $d^2/2$-convexity.

Theorem 2.9.1 ([25])**.** *Suppose (M, g) is a smooth compact manifold without boundary. Then there is a constant c depending on M such that if $\varphi\colon M \to \mathbb{R}$ is C^2 and*

$$\|\nabla\varphi\|_\infty + \|\nabla^2\varphi\|_\infty \leqslant c,$$

then φ is $d^2/2$-convex.

With this theorem in hand we can state a new quantitative stability result.

Theorem 2.9.2. *Suppose (M, g) is a smooth compact manifold with non-negative sectional curvatures and no boundary. Fix $\nu, \mu_1, \mu_2 \in \mathcal{P}_2(M)$. Let $\nu \ll d\,\mathrm{vol}_g$. Let T_1, T_2 be the optimal maps between ν and μ_1, μ_2 respectively and assume $T_1^{-1} = \exp(\nabla\varphi)$ for some $d^2/2$-convex $\varphi\colon M \to \mathbb{R}$. Then there is a constant c depending on the manifold M such that whenever $\|\nabla\varphi\|_\infty + \|\nabla^2\varphi\|_\infty \leqslant c$,*

$$\int_M d(T_1, T_2)^2 d\nu \leqslant \frac{c}{c - (\|\nabla\varphi\|_\infty + \|\nabla^2\varphi\|_\infty)} W_2^2(\mu_1, \nu) + W_2^2(\mu_1, \mu_2).$$

Remark 2.9.3. *Contrast with [8], where they get the estimate*

$$\int_M d(T_1, T_2)^2 d\nu \lesssim W_2^2(\mu_1, \mu_2) + W_2(\mu_1, \mu_2) W_2(\nu, \mu_2).$$

Proof of Theorem 2.9.2. If the denominator on the left term is 0 then the bound is trivial. So suppose it isn't. Since $W_2(M)$ is non-negatively curved, we may apply theorem 2.7.11 to see that, if the geodesic between ν and μ_1 is $(\lambda_{\text{in}}, \lambda_{\text{out}})$ extendible, then

$$k(\mu_1, \mu_2, \nu) \geqslant \frac{\lambda_{\text{out}}}{1 + \lambda_{\text{out}}} - \frac{1}{\lambda_{\text{in}}}.$$

Now, by the sufficient condition in 2.9.1 we know

$$\lambda_{\text{in}} \geqslant \frac{c}{\|\nabla\varphi\|_\infty + \|\nabla^2\varphi\|_\infty} - 1.$$

Plugging this in to equation 2.9.0.1, we get

$$\begin{aligned}
\int_M d(T_{\nu\to\mu_1}, T_{\nu\to\mu_2})^2 d\nu &= (1 - k(\mu_1, \mu_2, \nu))W_2^2(\mu_2, \nu) + W_2^2(\mu_1, \mu_2) \\
&\leqslant \left(1 + \frac{1}{\lambda_{\text{in}}} - \frac{\lambda_{\text{out}}}{1 + \lambda_{\text{out}}}\right) W_2^2(\nu, \mu_1) + W_2^2(\mu_1, \mu_2) \\
&\leqslant \left(1 + \frac{1}{\lambda_{\text{in}}}\right) W_2^2(\nu, \mu_1) + W_2^2(\mu_1, \mu_2) \\
&\leqslant \frac{c}{c - (\|\nabla\varphi\|_\infty + \|\nabla^2\varphi\|_\infty)} W_2^2(\mu_1, \nu) + W_2^2(\mu_1, \mu_2).
\end{aligned}$$

□

Chapter 3

Averaging probability distributions: statistics

3.1 Introduction to the problem

In this chapter we develop the first major application of the theory in Chapter 2. We are interested here in perhaps the most basic statistical question of all: how quickly does the sample average converge to the population average? Specifically, let $P \in \mathcal{P}_2(X)$ for X a non-negatively curved metric space. Assume P has a barycenter b^*. Given samples $x_1, \dots, x_n \sim P$, we shall let P_n be the **empirical measure** defined as

$$P_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}.$$

Assume P_n has a barycenter $\hat{b}_n$. Then we wish to understand the quantity

$$\mathbb{E}[d_X^2(\hat{b}_n, b^*)],$$

where the expectation is over samples of size n. We note that in the case where $X = \mathbb{R}^d$, we can use the fact that barycenters are averages to see that this is precisely

$$\mathbb{E}[d_X^2(\hat{b}_n, b^*) = \mathbb{E}[\|\hat{b}_n - b^*\|_2^2] = \frac{\sigma^2(P)}{n}.$$

We shall call a rate of the form c/n for **parametric**. This is the model case that we will try and emulate in our results.

Remark 3.1.1. *We shall assume throughout that barycenters exist. In all cases of interest this is know to hold [38]. Uniqueness of the population barycenter b^* will hold under the hypotheses we consider as well. We note that our results do not require uniqueness of the empirical barycenter.*

3.2 Parametric rates under bi-extendibility

We now state the main result of this section.

Theorem 3.2.1. *Suppose (X, d_X) is non-negatively curved and $P \in \mathcal{P}_2(X)$ with barycenter b^*. If there exists $k_{\min} > 0$ such that $k_{b^*}^{b}(x) > k_{\min}$ (cf. Definition 2.7.7) for all $b, x \in X$, then b^* is unique and any empirical barycenter $\hat{b}_n$ obeys*

$$\mathbb{E}[d_X^2(\hat{b}_n, b^*)] \leqslant \frac{4\sigma^2}{nk_{\min}^2}$$

For a moment taking this statement as granted we immediately obtain the following corollaries courtesy of Section 2.7.

Corollary 3.2.2. *Suppose (X, d_X) is non-negatively curved and $P \in \mathcal{P}_2(X)$ with barycenter b^*. Suppose for each $x \in \operatorname{supp}(P)$ the geodesic from b^* to x is $(\lambda_{\text{in}}, \lambda_{\text{out}})$-extendible such that*

$$k_0 := \frac{\lambda_{\text{out}}}{1 + \lambda_{\text{out}}} - \frac{1}{\lambda_{\text{in}}}$$

is positive. Then

$$\mathbb{E}[d_X^2(\hat{b}_n, b^*)] \leqslant \frac{4\sigma^2(P)}{nk_0^2}.$$

In case $(X, d_X) = W_2(\mathbb{R}^d)$, suppose P has a barycenter $b^* \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ such that for each $\mu \in \operatorname{supp}(P)$, the optimal Kantorovich potential $\varphi_{b^* \to \mu}$ is α-strongly convex and β-smooth for $\beta - \alpha < 1$. Then

$$\mathbb{E}[W_2^2(\hat{b}_n, b^*)] \leqslant \frac{4\sigma^2(P)}{n(1 - (\beta - \alpha))^2}.$$

Remark 3.2.3. *This result significantly advances the state of the art. Early works on empirical barycenters focused on consistency and limit distributions [10, 11, 33]. Asymptotics in the Wasserstein case were considered in [38]. Finite sample rates have been established in the non-positively* *curved regime and mild generalizations thereof [58, 13, 54]. The most significant result valid on spaces of non-negative curvature is from [3]. There, the authors show dimension dependent rates under a variance inequality. In case $(X, d_X) = W_2(\mathbb{R}^d)$, their convergence rates are of the form $n^{-1/d}$.*

Proof of Theorem 3.2.1. Fix n elements $x_1, \dots, x_n \in \operatorname{supp}(P)$ and let $\hat{b}_n$ be their barycenter. We apply our assumption on the hugging function to observe that for each x_i,

$$\| \log_{b^*}(\hat{b}_n) - \log_{b^*}(x_i) \|_{b^*}^2 \leqslant k_{\min} d_X^2(b^*, \hat{b}_n) + d_X^2(\hat{b}_n, x_i).$$

On the other hand, by definition of the tangent cone

$$\| \log_{b^*}(\hat{b}_n) - \log_{b^*}(x_i) \|_{b^*}^2 = d_X^2(b^*, \hat{b}_n) - 2\langle \log_{b^*}(\hat{b}_n) - \mathrm{id}, \log_{b^*}(x_i) - \mathrm{id} \rangle_{b^*} + d_X^2(b^*, x_i).$$

Whence

$$(1 - k_{\min}) d_X^2(b^*, \hat{b}_n) \leqslant 2\langle \log_{b^*}(\hat{b}_n), \log_{b^*}(x_i) \rangle_{b^*} + d_X^2(\hat{b}_n, x_i) - d_X^2(b^*, x_i).$$

If we sum over i and divide by n then the difference on the right hand side is non-positive, so we get

$$(1 - k_{\min}) d_X^2(b^*, \hat{b}_n) \leqslant 2\langle \log_{b^*}(\hat{b}_n), \bar{b}_n \rangle_{b^*}.$$

where we set $\bar{b}_n := \frac{1}{n} \sum_{i=1}^n \log_{b^*}(x_i)$, Applying Cauchy-Schwarz and simplifying we find

$$(1 - k_{\min})^2 d_X^2(b^*, \hat{b}_n) \leqslant 4 \| \bar{b}_n \|_{b^*}^2$$

We observe that by theorem 2.7.5-(3) the $\log_{b^*}(x_i)$ are in a Hilbert space, so that

$$\| \bar{b}_n \|_{b^*}^2 = \frac{1}{n^2} \sum_{i,j=1}^n \langle \log_{b^*}(x_i) - \mathrm{id}, T_{b^* \to x_j} - \mathrm{id} \rangle_{b^*}.$$

Taking each x_i independently distributed according to P and applying theorem 2.7.5-(2) we obtain

$$(1 - k_{\min})^2 \mathbb{E}[d_X^2(b^*, \hat{b}_n)] \leqslant \frac{4}{n} \mathbb{E}[d_X^2(b^*, x)] = \frac{4\sigma^2}{n}.$$

This proves the result. □

Chapter 4

Averaging probability distributions: optimization

In this chapter, we turn to the question of computing Wasserstein barycenters. We develop a general machinery to study first-order optimization methods for this purpose, and apply it to the case of distributions supported on Gaussians known as the **Bures-Wasserstein** manifold. We provide the first analysis of exponential convergence of gradient descent in this setting, resolving an open question of [6].

4.1 Introduction and main results

Establishing fast convergence of first-order methods is usually intimately related to convexity. As we well know by now, however, the barycenter functional on $W_2(\mathbb{R}^d)$ cannot be expected to be (geodesically) convex. Indeed, as we can see in Figure 4.1, the barycenter functional may even be *concave* along geodesics.

Fortunately, the optimization literature describes conditions for global convergence of first order algorithms even for non-convex objectives. We shall employ one such condition, a Polyak-Łojasiewicz (PL) inequality of the form (4.3.2.1), which is known to yield linear convergence for a variety of gradient methods on flat spaces even in absence of convexity [32].

The main work of this chapter is in two parts. The first is to develop general machinery towards a PL inequality for the barycenter functional, and the second is to instantiate it for

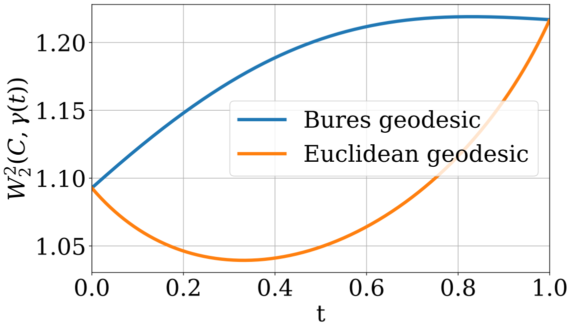

Figure 4-1: Example of the non-geodesic convexity of W_2^2. Displayed is the squared Bures distance along a Wasserstein geodesic and a Euclidean geodesic. Details are given in Appendix B.0.2

the Bures-Wasserstein case. The consequences for distributions supported on Gaussians are given below.

Theorem 4.1.1. *Fix $\zeta \in (0, 1]$ and let P be a distribution supported on mean-zero Gaussian measures whose covariance matrices Σ satisfy $\|\Sigma\|_{\text{op}} \leqslant 1$ and $\det \Sigma \geqslant \zeta$. Then, P has a unique barycenter b^*, and Gradient Descent (Algorithm 1) initialized at $b_0 \in \text{supp}(P)$ yields a sequence $(b_T)_{T \geqslant 1}$ such that*

$$W_2^2(b_T, b^*) \leqslant \frac{2}{\zeta}\Big(1 - \frac{\zeta^2}{4}\Big)^T [F(b_0) - F(b^*)] \,.$$

The above theorem establishes a linear rate of convergence for gradient descent and answers a question left open in [6]. Moreover, when P is an empirical distribution, combined with the existing results of [3, 36], it yields a procedure to estimate Bures-Wasserstein barycenters at the parametric rate after a number of iterations that is logarithmic in the sample size n.

Still in the Gaussian case, we also show that a stochastic gradient descent (SGD) algorithm converges to the true barycenter at a parametric rate.

Theorem 4.1.2. *Fix $\zeta \in (0, 1]$ and let P be a distribution supported on mean-zero Gaussian measures whose covariance matrices Σ satisfy $\|\Sigma\|_{\text{op}} \leqslant 1$ and $\det \Sigma \geqslant \zeta$. Then, P has a unique barycenter b^*, and Stochastic Gradient Descent (Algorithm 2) run on a sample of size $n + 1$ from P returns a Gaussian measure b_n such that*

$$\mathbb{E} W_2^2(b_n, b^*) \leqslant \frac{96\sigma^2(P)}{n\zeta^5}.$$

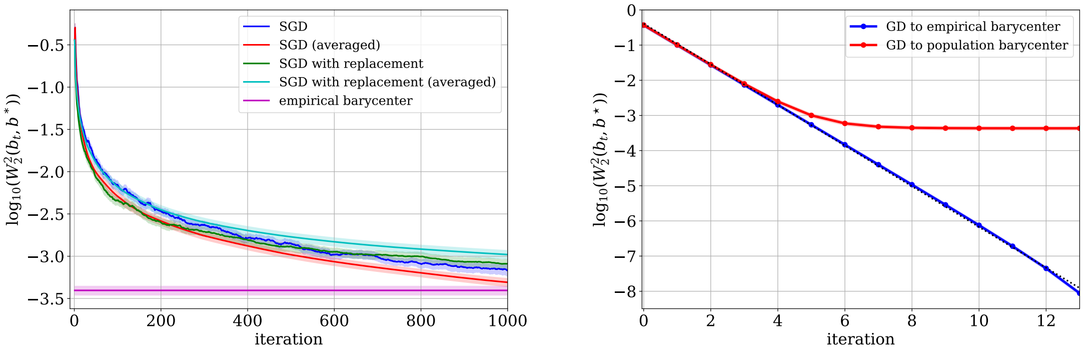

Figure 4-2: Left. Convergence of SGD on Bures manifold for $n = 1000$, $d = 3$, and $b^\star = \gamma_{0,I_3}$. Right: linear convergence of GD on the same problem.

This theorem shows that SGD yields an estimator b_n, different from the empirical barycenter $\hat{b}_n$, that also converges at the parametric rate to $b^\star$. Hence we actually have two ways to estimate an empirical barycenter: we can either apply theorem 4.1.1 to the empirical distribution or theorem 4.1.2 to the population distribution. The latter exhibits slower convergence but has much cheaper iterations.

As far as we are aware, these results provide the first non-asymptotic rates of convergence for first-order methods on the Bures-Wasserstein manifold. The problem of Bures-Wasserstein barycenters over Gaussians has been studied since the 1990's [35]. Previous works have focused primarily on empirical and asymptotic explorations, as well as connections with matrix analysis [6, 9].

Remark 4.1.3. *The assumption $\|\Sigma\|_{\mathrm{op}} \leqslant 1$ is simply a normalization and changes nothing. A natural sufficient condition for $\det \Sigma \geqslant \zeta$ to be satisfied is when all the eigenvalues of the covariance matrix Σ are lower bounded by a constant $\lambda_{\min} > 0$. In this case, the parameter $\zeta \geqslant \lambda_{\min}^d$ can be exponentially small in the dimension. Note, however, that in this case the Gaussian measure is quite degenerate in the sense that the density of $\gamma_{0,\Sigma}$ is exponentially large at 0.*

In Figure 4-2, we present an experiment confirming these two results; see Appendix B for more details and further numerical results.

4.2 Local to global phenomena for Wasserstein barycenters

The gradient of the barycenter functional is defined as

$$\nabla F(b) := \mathbb{E}_{\mu\sim P}[\mathrm{id} - T_{b\to\mu}] \qquad \forall b \in \mathcal{P}_{2,\mathrm{ac}}.$$

This formula can be justified rigorously, but we take it simply as a convention [7]. The early work of Agueh and Carlier in 2011 [1] showed that the norm of this gradient is closely related to the barycenter problem. Specifically, they proved:

Theorem 4.2.1 ([1] Prop. 3.8)**.** *Suppose P has finite support in $\mathcal{P}_{2,\mathrm{ac}}$. Fix $b \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$. For each $\mu \in \mathrm{supp}(P)$, let the Monge-Kantorovich potentials generating the optimal coupling from b to μ be $\varphi_{b\to\mu}$ so that $T_{b\to\mu} = \nabla\varphi_{b\to\mu}$. Then b is the barycenter of P if and only if there exists a constant C such that*

$$\mathbb{E}_{\mu\sim P}[\varphi_{b\to\mu_i}(x)] \leqslant C + \frac{1}{2}\|x\|_2^2 \qquad \forall x \in \mathbb{R}^d$$

and such that equality holds for b-a.e. $x \in \mathbb{R}^d$.

The crucial point here is that the statement that $\mathbb{E}_{\mu\sim P}[\varphi_{b\to\mu_i}(x)] = \frac{1}{2}\|x\|_2^2$ b-a.e. is equivalent to b being a critical point for F:

$$\|\nabla F(b)\|_{L^2(b)} = 0.$$

Indeed, under some mild regularity and support assumptions on b and the support of P, their theorem can be strengthened to an unconditional if and only if: $\|\nabla F(b)\|_{L^2(b)} = 0$ if and only if $b = b^*$ [48].

Taking this interpretation we have found a promising fact. The positive curvature of $W_2(\mathbb{R}^d)$ *a priori* blocks convexity of $F(b)$, and yet, $F(b)$ still has one of the local-to-global attributes of convex functions: critical points are global optima. It is this fact which motivates the previous and present study of first-order methods for computing barycenters.

4.3 Sufficient conditions for first-order methods to converge on $W_2(\mathbb{R}^d)$

In this section we describe an approach to combine the 1-smoothness of $F(b)$ (cf. proposition 2.3.4) with a certain quantitative weakening of strong convexity to obtain fast rates of convergence for first-order algorithms for Wasserstein barycenters. To begin, we shall describe the methods in consideration.

Given a sequence of step-sizes $(\eta_t)_{t\geqslant 1}$ and an initial point $b_0 \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ we consider the **gradient descent (GD) dynamics**:

$$b_{t+1} := (\text{id} - \eta_t \nabla F(b_t))_{\#} b_t \qquad t = 1, 2, \dots \tag{4.3.0.1}$$

We shall also consider the **stochastic gradient descent (SGD) dynamics**: given a sequence of step sizes $(\eta_t)_{t=1}^{n-1}$, an initial point $b_0 \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$, and n samples $\mu_t \sim P$, $t = 1, \dots, n$, the iterates are

$$b_{t+1} := (\text{id} + \eta_t (T_{b_t \to \mu_{t+1}} - \text{id}))_{\#} b_t \qquad t = 0, \dots, n-1. \tag{4.3.0.2}$$

To analyze these dynamics we shall first need a calculus version of the 1-smoothness property of F.

Proposition 4.3.1. *For any $b_0, b_1 \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ such that $F(b_0) < \infty$, the barycenter functional satisfies the smoothness inequality*

$$F(b_1) \leqslant F(b_0) + \langle \nabla F(b_0), T_{b_0 \to b_1} - \text{id} \rangle_{b_0} + \frac{1}{2} W_2^2(b_0, b_1). \tag{4.3.1.1}$$

Moreover, for any $b \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ and $b^+ := [\text{id} - \nabla F(b)]_{\#} b$, it holds.

$$F(b^+) - F(b) \leqslant -\frac{1}{2} \|\nabla F(b)\|_b^2. \tag{4.3.1.2}$$

Proof. Let $(b_s)_{s\in[0,1]}$ be the constant-speed geodesic between arbitrary $b_0, b_1 \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$. From the non-negative curvature inequality (2.2.1.2), it holds that for any $s \in (0, 1]$,

$$\int \frac{W_2^2(b_s, \mu) - W_2^2(b_0, \mu)}{s} \, \mathrm{d}P(\mu) \geqslant \int [W_2^2(b_1, \mu) - W_2^2(b_0, \mu)] \, \mathrm{d}P(\mu) - (1 - s) W_2^2(b_0, b_1).$$

We will apply the dominated convergence theorem to the left-hand side. To do this, we observe that

$$\begin{aligned}\frac{1}{s}\left|W_2^2(b_s,\mu)-W_2^2(b_0,\mu)\right| &= \frac{1}{s}(W_2(b_s,\mu)+W_2(b_0,\mu))(W_2(b_s,\mu)-W_2(b_0,\mu)) \\ &\leqslant \frac{1}{s}(W_2(b_0,b_s)+2W_2(b_0,\mu))W_2(b_0,b_s) \\ &\leqslant (W_2(b_0,b_1)+2W_2(b_0,\mu))W_2(b_0,b_1).\end{aligned}$$

By our assumption that $F(b_0)<\infty$, the result is integrable. Hence, the LHS converges to

$$\begin{aligned}\int \frac{\mathrm{d}}{\mathrm{d}s}W_2^2(b_s,\mu)\big|_{s=0_+}\,\mathrm{d}P(\mu) &= -2\int \langle T_{b_0\to\mu}-\mathrm{id}, T_{b_0\to b_1}-\mathrm{id}\rangle_{L_2(b_0)}\,\mathrm{d}P(\mu) \\ &= 2\langle \nabla F(b_0), T_{b_0\to b_1}-\mathrm{id}\rangle_{b_0}\,,\end{aligned}$$

where in the first identity, we used the characterization of [7, Prop. 7.3.6]. Rearranging terms yields (4.3.1.1).

Noticing that $W_2^2(b,b^+)=\|-\nabla F(b)\|_b^2$, the second equation follows from the first. □

We now introduce a certain quantitative weakening of strong convexity which shall be useful in the sequel.

Definition 4.3.2. *We say that P satisfies a Polyak-Łojasiewicz (PL) inequality at $b\in\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ if*

$$2C_{\mathsf{PL}}(F(b)-F(b^*))\leqslant\|\nabla F(b)\|_{L^2(b)}^2. \tag{4.3.2.1}$$

Remark 4.3.3. *PL inequalities are a standard item in the theory of optimization, for they are essentially a minimal requirement to derive typical rates of convergence for optimization methods [32]. We note as well that a PL inequality for the barycenter functional can be viewed as a quantitative strengthening of the "critical points are optimal" statement from the previous section.*

As the following theorems evidence, if F satisfies a PL inequality at each iterate of gradient descent or stochastic gradient descent then we can recover the usual rates from Euclidean optimization.

Theorem 4.3.4 (Exponential convergence for GD)**.** *Suppose F satisfies a PL inequality at each iterate of the gradient descent dynamics $(b_t)_{t<T}$. Then*

$$F(b_T)-F(b^*)\leqslant(1-C_{\mathsf{PL}})^T(F(b_0)-F(b^*)).$$

Proof of Theorem 4.3.4. The PL inequality implies $F(b_t) < \infty$ for all t. We may thus apply the smoothness (4.3.1.2) and PL (4.3.2.1) inequalities, to see that

$$F(b_{t+1}) - F(b_t) \leqslant -C_{\mathsf{PL}}[F(b_t) - F(b^*)].$$

It yields $F(b_{t+1}) - F(b^*) \leqslant (1 - C_{\mathsf{PL}})[F(b_t) - F(b^*)]$, which gives the result. □

We also get a $1/n$ rate for the SGD dynamics.

Theorem 4.3.5 ($1/n$ convergence for SGD)**.** *Suppose that there exists a constant $C_{\mathsf{PL}} > 0$ such that F satisfies the PL inequality (4.3.2.1) at all iterates $(b_t)_{0\leqslant t\leqslant n}$ of SGD run with step size*

$$\eta_t = C_{\mathsf{PL}}\Big(1 - \sqrt{1 - \frac{2(t+k)+1}{C_{\mathsf{PL}}^2(t+k+1)^2}}\Big) \leqslant \frac{2}{C_{\mathsf{PL}}(t+k+1)}, \tag{4.3.5.1}$$

where we take $k = 2/C_{\mathsf{PL}}^2 - 1 \geqslant 0$. Then,

$$\mathbb{E}F(b_n) - F(b^*) \leqslant \frac{3\sigma^2(P)}{C_{\mathsf{PL}}^2 n}.$$

The proof is relegated to the appendix subsection A.2.1. Although mostly standard, there is a critical usage of the non-negative curvature of $W_2(\mathbb{R}^d)$.

4.4 An integrated Polyak-Łojasiewicz inequality

The aim of this section is to prove the following "integrated PL inequality".

Lemma 4.4.1 (Integrated PL)**.** *Let P satisfy a variance inequality with constant C_{var} and let $b \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ be such that the barycenter b^* of P is absolutely continuous w.r.t. b. Assume further the following measurability conditions: there exists a $P \otimes b^*$, $P \otimes b$-integrable mapping $\varphi\colon \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, $(\mu, x) \mapsto \varphi_{b\to\mu}(x)$, such that for P-almost every $\mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, $\varphi_{b\to\mu}\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is a Kantorovich potential for the optimal transport map from b to μ. Then,*

$$F(b) - F(b^*) \leqslant \frac{2}{C_{\mathsf{var}}} \left(\int_0^1 \|\nabla F(b)\|_{L^2(b_s)} ds\right)^2,$$

where $(b_s)_{s\in[0,1]}$ is the constant-speed W_2-geodesic joining b to b^*.

Remark 4.4.2. *We comment on deriving a full PL inequality from the above. If we know that $\|db_s/db\|_\infty \leqslant C$, then this immediately yields a full PL inequality. The difficulty with such an approach lies in the fact that we must guarantee it to be true when b is an iterate of the gradient descent trajectory. In fact, sup norm upper bounds on the iterates can be easily maintained throughout the trajectory as they hold along W_2-geodesics. To complete such an argument we would need a lower bound of the form $db \geqslant c$. However, ensuring lower bounds along W_2-geodesics is a difficult open problem [53].*

The following lemma will be useful for us. It appears in [39, Lem. A.1] in the case of Lipschitz functions. A minor modification of their proof allows us to handle locally Lipschitz rather than only Lipschitz functions, which we include in Appendix A.2.2.

Lemma 4.4.3. *Let $(b_s)_{s\in[0,1]}$ be a Wasserstein geodesic in $\mathcal{P}_2(\mathbb{R}^d)$. Let $\Omega \subseteq \mathbb{R}^d$ be a convex open subset for which $b_0(\Omega) = b_1(\Omega) = 1$. Then, for any function $f : \mathbb{R}^d \to \mathbb{R}$ which is locally Lipschitz on Ω, it holds that*

$$\Big|\int f\,\mathrm{d}b_0 - \int f\,\mathrm{d}b_1\Big| \leqslant W_2(b_0, b_1)\int_0^1 \|\nabla f\|_{L^2(b_s)}\,\mathrm{d}s.$$

Proof of Lemma 4.4.1. By Kantorovich duality (cf. theorem 2.1.1),

$$\begin{aligned}\frac{1}{2}W_2^2(b,\mu) &= \int\Big(\frac{\|\cdot\|^2}{2} - \varphi_{\mu\to b}\Big)\,\mathrm{d}\mu + \int\Big(\frac{\|\cdot\|^2}{2} - \varphi_{b\to\mu}\Big)\,\mathrm{d}b,\\ \frac{1}{2}W_2^2(b^*,\mu) &\geqslant \int\Big(\frac{\|\cdot\|^2}{2} - \varphi_{\mu\to b}\Big)\,\mathrm{d}\mu + \int\Big(\frac{\|\cdot\|^2}{2} - \varphi_{b\to\mu}\Big)\,\mathrm{d}b^*.\end{aligned}$$

This yields the inequality

$$F(b) - F(b^*) \leqslant \int\Big(\frac{\|\cdot\|^2}{2} - \int \varphi_{b\to\mu}\,\mathrm{d}P(\mu)\Big)\,\mathrm{d}(b - b^*),$$

where we have used the integrability assumption to swap integrals. Let $\overline{\varphi} := \int \varphi_{b\to\mu}\,\mathrm{d}P(\mu)$; this is a proper LSC convex function $\mathbb{R}^d \to \mathbb{R}\cup\{\infty\}$. We apply lemma 4.4.3 with $\Omega = \operatorname{int}\operatorname{dom}\overline{\varphi}$. Since $\overline{\varphi}$ is locally Lipschitz on the interior of its domain and $b^* \ll b$, then $b(\Omega) = b^*(\Omega) = 1$, whence

$$F(b) - F(b^*) \leqslant W_2(b, b^*)\int_0^1 \|\nabla\overline{\varphi} - \mathrm{id}\|_{L^2(b_s)}\,\mathrm{d}s \leqslant \sqrt{\frac{2[F(b) - F(b^*)]}{C_{\mathsf{var}}}}\int_0^1 \|\nabla\overline{\varphi} - \mathrm{id}\|_{L^2(b_s)}\,\mathrm{d}s.$$

Square and rearrange to yield

$$F(b) - F(b^*) \leqslant \frac{2}{C_{\mathsf{var}}} \Big(\int_0^1 \|\nabla\overline{\varphi} - \mathrm{id}\|_{L^2(b_s)} \Big)^2 \,\mathrm{d}s.$$

Recognizing that $\nabla F(b) = \mathrm{id} - \nabla\overline{\varphi}$ yields the result. □

4.5 Specializing to the Bures-Wasserstein case

As discussed in Remark 4.4.2, obtaining a full PL inequality from the integrated PL in general is not known. In this section, we apply the integrated PL inequality to obtain a bona fide PL for distributions supported on Gaussians.

Identifying a centered non-degenerate Gaussian measure with its covariance matrix, the Wasserstein geometry induces a Riemannian structure on the space of positive definite matrices, known as the Bures geometry. Accordingly, we now refer to the barycenter of P as the *Bures-Wasserstein barycenter* [20, 9].

4.5.1 Bures-Wasserstein gradient descent algorithms

We now specialize both GD and SGD when P is supported on mean-zero Gaussian measures. In this case, the updates of both algorithms take a remarkably simple form. To see this, for $m \in \mathbb{R}^d$, $\Sigma \in \mathbb{S}^d_{++}$, let $\gamma_{m,\Sigma}$ denote the Gaussian measure on $\mathbb{R}^d$ with mean m and covariance matrix Σ. In particular, the optimal coupling between γ_{m_0,Σ_0} and γ_{m_1,Σ_1} has the explicit form

$$x \mapsto T_{\gamma_{\mu_0,\Sigma_0} \to \gamma_{\mu_1,\Sigma_1}}(x) := m_1 + \Sigma_0^{-1/2} (\Sigma_0^{1/2} \Sigma_1 \Sigma_0^{1/2})^{1/2} \Sigma_0^{-1/2} (x - m_0). \qquad (4.5.0.1)$$

Observe that $T_{\gamma_{\mu_0,\Sigma_0} \to \gamma_{\mu_1,\Sigma_1}}$ is affine, and thus $\int T_{\gamma_{\mu_0,\Sigma_0} \to \gamma} \,\mathrm{d}P(\gamma)$ is affine.

This means that all of the GD (or SGD) iterates are Gaussian measures, so it suffices to keep track of the mean and covariance matrix of the current iterate. For both GD and SGD, the update equation for the descent step decomposes into two decoupled equations: an update equation for the mean, and an update equation for the covariance matrix. Moreover, the update equation for the mean is trivial, corresponding to a simple GD or SGD procedure on the objective function $m \mapsto \int \|m - m(\mu)\|^2 \,\mathrm{d}P(\mu)$. Therefore, for simplicity and without loss of generality, we consider only mean-zero Gaussians throughout this sec-

Algorithm 1 Bures-Wasserstein GD

```
procedure Bures-GD($\Sigma_0, P, T$)
    for $t = 1, \dots, T$ do
        $S_t \leftarrow \int \Sigma_{t-1}^{-1/2}\{\Sigma_{t-1}^{1/2}\Sigma(\mu)\Sigma_{t-1}^{1/2}\}^{1/2}\Sigma_{t-1}^{-1/2}\,\mathrm{d}P(\mu)$
        $\Sigma_t \leftarrow S_t\Sigma_{t-1}S_t$
    end for
    return $\Sigma_T$
end procedure
```

Algorithm 2 Bures-Wasserstein SGD

```
procedure Bures-SGD($\Sigma_0, (\eta_t)_{t=1}^T, (K_t)_{t=1}^T$)
    for $t = 1, \dots, T$ do
        $\hat{S}_t \leftarrow \Sigma_{t-1}^{-1/2}\{\Sigma_{t-1}^{1/2}K_t\Sigma_{t-1}^{1/2}\}^{1/2}\Sigma_{t-1}^{-1/2}$
        $\Sigma_t \leftarrow ((1-\eta_t)I_D + \eta_t\hat{S}_t)\Sigma_{t-1}((1-\eta_t)I_D + \eta_t\hat{S}_t)$
    end for
    return $\Sigma_T$
end procedure
```

tion and we simply have to write down the update equations for the covariance matrix Σ_t of the iterate. They are summarized in Algorithms 1 and 2.

In the rest of this section, we prove the guarantees for GD and SGD on the Bures-Wasserstein manifold given in theorems 4.1.1 and 4.1.2.

4.5.2 Proof of the main results

For simplicity, we make the following reductions: we assume that the Gaussians are centered (see previous subsection) and that the eigenvalues of the covariance matrices of the Gaussians are uniformly bounded above by 1. The latter assumption is justified by the observation that if there is a uniform upper bound on the eigenvalues of the covariance matrices, then we can simply rescale them so that the bound becomes 1. As can be easily checked, the barycenter thus obtained will be the scaled version of the original barycenter, so there is no harm in doing this.

While the centering and scaling assumptions stated above can be made without loss of generality, our results require the following regularity condition. Note that it is equivalent to a uniform upper bound on the densities of the Gaussians.

Definition 4.5.1 (ζ-regular)**.** *Fix $\zeta \in (0, 1]$. A distribution $P \in \mathcal{P}_2(\mathbb{R}^d)$ is said to be ζ-regular if its support is contained in*

$$\mathcal{S}_\zeta = \left\{\gamma_{0,\Sigma} \,:\, \Sigma \in \mathbb{S}^d_{++},\ \|\Sigma\|_{\mathrm{op}} \leqslant 1,\ \det \Sigma \geqslant \zeta\right\}. \tag{4.5.1.1}$$

We use this definition to prove our main theorems with the following four steps. We show that, for any ζ-regular distribution P,

1. Corollary 4.5.5: P has a barycenter in $\mathcal{S}_\zeta$.

2. Theorem 4.5.7: P obeys a variance inequality with $C_{\mathsf{var}} = \zeta$.

3. Theorem 4.5.8: P obeys a PL-inequality uniformly over $\mathcal{S}_\zeta$ with $C_{\mathsf{PL}} = \zeta^2/4$.

4. Corollary 4.5.6: GD and SGD initialized in $\mathcal{S}_\zeta$ remain in $\mathcal{S}_\zeta$.

We thus obtain a PL inequality throughout the optimization trajectories and can apply our optimizations results from Section 4.3 to conclude. For the rest of this section, we execute this proof plan.

We lay the groundwork for our proof by defining the following strong form of geodesic convexity for $\mathcal{S}_\zeta$.

Definition 4.5.2 (Defn. 7.4 [1])**.** *Fix $\mu, \nu_0, \nu_1 \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$. Let T_i be the optimal coupling from μ to ν_i, $i = 0, 1$. Then the* **generalized geodesic** *from ν_0 to ν_1 with base μ is the interpolated curve*

$$\nu_t^\mu := ((1-t)T_0 + tT_1)_\# \mu.$$

We say that a functional $G \colon \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) \to \mathbb{R}$ is **convex along generalized geodesics** *if for all generalized geodesics ν_t^μ, the function $G \circ \nu_t^\mu \colon [0, 1] \to \mathbb{R} \cup \{\infty\}$ is convex. A set $S \subset \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ is said to be convex along generalized geodesics if the convex indicator*

$$\iota_S(\mu) := \begin{cases} 0 & \mu \in S \\ \infty & \mu \notin S \end{cases}$$

is convex along generalized geodesics.

We next demonstrate two important functionals which are convex along generalized geodesics.

Lemma 4.5.3. *For a measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, let $M(\mu) := \int x \otimes x \, \mathrm{d}\mu(x)$. Then, the functional $\mu \mapsto \|M(\mu)\|_{\mathrm{op}} = \lambda_{\max}(M(\mu))$ is convex along generalized geodesics on $\mathcal{P}_2(\mathbb{R}^d)$.*

Proof. Let S^{d-1} denote the unit sphere of $\mathbb{R}^d$ and observe that for any $e \in S^{d-1}$ the function $x \mapsto \langle x, e\rangle^2$ is convex on $\mathbb{R}^d$. By known results for geodesic convexity in Wasserstein space (see [7, Prop. 9.3.2]), the functional $\mu \mapsto \int \langle \cdot, e\rangle^2 \, \mathrm{d}\mu = \langle e, M(\mu)e\rangle$ is convex along generalized geodesics in $\mathcal{P}_2(\mathbb{R}^d)$; hence, so is the functional $\mu \mapsto \max_{e \in S^{d-1}} \langle e, M(\mu)e\rangle = \|M(\mu)\|_{\mathrm{op}}$. □

The next lemma establishes convexity along generalized geodesics of $\mu \mapsto -\ln\det\Sigma(\mu)$. It follows from specializing lemma A.2.3 in the Appendix to the Bures-Wasserstein manifold.

Lemma 4.5.4. *The functional $\gamma_{0,\Sigma} \mapsto -\sum_{i=1}^{d} \ln \lambda_i(\Sigma)$ is convex along generalized geodesics on the space of non-degenerate Gaussian measures.*

It follows readily from lemmas 4.5.3 and 4.5.4 that the set $\mathcal{S}_\zeta$ is convex along generalized geodesics. Combining this with the fact that ζ-regular distributions have barycenters (proposition A.2.1 in the Appendix) and a result of Agueh and Carlier [1, Prop. 7.6] we get the following.

Corollary 4.5.5. *If P is ζ-regular then it has a barycenter $b^* \in \mathcal{S}_\zeta$.*

Moreover since SGD moves along geodesics and is initialized at $b_0 \in \operatorname{supp} P \subset \mathcal{S}_\zeta$, then all the iterates of SGD stay in $\mathcal{S}_\zeta$. To show that the same holds for GD, observe that the set $\log_{b_t}(\mathcal{S}_\zeta)$ is convex. Therefore, $-\nabla F(b_t) = \int (T_{b_t \to \mu} - \mathrm{id}) \, \mathrm{d}P(\mu) \in \log_{b_t}(\mathcal{S}_\zeta)$ as a convex combination of elements in this set. This is equivalent to $b_{t+1} = \exp_{b_t}(-\nabla F(b_t)) \in \mathcal{S}_\zeta$. These observations yield the following corollary.

Corollary 4.5.6. *The set $\mathcal{S}_\zeta$ is convex along generalized geodesics and when initialized in $\operatorname{supp} P$, the iterates of both GD and SGD remain in $\mathcal{S}_\zeta$.*

The next result establishes uniqueness and a variance inequality.

Theorem 4.5.7. *Suppose P is a ζ-regular distribution. Then P has a unique barycenter $b^* \in \mathcal{S}_\zeta$ and obeys a variance inequality with $C_{\mathsf{var}} = \zeta$.*

Proof. By lemma A.2.2 we know that the Jacobian of the transport map between any two elements of $\mathcal{S}_\zeta$ has eigenvalues between ζ and $1/\zeta$. We can thus apply theorem 2.3.7 to get a variance inequality with $C_{\mathsf{var}} = \zeta$. This implies uniqueness. □

We can now show a PL inequality over all of $\mathcal{S}_\zeta$.

Theorem 4.5.8. *Fix $\zeta \in (0, 1]$, and let P be a ζ-regular distribution. Then, the barycenter functional F satisfies the PL inequality with constant $C_{\mathsf{PL}} = \zeta^2/4$ uniformly at all $b \in \mathcal{S}_\zeta$:*

$$F(b) - F(b^*) \leqslant \frac{2}{\zeta^2} \|\nabla F(b)\|_b^2 .$$

Proof. For any $\gamma_{0,\Sigma} \in \mathcal{S}_\zeta$, the eigenvalues of Σ are in $[\zeta, 1]$. Let $(\tilde{b}_s)_{s\in[0,1]}$ be the constant-speed geodesic between $\tilde{b}_0 := b := \gamma_{0,\Sigma}$ and $\tilde{b}_1 := b^* := \gamma_{0,\Sigma^*}$. Combining lemma 4.4.1 (with an additional use of the Cauchy-Schwarz inequality) and theorem 4.5.7, we get

$$F(b) - F(b^*) \leqslant \frac{2}{\zeta} \int_0^1 \int \|\nabla F(b)\|_2^2 \, \mathrm{d}\tilde{b}_s \, \mathrm{d}s. \tag{4.5.8.1}$$

Define a random variable $X_s \sim \tilde{b}_s$ and observe that

$$\int \|\nabla F(b)\|_2^2 \, \mathrm{d}\tilde{b}_s = \mathbb{E}\|(\tilde{M} - I_D)X_s\|_2^2 , \quad \text{where } \tilde{M} = \int \Sigma^{-1/2} (\Sigma^{1/2} S \Sigma^{1/2})^{1/2} \Sigma^{-1/2} \, \mathrm{d}P(\gamma_{0,S}).$$

Moreover, recall that $X_s = sX_1 + (1-s)X_0$ where $X_0 \sim \tilde{b}_0$ and $X_1 \sim \tilde{b}_1$ are optimally coupled. Therefore, by Jensen's inequality, we have for all $s \in [0, 1]$,

$$\mathbb{E}\|(\tilde{M} - I_D)X_s\|_2^2 \leqslant s\mathbb{E}\|(\tilde{M} - I_D)X_1\|_2^2 + (1-s)\mathbb{E}\|(\tilde{M} - I_D)X_0\|_2^2 \leqslant \frac{1}{\zeta}\mathbb{E}\|(\tilde{M} - I_D)X_0\|_2^2 ,$$

where in the second inequality, we used the fact that

$$\mathbb{E}\|(\tilde{M} - I_D)X_1\|_2^2 = \operatorname{tr}\big(\Sigma^*(\tilde{M} - I_D)^2\big) \leqslant \|\Sigma^*\Sigma^{-1}\|_{\mathrm{op}} \operatorname{tr}\big(\Sigma(\tilde{M} - I_D)^2\big) \leqslant \frac{1}{\zeta}\mathbb{E}\|(\tilde{M} - I_D)X_0\|_2^2 .$$

Together with (4.5.8.1), it yields

$$F(b) - F(b^*) \leqslant \frac{2}{\zeta^2}\mathbb{E}\|(\tilde{M} - I_D)X_0\|_2^2 = \frac{2}{\zeta^2}\|\nabla F(b)\|_b^2 .$$

□

By combining corollary 4.5.6 with theorem 4.5.8, we can instantiate theorems 4.3.4 and 4.3.5 to obtain our main results for this chapter, theorem 4.1.1 and theorem 4.1.2 respectively.

Appendix A

Additional results and omitted proofs

A.1 Additional results and omitted proofs from Chapter 2

Lemma A.1.1. *Suppose $\varphi\colon \mathbb{R}^d \to \mathbb{R}$ is an α-strongly convex proper lower-semi continuous function. Then for arbitrary $x \in \mathbb{R}^d$ and almost every $y \in \mathbb{R}^d$*

$$\varphi(x) + \varphi^*(x) \geqslant \langle x, y\rangle + \frac{\alpha}{2}\|x - \nabla\varphi^*(y)\|_2^2.$$

Proof. Fix $x, y \in \mathbb{R}^d$, and let $x_0 \in \partial\varphi^*(y)$. Then $y \in \partial\varphi(x_0)$ (by e.g. Prop. 2.4 [60]), so

$$\varphi((1-t)x_0 + tx) \geqslant t\langle x - x_0, y\rangle + \varphi(x_0).$$

Using this, we have

$$\begin{aligned}
\varphi^*(y) &\geqslant \langle x_0, y\rangle - \varphi(x_0) \\
&\geqslant \langle x, y\rangle + \frac{1}{t}(\varphi(x_0) - \varphi((1-t)x_0 + tx)) - \varphi(x_0) \\
&\geqslant \langle x, y\rangle + \varphi(x_0) - \varphi(x) + \frac{\alpha}{2}(1-t)\|x - x_0\|_2^2 - \varphi(x_0),
\end{aligned}$$

where the last inequality follows from α-strong convexity. Rearranging and taking $t \to 0$ yields

$$\varphi(x) + \varphi(y) \geqslant \langle x, y\rangle + \frac{\alpha}{2}\|x - x_0\|_2^2.$$

We know that φ^* is differentiable at almost every $y \in \mathbb{R}^d$, in which case $x_0 = \nabla\varphi^*(y)$. This gives the result. □

A.2 Additional results and omitted proofs from Chapter 4

A.2.1 Proof of Theorem 4.3.5

We begin by noting that the step size η_t is chosen to solve the equation

$$1 - 2C_{\mathsf{PL}}\eta_t + \eta_t^2 = \Big(\frac{t+k}{t+k+1}\Big)^2.$$

We use the definition of Wasserstein distance to calculate:

$$\begin{aligned} W_2^2(b_{t+1}, \mu) &\le \|\log_{b_t} b_{t+1} - \log_{b_t} \mu\|_{b_t}^2 = \|\eta_t \log_{b_t} \mu_{t+1} - \log_{b_t} \mu\|_{b_t}^2 \\ &= \|\log_{b_t} \mu\|_{b_t}^2 + \eta_t^2 \|\log_{b_t} \mu_{t+1}\|_{b_t}^2 - 2\eta_t \langle \log_{b_t} \mu, \log_{b_t} \mu_{t+1}\rangle_{b_t}. \end{aligned}$$

Taking the expectation with respect to $(\mu, \mu_{t+1}) \sim P^{\otimes 2}$ (conditioning appropriately on the increasing sequence of σ-fields), we have

$$\mathbb{E}F(b_{t+1}) \leqslant \mathbb{E}[(1+\eta_t^2)F(b_t) - \eta_t \|\nabla F(b_t)\|_{L^2(b_t)}^2].$$

Using the PL inequality (4.3.2.1),

$$\mathbb{E}F(b_{t+1}) \leqslant \mathbb{E}\big[(1+\eta_t^2)F(b_t) - 2C_{\mathsf{PL}}\eta_t[F(b_t) - F(b^*)]\big].$$

Subtracting $F(b^*)$ and rearranging,

$$\mathbb{E}F(b_{t+1}) - F(b^*) \leqslant (1 - 2C_{\mathsf{PL}}\eta_t + \eta_t^2)[\mathbb{E}F(b_t) - F(b^*)] + \frac{\eta_t^2}{2}\sigma^2(P),$$

With the chosen step size, we find

$$\mathbb{E}F(b_{t+1}) - F(b^*) \leqslant \Big(\frac{t+k}{t+k+1}\Big)^2 [\mathbb{E}F(b_t) - F(b^*)] + \frac{2\sigma^2(P)}{C_{\mathsf{PL}}^2 (t+k+1)^2}.$$

Or equivalently,

$$(t+k+1)^2[\mathbb{E}F(b_{t+1}) - F(b^*)] \leqslant (t+k)^2[\mathbb{E}F(b_t) - F(b^*)] + \frac{2\sigma^2(P)}{C_{\mathsf{PL}}^2}.$$

Unrolling over $t = 0, 1, \dots, n-1$ yields

$$(n+k)^2[\mathbb{E}F(b_n) - F(b^*)] \leqslant k^2[\mathbb{E}F(b_0) - F(b^*)] + \frac{2n\sigma^2(P)}{C_{\mathsf{PL}}^2},$$

or, equivalently,

$$\mathbb{E}F(b_n) - F(b^*) \leqslant \frac{k^2}{(n+k)^2}[\mathbb{E}F(b_0) - F(b^*)] + \frac{2\sigma^2(P)}{C_{\mathsf{PL}}^2(n+k)}. \tag{A.2.0.1}$$

To conclude the proof, recall that from (4.3.1.1), we have

$$F(b_0) - F(b^*) \leqslant \frac{1}{2}W_2^2(b_0, b^*).$$

Taking the expectation over $b_0 \sim P$ we find

$$\mathbb{E}F(b_0) - F(b^*) \leqslant F(b^*) = \frac{1}{2}\sigma^2(P),$$

as claimed. Together with (A.2.0.1), it yields

$$\mathbb{E}F(b_n) - F(b^*) \leqslant \frac{\sigma^2(P)}{n+k}\Big(\frac{k^2}{2(n+k)} + \frac{2}{C_{\mathsf{PL}}^2}\Big) \leqslant \frac{\sigma^2(P)}{n}\Big(\frac{k+1}{2} + \frac{2}{C_{\mathsf{PL}}^2}\Big).$$

Plugging in the value of k completes the proof.

A.2.2 Proof of Lemma 4.4.3

According to [39, Corollary 7.22], there exists a probability measure Π on the space of constant-speed geodesics in $\mathbb{R}^d$ such that $\gamma \sim \Pi$ and b_s is the law of $\gamma(s)$. In particular, it yields

$$\int f \,\mathrm{d}b_0 - \int f \,\mathrm{d}b_1 = \int \big[f\big(\gamma(0)\big) - f\big(\gamma(1)\big)\big] \,\mathrm{d}\Pi(\gamma).$$

We can cover the geodesic $(\gamma(s))_{s\in[0,1]}$ by finitely many open neighborhoods contained in Ω so that f is Lipschitz on each such neighborhood; thus, the mapping $t \mapsto f(\gamma(s))$ is Lipschitz and we may apply the fundamental theorem of calculus, the Fubini-Tonelli theorem, and Cauchy-Schwarz:

$$\begin{aligned}\int f \, \mathrm{d}b_0 - \int f \, \mathrm{d}b_1 &= \int \int_0^1 \big\langle \nabla f\big(\gamma(s)\big), \dot{\gamma}(s) \big\rangle \, \mathrm{d}s \, \mathrm{d}\Pi(\gamma) \\ &\leqslant \int_0^1 \int \operatorname{len}(\gamma) \big\| \nabla f\big(\gamma(s)\big) \big\| \, \mathrm{d}\Pi(\gamma) \, \mathrm{d}s \\ &\leqslant \int_0^1 \Big(\int \operatorname{len}(\gamma)^2 \, \mathrm{d}\Pi(\gamma) \Big)^{1/2} \Big(\int \big\| \nabla f\big(\gamma(s)\big) \big\|^2 \, \mathrm{d}\Pi(\gamma) \Big)^{1/2} \mathrm{d}s \\ &= W_2(b_0, b_1) \int_0^1 \| \nabla f \|_{L^2(b_s)} \, \mathrm{d}s. \end{aligned}$$

It yields the result.

A.2.3 Results for Bures-Wasserstein barycenters

Proposition A.2.1 (Gaussian barycenter). *Fix $0 < \lambda_{\min} \leqslant \lambda_{\max} < \infty$. Let $P \in \mathcal{P}_2(\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d))$ be such that for all $\mu \in \operatorname{supp} P$, $\mu = \gamma_{m(\mu),\Sigma(\mu)}$ is a Gaussian with $\lambda_{\min} I_D \preceq \Sigma(\mu) \preceq \lambda_{\max} I_D$. Let γ_{m^*,Σ^*} be the Gaussian measure with mean $m^* := \int m(\mu) \, \mathrm{d}P(\mu)$ and covariance matrix Σ^* which is a fixed point of the mapping $S \mapsto F(S) := \int \big(S^{1/2} \Sigma(\mu) S^{1/2}\big)^{1/2} \, \mathrm{d}P(\mu)$. Then, γ_{m^*,Σ^*} is a barycenter of P.*

Proof. To show that there exists a fixed point for the mapping F, apply Brouwer's fixed-point theorem as in [1, Theorem 6.1]. To see that γ_{m^*,Σ^*} is indeed a barycenter, we observe the mapping

$$\varphi : (\mu, x) \mapsto \varphi_\mu(x) := \langle x, m(\mu) \rangle + \frac{1}{2} \langle x - m^*, (\Sigma^*)^{-1/2} \big[(\Sigma^*)^{1/2} \Sigma(\mu) (\Sigma^*)^{1/2}\big]^{1/2} (\Sigma^*)^{-1/2} (x - m^*) \rangle$$

satisfies the following condition:

$$\mathbb{E}_{\mu \sim P}[\varphi_\mu(x)] = \frac{1}{2} \|x\|_2^2 + \frac{1}{2} \|m^*\|_2^2.$$

This implies optimality by simply applying the dual definition of optimal transport to any alternative $\gamma_{m,\Sigma}$:

$$\begin{aligned}
F(\gamma_{m,\Sigma}) &\geqslant \iint \left(\frac{1}{2}\|x\|_2^2 - \varphi_\mu\right) \mathrm{d}\gamma_{m,\Sigma}\mathrm{d}Q(\mu) + \iint \left(\frac{1}{2}\|x\|_2^2 - \varphi_\mu^*\right) \mathrm{d}\mu\mathrm{d}Q(\mu) \\
&= \iint \left(\frac{1}{2}\|x\|_2^2 - \varphi_\mu^*\right) \mathrm{d}\mu\mathrm{d}Q(\mu) \\
&= F(\gamma_{m^*,\Sigma^*}),
\end{aligned}$$

where we swapped integrals with the same justification as in the proof of Theorem 2.3.7. Hence γ_{m^*,Σ^*} is a barycenter of P. □

Lemma A.2.2. *Suppose there exist constants $0 < \lambda_{\min} \leqslant \lambda_{\max} < \infty$ such that all of the eigenvalues of $\Sigma, \Sigma' \in \mathbb{S}_{++}^D$ are bounded between $\lambda_{\min}$ and $\lambda_{\max}$ and define $\kappa = \lambda_{\max}/\lambda_{\min}$. Then, the transport map from $\gamma_{0,\Sigma}$ to $\gamma_{0,\Sigma'}$ is (κ^{-1}, κ)-regular.*

Proof. The transport map from $\gamma_{0,\Sigma}$ to $\gamma_{0,\Sigma'}$ is the map $x \mapsto \Sigma^{-1/2}(\Sigma^{1/2}\Sigma'\Sigma^{1/2})^{1/2}\Sigma^{-1/2}x$. Throughout this proof, we write $\|\cdot\| = \|\cdot\|_{\mathrm{op}}$ for simplicity. We have the trivial bound

$$\|\Sigma^{-1/2}(\Sigma^{1/2}\Sigma'\Sigma^{1/2})^{1/2}\Sigma^{-1/2}\| \leqslant \sqrt{\|\Sigma^{-1}\|\|\Sigma^{1/2}\Sigma'\Sigma^{1/2}\|\|\Sigma^{-1}\|}.$$

Moreover $\|\Sigma^{-1}\| \leqslant \lambda_{\min}^{-1}$ and $\|\Sigma^{1/2}\Sigma'\Sigma^{1/2}\| \leqslant \lambda_{\max}^2$, so that the smoothness is bounded by

$$\|\Sigma^{-1/2}(\Sigma^{1/2}\Sigma'\Sigma^{1/2})^{1/2}\Sigma^{-1/2}\| \leqslant \frac{\lambda_{\max}}{\lambda_{\min}}.$$

We can take advantage of the fact that Σ, Σ' are interchangeable and infer that the strong convexity parameter of the transport map from Σ to Σ' is the inverse of the smoothness parameter of the transport map from Σ' to Σ. In other words,

$$\min_{1\leqslant j\leqslant D} \lambda_j\big(\Sigma^{-1/2}(\Sigma^{1/2}\Sigma'\Sigma^{1/2})^{1/2}\Sigma^{-1/2}\big) \geqslant \frac{\lambda_{\min}}{\lambda_{\max}}.$$

This concludes the proof. □

Lemma A.2.3. *Identify measures $\rho \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ with their densities, and let the $\|\cdot\|_{L^\infty}$ norm denote the L^∞-norm (essential supremum) w.r.t. the Lebesgue measure on $\mathbb{R}^d$. Then, for any*

$b, \mu_0, \mu_1 \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, any $s \in [0,1]$, and almost every $x \in \mathbb{R}^d$, it holds that

$$\ln \mu_s^b\big(\nabla\varphi_{b\to\mu_s^b}(x)\big) \leqslant (1-s)\ln\mu_0\big(\nabla\varphi_{b\to\mu_0}(x)\big) + s\ln\mu_1\big(\nabla\varphi_{b\to\mu_1}(x)\big).$$

In particular, taking the essential supremum over x on both sides, we deduce that the functional $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) \to (-\infty,\infty]$ given by $\rho \mapsto \ln\|\cdot\|_{L^\infty}$ is convex along generalized geodesics.

Proof. Let $\rho := [(1-s)T_{b\to\mu} + sT_{b\to\nu}]_{\#}b$ be a point on the generalized geodesic with base b connecting μ to ν. Let $\varphi_{b\to\mu}$, $\varphi_{b\to\nu}$ be the convex potentials whose gradients are $T_{b\to\mu}$ and $T_{b\to\nu}$ respectively. Then, for almost all $x \in \mathbb{R}^d$, the Monge-Ampère equation applied to the pairs (b,μ), (b,ν), and (b,ρ) respectively, yields

$$b(x) = \begin{cases} \mu\big(\nabla\varphi_{b\to\mu}(x)\big)\det D_{\mathrm{A}}^2\varphi_{b\to\mu}(x) \\ \nu\big(\nabla\varphi_{b\to\nu}(x)\big)\det D_{\mathrm{A}}^2\varphi_{b\to\nu}(x) \\ \rho\big((1-s)\nabla\varphi_{b\to\mu}(x) + s\nabla\varphi_{b\to\nu}(x)\big)\det\big((1-s)D_{\mathrm{A}}^2\varphi_{b\to\mu}(x) + sD_{\mathrm{A}}^2\varphi_{b\to\nu}(x)\big). \end{cases}$$

Here, $D_{\mathrm{A}}^2\varphi$ denotes the Hessian of φ in the Alexandrov sense; see [60, Theorem 4.8].

Fix x such that $b(x) > 0$. On the one hand, applying log-concavity of the determinant, it follows from the third Monge-Ampère equation that

$$\begin{aligned} \ln b(x) &= \ln\rho\big((1-s)\nabla\varphi_{b\to\mu}(x) + s\nabla\varphi_{b\to\nu}(x)\big) + \ln\det\big((1-s)D_{\mathrm{A}}^2\varphi_{b\to\mu}(x) + sD_{\mathrm{A}}^2\varphi_{b\to\nu}(x)\big) \\ &\geqslant \ln\rho\big((1-s)\nabla\varphi_{b\to\mu}(x) + s\nabla\varphi_{b\to\nu}(x)\big) + (1-s)\ln\det D_{\mathrm{A}}^2\varphi_{b\to\mu}(x) + s\ln\det D_{\mathrm{A}}^2\varphi_{b\to\nu}(x). \end{aligned}$$

On the other hand, it follows from the first two Monge-Ampère equations that

$$\begin{aligned} \ln b(x) = (1-s)\ln\mu\big(\nabla\varphi_{b\to\mu}(x)\big) &+ s\ln\nu\big(\nabla\varphi_{b\to\nu}(x)\big) \\ &+ (1-s)\ln\det D_{\mathrm{A}}^2\varphi_{b\to\mu}(x) + s\ln\det D_{\mathrm{A}}^2\varphi_{b\to\nu}(x). \end{aligned}$$

The above two displays yield

$$\ln\rho\big((1-s)\nabla\varphi_{b\to\mu}(x) + s\nabla\varphi_{b\to\nu}(x)\big) \leqslant (1-s)\ln\mu\big(\nabla\varphi_{b\to\mu}(x)\big) + s\ln\nu\big(\nabla\varphi_{b\to\nu}(x)\big)$$

It yields the result. □

Appendix B

Experiments for Bures-Wasserstein barycenters

In this section, we demonstrate the linear convergence of GD, the fast rate of estimation for SGD, and some potential advantages of averaging stochastic gradient by way of numerical experiments. In evaluating SGD, we also include a variant that involves sampling with replacement from the empirical distribution.

B.0.1 Simulations for the Bures manifold

First, we begin by illustrating how SGD indeed achieves the fast rate of convergence to the true barycenter on the Bures manifold, as indicated by Theorem 4.1.2.

To generate distributions with a known barycenter, we use the following fact. If the mean of the distribution $(\log_{b^\star})_{\#}P$ is 0, then $b^\star$ is a barycenter of P. This fact follows from our PL inequality (Theorem 4.5.8) or also from general arguments in [48, Theorem 2]. We also use the fact that the tangent space of the Bures manifold is given by the set of all symmetric matrices [9].

Figure 4-2 shows convergence of SGD for distributions on the Bures manifold. To generate a sample, we let A_i be a matrix with i.i.d. γ_{0,σ^2} entries. Our random sample on the Bures manifold is then given by

$$\Sigma_i = \exp_{\gamma_{0,I_D}}\Big(\frac{A_i + A_i^\top}{2}\Big), \tag{B.0.0.1}$$

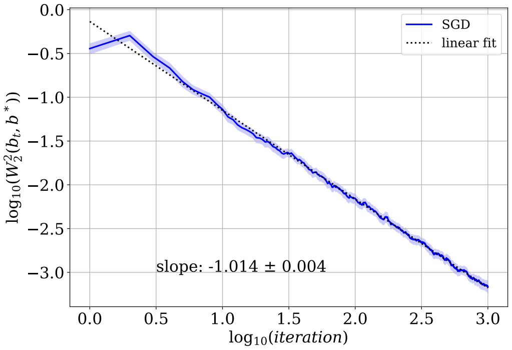

Figure B-1: Log-log plot of convergence for SGD on Bures manifold for $n = 1000$, $d = 3$, and and $b^\star = \gamma_{0,I_3}$. This corresponds to the experiment on the left in Figure 4-2.

which has population barycenter $b^\star = \gamma_{0,I_D}$. An explicit form of this exponential map is derived in [40]. We run two versions of SGD. The first variant uses each sample only once, and passes over the data once. The second variant samples from $\Sigma_1, \ldots, \Sigma_n$ with replacement at each iteration and takes the stochastic gradient step towards the selected matrix. For the resulting sequences, we also show the results of averaging the iterates. Specifically, if $(b_t)_{t\in\mathbb{N}}$ is the sequence generated by SGD, then the averaged sequence is given by $\tilde{b}_0 = b_0$ and

$$\tilde{b}_{t+1} = \left[\frac{t}{t+1}\,\mathrm{id} + \frac{1}{t+1} T_{\tilde{b}_t \to b_{t+1}}\right]_{\#} \tilde{b}_t.$$

On Riemannian manifolds, averaged SGD is known to attain optimal statistical rates under smoothness and geodesic convexity assumptions [59].

Here, we generate 100 datasets of size $n = 1000$ in the way specified above and set $\sigma^2 = 0.25$. In this experiment, the SGD step size is chosen to be $\eta_t = 2/[0.7 \cdot (t + 2/0.7 + 1)]$. The results from these 100 datasets are then averaged for each algorithm, and we also display 95% confidence bands for the resulting sequences. As is clear from the log-log plot in Figure B-1, SGD achieves the fast $O(n^{-1})$ statistical rate on this dataset.

The right of Figure 4-2 shows convergence of GD to the empirical barycenter and true barycenter. We generate samples in the same way as before. This linear convergence was observed previously by [6].

In Figure B-2, we repeat the same experiment, except this time the barycenter has co-

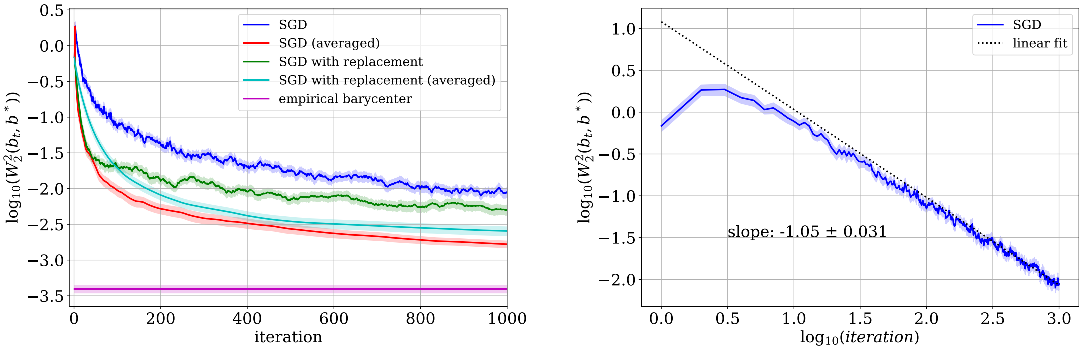

Figure B-2: Convergence of SGD on Bures manifold. Here, $n = 1000$, $d = 3$, and barycenter given by $\mathrm{diag}(20, 1, 1)$. The result displays the average over 100 randomly generated datasets.

variance matrix

$$\Sigma^\star = \begin{pmatrix} 20 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

and the entries of A_i are drawn i.i.d. from $\gamma_{0,1}$. In this situation, the condition numbers of the matrices generated according to this distribution are typically much larger than those centered around γ_{0,I_3}. To account for a potentially smaller PL constant, we chose $\eta_t = 2/[0.1 \cdot (t + 2/0.1 + 1)]$. It is again clear from the right pane in Figure B-2 that SGD achieves the fast $O(n^{-1})$ statistical rate on this dataset. To account for the slow convergence initially, we only fit this line to the last 500 iterations. We also note that averaging yields drastically better performance in this case, which we are currently unable to theoretically justify.

Figure B-3 shows convergence of SGD with replacement to the empirical barycenter. We generate $n = 500$ samples in the same way as in Figure 4-2, where the true barycenter is I_3 and $\sigma^2 = 0.25$. We calculate the error obtained by the empirical barycenter by running GD on this dataset until convergence, which is displayed with the green line. We also calculate the error obtained by a single pass of SGD, which is given by the blue line. SGD with replacement is then run for 5000 iterations, and we observe that it does indeed achieve better error than single pass SGD if run for long enough. SGD with replacement converges to the empirical barycenter, albeit at a slow rate.

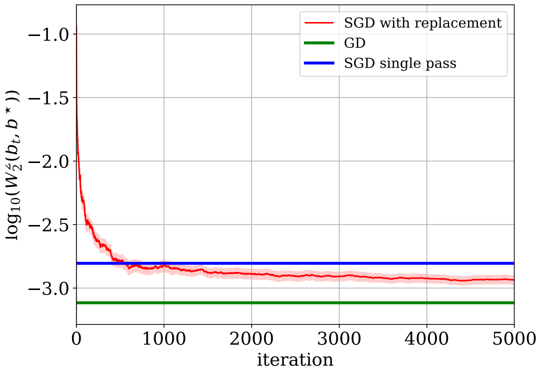

Figure B-3: Convergence of SGD on Bures manifold. Here, $n = 500$, $d = 3$, and the distribution is given by (B.0.0.1) with $\Sigma^\star = I_3$ and $\sigma^2 = 0.25$. The result displays the average over 100 randomly generated datasets.

B.0.2 Details of the non-convexity example

We consider the example of the Wasserstein metric restricted to centered Gaussian measures, which induces the Bures metric on positive definite matrices. Even restricted to such Gaussian measures, the Wasserstein barycenter objective is geodesically non-convex, despite the fact that it is Euclidean convex [62]. Figure 4.1 gives a simulated example of this fact. This figure plots the Bures distance squared between a positive definite matrix C and points along some geodesic γ, which runs between two matrices A and B. The matrices used in this example are

$$A = \begin{pmatrix} 0.8 & -0.4 \\ -0.4 & 0.3 \end{pmatrix}, \qquad B = \begin{pmatrix} 0.3 & -0.5 \\ -0.5 & 1.0 \end{pmatrix}, \qquad C = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.6 \end{pmatrix},$$

and $\gamma(t)$, $t \in [0, 1]$, is taken to be the Bures or Euclidean geodesic from A to B (the Euclidean geodesic is given by $t \mapsto (1 - t)A + tB$). This function is clearly non-convex, and therefore we cannot assume that there is some underlying strong convexity (although the Bures distance is in fact strongly geodesically convex for sufficiently small balls [29]).

Bibliography

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

[2] Martial Agueh and Guillaume Carlier. Vers un théorème de la limite centrale dans l'espace de wasserstein? *Comptes Rendus Mathématique*, 355(7):812–818, 2017.

[3] Adil Ahidar-Coutrix, Thibaut Le Gouic, and Quentin Paris. Convergence rates for empirical barycenters in metric spaces: curvature, convexity and extendable geodesics. *Probability Theory and Related Fields*, pages 1–46.

[4] A. D. Aleksandrov. A theorem on triangles in a metric space and some of its applications. 38:5–23, 1951.

[5] Stephanie Alexander, Vitali Kapovitch, and Anton Petrunin. Alexandrov geometry: preliminary version no. 1. *arXiv preprint arXiv:1903.08539*, 2019.

[6] Pedro C Álvarez-Esteban, E Del Barrio, JA 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.

[7] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[8] Luigi Ambrosio, Federico Glaudo, and Dario Trevisan. On the optimal map in the 2-dimensional random matching problem. *arXiv preprint arXiv:1903.12153*, 2019.

[9] Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the bures–wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 2018.

[10] Rabi Bhattacharya, Vic Patrangenaru, et al. Large sample theory of intrinsic and extrinsic sample means on manifolds. *The Annals of Statistics*, 31(1):1–29, 2003.

[11] Rabi Bhattacharya, Vic Patrangenaru, et al. Large sample theory of intrinsic and extrinsic sample means on manifolds—ii. *The Annals of Statistics*, 33(3):1225–1259, 2005.

[12] Jérémie Bigot, Elsa Cazelles, and Nicolas Papadakis. Data-driven regularization of wasserstein barycenters with an application to multivariate density registration. *Information and Inference: A Journal of the IMA*, 8(4):719–755, 2019.

[13] Jérémie Bigot, Raúl Gouet, Thierry Klein, Alfredo Lopez, et al. Upper and lower risk bounds for estimating the wasserstein barycenter of random measures on the real line. *Electronic Journal of Statistics*, 12(2):2253–2289, 2018.

[14] Emmanuel Boissard, Thibaut Le Gouic, Jean-Michel Loubes, et al. Distribution's template estimate with wasserstein metrics. *Bernoulli*, 21(2):740–759, 2015.

[15] Nicolas Bonneel, Gabriel Peyré, and Marco Cuturi. Wasserstein barycentric coordinates: Histogram regression using optimal transport. *ACM Transactions on Graphics*, 35(4), 2016.

[16] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

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

[18] Dmitri Burago, Iu D Burago, Yuri Burago, Sergei A Ivanov, and Sergei Ivanov. *A course in metric geometry*, volume 33. American Mathematical Soc., 2001.

[19] Yu Burago, Mikhail Gromov, and Gregory Perel'man. Ad alexandrov spaces with curvature bounded below. *Russian mathematical surveys*, 47(2):1, 1992.

[20] Donald Bures. An extension of Kakutani's theorem on infinite product measures to the tensor product of semifinite w*-algebras. *Transactions of the American Mathematical Society*, 135:199–212, 1969.

[21] Guillaume Carlier and Ivar Ekeland. Matching for teams. *Economic theory*, 42(2):397–418, 2010.

[22] Sinho Chewi, Tyler Maunu, Philippe Rigollet, and Austin J Stromme. Gradient descent algorithms for bures-wasserstein barycenters. *arXiv preprint arXiv:2001.01700*, 2020.

[23] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

[24] Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *International Conference on Machine Learning*, pages 685–693, 2014.

[25] Federico Glaudo. On the c-concavity with respect to the quadratic cost on a manifold. *Nonlinear Analysis*, 178:145–151, 2019.

[26] Thibaut Le Gouic, Quentin Paris, Philippe Rigollet, and Austin J Stromme. Fast convergence of empirical barycenters in alexandrov spaces and the wasserstein space. *arXiv preprint arXiv:1908.00828*, 2019.

[27] Alexandre Gramfort, Gabriel Peyré, and Marco Cuturi. Fast optimal transport averaging of neuroimaging data. In *International Conference on Information Processing in Medical Imaging*, pages 261–272. Springer, 2015.

[28] Mikhael Gromov. Sign and geometric meaning of curvature. *Rendiconti del Seminario Matematico e Fisico di Milano*, 61(1):9–123, 1991.

[29] W. Huang, K. A. Gallivan, and P.-A. Absil. A Broyden class of quasi-Newton methods for Riemannian optimization. *SIAM Journal on Optimization*, 25(3):1660–1685, 2015.

[30] Jan-Christian Hütter and Philippe Rigollet. Minimax rates of estimation for smooth optimal transport maps. *arXiv preprint arXiv:1905.05828*, 2019.

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

[32] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

[33] Wilfrid S Kendall, Huiling Le, et al. 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.

[34] Martin Knott and Cyril S Smith. On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43(1):39–49, 1984.

[35] Martin Knott and Cyril S Smith. On a generalization of cyclic monotonicity and distances among random vectors. *Linear algebra and its applications*, 199:363–371, 1994.

[36] Alexey Kroshnin, Vladimir Spokoiny, and Alexandra Suvorikova. Statistical inference for bures-wasserstein barycenters. *arXiv preprint arXiv:1901.00226*, 2019.

[37] Thibaut Le Gouic. Dual and multimarginal problems for the wasserstein barycenter. 2020. unpublished.

[38] Thibaut Le Gouic and Jean-Michel Loubes. Existence and Consistency of Wasserstein Barycenters. *Probability Theory and Related Fields*, August 2017.

[39] John Lott and Cédric Villani. Ricci curvature for metric-measure spaces via optimal transport. *Annals of Mathematics*, pages 903–991, 2009.

[40] L Malagò, L. Montrucchio, and G. Pistone. Wasserstein Riemannian geometry of Gaussian densities. *Information Geometry*, 1(2):137–179, 2018.

[41] Robert J McCann. A convexity principle for interacting gases. *Advances in mathematics*, 128(1):153–179, 1997.

[42] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

[43] Shin-ichi Ohta. Convexities of metric spaces. *Geometriae Dedicata*, 125(1):225–250, 2007.

[44] Shin-ichi Ohta. Barycenters in alexandrov spaces of curvature bounded below. *Advances in geometry*, 12(4):571–587, 2012.

[45] Ingram Olkin and Svetlozar T Rachev. Maximum submatrix traces for positive definite matrices. *SIAM journal on matrix analysis and applications*, 14(2):390–397, 1993.

[46] Victor M Panaretos and Yoav Zemel. Statistical aspects of wasserstein distances. *Annual review of statistics and its application*, 6:405–431, 2019.

[47] Victor M Panaretos, Yoav Zemel, et al. Amplitude and phase variation of point processes. *The Annals of Statistics*, 44(2):771–812, 2016.

[48] VM Panaretos and Y Zemel. Fréchet means and procrustes analysis in wasserstein space. *Bernoulli, To be published*, 2017.

[49] G. Peyré and M. Cuturi. Computational Optimal Transport. *ArXiv e-prints*, March 2018.

[50] Julien Rabin and Nicolas Papadakis. Convex color image segmentation with optimal transport distances. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 256–269. Springer, 2015.

[51] Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer, 2011.

[52] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*. 2015.

[53] Filippo Santambrogio and Xu-Jia Wang. Convexity of the support of the displacement interpolation: Counterexamples. *Applied Mathematics Letters*, 58:152–158, 2016.

[54] Christof Schötz et al. Convergence rates for the generalized fréchet mean via the quadruple inequality. *Electronic Journal of Statistics*, 13(2):4280–4345, 2019.

[55] Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):66, 2015.

[56] Sanvesh Srivastava, Cheng Li, and David B Dunson. Scalable bayes via barycenter in wasserstein space. *The Journal of Machine Learning Research*, 19(1):312–346, 2018.

[57] Karl-Theodor Sturm. Metric spaces of lower bounded curvature. *Expositiones Mathematicae*, 17, 01 1999.

[58] Karl-Theodor Sturm. Probability measures on metric spaces of nonpositive. *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.

[59] Nilesh Tripuraneni, Nicolas Flammarion, Francis Bach, and Michael I. Jordan. Averaging stochastic gradient descent on Riemannian manifolds. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 650–687, 2018.

[60] Cédric Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.

[61] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[62] Melanie Weber and Suvrit Sra. Nonconvex stochastic optimization on manifolds via riemannian frank-wolfe methods. *arXiv preprint arXiv:1910.04194*, 2019.