Proceedings of Machine Learning Research vol 125:1–29, 2020 33rd Annual Conference on Learning Theory

Gradient descent algorithms for Bures-Wasserstein barycenters

Sinho Chewi SCHEWI@MIT.EDU
Tyler Maunu MAUNUT@MIT.EDU
Philippe Rigollet RIGOLLET@MATH.MIT.EDU
Austin J. Stromme ASTROMME@MIT.EDU
Massachusetts Institute of Technology
77 Massachusetts Avenue,
Cambridge, MA 02139-4307, USA

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

We study first order methods to compute the barycenter of a probability distribution P over the space of probability measures with finite second moment. We develop a framework to derive global rates of convergence for both gradient descent and stochastic gradient descent despite the fact that the barycenter functional is not geodesically convex. Our analysis overcomes this technical hurdle by employing a Polyak-Łojasiewicz (PL) inequality and relies on tools from optimal transport and metric geometry. In turn, we establish a PL inequality when P is supported on the Bures-Wasserstein manifold of Gaussian probability measures. It leads to the first global rates of convergence for first order methods in this context.

Keywords: geodesic optimization, optimal transport, Wasserstein barycenters

1. Introduction

We consider the following statistical problem. We observe n independent copies $\mu_1, \dots, \mu_n$ of a random probability measure μ over $\mathbb{R}^D$. Assume furthermore that $\mu \sim P$, where P is an unknown distribution over probability measures. We wish to output a single probability measure on $\mathbb{R}^D$, $\bar{\mu}_n$, which represents the *average* measure under P in a suitable sense. For example, the measures $\mu_1, \dots, \mu_n$ may arise as representations of images, in which case the average of the measures with respect to the natural linear structure on the space of signed measures is unsuitable for many applications (Cuturi and Doucet, 2014). Instead, we study the *Wasserstein barycenter* (Agueh and Carlier, 2011) which has been proposed in the literature as a more desirable notion of average because it incorporates the geometry of the underlying space. Wasserstein barycenters have been applied in many areas, e.g. graphics, neuroscience, statistics, economics, and algorithmic fairness (Carlier and Ekeland, 2010; Rabin et al., 2011; Rabin and Papadakis, 2015; Solomon et al., 2015; Gramfort et al., 2015; Bonneel et al., 2016; Srivastava et al., 2018; Le Gouic and Loubes, 2020).

To formally set up the situation, let $\mathcal{P}_2(\mathbb{R}^D)$ be the set of all (Borel) probability measures on $\mathbb{R}^D$ with finite second moment, and let $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$ be the subset of those measures in $\mathcal{P}_2(\mathbb{R}^D)$ that are absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^D$ and thus admit a density. When endowed with the *2-Wasserstein metric*, W_2, this set forms a geodesic metric space $(\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D), W_2)$. We denote by P_n the empirical distribution of the sample $\mu_1, \dots, \mu_n$.

A *barycenter* of P, denoted $b^\star$, is defined to be a minimizer of the functional

$$F(b) := \frac{1}{2} P W_2^2(b, \cdot) = \frac{1}{2} \int W_2^2(b, \cdot) \, \mathrm{d}P.$$

A natural estimator of $b^\star$ is the *empirical barycenter* $\hat{b}_n$, defined as a minimizer of

$$F_n(b) := \frac{1}{2} P_n W_2^2(b, \cdot) = \frac{1}{2n} \sum_{i=1}^{n} W_2^2(b, \mu_i).$$

Statistical consistency of the empirical barycenter in a general context was first established in (Le Gouic and Loubes, 2017) and further work has focused on providing effective rates of convergence for the quantity $W_2^2(\hat{b}_n, b^\star)$. A first step towards this goal was made in (Ahidar-Coutrix et al., 2020) by deriving nonparametric rates of the form $W_2^2(\hat{b}_n, b^\star) \lesssim n^{-1/D}$ when $D \geqslant 3$. Moreover, in the same paper (Ahidar-Coutrix et al., 2020), the authors establish parametric rates of the form $W_2^2(\hat{b}_n, b^\star) \lesssim n^{-1}$ when P is supported on a space of finite doubling dimension.

An important example with this property arises when P is supported on centered non-degenerate Gaussian measures, first studied by Knott and Smith in 1994 (Knott and Smith, 1994). In this case, Gaussians can be identified with their covariance matrices, and the Wasserstein metric induces a metric on the space of positive definite matrices. This metric, known as the *Bures* or *Bures-Wasserstein* metric is the distance function for a Riemannian metric on the manifold of positive definite matrices, known as the *Bures manifold* (Modin, 2017; Bhatia et al., 2019). The name of the Bures manifold originates from quantum physics and quantum information theory, where it is used to model the space of density matrices (Bures, 1969). In the Bures case of the barycenter problem, more precise statistical results, including central limit theorems, are known (Martial and Carlier, 2017; Kroshnin et al., 2019).

It is worth noting that parametric rates are also achievable in the infinite-dimensional case under additional conditions. First, it is not surprising that such rates are achievable over $(\mathcal{P}_2(\mathbb{R}), W_2)$ since this space can be isometrically embedded in a Hilbert space (Panaretos and Zemel, 2016; Bigot et al., 2018). Moreover, it was shown that, under additional geometric conditions, such rates are achievable for much more general infinite-dimensional spaces (Le Gouic et al., 2019), including $(\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D), W_2)$ for any $D \geqslant 2$.

While these results are satisfying from a statistical perspective, they do not provide guidelines for the *computation* of the empirical barycenter $\hat{b}_n$. In practice, Wasserstein barycenters are estimated using iterative, first order algorithms (Cuturi and Doucet, 2014; Álvarez Esteban et al., 2016; Backhoff-Veraguas et al., 2018; Claici et al., 2018; Zemel and Panaretos, 2019) but often lack theoretical guarantees. Recently, this line of work has provided rates of convergence for first order algorithms employed to compute the Wasserstein barycenter of distributions with a *common discrete support* (Guminov et al., 2019; Kroshnin et al., 2019; Dvinskikh, 2020; Lin et al., 2020). In this framework, the computation of Wasserstein barycenters is a convex optimization problem with additional structure. However, first order methods can also be envisioned beyond this traditional framework by adopting a non-Euclidean perspective on optimization. This approach is supported by the influential work of Otto (Otto, 2001) who established that Wasserstein space bears resemblance to a Riemannian manifold. In particular, one can define the Wasserstein gradient of the functional F, so it does indeed make sense to consider an intrinsic *gradient descent*-based approach towards estimating $b^\star$. However, the convergence guarantees for such first order methods are largely unexplored.

When the distribution P is supported on the Bures-Wasserstein manifold of Gaussian probability measures, gradient descent takes the form of a concrete and tractable update equation on the mean and covariance matrix of the candidate barycenter. In the population setting (where the distribution P is known), such an algorithm was proposed in Álvarez-Esteban et al. (Álvarez Esteban et al., 2016), where it is described as a fixed-point algorithm. Álvarez-Esteban et al. prove that the fixed-point algorithm converges to the true barycenter as the number of iterations goes to infinity. The consistency results were further generalized in (Backhoff-Veraguas et al., 2018; Zemel and Panaretos, 2019) and extended to the non-population and stochastic gradient case. However, the literature currently does not provide any rates of convergence for these first order methods. In fact, Álvarez-Esteban et al. empirically observed a linear rate of convergence for the gradient descent algorithm in the Gaussian setting and left open the theoretical study of this phenomenon for future study. One contribution of this paper is to establish this rate of convergence (Theorem 1), and we also provide multiple extensions including the first rate of convergence for stochastic gradient descent in this context.

On our way to proving rates of convergence in the Bures-Wasserstein case, we also establish results that apply to the more general setting where P may not be supported on Gaussian probability measures. In particular, we establish an integrated Polyak-Łojasiewicz inequality (Lemma 8) and a new variance inequality (Theorem 6) that are of independent interest.

NOTATION. We denote the set of positive definite matrices by $\mathbb{S}^D_{++}$, and the set of positive semidefinite matrices by $\mathbb{S}^D_+$. We denote by $\lambda_1(\Sigma), \dots, \lambda_D(\Sigma) \geqslant 0$ the eigenvalues of a matrix $\Sigma \in \mathbb{S}^D_+$. The Gaussian measure on $\mathbb{R}^D$ with mean $m \in \mathbb{R}^D$ and covariance matrix $\Sigma \in \mathbb{S}^D_+$ is denoted $\gamma_{m,\Sigma}$. We reserve the notation $\log$ for the inverse of the Riemannian exponential map (which we review in 3.1) and use instead $\ln(\cdot)$ to denote the natural logarithm. The (convex analysis) indicator function $\iota_{\mathcal{C}}$ of a set $\mathcal{C}$ is defined by $\iota_{\mathcal{C}}(x) = 0$ if $x \in \mathcal{C}$ and $\iota_{\mathcal{C}}(x) = +\infty$ otherwise. We denote by id the identity map of $\mathbb{R}^D$.

2. Main results

In this paper, we develop a general machinery to study first order methods for optimizing the barycenter functional on Wasserstein space. Establishing fast convergence of first order methods is usually intimately related to convexity. Since our setting is on the curved Wasserstein space, we talk about *geodesic convexity* rather than the usual notion convexity employed in flat, Euclidean spaces. Geodesic convexity has been used to study statistical efficiency in manifold constrained estimation (Auderset et al., 2005; Wiesel, 2012) and, more recently, in optimization (Bonnabel, 2013; Bacak, 2014; Zhang and Sra, 2016).

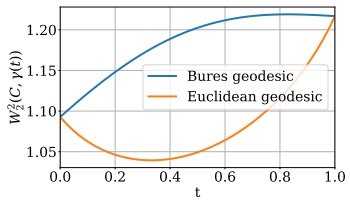

Figure 1: Example of the non-geodesic convexity of W_2^2. Details are given in Appendix B.2.

Barring a direct approach to establishing quantitative convergence guarantees, the barycenter functional is actually not geodesically convex on Wasserstein space. In fact, the barycenter functional may even be *concave* along geodesics; see Figure 1. As such, it does not lend itself to the general techniques of geodesically convex optimization. This non-convexity is a manifestation of the non-negative curvature of $(\mathcal{P}_2(\mathbb{R}^D), W_2)$ (cf. subsection 3.1) (Sturm, 2003).

Fortunately, the optimization literature describes conditions for global convergence of first order algorithms even for non-convex objectives. In this work, we employ a Polyak-Łojasiewicz (PL) inequality of the form (6), which is known to yield linear convergence for a variety of gradient methods on flat spaces even in absence of convexity (Karimi et al., 2016).

In this paper, we study the barycenter functional

$$G(b) := \frac{1}{2} Q W_2^2(b, \cdot) = \frac{1}{2} \int W_2^2(b, \cdot) \, \mathrm{d}Q \,, \tag{1}$$

for some generic distribution Q with barycenter $\bar{b}$. This notation allows us to treat simultaneously the cases where $Q = P$ and $Q = P_n$, which are the situations of interest for statisticians. The case when Q is an arbitrary discrete distribution supported on Gaussians has also been studied in the geodesic optimization literature (Agueh and Carlier, 2011; Álvarez Esteban et al., 2016; Weber and Sra, 2017; Bhatia et al., 2019; Weber and Sra, 2019; Zemel and Panaretos, 2019). Our main theorems, for gradient descent and stochastic gradient descent respectively, are stated below.

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

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

The above theorem establishes a linear rate of convergence for gradient descent and answers a question left open in (Álvarez Esteban et al., 2016). Moreover, when $Q = P_n$, combined with the existing results of (Kroshnin et al., 2019; Ahidar-Coutrix et al., 2020), it yields a procedure to estimate 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 2 *Fix $\zeta \in (0, 1]$ and let Q be a distribution supported on mean-zero Gaussian measures whose covariance matrices Σ satisfy $\|\Sigma\|_{\mathrm{op}} \leqslant 1$ and $\det \Sigma \geqslant \zeta$. Then, Q has a unique barycenter $\bar{b}$, and Stochastic Gradient Descent (Algorithm 2) run on a sample of size $n + 1$ from Q returns a Gaussian measure b_n such that*

$$\mathbb{E} W_2^2(b_n, \bar{b}) \leqslant \frac{96 \, \mathrm{var}(Q)}{n \zeta^5} \,, \quad \textit{where} \quad \mathrm{var}(Q) = \int W_2^2(\cdot, \bar{b}) \, \mathrm{d}Q \,.$$

When applied to $Q = P$, Theorem 2 shows that SGD yields an estimator b_n different from the empirical barycenter $\hat{b}_n$ that also converges at the parametric rate to $b^\star$. When applied to $Q = P_n$, this leads an alternative to gradient descent to estimate the empirical barycenter $\hat{b}_n$ that exhibits a slower convergence but that has much cheaper iterations and lends itself better to parallelization.

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.

Remark 3 *A natural sufficient condition of $\det \Sigma \geqslant \zeta$ 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.*

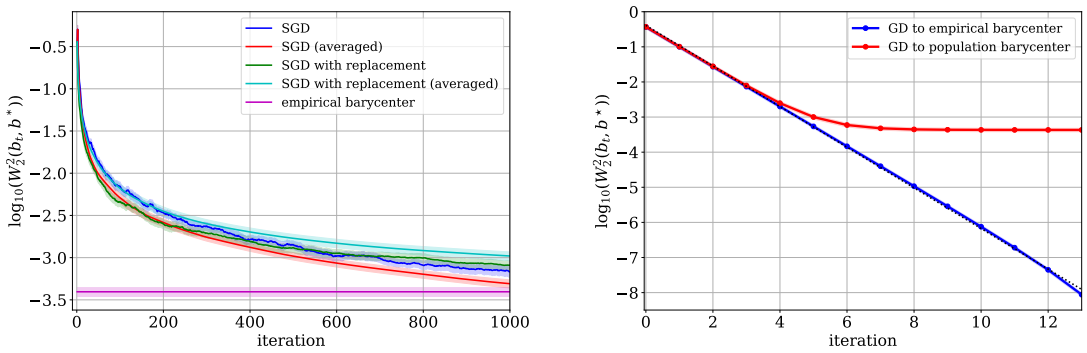

Figure 2: Left. Convergence of SGD on Bures manifold for $n = 1000$, $d = 3$, and and $b^\star = \gamma_{0,I_3}$. Right: linear convergence of GD on the same problem.

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

3. Gradient descent on Wasserstein Space

In this section, we first review some background on optimal transport and describe first order algorithms on Wasserstein space. Then, we derive rates of convergence assuming a Polyak-Łojasiewicz (PL) inequality. Theorems 4 and 5 below are proved using modifications of the usual proofs in the optimization literature. Their proofs make critical use of the non-negative curvature of the Wasserstein space and are deferred to Appendix C.

3.1. Notation and background on optimal transport

In this section, we give a quick overview of the background and notation for optimal transport that is relevant for the paper. We provide a more thorough review of Riemannian geometry and the geometry of Wasserstein space in Appendix A. For each topic below, we also provide a reference to a useful presentation.

Wasserstein distance. (Villani, 2003, Chapter 1). Given a Polish space (E, d), we denote by $\mathcal{P}_2(E)$ the collection of all (Borel) probability measures μ on E such that $\mathbb{E}_{X\sim\mu}[d(X, y)^2] < \infty$ for some $y \in E$. For two measures $\mu, \nu \in \mathcal{P}_2(E)$, let $\Pi_{\mu,\nu}$ be the set of couplings between μ and ν, that is, the collection of probability measures π on $E \times E$ such that if $(X, Y) \sim \pi$, then $X \sim \mu$ and $Y \sim \nu$. The 2-Wasserstein distance between μ and ν is then defined as

$$W_2^2(\mu, \nu) := \inf_{\pi\in\Pi_{\mu,\nu}} \mathbb{E}_{(X,Y)\sim\pi}[d(X, Y)^2]. \tag{2}$$

We are primarily interested in the cases $E = \mathbb{R}^D$ equipped with the standard Euclidean metric, and $E = \mathcal{P}_2(\mathbb{R}^D)$ equipped with the Wasserstein metric. Thus, $\mathcal{P}_2(\mathbb{R}^D)$ denotes the space of probability measures on $\mathbb{R}^D$ with finite second moment, and $\mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^D))$ denotes the space of measures P on $\mathcal{P}_2(\mathbb{R}^D)$ such that $\mathbb{E}_{\nu\sim P} W_2^2(\mu_0, \nu) < \infty$ for some, and therefore any, $\mu_0 \in \mathcal{P}_2(\mathbb{R}^D)$.

If $\mu \in \mathcal{P}_2(\mathbb{R}^D)$ is absolutely continuous w.r.t. the Lebesgue measure, we write $\mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$, and we similarly define the space $\mathcal{P}_2(\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D))$.

Transport map. (Villani, 2003, Chapter 2). Given a measure μ and a map $T\colon \mathbb{R}^D \to \mathbb{R}^D$, the pushforward $T_{\#}\mu$ is the law of $T(X)$ when $X \sim \mu$. For $\mu, \nu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$, Brenier's theorem tells us that there exists a unique optimal coupling $\pi^\star \in \Pi_{\mu,\nu}$ that achieves the minimum in (2) and furthermore that it is induced by a mapping $T_{\mu\to\nu}$, in the sense that if $X \sim \mu$ then $(X, T_{\mu\to\nu}(X)) \sim \pi^\star$. Moreover, $T_{\mu\to\nu}$ is the (μ-a.e. unique) gradient of a convex function $\varphi_{\mu\to\nu}$ such that

$$(\nabla\varphi_{\mu\to\nu})_{\#}\mu = \nu.$$

Kantorovich potential. (Villani, 2003, Chapter 2). The $\varphi_{\mu\to\nu} : \mathbb{R}^D \to \mathbb{R}$ specified in this way is called the *Kantorovich potential* for the optimal transport from μ to ν. For $\alpha, \beta > 0$, if $\varphi_{\mu\to\nu}$ is α-strongly convex and β-smooth, in the sense that for all $x, y \in \mathbb{R}^D$,

$$\frac{\alpha}{2}\|y-x\|^2 \leqslant \varphi_{\mu\to\nu}(y) - \varphi_{\mu\to\nu}(x) - \langle \nabla\varphi_{\mu\to\nu}(x), y-x\rangle \leqslant \frac{\beta}{2}\|y-x\|^2, \tag{3}$$

then we say that the potential $\varphi_{\mu\to\nu}$ is (α,β)-*regular*.

Geodesics. (Villani, 2003, Section 5.1). The space $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$ space is a geodesic space, where the geodesics are given by McCann's interpolation. Defining $\mu_s := ((1-s)\,\mathrm{id} + sT_{\mu_0\to\mu_1})_{\#}\mu_0$, then $(\mu_s)_{s\in[0,1]}$ is a constant-speed geodesic in Wasserstein space which connects μ_0 to μ_1. For any $\nu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$, define the *generalized geodesic* with base ν and connecting μ_0 to μ_1 by $(\mu_s^\nu)_{s\in[0,1]}$ where $\mu_s^\nu := [(1-s)T_{\nu\to\mu_0} + sT_{\nu\to\mu_1}]_{\#}\nu$.

Tangent bundle. (Ambrosio et al., 2008, Chapter 8). For $b \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$ define the "tangent space" at b by

$$T_b\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D) := \overline{\{\lambda(\nabla\varphi - \mathrm{id})\colon \lambda > 0,\ \varphi \in C_c^\infty(\mathbb{R}^D),\ \varphi \text{ convex}\}}^{L^2(b)}.$$

For $v \in T_b\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$ we write $\|v\|_b := \|v\|_{L^2(b)}$. Moreover, for any $b, b' \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$, define the map $\log_b : \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D) \to T_b\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$ by $\log_b(b') := T_{b\to b'} - \mathrm{id}$. Reciprocally, we define the map $\exp_b : T_b\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D) \to \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$ by $\exp_b(v) = (\mathrm{id}+v)_{\#}b$.

Convexity. (Agueh and Carlier, 2011, Section 7). We are now in a position to define two notions of convexity in Wasserstein space. Consider any functional $\mathcal{F} : \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D) \to (-\infty, \infty]$ on Wasserstein space. We say that $\mathcal{F}$ is *geodesically convex* if for all $\mu_0, \mu_1 \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$, the constant-speed geodesic $(\mu_s)_{s\in[0,1]}$ from μ_0 to μ_1 satisfies $\mathcal{F}(\mu_s) \leqslant (1-s)\mathcal{F}(\mu_0) + s\mathcal{F}(\mu_1)$ for all $s \in [0,1]$. We say that $\mathcal{F}$ is *convex along generalized geodesics* if for all choices $\nu, \mu_0, \mu_1 \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$, it holds that $\mathcal{F}(\mu_s^\nu) \leqslant (1-s)\mathcal{F}(\mu_0) + s\mathcal{F}(\mu_1)$ for all $s \in [0,1]$. Observe that the notion of generalized geodesic reduces to that of a geodesic when $\nu = \mu_0$, so that convexity along generalized geodesic is a stronger notion than convexity along geodesics. We say that a set $\mathcal{C} \subset \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$ is convex along geodesics (resp. generalized geodesics) if its indicator function $\iota_{\mathcal{C}}$ is convex along geodesics (resp. generalized geodesics). Note that a set $\mathcal{C}$ is convex along generalized geodesics with base b if and only if the set $\log_b(\mathcal{C})$ is convex in the usual sense.

Curvature. (Ambrosio et al., 2008, Theorem 7.3.2). Lastly, we often use the fact that $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$ is non-negatively curved in the sense of Alexandrov. More specifically, we use the fact that for $\mu_0, \mu_1, \nu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$, if $(\mu_s)_{s\in[0,1]}$ denotes the constant-speed geodesic connecting μ_0 to μ_1, then for all $s \in [0,1]$,

$$W_2^2(\mu_s, \nu) \geqslant (1-s)W_2^2(\mu_0,\nu) + sW_2^2(\mu_1,\nu) - s(1-s)W_2^2(\mu_0,\mu_1). \tag{4}$$

Moreover, for any $\mu, \nu, b \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$ the definition of Wasserstein distance implies

$$W_2(\mu,\nu) \leqslant \|T_{b\to\nu} \circ T_{\mu\to b} - \mathrm{id}\|_{L^2(\mu)} = \|T_{b\to\nu} - T_{b\to\mu}\|_{L^2(b)} = \|\log_b(\mu) - \log_b(\nu)\|_b. \quad (5)$$

3.2. Gradient descent algorithms over Wasserstein space

3.2.1. GRADIENT DESCENT.

Let Q be a probability distribution over $(\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D), W_2)$. In the sequel, we focus on the cases where $Q = P$, $Q = P_n$, or Q is a weighted atomic distribution, but our results apply generically to any Q that satisfy the conditions stated in the theorems below.

Using the techniques of (Ambrosio et al., 2008), the *gradient* of the barycenter functional G defined in (1) may be easily computed (Zemel and Panaretos, 2019). Analogous to the Riemannian formula (reviewed in Appendix A), the Wasserstein gradient of G at $b \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$ is the mapping $\nabla G(b) : \mathbb{R}^D \to \mathbb{R}^D$ defined by

$$\nabla G(b) := -Q \log_b(\cdot) = -\int (T_{b\to\mu} - \mathrm{id})\, \mathrm{d}Q(\mu).$$

Denote by $\bar{b}$ any minimizer of G.

The primary assumption we work with is common in the optimization literature. We say that G satisfies a *Polyak-Łojasiewicz (PL)* inequality at b if

$$\|\nabla G(b)\|_b^2 \geqslant 2C_{\mathsf{PL}}[G(b) - G(\bar{b})] \quad \text{for some } C_{\mathsf{PL}} > 0. \quad (6)$$

It follows from (13) below that $C_{\mathsf{PL}} \leqslant 1$ for any such Q.

The *gradient descent (GD)* iterates on G are defined as

$$b_0 \in \operatorname{supp} Q, \qquad b_{t+1} := \exp_{b_t}\big(-\nabla G(b_t)\big) = [\mathrm{id} - \nabla G(b_t)]_{\#} b_t \quad \text{for } t \geqslant 1. \quad (7)$$

Note that this method employs a unit step size. This is in agreement with the observation made in (Zemel and Panaretos, 2019, Lemma 2) that it leads to the best decrement in G with respect to the smoothness upper bound, see Theorem 7 below.

The following theorem shows that a PL inequality yields a linear rate of convergence.

Theorem 4 (Rate of convergence for gradient descent) *If G satisfies the PL inequality (6) at all the iterates $(b_t)_{t<T}$, then*

$$G(b_T) - G(\bar{b}) \leqslant (1 - C_{\mathsf{PL}})^T [G(b_0) - G(\bar{b})].$$

3.2.2. STOCHASTIC GRADIENT DESCENT.

PL inequalities are also useful in the stochastic setting where we observe n independent copies $\mu_1, \ldots, \mu_n$ of $\mu \sim Q$. In this case, we consider the natural *stochastic gradient descent (SGD)* iterates defined by $b_0 := \mu_0$, and for $t = 0, 1, \ldots, n-1$,

$$b_{t+1} := \exp_{b_t}\big(-\eta_t \log_{b_t}(\mu_{t+1})\big) = [\mathrm{id} + \eta_t(T_{b_t\to\mu_{t+1}} - \mathrm{id})]_{\#} b_t\,, \quad (8)$$

where $\eta_t \in (0,1)$ denotes the step size. At each iteration, SGD moves the iterate along the geodesic between b_t and μ_{t+1} for step size η_t. Under the assumption of a PL inequality, we show that SGD achieves a parametric rate of convergence.

In the following result, we recall that the *variance* of Q is defined as

$$\operatorname{var}(Q) := \int W_2^2(\bar{b}, \cdot)\, \mathrm{d}Q = 2G(\bar{b}).$$

Theorem 5 (Rates of convergence for SGD) *Assume that there exists a constant $C_{\mathsf{PL}} > 0$ such that the following holds: G satisfies the PL inequality (6) at all the iterates $(b_t)_{0 \leqslant t \leqslant n}$ of SGD run with step size*

$$\eta_t = C_{\mathsf{PL}}\Big(1 - \sqrt{1 - \frac{2(t+k)+1}{C_{\mathsf{PL}}^2 (t+k+1)^2}}\Big) \leqslant \frac{2}{C_{\mathsf{PL}}(t+k+1)}, \tag{9}$$

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

$$\mathbb{E}G(b_n) - G(\bar{b}) \leqslant \frac{3 \operatorname{var}(Q)}{C_{\mathsf{PL}}^2 n}.$$

The parameter k in (9) ensures that the step size is well-defined and less than 1.

3.3. Properties of the barycenter functional

Unlike results in generic optimization, this paper focuses on a specific function to optimize: the barycenter functional. In fact, this is a vast family of functionals, each indexed by the distribution Q in (1). However, some structure is shared across this family. In the rest of this section, we extract properties that are relevant to our optimization questions: a variance inequality, smoothness, as well as an integrated PL inequality. These properties are valid for general distributions Q over $\mathcal{P}_2(\mathbb{R}^D)$ and are specialized to the Bures manifold in the next section.

3.3.1. VARIANCE INEQUALITY.

Variance inequalities indicate quadratic growth of the barycenter functional around its minimum. More specifically, we say that Q satisfies a *variance inequality* with constant $C_{\mathsf{var}} > 0$ if

$$G(b) - G(\bar{b}) \geqslant \frac{C_{\mathsf{var}}}{2} W_2^2(b, \bar{b}), \qquad \forall b \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D). \tag{10}$$

In particular, (10) implies uniqueness of $\bar{b}$. The importance of variance inequalities for obtaining statistical rates of convergence for the empirical barycenter was emphasized in (Ahidar-Coutrix et al., 2020). In (Le Gouic et al., 2019), it is shown that an assumption on the regularity of the transport maps from the barycenter $\bar{b}$ implies a variance inequality. Specifically, suppose that all of the Kantorovich potentials $\varphi_{\bar{b}\to\mu}$ for $\mu \in \operatorname{supp} Q$ are (α, β)-regular in the sense of (3). Then, a variance inequality holds with $C_{\mathsf{var}} = 1 - (\beta - \alpha)$.

It turns out that a variance inequality holds without needing to assume smoothness of $\varphi_{\bar{b}\to\mu}$: assuming that the potential $\varphi_{\bar{b}\to\mu}$ is $(\alpha(\mu), \infty)$-regular for each $\mu \in \operatorname{supp} Q$ yields a variance inequality with $C_{\mathsf{var}} = \int \alpha(\mu)\, \mathrm{d}Q(\mu)$. The improvement here is critical for achieving global results on the Bures manifold. Moreover, when combined with the work of (Ahidar-Coutrix et al., 2020) it yields improved statistical guarantees for the empirical barycenter. To formally state this result, we need the notion of an *optimal dual solution* for the barycenter problem. A discussion of this concept, along with a proof of the following theorem, is given in Appendix C.2. We verify that the hypotheses of the theorem hold in the case when Q is supported on non-degenerate Gaussian measures in Appendix C.5.

Theorem 6 (Variance inequality) *Fix $Q \in \mathcal{P}_2(\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D))$ be a distribution with barycenter $\bar{b} \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$. Assume that there exists an optimal dual solution φ for the barycenter problem w.r.t. $\bar{b}$ such that, for Q-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 $\alpha : \mathcal{P}_2(\mathbb{R}^D) \to \mathbb{R}_+$. Then, Q satisfies a variance inequality (10) with constant*

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

3.3.2. SMOOTHNESS.

Recall that a convex differentiable function $f : \mathbb{R}^D \to \mathbb{R}$ is β-smooth if

$$f(y) \leqslant f(x) + \langle \nabla f(x), y - x\rangle + \frac{\beta}{2}\|y - x\|^2, \qquad \forall x, y \in \mathbb{R}^D. \tag{11}$$

A consequence of β-smoothness is the following inequality, which measures how much progress gradient descent makes in a single step (Bubeck, 2015).

$$f\big(x - \beta^{-1}\nabla f(x)\big) - f(x) \leqslant -\frac{1}{2\beta}\|\nabla f(x)\|^2. \tag{12}$$

In fact, only the latter inequality (12) is needed for the analysis of gradient descent methods. It was noted, first in (Álvarez Esteban et al., 2016, Proposition 3.3) and then in (Zemel and Panaretos, 2019, Lemma 2), that an analogue of (12) holds in Wasserstein space for the barycenter functional. Below, we provide a different, more geometric proof of this fact that emphasizes the collective role of smoothness and curvature. On the way, we also establish a smoothness inequality (13) that is used in the proof of Theorem 4 and also ensures that $C_{\mathsf{PL}} \leqslant 1$ for any distribution Q supported on $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$.

Theorem 7 *For any $b_0, b_1 \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$ the barycenter functional satisfies the smoothness inequality*

$$G(b_1) \leqslant G(b_0) + \langle \nabla G(b_0), \log_{b_0} b_1\rangle_{b_0} + \frac{1}{2}W_2^2(b_0, b_1). \tag{13}$$

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

$$G(b^+) - G(b) \leqslant -\frac{1}{2}\|\nabla G(b)\|_b^2. \tag{14}$$

Proof Let $(b_s)_{s\in[0,1]}$ be the constant-speed geodesic between arbitrary $b_0, b_1 \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$. From the non-negative curvature inequality (4), it holds that for any $s \in (0, 1]$,

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

By dominated convergence, the left-hand side converges to

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

where in the first identity, we used the characterization of (Ambrosio et al., 2008, Proposition 7.3.6). Rearranging terms yields (13).

Noticing that $W_2^2(b, b^+) = \|-\nabla G(b)\|_b^2$, Theorem 7 is now an immediate consequence of (13) applied to $b_0 = b$ and $b_1 = b^+$. ■

3.3.3. AN INTEGRATED PL INEQUALITY.

The main technical hurdle of this work is to provide sufficient conditions under which the PL inequality holds. The following lemma, proved in Appendix C.3, is our main device to establish PL inequalities.

Lemma 8 *Let Q 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 $\bar{b}$ of Q is absolutely continuous w.r.t. b. Assume further the following measurability conditions: there exists a measurable mapping $\varphi : \mathcal{P}_2(\mathbb{R}^D) \times \mathbb{R}^D \to \mathbb{R} \cup \{\infty\}$, $(\mu, x) \mapsto \varphi_{b\to\mu}(x)$, such that, for Q-almost every $\mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$, $\varphi_{b\to\mu} : \mathbb{R}^D \to \mathbb{R} \cup \{\infty\}$ is a Kantorovich potential for the optimal transport from b to μ. Then,*

$$G(b) - G(\bar{b}) \leqslant \frac{2}{C_{\mathsf{var}}} \Big(\int_0^1 \|\nabla G(b)\|_{L^2(b_s)} \, \mathrm{d}s \Big)^2 ,$$

where $(b_s)_{s\in[0,1]}$ is the constant-speed W_2-geodesic beginning at $b_0 := b$ and ending at $b_1 := \bar{b}$.

This lemma can yield a PL inequality in quite general situations, but the crucial issue is whether these conditions hold uniformly for each iterate in the optimization trajectory. In the next section, we show how to turn an integrated PL inequality into a bona fide PL inequality when Q is supported on certain Gaussian measures.

4. Gradient descent on the Bures-Wasserstein manifold

Upon identifying a centered non-degenerate Gaussian 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 Q as the *Bures-Wasserstein barycenter*.

4.1. Bures-Wasserstein gradient descent algorithms

We now specialize both GD and SGD when Q 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. The set of non-degenerate Gaussians constitutes a well-behaved subset of Wasserstein space, called the *Bures-Wasserstein* manifold (Bures, 1969; Bhatia et al., 2019). In particular, the optimal coupling between γ_{m_0,Σ_0} and γ_{m_1,Σ_1} has the explicit form

$$x \mapsto T_{\gamma_{\mu_0,\Sigma_0}\to\gamma_{\mu_1,\Sigma_1}}(x) := m_1 + \Sigma_0^{-1/2}(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2})^{1/2}\Sigma_0^{-1/2}(x - m_0). \tag{15}$$

Observe that $T_{\gamma_{\mu_0,\Sigma_0}\to\gamma_{\mu_1,\Sigma_1}}$ is affine, and thus $\int T_{\gamma_{\mu_0,\Sigma_0}\to\gamma} \, \mathrm{d}Q(\gamma)$ 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 $m \mapsto \int \|m - m(\mu)\|^2 \, \mathrm{d}Q(\mu)$, which is just mean estimation in $\mathbb{R}^D$. Therefore, for simplicity and without loss of generality, we consider only mean-zero Gaussians throughout this paper and we simply have to write down the update equations for the covariance matrix Σ_t of the iterate. The resulting update equations are summarized in Algorithms 1 and 2 below.

Algorithm 1: Bures-Wasserstein GD

Input: Σ_0, Q, T
for $t = 1, \dots, T$ **do**
  $S_t \leftarrow \int \{\Sigma_{t-1}^{1/2} \Sigma(\mu) \Sigma_{t-1}^{1/2}\}^{1/2} \, \mathrm{d}Q(\mu)$
  $\Sigma_t \leftarrow \Sigma_{t-1}^{-1/2} S_t^2 \Sigma_{t-1}^{-1/2}$
end
return Σ_T

Algorithm 2: Bures-Wasserstein SGD

Input: $\Sigma_0, (\eta_t)_{t=1}^T, (K_t)_{t=1}^T$
for $t = 1, \dots, T$ **do**
  $\hat{S}_t \leftarrow \Sigma_{t-1}^{-1/2} \{\Sigma_{t-1}^{1/2} K_t \Sigma_{t-1}^{1/2}\}^{1/2} \Sigma_{t-1}^{-1/2}$
  $\hat{G}_t \leftarrow (1 - \eta_t) I_D + \eta_t \hat{S}_t$
  $\Sigma_t \leftarrow \hat{G}_t \Sigma_{t-1} \hat{G}_t$
end
return Σ_T

In the rest of this section, we prove the guarantees for GD and SGD on the Bures-Wasserstein manifold given in Theorems 1 and 2.

4.2. 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 apply a simple rescaling argument (Lemma 14 in the Appendix).

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

$$\mathcal{S}_\zeta = \left\{ \gamma_{0,\Sigma} \,:\, \Sigma \in \mathbb{S}_{++}^D,\ \|\Sigma\|_{\mathrm{op}} \leqslant 1,\ \det \Sigma \geqslant \zeta \right\}. \tag{16}$$

Hereafter, we always assume that Q is ζ-regular for some $\zeta > 0$. Under this condition, it can be shown that the barycenter of Q exists and is unique (Proposition 15 in the Appendix).

We begin with a brief outline of the proof.

(i) If we initialize gradient descent (or stochastic gradient descent) at one of the elements of the support of Q, then all of the iterates, all of the elements of $\operatorname{supp} Q$, the barycenter $\bar{b}$, and all of elements of geodesics between these measures are non-denegerate Gaussians $\gamma_{0,\Sigma} \in \mathcal{S}_\zeta$.

(ii) Using Lemma 8, we establish a PL inequality holds with a uniform constant over $\mathcal{S}_\zeta$.

(iii) The guarantees for GD and SGD on the Bures manifold follow immediately from the PL inequality and our general convergence results (Theorems 4, 5).

In the sequel, we use *geodesic convexity* as a key tool to control the iterates of the gradient descent algorithm. We note that this discussion is not about proving some sort of geodesic convexity for our objective, which cannot hold in general. Our main interest in geodesic convexity comes from the following fact: if all of the elements of the support of Q lie in a geodesically convex set $\mathcal{S}_\zeta$, and we initialize the algorithm at an element of $\mathcal{S}_\zeta$, then all of the iterates of stochastic gradient descent are simply moving along geodesics within this set, and so remain in $\mathcal{S}_\zeta$. The same is true for the iterates of gradient descent, provided that we replace geodesic convexity with *convexity along generalized geodesics*. Refer to Section 3.1 for definitions of these terms. We begin with the following fact.

Lemma 10 *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 (Ambrosio et al., 2008, Proposition 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 18 in the Appendix to the Bures-Wasserstein manifold.

Lemma 11 *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 10 and 11 that the set $\mathcal{S}_\zeta$ is convex along generalized geodesics. Moreover since SGD moves along geodesics and is initialized at $b_0 \in \operatorname{supp} Q \subset \mathcal{S}_\zeta$, 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, $-\nabla G(b_t) = \int (T_{b_t\to\mu} - \mathrm{id})\,\mathrm{d}Q(\mu) \in \log_{b_t}(\mathcal{S}_\zeta)$ as a convex combination of elements in this set. This is equivalent to $b_{t+1} = \exp_{b_t}(-\nabla G(b_t)) \in \mathcal{S}_\zeta$. These observations yield the following corollary.

Corollary 12 *The set $\mathcal{S}_\zeta$ is convex along generalized geodesics and when initialized in $\operatorname{supp} Q$, the iterates of both GD and SGD remain in $\mathcal{S}_\zeta$.*

This completes the first step (i) of the proof. Moving on to step (ii), we get from Theorem 19 that G satisfies a PL inequality with constant $C_{\mathsf{PL}} = \zeta^2/4$ at all $b \in \mathcal{S}_\zeta$ and in particular at all the iterates of both GD and SGD.

Combined with the general bound in Theorems 4 and the variance inequality in Theorem 17, this completes the proof of Theorems 1 for GD. To prove Theorem 2, take $k = 1/C_{\mathsf{PL}} = 4/\zeta^2$ so that Theorem 5 yields

$$\mathbb{E}G(b_n) - G(\bar{b}) \leqslant \frac{48\operatorname{var}(Q)}{n\zeta^4}\,.$$

Combining this bound with the variance inequality in Theorem 17 completes the proof of Theorem 2.

Acknowledgments

PR was supported by NSF awards IIS-1838071, DMS-1712596, DMS-TRIPODS-1740751, and ONR grant N00014-17- 1-2147. Sinho Chewi and Austin J. Stromme were supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program. We thank the anonymous reviewers for helpful comments.

References

M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM J. Math. Anal.*, 43(2): 904–924, 2011.

A. Ahidar-Coutrix, T. Le Gouic, and Q. Paris. Convergence rates for empirical barycenters in metric spaces: curvature, convexity and extendable geodesics. *Probab. Theory Related Fields*, 177(1-2): 323–368, 2020.

P. C. Álvarez Esteban, E. del Barrio, J. A. Cuesta-Albertos, and C. Matrán. A fixed-point approach to barycenters in Wasserstein space. *J. Math. Anal. Appl.*, 441(2):744–762, 2016.

L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008.

C. Auderset, C. Mazza, and E. A. Ruh. Angular Gaussian and Cauchy estimation. *Journal of Multivariate Analysis*, 93(1):180–197, 2005.

M. Bacak. *Convex analysis and optimization in Hadamard spaces*. De Gruyter Series in Nonlinear Analysis and Applications. De Gruyter, Berlin, 2014.

J. Backhoff-Veraguas, J. Fontbona, G. Rios, and F. Tobar. Bayesian learning with Wasserstein barycenters. *arXiv e-prints*, art. arXiv:1805.10833, May 2018.

R. Bhatia, T. Jain, and Y. Lim. On the Bures–Wasserstein distance between positive definite matrices. *Expo. Math.*, 37(2):165–191, 2019.

J. Bigot, R. Gouet, T. Klein, and A. López. Upper and lower risk bounds for estimating the Wasserstein barycenter of random measures on the real line. *Electron. J. Stat.*, 12(2):2253–2289, 2018.

S. Bonnabel. Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013.

N. Bonneel, G. Peyré, and M. Cuturi. Wasserstein barycentric coordinates: Histogram regression using optimal transport. *ACM Transactions on Graphics*, 35(4), 2016.

J. M. Borwein and A. S. Lewis. *Convex analysis and nonlinear optimization*, volume 3 of *CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC*. Springer, New York, second edition, 2006. Theory and examples.

S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

D. Burago, Y. Burago, and S. Ivanov. *A course in metric geometry*, volume 33 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2001.

D. 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.

G. Carlier and I. Ekeland. Matching for teams. *Economic Theory*, 42(2):397–418, 2010.

S. Claici, E. Chien, and J. Solomon. Stochastic Wasserstein barycenters. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 999–1008, Stockholmsmssan, Stockholm Sweden, 2018.

M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 685–693, Bejing, China, 2014.

M. P. a. do Carmo. *Riemannian geometry*. Mathematics: Theory & Applications. Birkhäuser Boston, Inc., Boston, MA, 1992. Translated from the second Portuguese edition by Francis Flaherty.

D. Dvinskikh. Stochastic approximation versus sample average approximation for population Wasserstein barycenter calculation. *arXiv e-prints*, art. arXiv:2001.07697, January 2020.

A. Gramfort, G. Peyré, and M. Cuturi. Fast optimal transport averaging of neuroimaging data. In *International Conference on Information Processing in Medical Imaging*, pages 261–272. Springer, 2015.

S. Guminov, P. Dvurechensky, N. Tupitsa, and A. Gasnikov. Accelerated alternating minimization, accelerated Sinkhorn's algorithm and accelerated iterative Bregman projections. *arXiv e-prints*, art. arXiv:1906.03622, June 2019.

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.

H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Lojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

M. Knott and C. S. Smith. On a generalization of cyclic monotonicity and distances among random vectors. *Linear Algebra and its Applications*, 199:363–371, 1994.

A. Kroshnin, V. Spokoiny, and A. Suvorikova. Statistical inference for Bures-Wasserstein barycenters. *arXiv e-prints*, art. arXiv:1901.00226, January 2019.

A. Kroshnin, N. Tupitsa, D. Dvinskikh, P. Dvurechensky, A. Gasnikov, and C. Uribe. On the complexity of approximating Wasserstein barycenters. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3530–3540, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

T. Le Gouic. Dual and multimarginal problems for the Wasserstein barycenter. 2020. Unpublished.

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

T. Le Gouic and J.-M. Loubes. The price for fairness in a regression framework. *arXiv e-prints*, art. arXiv:2005.11720, May 2020.

T. Le Gouic, Q. Paris, P. Rigollet, and A. J. Stromme. Fast convergence of empirical barycenters in Alexandrov spaces and the Wasserstein space. *arXiv e-prints*, art. arXiv:1908.00828, August 2019.

T. Lin, N. Ho, X. Chen, M. Cuturi, and M. I. Jordan. Fixed-support Wasserstein barycenters: Computational hardness and fast algorithm. *arXiv e-prints*, art. arXiv:2002.04783, February 2020.

J. Lott and C. Villani. Ricci curvature for metric-measure spaces via optimal transport. *Ann. of Math. (2)*, 169(3):903–991, 2009.

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

A. Martial and G. Carlier. Vers un théorème de la limite centrale dans l'espace de Wasserstein? *C. R. Math. Acad. Sci. Paris*, 355(7):812–818, 2017.

K. Modin. Geometry of matrix decompositions seen through optimal transport and information geometry. *J. Geom. Mech.*, 9(3):335–390, 2017.

F. Otto. The geometry of dissipative evolution equations: the porous medium equation. *Comm. Partial Differential Equations*, 26:101–174, 2001.

V. M. Panaretos and Y. Zemel. Amplitude and phase variation of point processes. *Ann. Statist.*, 44 (2):771–812, 2016.

J. Rabin and N. 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.

J. Rabin, G. Peyré, J. Delon, and M. 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.

R. T. Rockafellar. *Convex analysis*. Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ, 1997. Reprint of the 1970 original, Princeton Paperbacks.

F. Santambrogio. *Optimal transport for applied mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and their Applications*. Birkhäuser/Springer, Cham, 2015.

J. Solomon, F. De Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):1–11, 2015.

S. Srivastava, C. Li, and D. B. Dunson. Scalable Bayes via barycenter in Wasserstein space. *The Journal of Machine Learning Research*, 19(1):312–346, 2018.

K.-T. Sturm. Probability measures on metric spaces of nonpositive curvature. In *Heat kernels and analysis on manifolds, graphs, and metric spaces (Paris, 2002)*, volume 338 of *Contemp. Math.*, pages 357–390. Amer. Math. Soc., Providence, RI, 2003.

N. Tripuraneni, N. Flammarion, F. Bach, and M. I. Jordan. Averaging stochastic gradient descent on Riemannian manifolds. In S. Bubeck, V. Perchet, and P. Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 650–687, 2018.

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

C. Villani. *Optimal transport: old and new*, volume 338 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009.

M. Weber and S. Sra. Riemannian optimization via Frank-Wolfe methods. *arXiv e-prints*, art. arXiv:1710.10770, October 2017.

M. Weber and S. Sra. Projection-free nonconvex stochastic optimization on Riemannian manifolds. *arXiv e-prints*, art. arXiv:1910.04194, October 2019.

A. Wiesel. Geodesic convexity and covariance estimation. *IEEE Transactions on Signal Processing*, 60(12):6182–6189, 2012.

Y. Zemel and V. M. Panaretos. Fréchet means and Procrustes analysis in Wasserstein space. *Bernoulli*, 25(2):932–976, 2019.

H. Zhang and S. Sra. First-order methods for geodesically convex optimization. In V. Feldman, A. Rakhlin, and O. Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1617–1638, Columbia University, New York, New York, USA, 2016.

Appendix A. Geometry and Wasserstein space

In this section we give a more detailed introduction to Riemannian manifolds and discuss analogies to Wasserstein space which are present throughout the paper. We refer readers to (do Carmo, 1992) for a standard introduction to Riemannian geometry.

A.1. Riemannian geometry

An n-dimensional manifold M is a topological space which is Hausdorff, second countable, and locally homeomorphic to $\mathbb{R}^n$. A smooth atlas is a collection of smooth charts $\{\psi_\alpha\}_{\alpha\in\mathcal{A}}$ so that each $\psi_\alpha \colon U_\alpha \subset M \to \mathbb{R}^n$ is a homeomorphism from an open set U_α in M, $M = \bigcup_{\alpha\in\mathcal{A}} U_\alpha$, and such that for all $\alpha, \alpha' \in \mathcal{A}$, $\psi_\alpha \circ \psi_{\alpha'}^{-1}$ is smooth wherever defined. For a fixed choice of smooth atlas, we

declare a function $f\colon M \to \mathbb{R}$ to be smooth if $f \circ \psi_\alpha^{-1}$ is for each $\alpha \in \mathcal{A}$. The manifold together with a smooth atlas defines a smooth n-dimensional manifold, and we shall always suppress mention of the atlas. A map $f\colon M \to N$ between two smooth manifolds is said to be smooth if its composition with smooth charts is.

Given a smooth n-dimensional manifold M and a point $p \in M$, the tangent space T_pM is the equivalence class of all smooth curves $\gamma\colon (-\varepsilon, \varepsilon) \to M$ such that $\gamma(0) = p$, where two such curves γ_0, γ_1 are equivalent if, with respect to every coordinate chart ψ defined in a neighborhood of p, $(\psi \circ \gamma_0)'(0) = (\psi \circ \gamma_1)'(0)$. As such, T_pM is a real n-dimensional vector space for each $p \in M$. The cotangent space at $p \in M$ is then the dual to T_pM, which we shall denote T_p^*M. The tangent bundle is the disjoint union $TM := \bigsqcup_{p \in M} T_pM$, and the cotangent bundle is similarly the disjoint union $T^*M := \bigsqcup_{p \in M} T_p^*M$. The smooth structure on M induces a smooth structure on TM and T^*M, so each is then a $2n$-dimensional smooth manifold in its own right.

A smooth vector field $X\colon M \to TM$ is then a smooth map $p \mapsto X_p$ such that $X_p \in T_pM$ for all $p \in M$, and similarly for a smooth covector field $\alpha\colon M \to T^*M$. Higher-order tensors are defined similarly: a (p, q)-tensor field is a smooth mapping $T\colon M \to (TM)^p \otimes (T^*M)^q$. The differential $df\colon M \to T^*M$ of a smooth function f on M is the smooth covector field such that $df_p\colon T_pM \to \mathbb{R}$ obeys $df_p(v) := (f \circ \gamma)'(0)$, where γ is any curve with tangent vector $v \in T_pM$ at $\gamma(0) = p$.

A Riemannian manifold (M, g) is a smooth n-dimensional manifold M with a smooth metric tensor $g\colon M \to T^*M \otimes T^*M$; at each point of M, this is a positive definite bilinear form. The metric tensor therefore defines a smoothly varying choice of inner product on the tangent spaces of M. In addition to giving rise to notions of length and geodesics, the metric tensor provides a canonical isomorphism (the Riesz isomorphism) between the tangent space and cotangent space: for a vector $v \in T_pM$ the covector $\alpha_v \in T_p^*M$ is defined by $\alpha_v(w) = g_p(v, w)$. For a covector $\alpha \in T_p^*M$ the vector $v_\alpha \in T_pM$ is defined as the unique solution of $\alpha(w) = g_p(v_\alpha, w)$ for all $w \in T_pM$. A smooth vector field X can be accordingly transformed into a smooth covector field denoted $X^\flat$, and a smooth covector field ω can be transformed into a smooth vector field $\omega^\#$. The gradient of a function $f\colon M \to \mathbb{R}$ is defined then as $\nabla f := (df)^\#$: in other words, for all $p \in M$ and $v \in T_pM$, $df_p(v) = g_p(\nabla f(p), v)$.

We typically write $\langle \cdot, \cdot \rangle_p$ instead of $g_p(\cdot, \cdot)$, and we write $\|\cdot\|_p$ for the norm induced by the metric tensor, i.e., $\|v\|_p := \sqrt{\langle v, v \rangle_p}$. In this notation, the distance between points $p, q \in M$ is defined as

$$d_M(p, q) := \inf_{\gamma \in \Gamma(p,q)} \int_0^1 \|\gamma'(t)\|_{\gamma(t)} \, \mathrm{d}t,$$

where $\Gamma(p, q)$ is the collection of all smooth (or piecewise continuous) curves $\gamma\colon [0, 1] \to M$ such that $\gamma(0) = p$ and $\gamma(1) = q$. If M is connected, then the distance d_M is indeed a metric. If we additionally assume that (M, d_M) is complete as a metric space then by the Hopf-Rinow theorem the value of the above minimization problem is attained by at least one curve $\gamma\colon [0, 1] \to M$ such that $t \mapsto \|\gamma'(t)\|_{\gamma(t)}$ is constant, which is said to be a constant-speed (minimizing) geodesic.

For any $p \in M$, there always exists an $\varepsilon > 0$ such that for any vector $v \in T_pM$ with $\|v\|_p < \varepsilon$, there is a unique constant-speed geodesic $\gamma_v\colon [0, 1] \to M$ obeying $\gamma_v(0) = p$ and $\gamma_v'(0) = v$.[1] On the ball $B_\varepsilon(0)$ with radius ε and center $0 \in T_pM$ (with respect to the norm $\|\cdot\|_p$), we can now

1. In fact, a stronger result holds: there exists a neighborhood U of p such that for any two points $q, q' \in U$, there is a unique constant-speed minimizing geodesic $\gamma\colon [0, 1] \to U$ joining q to q'. Such a neighborhood is called a totally normal neighborhood of p.

define the exponential map $\exp_p\colon B_\varepsilon(0) \to M$ by $v \in V_p \mapsto \gamma_v(1)$. The exponential map is a diffeomorphism onto its image, so we can define the inverse mapping $\log_p\colon \exp_p(B_\varepsilon(0)) \to T_pM$. If M is complete, the domain of definition of any constant-speed geodesic $\gamma\colon [0,1] \to \mathbb{R}$ can be extended to all of $\mathbb{R}$ such that at each time γ is locally a constant-speed minimizing geodesic; in this case, the exponential mapping can be extended to a mapping $\exp_p\colon T_pM \to M$. Note, however, that the mapping $\log_p$ is not necessarily defined everywhere.

We lastly recall that for fixed $q \in M$ and p which does not belong to the cut locus of q (the set of points for which there exists more than one constant-speed minimizing geodesic from p),

$$[\nabla d_M^2(\cdot, q)](p) = -2\log_p(q).$$

This statement has an intuitive meaning: it simply says that outside of the cut locus of q, the gradient of the squared distance points in the direction of maximum increase.[2]

A.2. Riemannian interpretation of Wasserstein space

In this section, we briefly explain the interpretation set out in (Otto, 2001) of the Wasserstein space of probability measures as a Riemannian manifold. For more introductory expositions of this subject, we refer to (Villani, 2003, Chapter 8) and (Santambrogio, 2015, Chapter 5). The task of putting this formal discussion on rigorous footing is undertaken in (Ambrosio et al., 2008, Chapter 8). We also note that many treatments view Wasserstein space as a length space using the framework of metric geometry; see (Burago et al., 2001) for an introduction to this approach.

Let $\mu_0 \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$ and consider a family $(v_t)_{t\in[0,1]}$ of smooth vector fields on $\mathbb{R}^D$, that is, $v_t : \mathbb{R}^D \to \mathbb{R}^D$ for each $t \in [0,1]$. Suppose we draw $X_0 \sim \mu_0$ and we evolve X_0 according to the ODE $\dot{X}_t = v_t(X_t)$ for $t \in [0,1]$, that is, we seek an integral curve of $(v_t)_{t\in[0,1]}$ with starting point X_0. If we let μ_t denote the law of X_t, we may compute the evolution of $(\mu_t)_{t\in[0,1]}$ as follows. Take any smooth test function ψ on $\mathbb{R}^D$, and (ignoring any issues of regularity) compute

$$\begin{aligned}\partial_t \int \psi \,\mathrm{d}\mu_t = \partial_t \mathbb{E}\psi(X_t) = \mathbb{E}\partial_t\psi(X_t) = \mathbb{E}\langle\nabla\psi(X_t), v_t(X_t)\rangle = \int \langle\nabla\psi, v_t\rangle \,\mathrm{d}\mu_t \\ = -\int \psi \,\mathrm{div}(v_t\mu_t).\end{aligned}$$

This suggests that the pair $(\mu_t)_{t\in[0,1]}$, $(v_t)_{t\in[0,1]}$ should solve the following PDE, which is known as the continuity equation:

$$\partial_t\mu_t + \mathrm{div}(v_t\mu_t) = 0. \tag{17}$$

This PDE can be interpreted in a suitable weak sense, e.g.: for any smooth test function ψ with compact support, the mapping $t \mapsto \int \psi \,\mathrm{d}\mu_t$ should be absolutely continuous and thus differentiable at almost every $t \in [0,1]$, and its derivative should satisfy $\partial_t \int \psi \,\mathrm{d}\mu_t = \int \langle\nabla\psi, v_t\rangle \,\mathrm{d}\mu_t$.

Since the vector fields $(v_t)_{t\in[0,1]}$ govern the evolution of the curve $(\mu_t)_{t\in[0,1]} \subseteq \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$, we would like to equip $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$ with the structure of a Riemannian manifold such that $(v_t)_{t\in[0,1]}$ is interpreted as the tangent vectors to the curve $(\mu_t)_{t\in[0,1]}$. However, a problem arises: given a

2. When there are multiple constant-speed minimizing geodesics joining p to q, then the following fact is still true: the squared distance function $d_M^2(\cdot, q)$ is superdifferentiable at p. Moreover, for any constant-speed minimizing geodesic $\gamma\colon [0,1] \to M$ joining p to q, the vector $-2\gamma'(0) \in T_pM$ is a supergradient of $d_M^2(\cdot, q)$ at p.

curve $(\mu_t)_{t\in[0,1]}$ in Wasserstein space, there are many choices for the vector fields $(v_t)_{t\in[0,1]}$ which solve (17) together with $(\mu_t)_{t\in[0,1]}$. Indeed, if we fix any pair $(\mu_t)_{t\in[0,1]}$, $(v_t)_{t\in[0,1]}$ solving (17), then we obtain another solution by replacing v_t with $v_t + w_t$, where w_t is any vector field satisfying $\operatorname{div}(w_t\mu_t) = 0$. So, what should we take as the tangent vectors to $(\mu_t)_{t\in[0,1]}$?

We can take a hint from optimal transport. Specifically, Brenier's theorem asserts that in the optimal transport problem of transporting a measure ν_0 to another measure ν_1, the optimal transport plan is induced by a transport map, which is the gradient of a convex function φ. In other words, if we interpret ν_0 as a collection of particles, then each particle initially moves along the vector field $\nabla\varphi - \mathrm{id}$. In particular, taking $\nu_0 = \mu_0$ and $\nu_1 = \mu_\varepsilon$ for a small $\varepsilon > 0$, we expect the tangent vector of $(\mu_t)_{t\in[0,1]}$ at time 0 to be of the form $\nabla\varphi - \mathrm{id}$ for a convex function φ.

This motivates the definition of the tangent space to $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$ at a measure b, given in (Ambrosio et al., 2008, Chapter 8) as

$$T_b\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D) := \overline{\{\lambda\,(\nabla\varphi - \mathrm{id})\colon \lambda > 0,\ \varphi \in C_c^\infty(\mathbb{R}^D),\ \varphi \text{ convex}\}}^{L^2(b)},$$

where the closure is with respect to the $L^2(b)$ distance. We equip $T_b\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$ with the $L^2(b)$ metric, that is, for vector fields $v, w \in T_b\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$ we define $\langle v, w\rangle_b := \langle v, w\rangle_{L^2(b)} := \int\langle v, w\rangle\,\mathrm{d}b$. The metric induced by this Riemannian structure recovers the Wasserstein distance, in the sense that

$$W_2(\mu_0,\mu_1) = \inf\Big\{\int_0^1 \|v_t\|_{\mu_t}\,\mathrm{d}t \;\Big|\; (\mu_t)_{t\in[0,1]}, (v_t)_{t\in[0,1]} \text{ solves (17)}\Big\}.$$

Given two measures $\mu_0, \mu_1 \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$, there is a unique constant-speed minimizing geodesic joining μ_0 to μ_1. It is given by $\mu_t := [(1-t)\,\mathrm{id} + tT]_\#\mu_0$, where T is the optimal transport mapping from μ_0 to μ_1; this is known as McCann's interpolation. It satisfies

$$W_2(\mu_0,\mu_t) = tW_2(\mu_0,\mu_1) \qquad \forall t\in[0,1]. \tag{18}$$

Moreover, it can be shown that any constant-speed geodesic in $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$, that is, any curve $(\mu_t)_{t\in[0,1]} \subseteq \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$ satisfying (18), is necessarily of the form $\mu_t = [(1-t)\,\mathrm{id} + tT]_\#\mu_0$. The tangent vector to $(\mu_t)_{t\in[0,1]}$ at time 0 is the vector field $T - \mathrm{id}$.

Given $\mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$ and $v \in T_\mu\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$, we may now define the exponential map to be $\exp_\mu v := (\mathrm{id} + v)_\#\mu$. Given any other $\nu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$, we also define the logarithmic map to be $\log_\mu \nu := T_{\mu\to\nu} - \mathrm{id}$, where $T_{\mu\to\nu}$ is the optimal transport map from μ to ν. Observe that $\log_\mu \nu$ is well-defined for *any* pair $\mu, \nu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$.

Appendix B. Experiments

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.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 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 19) or also from general arguments in (Zemel and Panaretos, 2019, Theorem 2). We also use the fact that the tangent space of the Bures manifold is given by the set of all symmetric matrices (Bhatia et al., 2019).

Figure 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

$$\Sigma_i = \exp_{\gamma_{0,I_D}}\Big(\frac{A_i + A_i^\top}{2}\Big), \tag{19}$$

which has population barycenter $b^\star = \gamma_{0,I_D}$. An explicit form of this exponential map is derived in (Malagò et al., 2018). 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 $\Sigma_1, \dots, \Sigma_n$ 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

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

On Riemannian manifolds, averaged SGD is known to attain optimal statistical rates under smoothness and geodesic convexity assumptions (Tripuraneni et al., 2018).

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 3, SGD achieves the fast $O(n^{-1})$ statistical rate on this dataset.

The right of Figure 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 (Álvarez Esteban et al., 2016).

In Figure 4, we repeat the same experiment, except this time the barycenter has covariance 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 4 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 5 shows convergence of SGD with replacement to the empirical barycenter. We generate $n = 500$ samples in the same way as in Figure 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

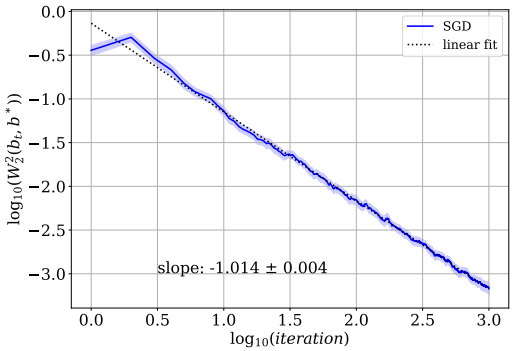

Figure 3: Log-log plot of convergence for SGD on Bures manifold for $n = 1000$, $d = 3$, and and $b^\star = \gamma_{0,I_3}$. This corresponds to the experiment on the left in Figure 2

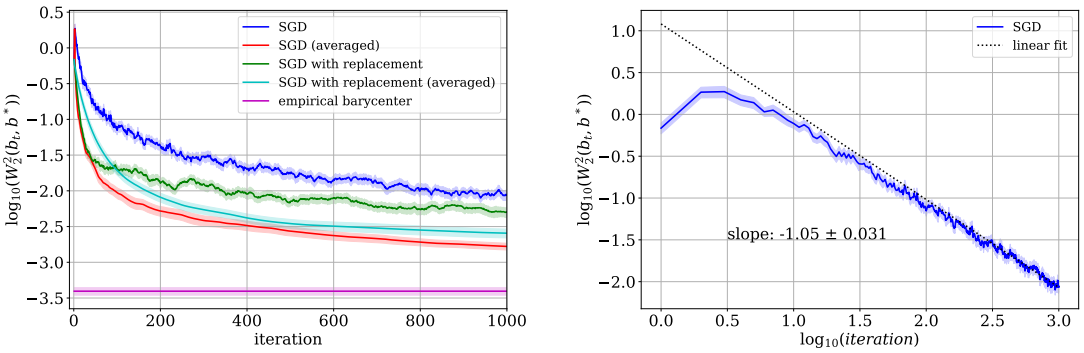

Figure 4: Convergence of SGD on Bures manifold. Here, $n = 1000$, $d = 3$, and barycenter given by $\mathrm{diag}(20, 1, 1)$. The result displays the average over 100 randomly generated datasets.

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.

B.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 (Weber and Sra, 2019). Figure 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,

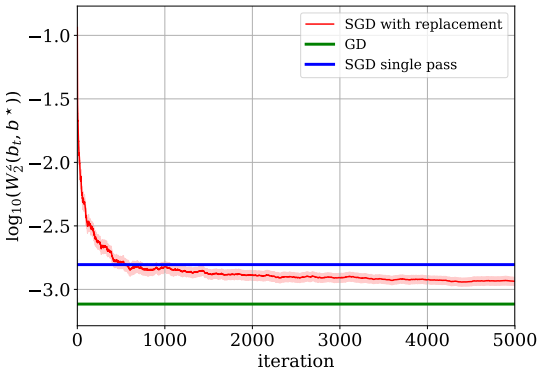

Figure 5: Convergence of SGD on Bures manifold. Here, $n = 500$, $d = 3$, and the distribution is given by (19) with $\Sigma^\star = I_3$ and $\sigma^2 = 0.25$. The result displays the average over 100 randomly generated datasets.

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 (Huang et al., 2015)).

Appendix C. Omitted proofs

C.1. Convergence bounds for GD and SGD under a PL inequality

This subsection gives proofs of the general convergence theorems for GD and SGD in the present paper. Both of these proofs use the non-negative curvature inequality (5). We note that the proof of Theorem 4 uses the non-negative curvature implicitly by invoking smoothness, while the use of non-negative curvature is explicit within the proof of Theorem 5.

C.1.1. PROOF OF THEOREM 4 FOR GD.

Using the smoothness (14) and the PL inequality (6), it holds that

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

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

C.1.2. PROOF OF THEOREM 5 FOR SGD.

Recall the SGD iterations on $n + 1$ observations:

$$b_0 := \mu_0, \qquad b_{t+1} := [(1 - \eta_t)\,\mathrm{id} + \eta_t T_{b_t \to \mu_{t+1}}]_{\#} b_t \quad \text{for } t = 0, \dots, n,$$

where the step size is given by

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

for some k such that $C_{\mathsf{PL}}^2(k+1)^2 \geqslant 2k+1$. We note that the step size η_t is chosen to solve the equation

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

Using the non-negative curvature (5), we get

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

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

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

Using the PL inequality (6),

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

Subtracting $G(\bar{b})$ and rearranging,

$$\mathbb{E}G(b_{t+1}) - G(\bar{b}) \leqslant (1 - 2C_{\mathsf{PL}}\eta_t + \eta_t^2)[\mathbb{E}G(b_t) - G(\bar{b})] + \frac{\eta_t^2}{2}\operatorname{var}(Q),$$

where we recall that $\operatorname{var}(Q) = 2G(\bar{b})$. With the chosen step size, we find

$$\mathbb{E}G(b_{t+1}) - G(\bar{b}) \leqslant \Big(\frac{t+k}{t+k+1}\Big)^2 [\mathbb{E}G(b_t) - G(\bar{b})] + \frac{2\operatorname{var}(Q)}{C_{\mathsf{PL}}^2(t+k+1)^2}.$$

Or equivalently,

$$(t+k+1)^2[\mathbb{E}G(b_{t+1}) - G(\bar{b})] \leqslant (t+k)^2[\mathbb{E}G(b_t) - G(\bar{b})] + \frac{2\operatorname{var}(Q)}{C_{\mathsf{PL}}^2}.$$

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

$$(n+k)^2[\mathbb{E}G(b_n) - G(\bar{b})] \leqslant k^2[\mathbb{E}G(b_0) - G(\bar{b})] + \frac{2n\operatorname{var}(Q)}{C_{\mathsf{PL}}^2},$$

or, equivalently,

$$\mathbb{E}G(b_n) - G(\bar{b}) \leqslant \frac{k^2}{(n+k)^2}[\mathbb{E}G(b_0) - G(\bar{b})] + \frac{2\operatorname{var}(Q)}{C_{\mathsf{PL}}^2(n+k)}. \tag{20}$$

To conclude the proof, recall that from (13), we have

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

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

$$\mathbb{E}G(b_0) - G(\bar{b}) \leqslant G(\bar{b}) = \frac{1}{2} \operatorname{var}(Q),$$

as claimed. Together with (20), it yields

$$\mathbb{E}G(b_n) - G(\bar{b}) \leqslant \frac{\operatorname{var}(Q)}{n+k} \Big(\frac{k^2}{2(n+k)} + \frac{2}{C_{\mathsf{PL}}^2} \Big) \leqslant \frac{\operatorname{var}(Q)}{n} \Big(\frac{k+1}{2} + \frac{2}{C_{\mathsf{PL}}^2} \Big).$$

Plugging-in the value of k completes the proof.

C.2. Variance inequality: Theorem 6

We begin this section with a review of Kantorovich duality, which we use to discuss the dual of the barycenter problem. Then, we present the proof of Theorem 6.

Given two measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^D)$ and maps $f \in L^1(\mu)$, $g \in L^1(\nu)$ such that $f(x) + g(y) \geqslant \langle x, y \rangle$ for μ-a.e. $x \in \mathbb{R}^D$ and ν-a.e. $y \in \mathbb{R}^D$, it is easy to see that

$$\frac{1}{2} W_2^2(\mu, \nu) \geqslant \int \Big(\frac{\|\cdot\|^2}{2} - f \Big) \, \mathrm{d}\mu + \int \Big(\frac{\|\cdot\|^2}{2} - g \Big) \, \mathrm{d}\nu.$$

Kantorovich duality (see e.g. (Villani, 2003)) says that equality holds for some pair $f = \varphi$, $g = \varphi^*$ where φ is a proper LSC convex function and φ^* denotes its convex conjugate, i.e.,

$$\frac{1}{2} W_2^2(\mu, \nu) = \int \Big(\frac{\|\cdot\|^2}{2} - \varphi \Big) \, \mathrm{d}\mu + \int \Big(\frac{\|\cdot\|^2}{2} - \varphi^* \Big) \, \mathrm{d}\nu.$$

The map φ is called a Kantorovich potential for (μ, ν).

Accordingly, given $\bar{b} \in \mathcal{P}_2(\mathbb{R}^D)$, we call a measurable mapping $\varphi : \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D) \to L^1(\bar{b})$, $\mu \mapsto \varphi_\mu$, an *optimal dual solution* for the barycenter problem if the following two conditions are met: (1) for Q-a.e. μ, the mapping φ_μ is a Kantorovich potential for $(\bar{b}, \mu)$; (2) it holds that

$$\int \Big(\frac{\|\cdot\|^2}{2} - \varphi_\mu \Big) \, \mathrm{d}Q(\mu) = 0. \tag{21}$$

It is easily seen that these conditions imply that $\bar{b}$ is the barycenter of Q:

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

The existence of an optimal dual solution for the barycenter problem is known in the finitely supported case (Agueh and Carlier, 2011), and existence can be shown for the general case under mild

conditions (Le Gouic, 2020). For completeness, we give a self-contained proof of the existence of an optimal dual solution in the case where Q is supported on Gaussian measures in Appendix C.5.

Proof [Proof of Theorem 6] By the strong convexity assumption, it holds for Q-a.e. $\mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$ and a.e. $x \in \mathbb{R}^D$,

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

which can be rearranged into

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

Integrating this w.r.t. the optimal transport plan γ_μ between μ and $b \in \mathcal{P}_2(\mathbb{R}^D)$, yields

$$\frac{1}{2}\Big(W_2^2(\mu, b) - \alpha(\mu)\int \|T_{\mu\to b} - T_{\mu\to\bar{b}}\|^2\,\mathrm{d}\mu\Big) \geqslant \int\Big(\frac{\|\cdot\|^2}{2} - \varphi_\mu^*\Big)\,\mathrm{d}\mu + \int\Big(\frac{\|\cdot\|^2}{2} - \varphi_\mu\Big)\,\mathrm{d}b.$$

Observe also that (5) implies $\|T_{\mu\to b} - T_{\mu\to\bar{b}}\|^2_{L^2(\mu)} \geqslant W_2^2(b, \bar{b})$. Integrating these inequalities with respect to Q yields

$$\begin{aligned} G(b) - \frac{1}{2}\Big(\int \alpha\,\mathrm{d}Q\Big)W_2^2(b, \bar{b}) &\geqslant \int\Big[\int\Big(\frac{\|\cdot\|^2}{2} - \varphi_\mu^*\Big)\,\mathrm{d}\mu + \int\Big(\frac{\|\cdot\|^2}{2} - \varphi_\mu\Big)\,\mathrm{d}b\Big]\,\mathrm{d}Q(\mu) \\ &= \iint\Big(\frac{\|\cdot\|^2}{2} - \varphi_\mu^*\Big)\,\mathrm{d}\mu\,\mathrm{d}Q(\mu) = G(\bar{b}). \end{aligned}$$

where in the last two identities, we used (21). It implies the variance inequality. ■

C.3. Integrated PL inequality

The following lemma appears in (Lott and Villani, 2009, Lemma A.1) in the case of Lipschitz functions. A minor modification of their proof allows to handle locally Lipschitz rather than only Lipschitz functions. We include the modified proof for completeness.

Lemma 13 *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 According to (Villani, 2009, 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:

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

It yields the result. ■

Proof [Proof of Lemma 8] By Kantorovich duality (Villani, 2003),

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

This yields the inequality

$$G(b) - G(\bar b) \leqslant \int\Big(\frac{\|\cdot\|^2}{2} - \int \varphi_{b\to\mu}\,\mathrm{d}Q(\mu)\Big)\,\mathrm{d}(b-\bar b).$$

Let $\bar\varphi := \int \varphi_{b\to\mu}\,\mathrm{d}Q(\mu)$; this is a proper LSC convex function $\mathbb{R}^D \to \mathbb{R}\cup\{\infty\}$. We apply Lemma 13 with $\Omega = \operatorname{int}\operatorname{dom}\bar\varphi$. Since $\bar\varphi$ is locally Lipschitz on the interior of its domain ((Rockafellar, 1997, Theorem 10.4) or (Borwein and Lewis, 2006, Theorem 4.1.3)) and $\bar b \ll b$, then $b(\Omega) = \bar b(\Omega) = 1$, whence

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

Square and rearrange to yield

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

Recognizing that $\nabla G(b) = \mathrm{id} - \nabla\bar\varphi$ yields the result. ■

C.4. Rescaling lemma

Lemma 14 *For any $\alpha > 0$ and $\mu \in \mathcal{P}_2(\mathbb{R}^D)$, let μ_α be the law of αX, where $X \sim \mu$. Let $\mu \sim Q$ be a random measure drawn from Q, and let Q_α be the law of μ_α. Then, $\bar b$ is a barycenter of Q if and only if $\bar b_\alpha$ is a barycenter of Q_α.*

Proof It is an easy calculation to see that for any $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^D)$,

$$W_2(\mu_\alpha, \nu_\alpha) = \alpha W_2(\mu, \nu)$$

(see, for instance, (Villani, 2003, Proposition 7.16)). Let

$$G_\alpha(b) := \frac{1}{2} \int W_2^2(\cdot, b) \, \mathrm{d}Q_\alpha(\mu).$$

By the previous reasoning, $G_\alpha(b_\alpha) = \alpha^2 G(b)$. In particular, the mapping $\bar{b} \mapsto \bar{b}_\alpha$ is a one-to-one correspondence between the minimizers of these two functionals. ■

C.5. Properties of the Bures-Wasserstein barycenter

Existence and uniqueness of the barycenter in the case where Q is finitely supported follows from the seminal work of Agueh and Carlier (Agueh and Carlier, 2011). We extend this result to the case where Q is not finitely supported.

Proposition 15 (Gaussian barycenter) *Fix $0 < \lambda_{\min} \leqslant \lambda_{\max} < \infty$. Let $Q \in \mathcal{P}_2(\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D))$ be such that for all $\mu \in \operatorname{supp} Q$, $\mu = \gamma_{m(\mu),\Sigma(\mu)}$ is a Gaussian with $\lambda_{\min} I_D \preceq \Sigma(\mu) \preceq \lambda_{\max} I_D$. Let $\gamma_{\bar{m},\bar{\Sigma}}$ be the Gaussian measure with mean $\bar{m} := \int m(\mu) \, \mathrm{d}Q(\mu)$ and covariance matrix $\bar{\Sigma}$ which is a fixed point of the mapping $S \mapsto G(S) := \int \left(S^{1/2} \Sigma(\mu) S^{1/2}\right)^{1/2} \mathrm{d}Q(\mu)$. Then, $\gamma_{\bar{m},\bar{\Sigma}}$ is the unique barycenter of Q.*

Proof To show that there exists a fixed point for the mapping G, apply Brouwer's fixed-point theorem as in (Agueh and Carlier, 2011, Theorem 6.1). To see that $\gamma_{\bar{m},\bar{\Sigma}}$ is indeed a barycenter, observe the mapping

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

satisfies the characterization (21) (so that φ is an optimal dual solution for the barycenter problem w.r.t. $\gamma_{\bar{m},\bar{\Sigma}}$) using the explicit form of the transport map (15), so $\gamma_{\bar{m},\bar{\Sigma}}$ is a barycenter of Q. Uniqueness follows from the variance inequality (Theorem 6) once we establish regularity of the optimal transport maps in Lemma 16. ■

Lemma 16 *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. ■

Theorem 6 readily yields the following variance inequality.

Theorem 17 *Fix $\zeta > 0$ and assume that Q is ζ-regular. Then Q has a unique barycenter $\bar{b}$ and it satisfes a variance inequality with constant $C_{\mathsf{var}} = \zeta$, that is, for any $b \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^D)$,*

$$G(b) - G(\bar{b}) \geqslant \frac{\zeta}{2} W_2^2(b, \bar{b})\,.$$

C.6. Generalized geodesic convexity of $\ln\|\cdot\|_{L^\infty}$

Lemma 18 *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 $\rho \mapsto \ln\|\rho\|_{L^\infty}$ 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 (Villani, 2003, 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

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

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

$$\ln b(x) = (1-s)\ln\mu\big(\nabla\varphi_{b\to\mu}(x)\big) + s\ln\nu\big(\nabla\varphi_{b\to\nu}(x)\big) \\ + (1-s)\ln\det D^2_{\mathrm{A}}\varphi_{b\to\mu}(x) + s\ln\det D^2_{\mathrm{A}}\varphi_{b\to\nu}(x).$$

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. ■

C.7. A PL inequality on the Bures-Wasserstein manifold

Theorem 19

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

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

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 $\tilde b_1 := \bar b := \gamma_{0,\bar\Sigma}$. Combining Lemma 8 (with an additional use of the Cauchy-Schwarz inequality) and Theorem 17, we get

$$G(b) - G(\bar b) \leqslant \frac{2}{\zeta}\int_0^1 \int \|\nabla G(b)\|_2^2 \,\mathrm{d}\tilde b_s \,\mathrm{d}s. \tag{22}$$

Define a random variable $X_s \sim \tilde b_s$ and observe that

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

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

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

Together with (22), it yields

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

■