Averaging on the Bures-Wasserstein manifold: dimension-free convergence of gradient descent

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study first-order optimization algorithms for computing the barycenter of Gaussian distributions with respect to the optimal transport metric. Although the objective is geodesically non-convex, Riemannian gradient descent empirically converges rapidly, in fact faster than off-the-shelf methods such as Euclidean gradient descent and SDP solvers. This stands in stark contrast to the best-known theoretical results, which depend exponentially on the dimension. In this work, we prove new geodesic convexity results which provide stronger control of the iterates, yielding a dimension-free convergence rate. Our techniques also enable the analysis of two related notions of averaging, the entropically-regularized barycenter and the geometric median, providing the first convergence guarantees for these problems.

1 Introduction

Averaging multiple data sources is among the most classical and fundamental subroutines in data science. However, a key modern challenge is that data is often more complicated than points in $\mathbb{R}^d$. In this paper, we study the task of averaging probability distributions on $\mathbb{R}^d$, a setting that commonly arises in machine learning and statistics [CD14; Ho+17; SLD18; Dog+19], computer vision and graphics [Rab+11; Sol+15], probability theory [KS94; RU02], and signal processing [Elv+20]; see also the surveys [PC+19; PZ19] and the references within.

The Wasserstein barycenter [AC11] has emerged as a particularly canonical notion of average. Formally, let $\mathcal{P}_2(\mathbb{R}^d)$ denote the space of probability measures on $\mathbb{R}^d$ with finite second moment, let P be a probability measure over $\mathcal{P}_2(\mathbb{R}^d)$, and let W_2 denote the 2-Wasserstein distance (i.e. the standard optimal transport distance). Then, the *Wasserstein barycenter* of P is a solution of

$$\underset{b \in \mathcal{P}_2(\mathbb{R}^d)}{\text{minimize}} \qquad \int W_2^2(b, \cdot)\, \mathrm{d}P\,. \tag{1}$$

A related notion of average is the *entropically-regularized Wasserstein barycenter* of P [Kro18; BCP19; CEK20], which is defined to be a solution of

$$\underset{b \in \mathcal{P}_2(\mathbb{R}^d)}{\text{minimize}} \qquad \int W_2^2(b, \cdot)\, \mathrm{d}P + \mathrm{ent}(b)\,, \tag{2}$$

where ent is an entropic penalty which allows for incorporating prior knowledge into the average. Lastly, a third related notion of average with better robustness properties (e.g., with a breakdown point of 50% [FVJ09]) is the *Wasserstein geometric median* of P, which is defined to be a solution of

$$\underset{b \in \mathcal{P}_2(\mathbb{R}^d)}{\text{minimize}} \qquad \int W_2(b, \cdot)\, \mathrm{d}P\,. \tag{3}$$

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

Importantly, while these three notions of average can be defined using other metrics in lieu of W_2, the Wasserstein distance is critical for many applications since it enables capturing geometric features of the distributions [CD14].

The many applications of Wasserstein barycenters and geometric medians (see e.g., [CE10; Rab+11; CD14; GPC15; RP15; Sol+15; BPC16; SLD18; GLR20]) have inspired significant research into their mathematical and statistical properties since their introduction roughly a decade ago [AC11; Rab+11]. For instance, on the mathematical side it is known that under mild conditions, the barycenter and geometric median exist, are unique, and admit dual formulations related to multimarginal optimal transport problems [CE10; AC11; COO15]. On the statistical side, [PZ16; AC17; LL17; Big+18; KSS19; Le +19; ALP20] provide finite-sample and asymptotic statistical guarantees for estimating the Wasserstein barycenter from samples.

However, computing these objects is challenging because of two fundamental obstacles. The first is that in general, barycenters and geometric medians can be complicated distributions which are much harder to (approximately) represent than the input distributions. The second is that generically, these problems are computationally hard in high dimensions. For instance, Wasserstein barycenters and geometric medians of discrete distributions are NP-hard to compute (even approximately) in high dimension [AB21b].

Algorithms for averaging on the Bures-Wasserstein manifold. Nevertheless, these computational obstacles can be potentially averted in parametric settings. This paper as well as most of the literature [Álv+16; Bac+18; ZP19; Che+20] on parametric settings focuses on the commonly arising setting where P is supported on Gaussian distributions.[1] As noted in [Álv+16], the Gaussian case also encompasses general location-scatter families.

There are two natural families of approaches for designing averaging algorithms in this setting. Both exploit the fact that modulo a simple re-centering of all distributions, the relevant space of probability distributions is isometric to the *Bures-Wasserstein manifold*, i.e. the cone of positive semidefinite matrices equipped with the Bures-Wasserstein metric (background is given in Section 2).

The first approach is simply to recognize the (regularized) Wasserstein barycenter problem as a convex optimization problem over the space of positive semidefinite matrices and apply off-the-shelf methods such as Euclidean gradient descent or semidefinite programming solvers. However, these methods have received little prior attention for good reason: they perform poorly due to severe scalability and parameter-tuning issues (see Section 3.3 for numerics). Briefly, the underlying issue is that these algorithms operate in the standard Euclidean geometry rather than the natural geometry of optimal transport. Moreover, this approach does not apply to the Wasserstein geometric median problem because even in one dimension, it is non-convex in the Euclidean geometry.

A much more effective approach in practice (see Section 3.3 for numerics) is to exploit the geometry of the Bures-Wasserstein manifold via geodesic optimization. Prior work has extensively pursued this direction, investigating the effectiveness of (stochastic) Riemannian gradient descent for computing Wasserstein barycenters, see e.g., [Álv+16; Bac+18; ZP19; Che+20].

Challenges for geodesic optimization over the Bures-Wasserstein manifold. Although geodesic optimization is natural for this problem, it comes with several important obstacles: the non-negative curvature of the Bures-Wasserstein manifold necessitates new tools for analysis, and moreover both the barycenter and geometric median problems are *non-convex* in the Bures-Wasserstein geometry. (These two issues are in fact intimately related, see Appendix A.4.) This prevents applying standard results in the geodesic optimization literature (see e.g., [ZS16; Bou20]) since in general it is only possible to prove local convergence guarantees for non-convex problems.

For the Wasserstein barycenter problem, it is possible to interpret Riemannian gradient descent (with step size one) as a fixed-point iteration, and through this lens establish asymptotic convergence [Álv+16; Bac+18; ZP19]. Obtaining non-asymptotic rates of convergence is more challenging because it requires developing quantitative proxies for the standard convexity inequalities needed to analyze gradient descent. The first such result was achieved by [Che+20]; showing that Riemannian gradient descent converges to the Wasserstein barycenter at a linear rate. Yet their convergence rate

[1] The Wasserstein barycenter of Gaussians was, in fact, first studied in the 1990s [OR93; KS94].

depends exponentially on the dimension d, and also their work does not extend to the Wasserstein geometric median or regularized Wasserstein barycenter.

1.1 Contributions

In this paper, we analyze first-order optimization algorithms on the Bures-Wasserstein manifold. We summarize our main results here and overview our techniques in the next section.

From exponential dimension dependence to dimension-free rates. In Section 3, we show that for the Wasserstein barycenter problem, Riemannian gradient descent enjoys dimension-free convergence rates (Theorem 2). This eliminates the exponential dimension dependence of state-of-the-art convergence rates [Che+20], which aligns with the empirical performance of this algorithm (see Figure 3). It also stands in sharp contrast to the setting of discrete distributions in which there are computational complexity barriers to achieving even polynomial dimension dependence [AB21b].

Moreover, in Theorem 3, we give a refinement of this result which replaces the worst-case assumption of uniform bounds on the matrices' eigenvalues with significantly weaker average-case assumptions.

Beyond barycenters. In Sections 4 and 5, we show how our analysis techniques also enable proving fast convergence of Riemannian gradient descent for computing regularized barycenters (Theorem 4) and geometric medians (Theorem 5). To the best of our knowledge, these are the first guarantees for notions of averaging on the Bures-Wasserstein manifold beyond the barycenter.

1.2 Techniques

Here we briefly sketch the specific technical challenges we face and how we address them to analyze Riemannian gradient descent for each of the three notions of Bures-Wasserstein average: barycenter, regularized barycenter, and median. Although each analysis necessarily exploits significant particularities of its own objective, their common structure may be of interest for studying other geodesically non-convex optimization problems.

Overcoming non-convexity. As we discuss in Appendix A.4, there is a close connection between the second-order behavior of our objective functionals and the non-negative curvature of the Bures-Wasserstein manifold. In particular, while non-negative curvature readily yields smoothness properties for the three functionals, the very same curvature presents fundamental obstructions to establishing convexity properties: all three functionals are geodesically non-convex. To circumvent this issue, we instead search for quantitative gradient domination conditions, also known as Polyak-Łojasiewicz inequalities [KNS16]. For each functional we employ specialized techniques, ranging from general optimization principles to arguments based on the theory of optimal transport. We ultimately show that these inequalities hold with constants depending on the conditioning of the iterates.

Overcoming ill-conditioned iterates. So long as smoothness and gradient domination inequalities hold at the current iterate, we can guarantee that the next iterate of gradient descent makes progress. However, the amount of progress degrades if the iterates of the algorithm are poorly-conditioned. The second major hurdle that must then be overcome is to control the regularity of the iterates. Here, the primary technical tool is shared across the analyses and guarantees that if the objective function is a sum of simpler functions, each of whose gradients point towards well-conditioned matrices, then the iterates of gradient descent remain well-conditioned too. Formally, this is captured by the following geodesic convexity result, which may be of independent interest. Below, $\mathbb{S}^d_{++}$ denotes the set of $d \times d$ positive definite matrices. See Appendix A.2 for a review of the relevant geometric concepts, and Appendix B for the proof, discussion of tightness, and complementary results.

Theorem 1. *The functionals* $-\sqrt{\lambda_{\min}}, \sqrt{\lambda_{\max}} : \mathbb{S}^d_{++} \to \mathbb{R}$ *are convex along generalized geodesics.*

Using this theorem in conjunction with careful analysis of the objective functions, we establish global convergence guarantees for first-order geodesic optimization.

1.3 Other related work

Averaging on curved spaces. Averages such as barycenters and medians on curved spaces have become popular due to the applications in machine learning, computer vision, analysis, radar signal processing [ABY13], and brain-computer interfaces [YBL16; CBB17]. While their mathematical properties such as existence and uniqueness are fairly well-understood [Afs11], their computation is an active area of research [Wei37; VZ00; Stu03; Yan10; BI13; Bac14; OP15]. For the Wasserstein barycenter problem in particular, there have been a multitude of approaches proposed for both the discrete setting (see e.g., [CD14; Ben+15; Bor17; Kro+19; Dvi20; Lin+20; AB21a; Gum+21]) and the continuous setting (see e.g., [CAD20; Li+20; Kor+21]).

Optimal transport and regularization. Our work on the entropically-regularized barycenter follows a long and fruitful interplay between optimal transport and entropic regularization. Entropic regularization yields computational speedups [Wil69; Cut13; ANR17; PC+19], brings fundamental connections to statistical problems such as the Schrödinger bridge [Sch31; Léo14] and maximum likelihood estimation [RW18], and enjoys much more regularity from a PDE perspective [Léo12; CEK20], which in turn has been useful for rigorously establishing functional inequalities [Led18; FGP20; Gen+20]. Entropic regularization of optimal transport specifically between Gaussians has also been extensively studied in the literature [BL20; Jan+20; MGM20].

2 Preliminaries

We write $\mathbb{S}^d$ for the space of symmetric $d \times d$ matrices, and $\mathbb{S}^d_{++}$ for the open subset of $\mathbb{S}^d$ consisting of positive definite matrices.

Given probability measures μ and ν on $\mathbb{R}^d$ with finite second moment, the 2-Wasserstein distance between μ and ν is defined as

$$W_2^2(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \int \|x - y\|^2 \, \mathrm{d}\pi(x, y) \,, \tag{4}$$

where $\Pi(\mu, \nu)$ denotes the set of couplings of μ and ν, i.e., the set of probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ whose marginals are respectively μ and ν. If μ and ν admit densities with respect to the Lebesgue measure on $\mathbb{R}^d$, then the infimum is attained, and the optimal coupling is supported on the graph of a map, i.e., there exists a map $T : \mathbb{R}^d \to \mathbb{R}^d$ such that for π-a.e. $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$, it holds that $y = T(x)$. The map T is called the *optimal transport map* from μ to ν.

We refer readers to [Vil03; San15] for an introduction to optimal transport, and to [Car92] and [Che+20, Appendix A.1] for background on Riemannian geometry. The Riemannian structure of optimal transport was introduced in the seminal work [Ott01]; detailed treatments can be found in [AGS08; Vil09], see also [Che+20, Appendix A.2] and Appendix A for a quick overview.

In this paper, we mainly work with centered Gaussians, which can be identified with their covariance matrices. (Extensions to the non-centered case are also discussed in the next sections.) We abuse notation via this identification: given $\Sigma, \Sigma' \in \mathbb{S}^d_{++}$, we write $W_2(\Sigma, \Sigma')$ for the 2-Wasserstein distance between centered Gaussians with covariance matrices Σ, Σ' respectively. Throughout, all Gaussians of interest are non-degenerate; that is, their covariances are non-singular.

The Wasserstein distance has a closed-form expression for Gaussians:

$$W_2^2(\Sigma, \Sigma') = \operatorname{tr}\left[\Sigma + \Sigma' - 2\left(\Sigma^{1/2}\Sigma'\Sigma^{1/2}\right)^{1/2}\right]. \tag{5}$$

Also, the optimal transport map from Σ to Σ' is the symmetric matrix

$$T_{\Sigma \to \Sigma'} = \Sigma^{-1/2}\left(\Sigma^{1/2}\Sigma'\Sigma^{1/2}\right)^{1/2}\Sigma^{-1/2} = \mathrm{GM}(\Sigma^{-1}, \Sigma'). \tag{6}$$

Above, $\mathrm{GM}(A, B) := A^{1/2}\left(A^{-1/2}BA^{-1/2}\right)^{1/2}A^{1/2}$ denotes the matrix geometric mean between two positive semidefinite matrices [Bha07, Ch. 4]. The Wasserstein distance on $\mathbb{S}^d_{++}$ in fact arises from a Riemannian metric, which was first introduced by Bures in [Bur69]. Hence, the Riemannian manifold $\mathbb{S}^d_{++}$ endowed with this Wasserstein distance is referred to as the *Bures-Wasserstein space*. The geometry of this space is studied in detail in [Mod17; BJL19]. We provide additional background on the Bures-Wasserstein manifold in Appendix A.

3 Barycenters

In this section, we consider the Bures-Wasserstein barycenter

$$\Sigma^\star \in \operatorname*{arg\,min}_{\Sigma \in \mathbb{S}^d_{++}} \int W_2^2(\Sigma, \cdot)\, \mathrm{d}P\,.$$

We refer to the introduction for a discussion of the past work on the Bures-Wasserstein barycenter. We also remark that the case when P is supported on possibly non-centered Gaussians is easily reduced to the centered case; see the discussion in [Che+20, §4].

3.1 Algorithms

We consider both Riemannian gradient descent (GD) and Riemannian stochastic gradient descent (SGD) algorithms for computing the Bures-Wasserstein barycenter, which are given as Algorithm 1 and Algorithm 2 respectively. GD is useful for high-precision solutions due to its linear convergence (Theorem 2), and SGD is useful for large-scale or online settings because of its cheaper updates. We refer to [ZP19; Che+20] for the derivation of the updates. Here, Σ_0 is the initialization, which can be taken to be any matrix in the support of P. For SGD, we also require a sequence $(\eta_t)_{t=1}^T$ of step sizes and a sequence $(K_t)_{t=1}^T$ of i.i.d. samples from P.

Algorithm 1 GD for Barycenters

procedure BARY-GD(Σ_0, P, T)
  for $t = 1, \dots, T$ **do**
    $S_t \leftarrow \int \mathrm{GM}(\Sigma_{t-1}^{-1}, \Sigma)\, \mathrm{d}P(\Sigma)$
    $\Sigma_t \leftarrow S_t \Sigma_{t-1} S_t$
  end for
  return Σ_T
end procedure

Algorithm 2 SGD for Barycenters

procedure BARY-SGD($\Sigma_0, (\eta_t)_{t=1}^T, (K_t)_{t=1}^T$)
  for $t = 1, \dots, T$ **do**
    $\hat{S}_t \leftarrow (1 - \eta_t) I_d + \eta_t\, \mathrm{GM}(\Sigma_{t-1}^{-1}, K_t)$
    $\Sigma_t \leftarrow \hat{S}_t \Sigma_{t-1} \hat{S}_t$
  end for
  return Σ_T
end procedure

3.2 Convergence guarantees

Denote the barycenter functional by $F(\Sigma) := \frac{1}{2} \int W_2^2(\Sigma, \cdot)\, \mathrm{d}P$, and denote the *variance* of P by $\operatorname{var} P := 2F(\Sigma^\star)$. We assume that P is supported on matrices whose eigenvalues lie in the range $[\lambda_{\min}, \lambda_{\max}]$, and we let $\kappa := \lambda_{\max}/\lambda_{\min}$ denote the condition number. Whereas the previous state-of-the-art convergence analysis for Algorithms 1 and 2 in [Che+20] suffered a dependence of κ^d, we show that the rates of convergence are in fact independent of the dimension d.

Theorem 2 (guarantees for Bures-Wasserstein GD/SGD)**.** *Assume that P is supported on covariance matrices whose eigenvalues lie in the range $[\lambda_{\min}, \lambda_{\max}]$, $0 < \lambda_{\min} \le \lambda_{\max} < \infty$. Let $\kappa := \lambda_{\max}/\lambda_{\min}$ denote the condition number. Assume that we initialize at $\Sigma_0 \in \operatorname{supp} P$.*

1. *(GD) Let Σ_n^{GD} denote the n-th iterate of GD (Algorithm 1). Then,*

$$\frac{1}{2\sqrt{\kappa}}\, W_2^2(\Sigma_n^{\mathrm{GD}}, \Sigma^\star) \le F(\Sigma_n^{\mathrm{GD}}) - F(\Sigma^\star) \le \exp\Big(-\frac{n}{4\kappa^{3/2}}\Big) \left\{F(\Sigma_0) - F(\Sigma^\star)\right\}.$$

2. *(SGD) Let Σ_n^{SGD} denote the n-th iterate of SGD (Algorithm 2). Then, with appropriate step sizes,*

$$\frac{1}{2\sqrt{\kappa}}\, \mathbb{E}\, W_2^2(\Sigma_n^{\mathrm{SGD}}, \Sigma^\star) \le \mathbb{E}\, F(\Sigma_n^{\mathrm{SGD}}) - F(\Sigma^\star) \le \frac{48\kappa^3 \operatorname{var} P}{n}\,.$$

In fact, using our new geodesic convexity results we can relax the conditioning assumption from requiring all matrices be *uniformly* well-conditioned, to being *individually* well-conditioned. This is a significant improvement when the eigenvalue ranges differ significantly between matrices.

Theorem 3. *Let $\kappa^\star := \sup_{\Sigma \in \operatorname{supp}(P)} \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$. The conclusions of Theorem 2 hold when replacing κ with $\kappa^\star$ everywhere.*

Actually, we deduce this from an even stronger statement: in Theorem 2, κ can be replaced everywhere by an average-case notion of conditioning, namely $\overline{\kappa} := \overline{\lambda_{\max}}/\overline{\lambda_{\min}}$ where $\overline{\lambda_{\min}}^{1/2} := \int \lambda_{\min}(\Sigma)^{1/2}\,\mathrm{d}P(\Sigma)$ and $\overline{\lambda_{\max}}^{1/2} := \int \lambda_{\max}(\Sigma)^{1/2}\,\mathrm{d}P(\Sigma)$.

We give the proofs of these results in Appendix C.1.

3.3 Numerical experiments

There are two natural competitors of Riemannian GD when minimizing the barycenter functional: (i) solving an SDP (Appendix C.3), and (ii) Euclidean GD (see Appendix C.2 for a description and analysis of the Euclidean gradient descent algorithm).

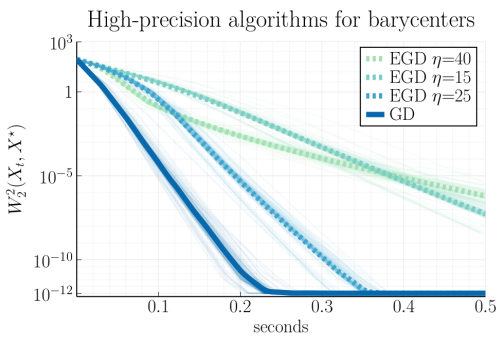

Figure 1: Riemannian vs. Euclidean GD.

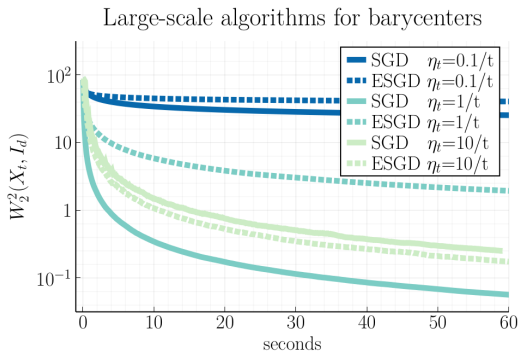

Figure 2: Riemannian vs. Euclidean stochastic GD.

In Figure 1 we compare Riemannian and Euclidean GD on a random dataset consisting of $n = 50$ covariance matrices of dimension $d = 50$, each with condition number $\kappa = 1000$. The eigenspaces of the matrices are independent Haar distributed, and their eigenvalues are equally spaced in the interval $[\lambda_{\min}, \lambda_{\max}] = [0.03, 30]$. Qualitatively similar results were observed for other input distributions; see Appendix F. We run 50 experiments and plot the average accuracy cut off at 10^{-12}; $X^\star$ denotes the best iterate. We omit SDP solvers from the plot because their runtime is orders of magnitude worse: using the Splitting Cone Solver (SCS) [ODo+16; ODo+19], the problem takes $\sim$15 seconds to solve, and MOSEK [MOS21] is even slower. We observed that Euclidean GD's rate of convergence is very sensitive to its step size, which depends heavily on the conditioning of the problem. Riemannian GD was the clear winner in our experiments, as its step size requires no tuning and it always performed no worse (in fact, often significantly better) than Euclidean GD.

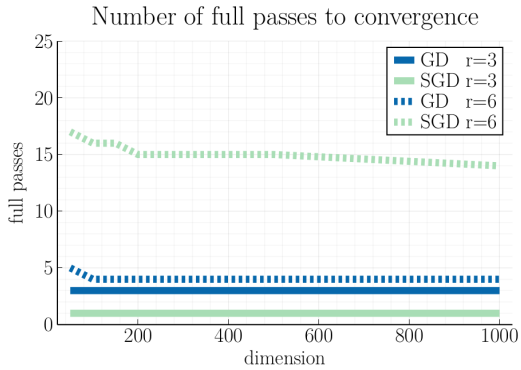

Figure 3: Passes before accuracy $W_2^2(X_t, X^\star) \leq 10^{-r}\,\mathrm{var}\,P$, for $r \in \{3, 6\}$. This is dimension independent for both Riemannian GD and SGD.

In Figure 2 we compare Riemannian and Euclidean SGD. We average 300×300 covariance matrices drawn from a distribution whose barycenter is known to be the identity, see Appendix F for details. We observe that Riemannian SGD typically outperforms Euclidean SGD, sometimes substantially.

In Figure 3 we illustrate the dimension independence of the two Riemannian algorithms, a main result of this paper. We use a dataset of $n = 50$ matrices, just as in Figure 1. To compare the algorithms on equal footing, the y-axis measures "full passes" over the 50 matrices: one pass constitutes one iteration of GD, or n iterations of SGD. Observe also the tradeoff between GD and SGD: SGD converges rapidly to low-precision solutions, but takes longer to converge to high-precision solutions.

4 Entropically-regularized barycenters

In this section, we consider the entropically-regularized barycenter $b^\star_{\mathrm{reg}}$ which minimizes

$$F_\gamma(b) := \frac{1}{2}\int W_2^2(b,\cdot)\,\mathrm{d}P + \gamma\,\mathrm{KL}\big(b \,\|\, \mathcal{N}(0,I_d)\big)\,,$$

where $\mathrm{KL}(\cdot\,\|\,\cdot)$ denotes the Kullback-Leibler (KL) divergence, and $\gamma > 0$ is a given regularization parameter. It suffices to consider the case when all of the measures are centered, see Remark 6. To justify considering this problem on the Bures-Wasserstein space, we provide the following proposition, proven in Appendix D.3.

Proposition 1. *Suppose P is supported on centered Gaussians whose covariance matrices have eigenvalues lying in the range $[1/\sqrt{\kappa}, \sqrt{\kappa}]$, for some $\kappa \geq 1$. Then, there exists a unique minimizer $b^\star_{\mathrm{reg}}$ of F_γ over $\mathcal{P}_2(\mathbb{R}^d)$, and this minimizer is a centered Gaussian distribution whose covariance matrix $\Sigma^\star$ also has eigenvalues in the range $[1/\sqrt{\kappa}, \sqrt{\kappa}]$.*

See the introduction for a discussion of the literature on this problem. Prior work focuses on a slightly different entropic penalty, the differential entropy $\int b \ln b$. Note that such a penalty encourages b to be diffuse over all of $\mathbb{R}^d$ (the minimizer blows up as $\gamma \to \infty$). Here, we focus on a KL divergence penalty which has the advantage of interpolating between two well-studied problems: the Wasserstein barycenter problem ($\gamma = 0$) and minimization of the KL divergence ($\gamma = \infty$). We take the standard Gaussian distribution as a canonical choice of reference distribution.

4.1 Algorithm

The Riemannian gradient descent algorithm for minimizing F_γ is given as Algorithm 3. See Appendix D for a derivation of the update rule.

Algorithm 3 GD for Regularized Barycenters

procedure RBARY-GD($\Sigma_0, P, T, \gamma, \eta$)
  for $t = 1, \ldots, T$ **do**
    $S_t \leftarrow \eta \int \mathrm{GM}(\Sigma_{t-1}^{-1}, \Sigma)\,\mathrm{d}P(\Sigma) + \eta\gamma\Sigma_{t-1}^{-1} + (1 - \eta\,(1+\gamma))I_d$
    $\Sigma_t \leftarrow S_t \Sigma_{t-1} S_t$
  end for
  return Σ_T
end procedure

4.2 Convergence guarantees

We provide two convergence guarantees for Algorithm 3. The first holds for all choices of the regularization parameter γ. However, this rate deteriorates with larger γ, and intuitively the optimization problem should become somewhat easier with larger regularization; hence, we prove a second rate of convergence to capture this behavior. The proof of each rate appears in Appendix D.

Theorem 4. *Fix $\gamma > 0$ and suppose that there is a $\kappa \geq 1$ such that P is supported on covariance matrices with eigenvalues in $[1/\sqrt{\kappa}, \sqrt{\kappa}]$. If Algorithm 3 is initialized at a point in $\operatorname{supp} P$ and run with step size $\eta = 1/(1 + 2\gamma\sqrt{\kappa})$, then the following hold.*

1. *For any choice of regularization parameter $\gamma > 0$ and any $t \geq 1$,*

$$F_\gamma(\Sigma_t) - F_\gamma(\Sigma^\star) \leq \exp\Big(-\frac{t}{\kappa^4\,(1 + 2\gamma\sqrt{\kappa})}\Big)\,\{F_\gamma(\Sigma_0) - F_\gamma(\Sigma^\star)\}\,.$$

2. *If $\gamma \geq 14\kappa^4$ is sufficiently large, then the following improved rate holds: for $t \geq 1$,*

$$F_\gamma(\Sigma_t) - F_\gamma(\Sigma^\star) \leq \exp\Big(-\frac{t}{6\sqrt{\kappa}}\Big)\,\{F_\gamma(\Sigma_0) - F_\gamma(\Sigma^\star)\}\,.$$

Similarly to Theorem 2, we can also provide guarantees in terms of the distance $W_2(\Sigma_t, \Sigma^\star)$ to the minimizer, but we omit this for brevity.

4.3 Numerical experiments

In Figure 4, we investigate the use of the regularization term $\gamma\,\mathrm{KL}(\cdot \,\|\, \mathcal{N}(0, I_d))$ to encode a prior belief of isotropy. In Figure 4, we generate $n = 100$ i.i.d. 20×20 covariance matrices from a distribution whose barycenter is the identity (see Appendix F). Then, for $\rho \in [0, 10]$ we compute the barycenter of a perturbed dataset obtained by adding $\rho e_1 e_1^\mathsf{T}$ to each matrix for different choices of γ. We see that for $\gamma = 0$ the barycenter quickly departs from isotropy, while for larger γ the regularization yields averages which are more consistent with our prior belief.

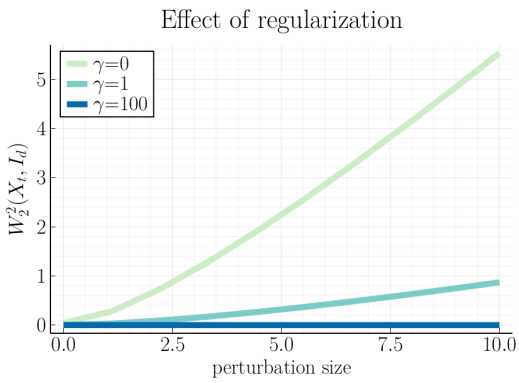

Figure 4: Effect of regularization for varying γ.

5 Geometric medians

In this section, we consider the Wasserstein geometric median

$$b^\star_{\mathrm{median}} \in \operatorname*{arg\,min}_{b \in \mathcal{P}_2(\mathbb{R}^d)} \int W_2(b, \cdot)\,\mathrm{d}P\,. \tag{7}$$

See the introduction for a discussion of the literature on this problem. Observe that, in contrast to the barycenter (1), here we are minimizing the average *unsquared* Wasserstein distance.

The following basic result justifies the consideration of the geometric median problem on the Bures-Wasserstein space.

Proposition 2. *Suppose that P is supported on centered non-degenerate Gaussians whose covariance matrices have eigenvalues lying in the range $[\lambda_{\min}, \lambda_{\max}]$, where $0 \le \lambda_{\min} \le \lambda_{\max} < \infty$. Then, there exists a solution to* (7) *which is also a centered non-degenerate Gaussian distribution; moreover, its covariance matrix $\Sigma^\star_{\mathrm{median}}$ can be taken to have eigenvalues in $[\lambda_{\min}, \lambda_{\max}]$.*

Proof. See Appendix E.1. □

Remark 1. Suppose now that P is supported on non-degenerate Gaussian distributions which are not necessarily centered. Then, the proof of Proposition 2 applies with minor modifications to show that the minimizer of the median functional is still attained at a Gaussian distribution. However, unlike the barycenter and entropically regularized barycenter, it is not the case that the mean of the Wasserstein geometric median is the Euclidean geometric median of the means, which means that it is not as straightforward to reduce to the centered case for this problem. Nevertheless, in Appendix E.2, we describe a reduction which allows the algorithm described in the next section to be applied in a black box manner to the non-centered case, with corresponding convergence guarantees.

5.1 Algorithm

Since the Wasserstein distance $W_2(\Sigma, \cdot)$ is neither geodesically convex nor geodesically smooth, nor Euclidean convex nor Euclidean smooth (see Remark 7), it poses challenges for optimization. We therefore smooth the objective before optimization. Given a desired target accuracy $\varepsilon > 0$, let

$$W_{2,\varepsilon} := \sqrt{W_2^2 + \varepsilon^2}\,, \qquad F_\varepsilon(b) := \int W_{2,\varepsilon}(b, \cdot)\,\mathrm{d}P\,.$$

The smoothed Riemannian gradient descent algorithm is given as Algorithm 4. See Appendix E for a derivation of the update rule.

Algorithm 4 Smoothed GD for Median

procedure MEDIAN-GD($\Sigma_0, P, T, \varepsilon$)
  for $t = 1, \ldots, T$ **do**
    $S_t \leftarrow I_d + \varepsilon \int \{\mathrm{GM}(\Sigma_{t-1}^{-1}, \Sigma) - I_d\} \, W_{2,\varepsilon}(\Sigma_{t-1}, \Sigma)^{-1} \, \mathrm{d}P(\Sigma)$
    $\Sigma_t \leftarrow S_t \Sigma_{t-1} S_t$
  end for
  return Σ_T
end procedure

5.2 Convergence guarantees

We now show that Algorithm 4 can find an $\mathcal{O}(\varepsilon)$-approximate minimizer for the geometric median functional in $\mathcal{O}(\kappa/\varepsilon^4)$ iterations. Let $F := F_0$ denote the unregularized functional. Note that since F typically does not have a unique minimizer, we only guarantee a small suboptimality gap.

Theorem 5. *Assume that P is supported on covariance matrices whose eigenvalues lie in $[\lambda_{\min}, \lambda_{\max}]$, $0 < \lambda_{\min} \leq \lambda_{\max} < \infty$. Let $\kappa := \lambda_{\max}/\lambda_{\min}$ denote the condition number, and let $0 < \varepsilon < 1$ denote a target accuracy. Assume that we initialize Algorithm 4 at $\Sigma_0 \in \operatorname{supp} P$. Then, Algorithm 4 outputs Σ_T satisfying $F(\Sigma_T) - F(\Sigma^\star_{\mathrm{median}}) \leq 3\varepsilon$ if*

$$T \geq \frac{8\kappa \, F_\varepsilon(\Sigma_0)^4}{\varepsilon^4} \, .$$

Proof. See Appendix E.1. □

5.3 Numerical experiments

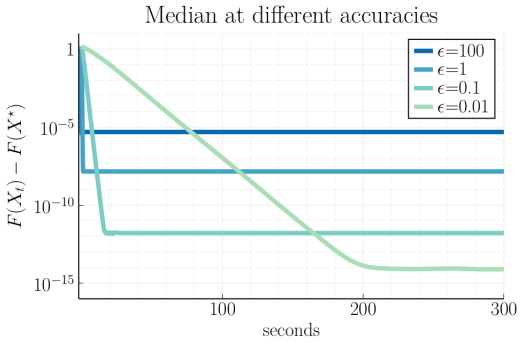

Figure 5: Evolution of median objective for varying ε. $X^\star$ denotes the best iterate.

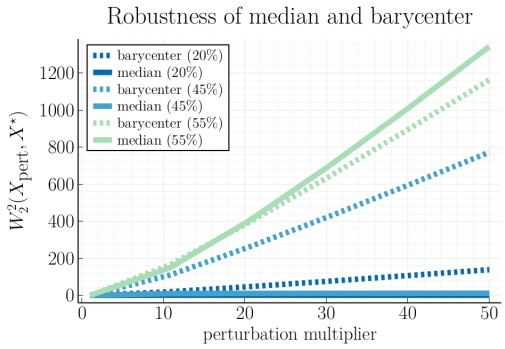

Figure 6: Robustness of the Wasserstein median.

In Figure 5 we plot the suboptimality gap for the unregularized objective $F = \int W_2(\cdot, \Sigma) \, \mathrm{d}P$ as we optimize F_ε using Algorithm 4 for varying choices of ε. There is a natural trade-off for choosing the regularization parameter ε: for smaller ε, the approximation to the true geometric median is better, but convergence is slower. The covariance matrices are generated as in Figure 1, with $n = d = 30$ and $[\lambda_{\min}, \lambda_{\max}] = [0.01, 10]$.

In Figure 6 we illustrate the robustness of the Wasserstein geometric median up to its breakdown point of 50% [FVJ09]. We take random input matrices as above, with $n = d = 20$ and $[\lambda_{\min}, \lambda_{\max}] = [1, 10]$, and compute their barycenter and approximate median ($\varepsilon = 1$). We then perturb a fraction (20%, 45%, and 55% for our figure) of the matrices by multiplying them by a constant greater than 1. The x-axis of the plot shows the size of the perturbation while the y-axis gives the distance of the original barycenter and median to the barycenter and median of this new, perturbed dataset.

We also implemented Euclidean GD for this geometric median problem; plots are omitted for brevity since the results are similar to those for the barycenter (c.f. Section 3.3). We note that Euclidean GD depends heavily on parameter tuning and does not come with convergence guarantees since the problem is non-convex in the Euclidean geometry.

References

[AB21a] J. M. Altschuler and E. Boix-Adsera. "Wasserstein barycenters can be computed in polynomial time in fixed dimension". In: *Journal of Machine Learning Research* 22.44 (2021), pp. 1–19.

[AB21b] J. M. Altschuler and E. Boix-Adsera. "Wasserstein barycenters are NP-hard to compute". In: *arXiv e-prints*, arXiv:2101.01100 (2021).

[ABY13] M. Arnaudon, F. Barbaresco, and L. Yang. "Riemannian medians and means with applications to radar signal processing". In: *IEEE Journal of Selected Topics in Signal Processing* 7.4 (2013), pp. 595–604.

[AC11] M. Agueh and G. Carlier. "Barycenters in the Wasserstein space". In: *SIAM J. Math. Anal.* 43.2 (2011), pp. 904–924.

[AC17] M. Agueh and G. Carlier. "Vers un théorème de la limite centrale dans l'espace de Wasserstein?" In: *C. R. Math. Acad. Sci. Paris* 355.7 (2017), pp. 812–818.

[Afs11] B. Afsari. "Riemannian L^p center of mass: existence, uniqueness, and convexity". In: *Proceedings of the American Mathematical Society* 139.2 (2011), pp. 655–673.

[AGS08] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Second. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 2008, pp. x+334.

[ALP20] A. Ahidar-Coutrix, T. Le Gouic, and Q. Paris. "Convergence rates for empirical barycenters in metric spaces: curvature, convexity and extendable geodesics". In: *Probab. Theory Related Fields* 177.1-2 (2020), pp. 323–368.

[Álv+16] P. C. Álvarez-Esteban et al. "A fixed-point approach to barycenters in Wasserstein space". In: *J. Math. Anal. Appl.* 441.2 (2016), pp. 744–762.

[ANR17] J. Altschuler, J. Niles-Weed, and P. Rigollet. "Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 1964–1974.

[Bac+18] J. Backhoff-Veraguas et al. "Bayesian learning with Wasserstein barycenters". In: *arXiv e-prints*, arXiv:1805.10833 (May 2018).

[Bac14] M. Bacák. "Computing medians and means in Hadamard spaces". In: *SIAM Journal on Optimization* 24.3 (2014), pp. 1542–1566.

[BCP19] J. Bigot, E. Cazelles, and N. Papadakis. "Penalization of barycenters in the Wasserstein space". In: *SIAM J. Math. Anal.* 51.3 (2019), pp. 2261–2285.

[Ben+15] J.-D. Benamou et al. "Iterative Bregman projections for regularized transportation problems". In: *SIAM Journal on Scientific Computing* 37.2 (2015), A1111–A1138.

[Bha07] R. Bhatia. *Positive definite matrices*. Princeton Series in Applied Mathematics. Princeton University Press, Princeton, NJ, 2007, pp. x+254.

[BI13] D. A. Bini and B. Iannazzo. "Computing the Karcher mean of symmetric positive definite matrices". In: *Linear Algebra and its Applications* 438.4 (2013), pp. 1700–1710.

[Big+18] J. Bigot et al. "Upper and lower risk bounds for estimating the Wasserstein barycenter of random measures on the real line". In: *Electron. J. Stat.* 12.2 (2018), pp. 2253–2289.

[BJL19] R. Bhatia, T. Jain, and Y. Lim. "On the Bures-Wasserstein distance between positive definite matrices". In: *Expo. Math.* 37.2 (2019), pp. 165–191.

[BL20] E. del Barrio and J.-M. Loubes. "The statistical effect of entropic regularization in optimal transportation". In: *arXiv e-prints*, arXiv:2006.05199 (2020).

[Bor17] S. Borgwardt. "Strongly polynomial 2-approximations of discrete Wasserstein barycenters". In: *arXiv preprint arXiv:1704.05491* (2017).

[Bou20] N. Boumal. "An introduction to optimization on smooth manifolds". In: *Available online, May* (2020).

[BPC16] N. Bonneel, G. Peyré, and M. Cuturi. "Wasserstein barycentric coordinates: histogram regression using optimal transport". In: *ACM Transactions on Graphics* 35.4 (2016).

[Bur69] D. Bures. "An extension of Kakutani's theorem on infinite product measures to the tensor product of semifinite w^*-algebras". In: *Trans. Amer. Math. Soc.* 135 (1969), pp. 199–212.

[CAD20] S. Cohen, M. Arbel, and M. P. Deisenroth. "Estimating barycenters of measures in high dimensions". In: *arXiv e-prints*, arXiv:2007.07105 (2020).

[Car92] M. P. do Carmo. *Riemannian geometry*. Mathematics: Theory & Applications. Translated from the second Portuguese edition by Francis Flaherty. Birkhäuser Boston, Inc., Boston, MA, 1992, pp. xiv+300.

[CBB17] M. Congedo, A. Barachant, and R. Bhatia. "Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review". In: *Brain-Computer Interfaces* 4.3 (2017), pp. 155–174.

[CD14] M. Cuturi and A. Doucet. "Fast computation of Wasserstein barycenters". In: *Proceedings of the 31st International Conference on Machine Learning*. Ed. by E. P. Xing and T. Jebara. Vol. 32. Proceedings of Machine Learning Research 2. Bejing, China, 2014, pp. 685–693.

[CE10] G. Carlier and I. Ekeland. "Matching for teams". In: *Economic Theory* 42.2 (2010), pp. 397–418.

[CEK20] G. Carlier, K. Eichinger, and A. Kroshnin. "Entropic-Wasserstein barycenters: PDE characterization, regularity and CLT". In: *arXiv e-prints*, arXiv:2012.10701 (2020).

[Che+20] S. Chewi et al. "Gradient descent algorithms for Bures-Wasserstein barycenters". In: *Proceedings of Thirty Third Conference on Learning Theory*. Ed. by J. Abernethy and S. Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, Sept. 2020, pp. 1276–1304.

[COO15] G. Carlier, A. Oberman, and E. Oudet. "Numerical methods for matching for teams and Wasserstein barycenters". In: *ESAIM: Mathematical Modelling and Numerical Analysis* 49.6 (2015), pp. 1621–1642.

[Cut13] M. Cuturi. "Sinkhorn distances: Lightspeed computation of optimal transport". In: *Advances in Neural Information Processing Systems* 26 (2013), pp. 2292–2300.

[Dog+19] P. Dognin et al. "Wasserstein barycenter model ensembling". In: *arXiv e-prints*, arXiv:1902.04999 (2019).

[Dvi20] D. Dvinskikh. "Stochastic approximation versus sample average approximation for population Wasserstein barycenters". In: *arXiv e-prints*, arXiv:2001.07697 (2020).

[Elv+20] F. Elvander et al. "Multi-marginal optimal transport using partial information with applications in robust localization and sensor fusion". In: *Signal Processing* 171 (2020), p. 107474.

[FGP20] M. Fathi, N. Gozlan, and M. Prod'homme. "A proof of the Caffarelli contraction theorem via entropic regularization". In: *Calc. Var. Partial Differential Equations* 59.3 (2020), Paper No. 96, 18.

[FVJ09] P. T. Fletcher, S. Venkatasubramanian, and S. Joshi. "The geometric median on Riemannian manifolds with application to robust atlas estimation". In: *NeuroImage* 45.1 (2009), S143–S152.

[Gen+20] I. Gentil et al. "An entropic interpolation proof of the HWI inequality". In: *Stochastic Process. Appl.* 130.2 (2020), pp. 907–923.

[GLR20] T. L. Gouic, J.-M. Loubes, and P. Rigollet. "Projection to fairness in statistical learning". In: *arXiv e-prints*, arXiv:2005.11720 (2020).

[GPC15] A. Gramfort, G. Peyré, and M. Cuturi. "Fast optimal transport averaging of neuroimaging data". In: *International Conference on Information Processing in Medical Imaging*. Springer. 2015, pp. 261–272.

[Gum+21] S. Guminov et al. "Accelerated alternating minimization, accelerated Sinkhorn's algorithm and accelerated iterative Bregman projections". In: *arXiv e-prints*, arXiv:1906.03622 (2021).

[Ho+17] N. Ho et al. "Multilevel clustering via Wasserstein means". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 1501–1509.

[Jan+20] H. Janati et al. "Entropic optimal transport between unbalanced Gaussian measures has a closed form". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 10468–10479.

[KNS16] H. Karimi, J. Nutini, and M. Schmidt. "Linear convergence of gradient and proximal-gradient methods under the Polyak-Lojasiewicz condition". In: *European Conference on Machine Learning and Knowledge Discovery in Databases - Volume 9851*. ECML PKDD 2016. Riva del Garda, Italy: Springer-Verlag, 2016, pp. 795–811.

[Kor+21] A. Korotin et al. "Continuous Wasserstein-2 barycenter estimation without minimax optimization". In: *arXiv e-prints*, arXiv:2102.01752 (2021).

[Kro+19] A. Kroshnin et al. "On the complexity of approximating Wasserstein barycenters". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by K. Chaudhuri and R. Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. Long Beach, California, USA: PMLR, Sept. 2019, pp. 3530–3540.

[Kro18] A. Kroshnin. "Fréchet barycenters in the Monge-Kantorovich spaces". In: *J. Convex Anal.* 25.4 (2018), pp. 1371–1395.

[KS94] M. Knott and C. S. Smith. "On a generalization of cyclic monotonicity and distances among random vectors". In: *Linear Algebra Appl.* 199 (1994), pp. 363–371.

[KSS19] A. Kroshnin, V. Spokoiny, and A. Suvorikova. "Statistical inference for Bures-Wasserstein barycenters". In: *arXiv e-prints*, arXiv:1901.00226 (Jan. 2019).

[Le +19] T. Le Gouic et al. "Fast convergence of empirical barycenters in Alexandrov spaces and the Wasserstein space". In: *arXiv e-prints*, arXiv:1908.00828 (Aug. 2019).

[Led18] M. Ledoux. *Remarks on some transportation cost inequalities*. 2018.

[Léo12] C. Léonard. "From the Schrödinger problem to the Monge-Kantorovich problem". In: *J. Funct. Anal.* 262.4 (2012), pp. 1879–1920.

[Léo14] C. Léonard. "A survey of the Schrödinger problem and some of its connections with optimal transport". In: *Discrete Contin. Dyn. Syst.* 34.4 (2014), pp. 1533–1574.

[Li+20] L. Li et al. "Continuous regularized Wasserstein barycenters". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 17755–17765.

[Lin+20] T. Lin et al. "Fixed-support Wasserstein barycenters: computational hardness and fast algorithm". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 5368–5380.

[LL17] T. Le Gouic and J.-M. Loubes. "Existence and consistency of Wasserstein barycenters". In: *Probab. Theory Related Fields* 168.3-4 (2017), pp. 901–917.

[MGM20] A. Mallasto, A. Gerolin, and H. Q. Minh. "Entropy-regularized 2-Wasserstein distance between Gaussian measures". In: *arXiv e-prints*, arXiv:2006.03416 (2020).

[Mod17] K. Modin. "Geometry of matrix decompositions seen through optimal transport and information geometry". In: *J. Geom. Mech.* 9.3 (2017), pp. 335–390.

[MOS21] MOSEK. *MOSEK Optimizer API for C*. 2021. URL: `https://docs.mosek.com/9.2/capi.pdf`.

[ODo+16] B. O'Donoghue et al. "Conic optimization via operator splitting and homogeneous self-dual embedding". In: *Journal of Optimization Theory and Applications* 169.3 (June 2016), pp. 1042–1068.

[ODo+19] B. O'Donoghue et al. *SCS: Splitting Conic Solver, version 2.1.3*. `https://github.com/cvxgrp/scs`. Nov. 2019.

[OP15] S.-i. Ohta and M. Pálfia. "Discrete-time gradient flows and law of large numbers in Alexandrov spaces". In: *Calculus of Variations and Partial Differential Equations* 54.2 (2015), pp. 1591–1610.

[OR93] I. Olkin and S. T. Rachev. "Maximum submatrix traces for positive definite matrices". In: *SIAM journal on matrix analysis and applications* 14.2 (1993), pp. 390–397.

[Ott01] F. Otto. "The geometry of dissipative evolution equations: the porous medium equation". In: *Comm. Partial Differential Equations* 26.1-2 (2001), pp. 101–174.

[PC+19] G. Peyré, M. Cuturi, et al. "Computational optimal transport: With applications to data science". In: *Foundations and Trends® in Machine Learning* 11.5-6 (2019), pp. 355–607.

[PZ16] V. M. Panaretos and Y. Zemel. "Amplitude and phase variation of point processes". In: *Ann. Statist.* 44.2 (2016), pp. 771–812.

[PZ19] V. M. Panaretos and Y. Zemel. "Statistical aspects of Wasserstein distances". In: *Annu. Rev. Stat. Appl.* 6 (2019), pp. 405–431.

[Rab+11] J. Rabin et al. "Wasserstein barycenter and its application to texture mixing". In: *International Conference on Scale Space and Variational Methods in Computer Vision*. Springer. 2011, pp. 435–446.

[RP15] 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*. Springer. 2015, pp. 256–269.

[RU02] L. Rüschendorf and L. Uckelmann. "On the n-coupling problem". In: *J. Multivariate Anal.* 81.2 (2002), pp. 242–258.

[RW18] P. Rigollet and J. Weed. "Entropic optimal transport is maximum-likelihood deconvolution". In: *C. R. Math. Acad. Sci. Paris* 356.11-12 (2018), pp. 1228–1235.

[San15] F. Santambrogio. *Optimal transport for applied mathematicians*. Vol. 87. Progress in Nonlinear Differential Equations and their Applications. Calculus of variations, PDEs, and modeling. Birkhäuser/Springer, Cham, 2015, pp. xxvii+353.

[Sch31] E. Schrödinger. *Über die umkehrung der naturgesetze*. Verlag der Akademie der Wissenschaften in Kommission bei Walter De Gruyter u ..., 1931.

[SLD18] S. Srivastava, C. Li, and D. B. Dunson. "Scalable Bayes via barycenter in Wasserstein space". In: *The Journal of Machine Learning Research* 19.1 (2018), pp. 312–346.

[Sol+15] J. Solomon et al. "Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains". In: *ACM Transactions on Graphics (TOG)* 34.4 (2015), pp. 1–11.

[Stu03] K.-T. Sturm. "Probability measures on metric spaces of nonpositive curvature". In: *Heat kernels and analysis on manifolds, graphs, and metric spaces (Paris, 2002)*. Vol. 338. Contemp. Math. Amer. Math. Soc., Providence, RI, 2003, pp. 357–390.

[Vil03] C. Villani. *Topics in optimal transportation*. Vol. 58. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2003, pp. xvi+370.

[Vil09] C. Villani. *Optimal transport*. Vol. 338. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Old and new. Springer-Verlag, Berlin, 2009, pp. xxii+973.

[VZ00] Y. Vardi and C.-H. Zhang. "The multivariate L_1-median and associated data depth". In: *Proc. Natl. Acad. Sci. USA* 97.4 (2000), pp. 1423–1426.

[Wei37] E. Weiszfeld. "Sur le point pour lequel la somme des distances de n points donnés est minimum". In: *Tohoku Mathematical Journal, First Series* 43 (1937), pp. 355–386.

[Wil69] A. G. Wilson. "The use of entropy maximising models, in the theory of trip distribution, mode split and route split". In: *Journal of Transport Economics and Policy* (1969), pp. 108–126.

[Yan10] L. Yang. "Riemannian median and its estimation". In: *LMS Journal of Computation and Mathematics* 13 (2010), pp. 461–479.

[YBL16] F. Yger, M. Berar, and F. Lotte. "Riemannian approaches in brain-computer interfaces: a review". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25.10 (2016), pp. 1753–1762.

[ZP19] Y. Zemel and V. M. Panaretos. "Fréchet means and Procrustes analysis in Wasserstein space". In: *Bernoulli* 25.2 (2019), pp. 932–976.

[ZS16] H. Zhang and S. Sra. "First-order methods for geodesically convex optimization". In: *29th Annual Conference on Learning Theory*. Ed. by V. Feldman, A. Rakhlin, and O. Shamir. Vol. 49. Proceedings of Machine Learning Research. Columbia University, New York, New York, USA: PMLR, 23–26 Jun 2016, pp. 1617–1638.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Computed on standard laptop
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]