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

Jason M. Altschuler
MIT
jasonalt@mit.edu

Sinho Chewi
MIT
schewi@mit.edu

Patrik Gerber
MIT
prgerber@mit.edu

Austin J. Stromme
MIT
astromme@mit.edu

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 GD empirically converges rapidly, in fact faster than off-the-shelf methods such as Euclidean GD and SDP solvers. This stands in stark contrast to the best-known theoretical results for Riemannian GD, which depend exponentially on the dimension. In this work, we prove new geodesic convexity results on auxiliary functionals; this provides strong control of the Riemannian GD iterates, ultimately 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 Riemannian GD for these problems.

1 Introduction

Averaging multiple data sources is among the most classical and fundamental subroutines in data science. However, a 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; Rab+11] 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

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

A related notion of average is the *entropically-regularized Wasserstein barycenter* of P , which is defined to be a solution of

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

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

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

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; LLR20]) 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]. And on the statistical side, [PZ16; AC17; LL17; Big+18; FLF19; ALP20; Le +21; KSS21] 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 represent (even approximately) 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 GD or semidefinite programming solvers. However, these methods have received little prior attention for good reason: they suffer from 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 GD 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 GD (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 GD. The first such result was achieved by [Che+20], showing that Riemannian GD converges to the Wasserstein barycenter at a linear rate. Yet their convergence rate depends exponentially on the

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

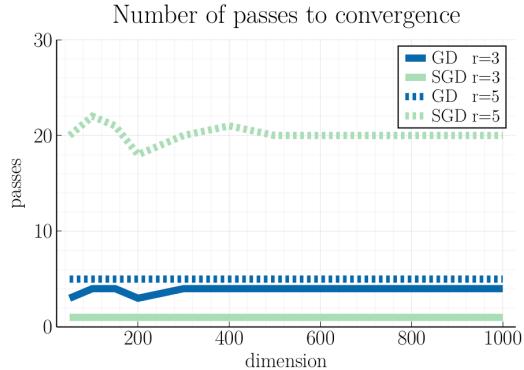


Figure 1: Passes until convergence error 10^{-r} to the barycenter, for $r \in \{3, 5\}$. This is *dimension independent* for Riemannian GD and SGD—consistent with our main results. Details in Section 3.

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 GD enjoys dimension-free convergence rates (Theorem 2). We make several comments to contextualize this result. First, our result 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 1). Second, our result stands in sharp contrast to the setting of discrete distributions in which there are computational complexity barriers to achieving even polynomial dimension dependence [AB21b]. Third, our result closes the gap between computation and statistical estimation for the Bures-Wasserstein barycenter, since dimension-free sample complexity bounds were recently proven in [FLF19].

Moreover, in Theorem 3, we further refine this result by replacing the worst-case assumption of uniform bounds on the matrices’ eigenvalues with a significantly weaker average-case assumption.

Beyond barycenters. In Sections 4 and 5, we show how our analysis techniques also enable proving fast convergence of Riemannian GD for computing regularized barycenters (Theorem 4) and geometric medians (Theorem 5). To the best of our knowledge, these are the first guarantees for Riemannian GD 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 GD for the three notions of Bures-Wasserstein average: barycenter, regularized barycenter, and geometric median. Although each analysis necessarily exploits particularities of its own objective, the common structure of our overarching analysis framework 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 these objective functionals and the non-negative curvature of the Bures-Wasserstein manifold. In particular, while non-negative curvature is used to prove smoothness properties for the three functionals, it also leads to them all being geodesically non-convex. To circumvent this issue, we establish gradient domination conditions known as Polyak-Łojasiewicz inequalities [Pol64; Loj63], which intuitively are quantitative proxies for strong convexity in non-convex settings (see e.g., [KNS16; Bol+17]). Proving such inequalities requires synthesizing general optimization principles with specialized arguments based on optimal transport theory. We ultimately

show that these inequalities hold with constants depending on the conditioning of the iterates, i.e., the ratio between the maximum and minimum eigenvalues of the corresponding covariance matrices.

Overcoming ill-conditioned iterates. So long as smoothness and gradient domination inequalities hold at the current iterate, standard optimization results guarantee that the next iterate of GD makes progress. However, the amount of progress degrades if the iterates are poorly conditioned, since then our PL inequality degrades. Thus the second major obstacle is to control the regularity of the iterates. Here, the primary technical tool is shared across the analyses. Informally, it states that if the objective is a sum of functions, each of whose gradients point towards well-conditioned matrices, then the GD iterates remain well-conditioned. 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 see Appendix B for the proof, discussion of tightness, and complementary results.

Theorem 1. *The functions $-\sqrt{\lambda_{\min}}, \sqrt{\lambda_{\max}} : \mathbb{S}_{++}^d \rightarrow \mathbb{R}$ 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

Averages such as barycenters and medians on general curved spaces have become popular due to far-ranging applications in domains such as 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 many approaches. These approaches vary significantly depending on if the setting is discrete or continuous. In the discrete setting, the problem is NP-hard in high dimension [AB21b]. In low dimension (or more precisely fixed dimension), reasonable approximations can be obtained using fixed-support approximations which reduce the problem to a large linear program [CD14; Ben+15; COO15; Kro+19; Lin+19; Lin+20; Dvi21; Gum+21; Haa+21], and it was recently shown that high-precision (or even exact) solutions can be computed in polynomial time using computational geometry techniques [AB21a].

In the continuous setting, the problem is in general intractable since the optimal barycenter is intractable even to represent, let alone to compute. Nevertheless, in certain settings it has been empirically effective to parameterize and solve using neural networks [CAD20; FTC21; Kor+21], stochastic gradient descent [Li+20], or Riemannian optimization [Álv+16; Bac+18; ZP19; Che+20].

However, for continuous settings, the theory currently lags far behind the empirics. This is true even in the seemingly simple setting of Gaussians, in which case the barycenter has a concise representation since it is also Gaussian. Riemannian GD for this problem was first proposed and demonstrated to be empirically effective in [Álv+16], where it was introduced as a fixed-point algorithm. Asymptotic convergence was proved in [Álv+16], and then extended to non-population and stochastic settings in [Bac+18; ZP19]. Non-asymptotic convergence rates were first shown in [Che+20], but there is a large gap between these theoretical rates and what is observed in practice. It particular, previous rates depend exponentially on the dimension. The present paper improves this to dimension-free.

For the other two problems we study, Bures-Wasserstein geometric medians and entropically-regularized barycenters, no convergence guarantees were previously known for the natural Riemannian GD algorithm.

Our work is in the midst of a flurry of exciting recent developments about entropically regularized barycenters. Several recent works have, simultaneously with each other, extensively studied these objects in the particular setting of Gaussians, leading to the establishment of fundamental results such as the fact that the regularized barycenter of Gaussians is itself Gaussian [BL20; Jan+20; MGM21]. Another recent and related line of work has established fundamental mathematical and statistical results for entropically regularized barycenters in the setting of general distributions, although with slightly different penalties than the KL divergence studied here, typically the differential entropy $\int b \ln b$ [Kro18; BCP19; CEK21] and sometimes even broader classes of regularizations [BCP19].

1.4 Organization

Section 2 briefly recalls relevant preliminaries. We analyze Riemannian GD for computing Bures-Wasserstein barycenters, regularized barycenters, and geometric medians in Sections 3, 4, and 5, respectively. We conclude in Section 6. We provide proofs as well as additional background and numerical results in the Appendix. Code reproducing all experiments in this paper can be found at <https://github.com/PatrikGerber/Bures-Barycenters>.

2 Preliminaries

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

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

where $\Pi(\mu, \nu)$ denotes the set of couplings of μ and ν , i.e., the 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 $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$ 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] for background on Riemannian geometry. The Riemannian structure of optimal transport was introduced in the seminal work [Ott01]—detailed treatments are in [AGS08; Vil09]; for completeness we also provide a quick overview in Appendix A.

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 below.) 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. Here, \mathbb{S}^d denotes the space of symmetric $d \times d$ matrices, and \mathbb{S}_{++}^d denotes the open subset of \mathbb{S}^d consisting of positive definite matrices. Throughout, all Gaussians are non-degenerate; that is, their covariances are non-singular.

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

$$W_2^2(\Sigma, \Sigma') = \text{tr}[\Sigma + \Sigma' - 2(\Sigma^{1/2}\Sigma'\Sigma^{1/2})^{1/2}]. \quad (5)$$

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

$$T_{\Sigma \rightarrow \Sigma'} = \Sigma^{-1/2} (\Sigma^{1/2}\Sigma'\Sigma^{1/2})^{1/2} \Sigma^{-1/2} = \text{GM}(\Sigma^{-1}, \Sigma'). \quad (6)$$

Above, $\text{GM}(A, B) := A^{1/2} (A^{-1/2}BA^{-1/2})^{1/2} A^{1/2}$ 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]. For completeness, we provide additional background on the Bures-Wasserstein manifold in Appendix A.

3 Barycenters

In this section, we consider the Bures-Wasserstein barycenter

$$\Sigma^* \in \arg \min_{\Sigma \in \mathbb{S}_{++}^d} \int W_2^2(\Sigma, \cdot) dP.$$

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 computing 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 . Note that whereas SGD requires choosing step sizes, GD simply uses step size 1, as justified in [ZP19].

Algorithm 1 GD for Barycenters

```

1: procedure BARY-GD( $\Sigma_0, P, T$ )
2:   for  $t = 1, \dots, T$  do
3:      $S_t \leftarrow \int \text{GM}(\Sigma_{t-1}^{-1}, \Sigma) dP(\Sigma)$ 
4:      $\Sigma_t \leftarrow S_t \Sigma_{t-1} S_t$ 
5:   return  $\Sigma_T$ 
```

Algorithm 2 SGD for Barycenters

```

1: procedure BARY-SGD( $\Sigma_0, (\eta_t)_{t=1}^T, (K_t)_{t=1}^T$ )
2:   for  $t = 1, \dots, T$  do
3:      $\hat{S}_t \leftarrow (1 - \eta_t) I_d + \eta_t \text{GM}(\Sigma_{t-1}^{-1}, K_t)$ 
4:      $\Sigma_t \leftarrow \hat{S}_t \Sigma_{t-1} \hat{S}_t$ 
5:   return  $\Sigma_T$ 
```

3.2 Convergence guarantees

Denote the barycenter functional by $F(\Sigma) := \frac{1}{2} \int W_2^2(\Sigma, \cdot) dP$, and denote the *variance* of P by $\text{var } P := 2F(\Sigma^*)$. 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. *Assume that P is supported on covariance matrices whose eigenvalues lie in the range $[\lambda_{\min}, \lambda_{\max}]$, $0 < \lambda_{\min} \leq \lambda_{\max} < \infty$. Let $\kappa := \lambda_{\max}/\lambda_{\min}$ denote the condition number. Assume that we initialize at $\Sigma_0 \in \text{supp } P$.*

1. (GD) Let Σ_T^{GD} denote the T -th iterate of GD (Algorithm 1). Then,

$$\frac{1}{2\sqrt{\kappa}} W_2^2(\Sigma_T^{\text{GD}}, \Sigma^*) \leq F(\Sigma_T^{\text{GD}}) - F(\Sigma^*) \leq \exp\left(-\frac{T}{4\kappa^{3/2}}\right) \{F(\Sigma_0) - F(\Sigma^*)\}.$$

2. (SGD) Let Σ_T^{SGD} denote the T -th iterate of SGD (Algorithm 2), where $(K_t)_{t=1}^T$ are i.i.d. from P . Then, with an appropriate choice of step sizes²,

$$\frac{1}{2\sqrt{\kappa}} \mathbb{E} W_2^2(\Sigma_T^{\text{SGD}}, \Sigma^*) \leq \mathbb{E} F(\Sigma_T^{\text{SGD}}) - F(\Sigma^*) \leq \frac{48\kappa^3 \text{var } P}{T}.$$

In fact, using our new geodesic convexity results we can also 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 $\kappa^* := \sup_{\Sigma \in \text{supp}(P)} \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$. The conclusions of Theorem 2 hold when replacing κ with κ^* 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 $\bar{\kappa} := \overline{\lambda_{\max}/\lambda_{\min}}$ where $\overline{\lambda_{\min}}^{1/2} := \int \lambda_{\min}(\Sigma)^{1/2} dP(\Sigma)$ and $\overline{\lambda_{\max}}^{1/2} := \int \lambda_{\max}(\Sigma)^{1/2} dP(\Sigma)$.

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

²Namely, $\eta_t = \frac{1}{4\kappa^{3/2}} \left(1 - \sqrt{1 - \frac{16\kappa^3 (2(t+t_0)+1)}{(t+t_0+1)^2}}\right)$ suffices, where $t_0 = 32\kappa^3 - 1$.

3.3 Numerical experiments

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

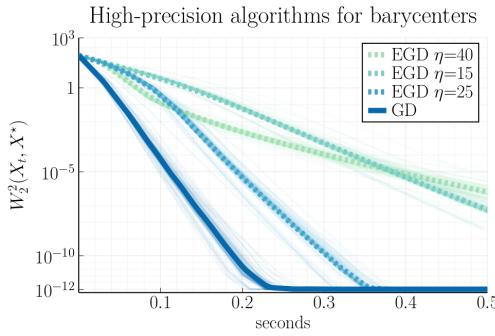


Figure 2: Riemannian vs. Euclidean GD.

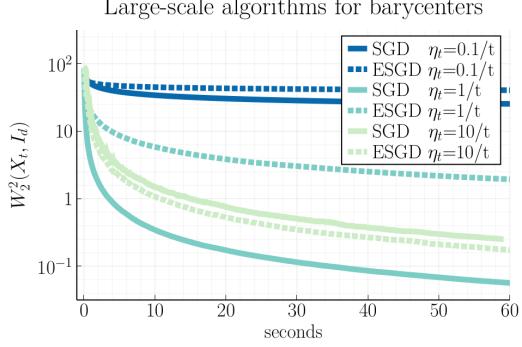


Figure 3: Riemannian vs. Euclidean SGD.

In Figure 2 we compare Riemannian and Euclidean GD on a dataset consisting of $n = 50$ covariance matrices of dimension $d = 50$, each with condition number $\kappa = 1000$. Their eigenspaces are independent Haar distributed, and their eigenvalues are equally spaced in the interval $[\lambda_{\min}, \lambda_{\max}] = [0.03, 30]$. Qualitatively similar results are observed for other inputs; see Appendix F. We run 50 experiments and plot the average accuracy cut off at 10^{-12} ; X^* denotes the best iterate. We omit SDP solvers from the plot because their runtime is orders of magnitude slower: using the Splitting Cone Solver (SCS) [ODo+16; ODo+19], the problem takes ~ 15 seconds to solve, and MOSEK [MOS21] is even slower. For completeness, we also compare GD to the Riemannian Frank-Wolfe algorithm [WS17] in Appendix F, and conclude that GD is superior. We observe that Euclidean GD’s convergence rate is quite 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.

In Figure 3 we observe that Riemannian SGD typically outperforms Euclidean SGD, sometimes substantially. We average 300×300 covariance matrices drawn from a distribution whose barycenter is known to be the identity, see Appendix F for details. As Figure 3 shows, in practice it can be helpful to tune the step sizes beyond the guidance given by our worst-case theoretical guarantees. In our experiments, Riemannian SGD was competitive on a wide range of problems with $\eta = 1/t$.

We comment on Figure 1, which illustrates the dimension independence of the two Riemannian algorithms, a main result of this paper. It plots the number of passes until convergence $W_2^2(X_t, X^*) \leq 10^{-r} \text{ var } P$ to the barycenter X^* , for $r \in \{3, 5\}$. To compare algorithms on equal footing, the y -axis measures “passes” over the $n = 50$ matrices: a pass constitutes one GD iteration, or n SGD iterations. The input is generated as in Figure 2. 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_{reg}^* which minimizes

$$F_\gamma(b) := \frac{1}{2} \int W_2^2(b, \cdot) dP + \gamma \text{KL}(b \parallel \mathcal{N}(0, I_d)),$$

where KL denotes the Kullback-Leibler divergence, and $\gamma > 0$ is a given regularization parameter. It suffices to consider the case when all measures are centered, see Remark 6. The following proposition justifies considering this problem on the Bures-Wasserstein space; proof 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_{reg}^* of F_γ over $\mathcal{P}_2(\mathbb{R}^d)$, and this minimizer is a centered Gaussian distribution whose covariance matrix Σ^* also has eigenvalues in the range $[1/\sqrt{\kappa}, \sqrt{\kappa}]$.*

As described in the introduction, prior work on the Wasserstein barycenter typically focuses on a slightly different entropic penalty, the differential entropy $\int b \ln b$. Note that the differential entropy penalty encourages b to be diffuse over all of \mathbb{R}^d (the minimizer blows up as $\gamma \rightarrow \infty$). 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 as a canonical choice of reference distribution, and note that our method of analysis can be extended to other reference measures at the cost of significant additional technical complexity. We thus choose to exclusively focus on the standard Gaussian case.

4.1 Algorithm

Algorithm 3 is Riemannian GD for minimizing F_γ . A derivation of the update rule is in Appendix D .

Algorithm 3 GD for Regularized Barycenters

```

1: procedure RBARY-GD( $\Sigma_0, P, T, \gamma, \eta$ )
2:   for  $t = 1, \dots, T$  do
3:      $S_t \leftarrow \eta \int GM(\Sigma_{t-1}^{-1}, \Sigma) dP(\Sigma) + \eta\gamma\Sigma_{t-1}^{-1} + (1 - \eta(1 + \gamma))I_d$ 
4:      $\Sigma_t \leftarrow S_t \Sigma_{t-1} S_t$ 
5:   return  $\Sigma_T$ 

```

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. We emphasize that as in §3, our convergence rates are dimension-independent. The proof of each rate appears in Appendix D.

Theorem 4. *Fix $\gamma > 0$ and suppose that P is supported on covariance matrices with eigenvalues in $[1/\sqrt{\kappa}, \sqrt{\kappa}]$. If Algorithm 3 is initialized at a point in $\text{supp } P$ 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 $T \geq 1$,*

$$F_\gamma(\Sigma_T) - F_\gamma(\Sigma^*) \leq \exp\left(-\frac{T}{\kappa^4(1 + 2\gamma\sqrt{\kappa})}\right) \{F_\gamma(\Sigma_0) - F_\gamma(\Sigma^*)\}.$$

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

$$F_\gamma(\Sigma_T) - F_\gamma(\Sigma^*) \leq \exp\left(-\frac{T}{6\sqrt{\kappa}}\right) \{F_\gamma(\Sigma_0) - F_\gamma(\Sigma^*)\}.$$

For brevity, we omit guarantees in terms of the distance $W_2(\Sigma_t, \Sigma^*)$ to the minimizer.

4.3 Numerical experiments

In Figure 4, we investigate the use of the regularization term $\gamma \text{KL}(\cdot \parallel \mathcal{N}(0, I_d))$ 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 $\rho e_1 e_1^\top$ 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.

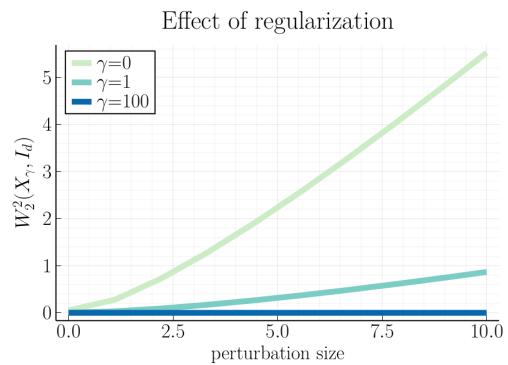


Figure 4: Effect of regularization for varying γ .

5 Geometric medians

In this section, we consider the Wasserstein geometric median

$$b_{\text{median}}^* \in \arg \min_{b \in \mathcal{P}_2(\mathbb{R}^d)} \int W_2(b, \cdot) dP. \quad (7)$$

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. It is proved in Appendix E.1.

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 $0 \leq \lambda_{\min} \leq \lambda_{\max} < \infty$. Then, there exists a solution to (7) which is also a centered non-degenerate Gaussian distribution; moreover, its covariance matrix Σ_{median}^* can be taken to have eigenvalues in $[\lambda_{\min}, \lambda_{\max}]$.*

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, thus 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

$$W_{2,\varepsilon} := \sqrt{W_2^2 + \varepsilon^2}, \quad F_\varepsilon(b) := \int W_{2,\varepsilon}(b, \cdot) dP.$$

Algorithm 4 provides pseudocode for running Riemannian GD on the smoothed functional F_ε . See Appendix E for a derivation of the update rule.

Algorithm 4 Smoothed GD for Median

```

1: procedure MEDIAN-GD( $\Sigma_0, P, T, \varepsilon$ )
2:   for  $t = 1, \dots, T$  do
3:      $S_t \leftarrow I_d + \varepsilon \int \{\text{GM}(\Sigma_{t-1}^{-1}, \Sigma) - I_d\} W_{2,\varepsilon}(\Sigma_{t-1}, \Sigma)^{-1} dP(\Sigma)$ 
4:      $\Sigma_t \leftarrow S_t \Sigma_{t-1} S_t$ 
5:   return  $\Sigma_T$ 

```

5.2 Convergence guarantees

We show that Algorithm 4 finds an $\mathcal{O}(\varepsilon)$ -approximate minimizer for the geometric median functional in $\mathcal{O}(\kappa/\varepsilon^4)$ iterations. We emphasize that as in our other results, this convergence is dimension-independent. Below, 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. The proof is in Appendix E.1.

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 $\Sigma_0 \in \text{supp } P$. Then, Algorithm 4 outputs Σ_T satisfying $F(\Sigma_T) - F(\Sigma_{\text{median}}^*) \leq 3\varepsilon$ if*

$$T \geq \frac{32\kappa F_\varepsilon(\Sigma_0)^4}{\varepsilon^4}.$$

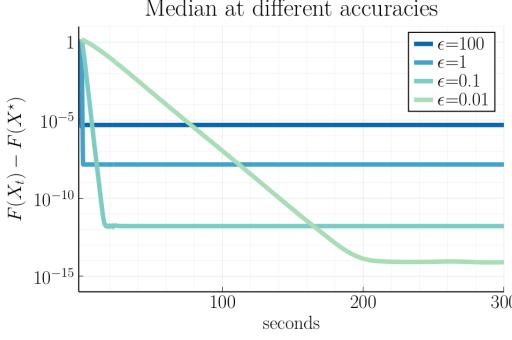


Figure 5: Evolution of median objective for varying ε . X^* denotes the best iterate.

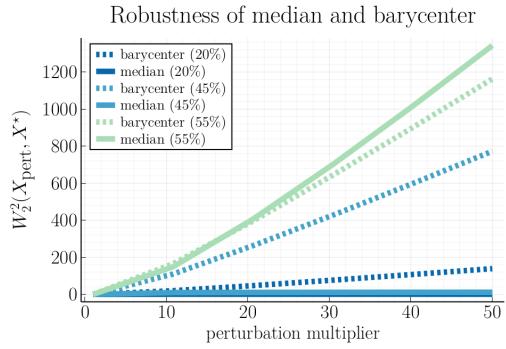


Figure 6: Robustness of the Wasserstein median.

5.3 Numerical experiments

In Figure 5 we plot the suboptimality gap for the unregularized objective $F = \int W_2(\cdot, \Sigma) dP$ as we optimize F_ε using Algorithm 4 for various ε . The regularization parameter ε has a natural trade-off: smaller ε results in better approximation to the (unregularized) geometric median, but slower convergence. The covariance matrices are generated as in Figure 2, with $n = d = 30$ and $[\lambda_{\min}, \lambda_{\max}] = [0.01, 10]$. The promising empirics in Figure 5 suggest that our algorithm performs even better in practice than our worst-case theoretical results guarantee: few iterations may suffice for convergence, and also moderate regularization may suffice for high-precision approximations.

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) in that Euclidean GD depends much more heavily on parameter tuning. Note also that Euclidean GD does not come with convergence guarantees for this problem since it is non-convex in the Euclidean geometry.

6 Discussion

In this paper we revisited the problem of computing Bures-Wasserstein barycenters and explained the empirical efficacy of Riemannian (S)GD by proving convergence rates that improve from exponential dimension dependence to dimension-free. An attractive feature of our analysis framework was that our tools were sufficiently general to prove similar dimension-free guarantees for related problems of interest, namely Bures-Wasserstein geometric medians and entropically-regularized barycenters.

Our results suggest several interesting directions for future research. The focus of this paper was dimension-dependence, and while we also improved the dependence on other parameters along the way, it is unclear if these other dependencies are optimal. Can the dependence on κ be improved via stronger PL inequalities? Is the dependence on ε improvable via alternate methods of smoothing in the case of geometric medians, or via fixed-point acceleration schemes such as Anderson acceleration in the case of barycenters?

More broadly, conventional wisdom from the now-established field of geodesic optimization tells us that whenever possible, one should recast a non-convex optimization problem as a convex one by changing the geometry. However, as demonstrated empirically in Section 3, for computing Bures-Wasserstein barycenters, it is significantly better to run GD in the non-convex geometry of optimal transport than in the convex geometry of Euclidean space. A full understanding of why and when non-convex geometry can be helpful in general optimization problems is an intriguing direction with potentially significant implications for both the theory and practice of non-convex optimization.

Acknowledgments and Disclosure of Funding

We are grateful to Victor-Emmanuel Brunel, Tyler Maunu, and Philippe Rigollet for stimulating conversations, to Pablo Parrilo for pointing out that Bures-Wasserstein barycenters have an SDP formulation (Appendix C.3), and to the anonymous reviewers for their thoughtful comments.

JA was supported by NSF Graduate Research Fellowship 1122374 and a TwoSigma PhD fellowship. SC and AS were supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program.

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: *SIAM Journal on Mathematics of Data Science, to appear* (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. Aguech and G. Carlier. “Barycenters in the Wasserstein space”. In: *SIAM Journal on Mathematical Analysis* 43.2 (2011), pp. 904–924.
- [AC17] M. Aguech and G. Carlier. “Vers un théorème de la limite centrale dans l'espace de Wasserstein?” In: *Comptes Rendus Mathématique. Académie des Sciences. Paris* 355.7 (2017), pp. 812–818.
- [AC20] K. Ahn and S. Chewi. “Efficient constrained sampling via the mirror-Langevin algorithm”. In: *arXiv e-prints*, arXiv:2010.16212 (Nov. 2020).
- [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: *Probability Theory and 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: *Journal of Mathematical Analysis and Applications* 441.2 (2016), pp. 744–762.
- [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.
- [BBI01] D. Burago, Y. Burago, and S. Ivanov. *A course in metric geometry*. Vol. 33. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2001, pp. xiv+415.
- [BCP19] J. Bigot, E. Cazelles, and N. Papadakis. “Penalization of barycenters in the Wasserstein space”. In: *SIAM Journal on Mathematical Analysis* 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.
- [BG12] R. Bhatia and P. Grover. “Norm inequalities related to the matrix geometric mean”. In: *Linear Algebra and its Applications* 437.2 (2012), pp. 726–733.
- [Bha07] R. Bhatia. *Positive definite matrices*. Princeton Series in Applied Mathematics. Princeton University Press, Princeton, NJ, 2007, pp. x+254.
- [Bha97] R. Bhatia. *Matrix analysis*. Vol. 169. Graduate Texts in Mathematics. Springer-Verlag, New York, 1997, pp. xii+347.

- [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: *Electronic Journal of Statistics* 12.2 (2018), pp. 2253–2289.
- [BJL19] R. Bhatia, T. Jain, and Y. Lim. “On the Bures-Wasserstein distance between positive definite matrices”. In: *Expositiones Mathematicae* 37.2 (2019), pp. 165–191.
- [BL06] J. M. Borwein and A. S. Lewis. *Convex analysis and nonlinear optimization*. Second. Vol. 3. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Theory and examples. Springer, New York, 2006, pp. xii+310.
- [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).
- [Bol+17] J. Bolte et al. “From error bounds to the complexity of first-order descent methods for convex functions”. In: *Mathematical Programming* 165.2 (2017), pp. 471–507.
- [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).
- [Bub15] S. Bubeck. “Convex optimization: algorithms and complexity”. In: *Foundations and Trends in Machine Learning* 8.3-4 (2015), pp. 231–357.
- [Bur69] D. Bures. “An extension of Kakutani’s theorem on infinite product measures to the tensor product of semifinite w^* -algebras”. In: *Transactions of the American Mathematical Society* 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: *International Conference on Machine Learning*. Vol. 32. 2. 2014, pp. 685–693.
- [CE10] G. Carlier and I. Ekeland. “Matching for teams”. In: *Economic Theory* 42.2 (2010), pp. 397–418.
- [CEK21] G. Carlier, K. Eichinger, and A. Kroshnin. “Entropic-Wasserstein barycenters: PDE characterization, regularity, and CLT”. In: *SIAM Journal on Mathematical Analysis* 53.5 (2021), pp. 5880–5914.
- [Che+20] S. Chewi et al. “Gradient descent algorithms for Bures-Wasserstein barycenters”. In: *Conference on Learning Theory*. Vol. 125. 2020, pp. 1276–1304.
- [CMT96] J. A. Cuesta-Albertos, C. Matrán-Bea, and A. Tuero-Díaz. “On lower bounds for the L^2 -Wasserstein metric in a Hilbert space”. In: *Journal of Theoretical Probability* 9.2 (1996), pp. 263–283.
- [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.
- [CT06] T. M. Cover and J. A. Thomas. *Elements of information theory*. Second. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2006, pp. xxiv+748.
- [CV21] J. A. Carrillo and U. Vaes. “Wasserstein stability estimates for covariance-preconditioned Fokker-Planck equations”. In: *Nonlinearity* 34.4 (Feb. 2021), pp. 2275–2295.
- [Dog+19] P. L. Dognin et al. “Wasserstein barycenter model ensembling”. In: *International Conference on Learning Representation*. 2019.
- [DS88] N. Dunford and J. T. Schwartz. *Linear operators, part 1: general theory*. Vol. 10. John Wiley & Sons, 1988.
- [Dvi21] D. Dvinskikh. “Stochastic approximation versus sample average approximation for Wasserstein barycenters”. In: *Optimization Methods and Software* 0.0 (2021), pp. 1–33.

- [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.
- [FLF19] R. Flamary, K. Lounici, and A. Ferrari. “Concentration bounds for linear Monge mapping estimation and optimal transport domain adaptation”. In: *arXiv preprint arXiv:1905.10155* (2019).
- [FTC21] J. Fan, A. Taghvaei, and Y. Chen. “Scalable computations of Wasserstein barycenter via input convex neural networks”. In: *International Conference on Machine Learning*. Vol. 139. 2021, pp. 1571–1581.
- [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.
- [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. “On a combination of alternating minimization and Nesterov’s momentum”. In: *International Conference on Machine Learning*. Vol. 139. 2021, pp. 3886–3898.
- [Haa+21] I. Haasler et al. “Multimarginal optimal transport with a tree-structured cost and the Schrödinger bridge problem”. In: *SIAM Journal on Control and Optimization* 59.4 (2021), pp. 2428–2453.
- [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: *International Conference on Learning Representations*. 2021.
- [Kro+19] A. Kroshnin et al. “On the complexity of approximating Wasserstein barycenters”. In: *International Conference on Machine Learning*. Vol. 97. 2019, pp. 3530–3540.
- [Kro18] A. Kroshnin. “Fréchet barycenters in the Monge-Kantorovich spaces”. In: *Journal of Convex Analysis* 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 and its Applications* 199 (1994), pp. 363–371.
- [KSS21] A. Kroshnin, V. Spokoiny, and A. Suvorikova. “Statistical inference for Bures-Wasserstein barycenters”. In: *The Annals of Applied Probability* 31.3 (2021), pp. 1264–1298.
- [Le +21] T. Le Gouic et al. “Fast convergence of empirical barycenters in Alexandrov spaces and the Wasserstein space”. In: *Journal of the European Math Society, to appear* (2021).
- [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+19] T. Lin et al. “On the complexity of approximating multimarginal optimal transport”. In: *arXiv preprint arXiv:1910.00152* (2019).
- [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.
- [LL01] J. D. Lawson and Y. Lim. “The geometric mean, matrices, metrics, and more”. In: *American Mathematical Monthly* 108.9 (2001), pp. 797–812.
- [LL17] T. Le Gouic and J.-M. Loubes. “Existence and consistency of Wasserstein barycenters”. In: *Probability Theory and Related Fields* 168.3–4 (2017), pp. 901–917.

- [LLR20] T. Le Gouic, J.-M. Loubes, and P. Rigollet. “Projection to fairness in statistical learning”. In: *arXiv e-prints*, arXiv:2005.11720 (2020).
- [Loj63] S. Lojasiewicz. “A topological property of real analytic subsets (in French)”. In: *Coll. du CNRS, Les équations aux dérivées partielles* 117.87-89 (1963), p. 2.
- [LSB12] S. Lacoste-Julien, M. Schmidt, and F. Bach. “A simpler approach to obtaining an $O(1/t)$ convergence rate for the projected stochastic subgradient method”. In: *arXiv e-prints*, arXiv:1212.2002 (2012).
- [MGM21] A. Mallasto, A. Gerolin, and H. Q. Minh. “Entropy-regularized 2-Wasserstein distance between Gaussian measures”. In: *Information Geometry* (Aug. 2021).
- [Mod17] K. Modin. “Geometry of matrix decompositions seen through optimal transport and information geometry”. In: *Journal of Geometric Mechanics* 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. Pálfa. “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: *Communications in 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.
- [Pol64] B. T. Polyak. “Gradient methods for solving equations and inequalities (in Russian)”. In: *USSR Computational Mathematics and Mathematical Physics* 4.6 (1964), pp. 17–32.
- [PZ16] V. M. Panaretos and Y. Zemel. “Amplitude and phase variation of point processes”. In: *The Annals of Statistics* 44.2 (2016), pp. 771–812.
- [PZ19] V. M. Panaretos and Y. Zemel. “Statistical aspects of Wasserstein distances”. In: *Annual Review of Statistics and its Application* 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: *Journal of Multivariate Analysis* 81.2 (2002), pp. 242–258.
- [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.
- [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* 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: *Proceedings of the National Academy of Sciences* 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.
- [WS17] M. Weber and S. Sra. “Riemannian optimization via Frank-Wolfe methods”. In: *arXiv e-prints*, arXiv:1710.10770 (2017).
- [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: *Conference on Learning Theory*. Vol. 49. 2016, pp. 1617–1638.

Appendix

Table of Contents

A Background on the Bures-Wasserstein manifold	16
A.1 Geometry	16
A.2 Geodesic convexity and generalized geodesic convexity	18
A.3 Geodesic optimization	19
A.4 Curvature and the barycenter functional	20
A.5 Additional facts about the Wasserstein metric	20
B Proofs for the geodesic convexity results	21
B.1 Proof of Theorem 1	21
B.2 Sharpness of Theorem 1	22
B.3 Eigenvalue clipping is a Bures-Wasserstein contraction	23
C Proofs for barycenters	24
C.1 Riemannian gradient descent	24
C.2 Euclidean gradient descent approach	26
C.3 SDP formulation	29
D Proofs for entropically-regularized barycenters	29
D.1 Trapping the iterates	30
D.2 Properties of the KL divergence	30
D.3 Existence and uniqueness of the minimizer	31
D.4 PL and smoothness inequalities	31
E Proofs for geometric medians	33
E.1 Convergence guarantee for smoothed Riemannian gradient descent	33
E.2 Reduction for non-zero means	37
F Further experiments and details	38

A Background on the Bures-Wasserstein manifold

In this section, we collect relevant background about Bures-Wasserstein geometry to make the paper more self-contained.

A.1 Geometry

We begin by describing the geometry of optimal transport, and then explain how to specialize the general concepts to the Bures-Wasserstein manifold. The books [AGS08; Vil09] are definitive references for the Riemannian structure of optimal transport. We attempt to convey the main relevant ideas, and in doing so do not attempt to be fully rigorous here.

Let $\mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ denote the space of all probability measures on \mathbb{R}^d which are absolutely continuous (i.e. admit a density w.r.t. the Lebesgue measure) and which have a finite second moment. When equipped with the 2-Wasserstein distance W_2 , it becomes a metric space. In fact, more is true: $(\mathcal{P}_{2,\text{ac}}(\mathbb{R}^d), W_2)$ admits a formal Riemannian structure which we now describe. Given $\mu_0, \mu_1 \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$, let T denote the optimal transport map from μ_0 to μ_1 ; thus, $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is a map satisfying $T_\# \mu_0 = \mu_1$.

Here, $\#$ denotes the pushforward operation, i.e. if $X \sim \mu_0$, then $T(X) \sim \mu_1$. The constant-speed geodesic $(\mu_t)_{t \in [0,1]}$ joining μ_0 to μ_1 is described via

$$\mu_t = [(1-t)\text{id} + tT]_\# \mu_0, \quad t \in [0,1].$$

This geodesic has the following interpretation: draw a ‘‘particle’’ $X_0 \sim \mu_0$, and move X_0 to $T(X_0)$ with constant speed for one unit of time along the Euclidean geodesic (i.e. straight line) joining these endpoints; thus, at time t , the particle is at position $X_t = (1-t)X_0 + tT(X_0)$. Then, μ_t is simply the law of X_t .

We take the tangent vector of the geodesic $(\mu_t)_{t \in [0,1]}$ at time 0 to be the mapping $T - \text{id}$; note that in the particle view, $T(X_0) - X_0$ represents the velocity of the particle at time 0. The tangent space $T_{\mu_0} \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ to $\mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ at μ_0 is then defined to consist of all possible tangent vectors to geodesics emanating from μ_0 . Actually, in order to make $T_{\mu_0} \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ formally into a (closed subset of a) Hilbert space, the definition is modified to read ([AGS08, Theorem 8.5.1])

$$T_{\mu_0} \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d) := \overline{\{\lambda(T_{\mu_0 \rightarrow \nu} - \text{id}) : \lambda > 0, \nu \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)\}}^{L^2(\mu_0)}, \quad (8)$$

where the overline denotes the $L^2(\mu_0)$ closure; we equip this tangent space with the $L^2(\mu_0)$ norm. Thus, for instance, we have $W_2^2(\mu_0, \mu_1) = \mathbb{E}[\|X_0 - T_{\mu_0 \rightarrow \mu_1}(X_0)\|^2] = \|T_{\mu_0 \rightarrow \mu_1} - \text{id}\|_{L^2(\mu_0)}^2$, which says that the squared norm of the tangent vector of the geodesic $(\mu_t)_{t \in [0,1]}$ equals the squared Wasserstein distance. We may write $\|\cdot\|_{\mu_0}$ as a shorthand for $\|\cdot\|_{L^2(\mu_0)}$.

The Riemannian exponential map \exp_μ is the mapping $T_\mu \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d) \rightarrow \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ which maps a tangent vector v to the constant-speed geodesic emanating from μ with velocity v , evaluated at time 1.³ From our description above, we see that $\exp_\mu v = (\text{id} + v)_\# \mu$, since the tangent vector joining μ to $T_\# \mu$ is $v = T - \text{id}$ (when T is an optimal transport map). It is also convenient to define the Riemannian logarithmic map $\log_\mu : \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d) \rightarrow T_\mu \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ to be the inverse of the exponential map \exp_μ ; in our context, $\log_\mu \nu = T_{\mu \rightarrow \nu} - \text{id}$.

In Riemannian geometry, it is common to localize the argument around a measure μ , which loosely means replacing a measure ν with its image $\log_\mu \nu$ in the tangent space at μ . This is convenient because the tangent space at μ is embedded in the Hilbert space $L^2(\mu)$, and we can leverage Hilbert space arguments (e.g. computing inner products). In order to do this one must quantify the distortion introduced by the map \log_μ , which is morally related to the curvature of the manifold.

We now specialize the above concepts to the Bures-Wasserstein manifold, in which non-degenerate centered Gaussians are identified with their covariance matrices; thus, the Bures-Wasserstein manifold is the space \mathbb{S}_{++}^d of positive-definite symmetric matrices equipped with a certain Riemannian metric.

The optimal transport problem between Gaussians is discussed in many places, e.g. [BJL19]. Given two covariance matrices $\Sigma, \Sigma' \in \mathbb{S}_{++}^d$, the optimal transport map between the corresponding centered Gaussians is the linear map $\mathbb{R}^d \rightarrow \mathbb{R}^d$ given by

$$T_{\Sigma \rightarrow \Sigma'} = \Sigma^{-1/2} (\Sigma^{1/2} \Sigma' \Sigma^{1/2})^{1/2} \Sigma^{-1/2}.$$

Note that this is a symmetric matrix. Since $AX \sim \mathcal{N}(0, A\Sigma A^\top)$ for $X \sim \mathcal{N}(0, \Sigma)$, the fact that $T_{\Sigma \rightarrow \Sigma'} X \sim \mathcal{N}(0, \Sigma')$ reduces to the matrix identity $T_{\Sigma \rightarrow \Sigma'} \Sigma T_{\Sigma \rightarrow \Sigma'} = \Sigma'$, which can be verified by hand. The above formula yields

$$\begin{aligned} W_2^2(\Sigma, \Sigma') &= \mathbb{E}[\|X - T_{\Sigma \rightarrow \Sigma'} X\|^2] = \mathbb{E}[\|X\|^2 + \|T_{\Sigma \rightarrow \Sigma'} X\|^2 - 2 \langle X, T_{\Sigma \rightarrow \Sigma'} X \rangle] \\ &= \text{tr}(\Sigma + \Sigma' - 2\Sigma T_{\Sigma \rightarrow \Sigma'}). \end{aligned} \quad (9)$$

From the general description of Wasserstein geodesics, the constant-speed geodesic $(\Sigma_t)_{t \in [0,1]}$ joining Σ to Σ' is given by

$$\Sigma_t = ((1-t)I_d + tT_{\Sigma \rightarrow \Sigma'})\Sigma((1-t)I_d + tT_{\Sigma \rightarrow \Sigma'}), \quad t \in [0,1]. \quad (10)$$

³Generally, in Riemannian geometry, the exponential map is not defined on the entire tangent space but rather a subset of it; this is also the case for Wasserstein space.

The tangent space $T_\Sigma \mathbb{S}_{++}^d$ is identified with the space \mathbb{S}^d of symmetric $d \times d$ matrices. Given $S \in T_\Sigma \mathbb{S}_{++}^d$, the tangent space norm of S is given by $\|S\|_{L^2(\mathcal{N}(0, \Sigma))} = \sqrt{\mathbb{E}[\|SX\|^2]} = \sqrt{\langle S^2, \Sigma \rangle}$, which we simply denote as $\|S\|_\Sigma$. More generally, given matrices A, B , we write $\langle A, B \rangle_\Sigma := \text{tr}(A^\top \Sigma B)$. The exponential map⁴ is $\exp_\Sigma S = (I_d + S) \Sigma (I_d + S)$, so that $\exp_\Sigma(T_{\Sigma \rightarrow \Sigma'} - I_d) = \Sigma'$. The inverse of the exponential map is then $\log_\Sigma \Sigma' = T_{\Sigma \rightarrow \Sigma'} - I_d$.

The description of the Bures-Wasserstein tangent space is in accordance with the general Riemannian structure of Wasserstein space (see [AGS08]) and agrees with the convention in [Che+20]. We now elaborate on other possible conventions, in order to dispel possible confusion.

The space \mathbb{S}_{++}^d is often studied as a manifold in other contexts, and the tangent space at any point is usually identified with \mathbb{S}^d . It is crucial to realize, however, that a tangent space is not simply a vector space (or inner product space); a tangent space also has the interpretation of describing velocities of curves. In other words, for each tangent vector S , we also need to prescribe which curves have velocity S . In the usual way of describing the manifold structure of \mathbb{S}_{++}^d , this prescription is given as follows. Given a curve $(\Sigma_t)_{t \in \mathbb{R}} \subseteq \mathbb{S}_{++}^d$, if $\dot{\Sigma}_0$ denotes the ordinary time derivative of this curve at time 0, then we declare $\dot{\Sigma}_0$ to be the tangent vector of the curve at time 0. Although this prescription is natural, observe that it conflicts with our description of the tangent space structure of the Bures-Wasserstein manifold; in particular, for the curve in (10), we have described the tangent vector to this curve (at time 0) to be $T_{\Sigma \rightarrow \Sigma'} - I_d$, but the ordinary time derivative of this curve is $(T_{\Sigma \rightarrow \Sigma'} - I_d)\Sigma + \Sigma(T_{\Sigma \rightarrow \Sigma'} - I_d)$.

To summarize the discussion in the preceding paragraph: although the usual description of the tangent space of \mathbb{S}_{++}^d at Σ and our description of the tangent space are formally the same, in that they are both identified with \mathbb{S}^d , they differ in that tangent vectors from the two descriptions give rise to different curves. Note that if we were to adopt the usual description of the tangent space of \mathbb{S}_{++}^d , then we would have to define the tangent space norm $\|\cdot\|_\Sigma$ differently from above. In this paper, we adopt the convention described earlier in this section in order to preserve the connection with the general setting of optimal transport.

A.2 Geodesic convexity and generalized geodesic convexity

Once we have geodesics, we can then define convex functions. A function $f : \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d) \rightarrow \mathbb{R}$ is said to be *geodesically convex* if for all constant-speed geodesics $(\mu_t)_{t \in [0,1]}$ (i.e., curves described by (8)), it holds that

$$f(\mu_t) \leq (1-t)f(\mu_0) + t f(\mu_1), \quad t \in [0, 1]. \quad (11)$$

It turns out, however, that many natural examples of geodesically convex functions on Wasserstein space are convex in a stronger sense, in that they satisfy the inequality (11) for a larger class of curves than geodesics. A *generalized geodesic* from μ_0 to μ_1 , with basepoint $\nu \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$, is defined to be the curve $(\mu_t^\nu)_{t \in [0,1]}$ defined by

$$\mu_t^\nu := [(1-t)T_{\nu \rightarrow \mu_0} + tT_{\nu \rightarrow \mu_1}]_\# \nu, \quad t \in [0, 1].$$

A function $f : \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d) \rightarrow \mathbb{R}$ is said to be *convex along generalized geodesics* if for every generalized geodesic $(\mu_t^\nu)_{t \in [0,1]}$, it holds that

$$f(\mu_t^\nu) \leq (1-t)f(\mu_0) + t f(\mu_1), \quad t \in [0, 1].$$

Note that the geodesic $(\mu_t)_{t \in [0,1]}$ joining μ_0 to μ_1 coincides with the generalized geodesic $(\mu_t^{\mu_0})_{t \in [0,1]}$, so that convexity along generalized geodesics is indeed stronger than geodesic convexity.

Generalized geodesics were studied in [AGS08] in order to rigorously study gradient flows on Wasserstein space. The added flexibility of generalized geodesics is sometimes important for applications [AC20]; in our work, as well as in [Che+20], generalized geodesics are needed to study the iterates of Riemannian GD.

The interpretation of generalized geodesics is that we linearize $\mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ on the tangent space $T_\nu \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$. This means that we replace μ_0 with its image $\log_\nu \mu_0 = T_{\nu \rightarrow \mu_0} - \text{id}$ in the tangent

⁴Technically the exponential map is only defined if $S + I_d \succeq 0$; this is because if $S + I_d$ is not positive semidefinite, then $S + I_d$ is not an optimal transport map due to Brenier's theorem.

space, and similarly for μ_1 . Since the tangent space is a subset of a Hilbert space, geodesics in the tangent space are described by straight lines, i.e.,

$$t \mapsto (1-t)T_{\nu \rightarrow \mu_0} + tT_{\nu \rightarrow \mu_1} - \text{id}.$$

If we translate back to $\mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$, we end up with the curve

$$t \mapsto \exp_\nu((1-t)T_{\nu \rightarrow \mu_0} + tT_{\nu \rightarrow \mu_1} - \text{id}) = [(1-t)T_{\nu \rightarrow \mu_0} + tT_{\nu \rightarrow \mu_1}]_\# \nu = \mu_t^\nu.$$

Thus, the property of being convex along generalized geodesics can be reformulated as requiring that

$$f \circ \exp_\nu : T_\nu \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d) \rightarrow \mathbb{R} \quad \text{is convex for every } \nu \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d). \quad (12)$$

In Euclidean space, convexity of a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is equivalent, via Jensen's inequality, to the following statement: for every probability measure P on \mathbb{R}^d , it holds that $f(\int x dP(x)) \leq \int f(x) dP(x)$. Since the Wasserstein barycenter is the Wasserstein analogue of the mean, we can write a similar definition on Wasserstein space. Given a probability measure P on $\mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$, let b_P denote its Wasserstein barycenter. We say that $f : \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d) \rightarrow \mathbb{R}$ is *convex along barycenters* if

$$f(b_P) \leq \int f(\mu) dP(\mu), \quad \text{for all } P \in \mathcal{P}_2(\mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)).$$

Similarly, via (12), we can define $f : \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d) \rightarrow \mathbb{R}$ to be convex along generalized barycenters if

$$f \circ \exp_\nu \left(\int v dP(v) \right) \leq \int f \circ \exp_\nu(v) dP(v) \quad \text{for all } \nu \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d) \text{ and } P \in \mathcal{P}_2(T_\nu \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)). \quad (13)$$

However, since the tangent space is embedded in a Hilbert space, there is no difference between (12) and (13).

To summarize the relationship between these four concepts:

$$\begin{aligned} \text{convex along generalized barycenters} &\iff \text{convex along generalized geodesics} \\ &\implies \text{convex along barycenters} \implies \text{geodesically convex}. \end{aligned}$$

For a justification of these facts and further discussion, see [AC11].

A.3 Geodesic optimization

Given a functional $F : \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d) \rightarrow \mathbb{R}$, we can define its Wasserstein gradient formally as follows. For any constant-speed geodesic $(\mu_t)_{t \in [0,1]}$, the gradient of F at μ_0 is the element $\nabla F(\mu_0) \in T_{\mu_0} \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ satisfying

$$\partial_t|_{t=0} F(\mu_t) = \langle \nabla F(\mu_0), T_{\mu_0 \rightarrow \mu_1} - \text{id} \rangle_{\mu_0}.$$

The Riemannian GD update for F with step size η starting at μ is

$$\mu^+ := \exp_\mu(-\eta \nabla F(\mu)) = [\text{id} - \eta \nabla F(\mu)]_\# \mu.$$

Note that the step size η should be chosen small enough that $-\eta \nabla F(\mu)$ lies in the domain of the exponential map. From the general description of the tangent space of Wasserstein space, $\nabla F(\mu)$ is the gradient of a mapping $\psi : \mathbb{R}^d \rightarrow \mathbb{R}$; then, $-\eta \nabla F(\mu)$ belongs to the domain of the exponential map if $\|\cdot\|^2/2 - \eta\psi$ is convex.

We say that F is α -strongly convex if

$$F(\mu_1) \geq F(\mu_0) + \langle \nabla F(\mu_0), \log_{\mu_0} \mu_1 \rangle_{\mu_0} + \frac{\alpha}{2} W_2^2(\mu_0, \mu_1), \quad \text{for all } \mu_0, \mu_1 \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d),$$

and β -smooth if

$$F(\mu_1) \leq F(\mu_0) + \langle \nabla F(\mu_0), \log_{\mu_0} \mu_1 \rangle_{\mu_0} + \frac{\beta}{2} W_2^2(\mu_0, \mu_1), \quad \text{for all } \mu_0, \mu_1 \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d).$$

These two properties are formally equivalent to the following statements: for any constant-speed geodesic $(\mu_t)_{t \in [0,1]}$, one has

$$\partial_t^2|_{t=0} F(\mu_t) \geq \alpha W_2^2(\mu_0, \mu_1) \quad \text{or} \quad \partial_t^2|_{t=0} F(\mu_t) \leq \beta W_2^2(\mu_0, \mu_1),$$

respectively.

A.4 Curvature and the barycenter functional

One of the interesting features of the barycenter problem is that, because it is defined in terms of the squared distance function, it captures key geometric features of the underlying space; in fact, this is arguably the reason for the success of the barycenter for geometric applications. To further discuss this connection, it is insightful to abstract the situation to computing barycenters on a metric space.

Given a metric space (X, d) and a probability measure P on X , a barycenter of P is a solution of

$$\underset{b \in X}{\text{minimize}} \quad F_P(b) := \frac{1}{2} \int d^2(b, \cdot) \, dP.$$

The basic structure required on X in order to study first-order optimization methods is the presence of geodesics. This is formalized by the notion of a *geodesic space*, which is studied in metric geometry; see [BBI01]. Then, we may define a function $F : X \rightarrow \mathbb{R}$ to be α -strongly convex if for all geodesics $(x_t)_{t \in [0,1]}$ in X , it holds that

$$F(x_t) \leq (1-t)F(x_0) + tF(x_1) - \frac{\alpha t(1-t)}{2} d^2(x_0, x_1), \quad \text{for all } t \in [0, 1].$$

It is known that the convexity properties of the barycenter functional F_P are related to the *curvature* of the space. Here, curvature is interpreted as the *Alexandrov curvature*, which is the generalization of sectional curvature to geodesic spaces, see [BBI01]. Then, the result is that F_P is 1-strongly convex for every probability measure P on X if and only if X has *non-positive curvature*; see [Stu03] for precise statements. In fact, the 1-strong convexity of barycenter functionals is essentially the definition of non-positive curvature in this context.

Consequently, much stronger results are known for barycenters in non-positively curved spaces, ranging from basic properties such as existence and uniqueness, to statistical estimation and optimization; for details see the nice article [Stu03].

In contrast, it is well-known that Wasserstein space $\mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ (and hence, the Bures-Wasserstein space) is *non-negatively curved* [AGS08, Theorem 7.3.2]. This means, for instance, that convexity and properties related to convexity (such as the PL inequality employed in Appendix C.1) are not automatic for the barycenter functional in Wasserstein space. On the other hand, as emphasized in [Che+20], this non-negative curvature is related to the *smoothness* of the barycenter functional.

A.5 Additional facts about the Wasserstein metric

Here we collect various facts about the Wasserstein metric for easy reference in the sequel.

1. Euclidean gradient vs. Bures-Wasserstein gradient.

Let $F : \mathbb{S}_{++}^d \rightarrow \mathbb{R}$ be a function. Throughout this paper, we denote by $D F$ the usual Euclidean gradient of F , and we reserve ∇F for the gradient with respect to the Bures-Wasserstein geometry. In fact, under our tangent space convention, these two quantities are related as follows: let $(\Sigma_t)_{t \in \mathbb{R}}$ denote a curve in \mathbb{S}_{++}^d . We temporarily denote the Euclidean tangent vector (i.e., ordinary time derivative) to this curve via $\dot{\Sigma}^E$, and the Bures-Wasserstein tangent vector via $\dot{\Sigma}^{BW}$, which are related via $\dot{\Sigma}^E = \dot{\Sigma}^{BW}\Sigma + \Sigma\dot{\Sigma}^{BW}$ (see the discussion in Appendix A.1). We can compute the time derivative of F in two ways:

$$\begin{aligned} \langle \nabla F(\Sigma_0), \dot{\Sigma}_0^{BW} \rangle_{\Sigma_0} &= \partial_t|_{t=0} F(\Sigma_t) = \langle D F(\Sigma_0), \dot{\Sigma}_0^E \rangle = \langle D F(\Sigma_0), \dot{\Sigma}_0^{BW}\Sigma_0 + \Sigma_0\dot{\Sigma}_0^{BW} \rangle \\ &= 2 \langle D F(\Sigma_0), \dot{\Sigma}_0^{BW} \rangle_{\Sigma_0}. \end{aligned}$$

From this we can conclude that

$$\nabla F(\Sigma_0) = 2 D F(\Sigma_0).$$

2. Gradient of the squared Wasserstein distance.

For any $\nu \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$, the gradient of the functional $W_2^2(\cdot, \nu)$ at μ is given by

$$\nabla W_2^2(\cdot, \nu)(\mu) = -2(T_{\mu \rightarrow \nu} - \text{id}) = -2 \log_\mu \nu.$$

This is derived in, e.g. [ZP19]. In the Bures-Wasserstein setting, it can be proven via matrix calculus; see the proof of Theorem 3.

3. Inverse of the transport map.

If $\Sigma, \Sigma' \in \mathbb{S}_{++}^d$, then the transport map $T_{\Sigma \rightarrow \Sigma'}$ is the inverse matrix for the transport map $T_{\Sigma' \rightarrow \Sigma}$. This can be verified from the formula (6) using the symmetry of the geometric mean. More generally, it is a special case of the convex conjugacy relation between optimal Kantorovich potentials.

4. Diagonal case.

If $\Sigma_0, \Sigma_1 \in \mathbb{S}_{++}^d$ are *diagonal matrices*, then $W_2^2(\Sigma_0, \Sigma_1) = \|\Sigma_0^{1/2} - \Sigma_1^{1/2}\|_{\text{F}}^2$ is the squared Frobenius norm between the square roots. This can be verified, e.g. from the explicit formula (5) using the fact that Σ_0 and Σ_1 commute. Note that in one dimension, all matrices are diagonal. More generally, these observations extend to when Σ_0 and Σ_1 commute.

Similarly, it can be seen from (10) that the geodesic is given by

$$\Sigma_t^{1/2} = (1-t)\Sigma_0^{1/2} + t\Sigma_1^{1/2}, \quad t \in [0, 1],$$

which says that the Bures-Wasserstein geodesic between diagonal (or commuting matrices) is simply the Euclidean geodesic after applying the square root map.

5. The case of non-zero means.

For any $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, suppose that the means of these distributions are m_μ and m_ν , respectively. Let $\bar{\mu}, \bar{\nu}$ denote the centered versions of these distributions. Then, it holds that

$$W_2^2(\mu, \nu) = \|m_\mu - m_\nu\|^2 + W_2^2(\bar{\mu}, \bar{\nu}).$$

This can be proven directly from the definition (4).

6. A lower bound on the Wasserstein distance.

Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$. If $\tilde{\mu}$ and $\tilde{\nu}$ are *Gaussian* measures with the same moments up to order two as μ and ν , respectively, then $W_2(\mu, \nu) \geq W_2(\tilde{\mu}, \tilde{\nu})$ [CMT96].

B Proofs for the geodesic convexity results

B.1 Proof of Theorem 1

See Appendix A.1 and A.2 for background on the relevant geometric concepts.

Proof of Theorem 1. We begin by proving that $-\sqrt{\lambda_{\min}}$ is convex (equivalently, $\sqrt{\lambda_{\min}}$ is concave) along generalized geodesics. Let Q be any finitely supported probability measure on \mathbb{S}_{++}^d , and let $\Sigma_0 \in \mathbb{S}_{++}^d$ denote the basepoint. It is equivalent to show that if Σ^* is the generalized barycenter of Q at Σ_0 , then $\sqrt{\lambda_{\min}(\Sigma^*)} \geq \int \sqrt{\lambda_{\min}(\Sigma)} dQ(\Sigma)$.

The generalized barycenter by definition is the matrix $\Sigma^* = \exp_{\Sigma_0}(\int T_{\Sigma_0 \rightarrow \Sigma} dQ(\Sigma) - I_d)$. If we write $\bar{T} := \int T_{\Sigma_0 \rightarrow \Sigma} dQ(\Sigma)$ for the average transport map, the statement we want to show is $\bar{T}\Sigma_0\bar{T} \succeq \alpha I_d$ where $\alpha := (\int \sqrt{\lambda_{\min}(\Sigma)} dQ(\Sigma))^2$. We observe that

$$\bar{T}\Sigma_0\bar{T} \succeq \alpha I_d \iff \bar{T}^{-1}\Sigma_0^{-1}\bar{T}^{-1} \preceq \alpha^{-1}I_d \iff \bar{T}^{-1} \preceq \text{GM}(\Sigma_0, \alpha^{-1}I_d),$$

where the first equivalence follows from the order preservation of inversion [BL06, Exercice 3.3.2] and the second from [LL01, Corollary 3.5]. In turn, this is equivalent to $\bar{T} \succeq \alpha^{1/2}\Sigma_0^{-1/2}$.

Since $\bar{T} = \int \Sigma_0^{-1/2} (\Sigma_0^{1/2} \Sigma \Sigma_0^{1/2})^{1/2} \Sigma_0^{-1/2} dQ(\Sigma)$, we want to prove $\int (\Sigma_0^{1/2} \Sigma \Sigma_0^{1/2})^{1/2} dQ(\Sigma) \succeq \alpha^{1/2}\Sigma_0^{1/2}$. To prove this, observe that $\Sigma \succeq \lambda_{\min}(\Sigma) I_d$, so $\Sigma_0^{1/2} \Sigma \Sigma_0^{1/2} \succeq \lambda_{\min}(\Sigma) \Sigma_0$. Since taking square roots preserves the PSD ordering (c.f. [BL06, Exercise 1.2.5]), upon taking square roots and integrating we deduce

$$\int (\Sigma_0^{1/2} \Sigma \Sigma_0^{1/2})^{1/2} dQ(\Sigma) \succeq \left(\int \sqrt{\lambda_{\min}(\Sigma)} dQ(\Sigma) \right) \Sigma_0^{1/2} = \alpha \Sigma_0^{1/2}.$$

Hence $-\sqrt{\lambda_{\min}}$ is convex along generalized geodesics.

The proof of convexity of $\sqrt{\lambda_{\max}}$ is similar. By [LL01, Corollary 3.5],

$$\bar{T}\Sigma_0\bar{T} \preceq \beta I_d \iff \bar{T} \preceq \text{GM}(\Sigma_0^{-1}, \beta I_d).$$

Since $\bar{T} = \int \Sigma_0^{-1/2} (\Sigma_0^{1/2} \Sigma \Sigma_0^{1/2})^{1/2} \Sigma_0^{-1/2} dQ(\Sigma)$, it thus suffices to show $\int (\Sigma_0^{1/2} \Sigma \Sigma_0^{1/2})^{1/2} dQ(\Sigma) \leq \beta^{1/2} \Sigma_0^{1/2}$ where $\beta := (\int \sqrt{\lambda_{\max}(\Sigma)} dQ(\Sigma))^2$ as desired. Noting that $\Sigma_0^{1/2} \Sigma \Sigma_0^{1/2} \leq \lambda_{\max}(\Sigma) \Sigma_0$, taking square roots and integrating yields

$$\int (\Sigma_0^{1/2} \Sigma \Sigma_0^{1/2})^{1/2} dQ(\Sigma) \leq \left(\int \sqrt{\lambda_{\max}(\Sigma)} dQ(\Sigma) \right) \Sigma_0^{1/2} = \beta \Sigma_0^{1/2}.$$

Hence the result. \square

Remark 2. This result implies for instance that the set of PSD matrices with eigenvalues lying in a certain range is convex along generalized geodesics.

Remark 3. There is a short proof of the weaker statement that the functionals $-\sqrt{\lambda_{\min}}$ and $\sqrt{\lambda_{\max}}$ are geodesically convex. The following argument is implicit in the proofs of [AC11, Theorem 6.1] and [BJL19, Theorem 8]. Let Q be a finitely supported probability measure on \mathbb{S}_{++}^d . The barycenter Σ^* of Q satisfies the fixed-point equation

$$\Sigma^* = \int (\Sigma^{*1/2} \Sigma \Sigma^{*1/2})^{1/2} dQ(\Sigma),$$

see [AC11, Theorem 6.1]. This implies

$$\lambda_{\min}(\Sigma^*) \geq \int \sqrt{\lambda_{\min}(\Sigma^{*1/2} \Sigma \Sigma^{*1/2})} dQ(\Sigma) \geq \sqrt{\lambda_{\min}(\Sigma^*)} \int \sqrt{\lambda_{\min}(\Sigma)} dQ(\Sigma).$$

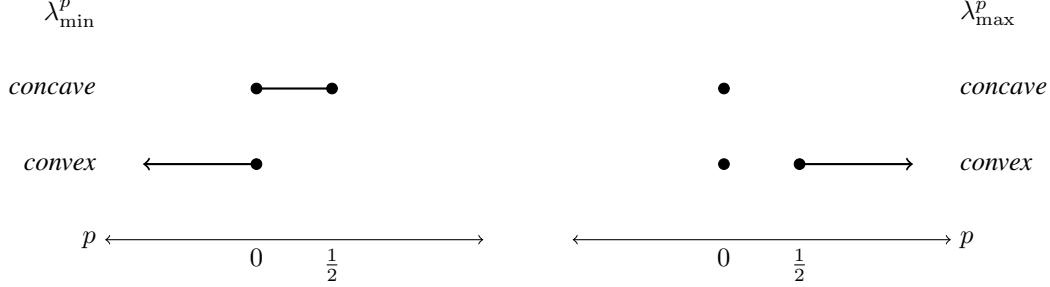
A similar argument applies for $\sqrt{\lambda_{\max}}$.

B.2 Sharpness of Theorem 1

We investigate the sharpness of this result in the following sense: for what exponents $p \geq 0$ is it true that the functionals $-\lambda_{\min}^p$, λ_{\max}^p are geodesically convex? For instance, the functional λ_{\max} was shown to be geodesically convex in [Che+20, Lemma 13].

In the following theorem, we show that the exponent $p = 1/2$ in Theorem 1 is optimal, in the sense that all possible geodesic convexity statements involving powers of λ_{\min} and λ_{\max} (except the trivial case $p = 0$) can be deduced from our result for $p = 1/2$.

Theorem 6. *The following diagrams depict the exponents $p \in \mathbb{R}$ for which λ_{\min}^p and λ_{\max}^p are concave or convex.*



The diagram is to be interpreted as follows. If part of the diagram is filled in with a solid black line, then the corresponding functional is concave/convex along generalized geodesics (and hence it is geodesically concave/convex). If part of the diagram is not filled in, then there exist counterexamples showing that the functional is not geodesically concave/convex.

Proof. First, we establish the positive results, which follow from composition rules:

- For $0 \leq p \leq 1/2$, λ_{\min}^p is the composition of the increasing concave function $(\cdot)^{2p}$ with the concave function $\sqrt{\lambda_{\min}}$, so it is concave.
- For $p \leq 0$, λ_{\min}^p is the composition of the decreasing convex function $(\cdot)^{2p}$ with the concave function $\sqrt{\lambda_{\min}}$, so it is convex.

- For $p \geq 1/2$, λ_{\max}^p is the composition of the increasing convex function $(\cdot)^{2p}$ with the convex function $\sqrt{\lambda_{\max}}$, so it is convex.

Next, we turn towards the negative results. First, recall from Fact 4 in Appendix A.5 that if Σ_0 and Σ_1 are one-dimensional, i.e., they are positive numbers, then the Bures-Wasserstein geodesic is

$$\Sigma_t = ((1-t)\Sigma_0^{1/2} + t\Sigma_1^{1/2})^2, \quad t \in [0, 1].$$

Also, in this case, λ_{\min} and λ_{\max} coincide and equal the identity; we thus abuse notation slightly in this paragraph by writing λ for both to handle the two cases simultaneously. Once we reparametrize by the square roots, it is seen that asking for concavity/convexity of λ^p is equivalent to asking for usual convexity of $(\cdot)^{2p}$ on \mathbb{R}_+ . This example rules out: (1) the concavity of λ^p for $p < 0$; (2) the convexity of λ^p for $0 < p < 1/2$; and (3) the concavity of λ^p for $p > 1/2$.

To rule out convexity of λ_{\min}^p for $p > 0$, consider $\Sigma = \text{diag}(\varepsilon, 1/\varepsilon)$ for small $\varepsilon > 0$. The transport map from Σ^{-1} to Σ is Σ , so from (10) the midpoint of this geodesic is $M := (\Sigma + \Sigma^{-1} + 2I_2)/4 = (\varepsilon + \varepsilon^{-1} + 2)I_2/4$. In particular, this implies that $\lambda_{\min}(M) \geq 1/(4\varepsilon) \gg \varepsilon = \max\{\lambda_{\min}(\Sigma), \lambda_{\min}(\Sigma^{-1})\}$. Thus λ_{\min}^p is not convex for any $p > 0$.

To rule out concavity of λ_{\max}^p for $p > 0$, note that for ε sufficiently small, in the previous example $\lambda_{\max}(M) \approx 1/(4\varepsilon) \ll 1/\varepsilon = \max\{\lambda_{\max}(\Sigma), \lambda_{\max}(\Sigma^{-1})\}$. Also, for any $p < 0$, the convexity of λ_{\max}^p would imply the concavity of λ_{\max}^{-p} due to the composition rules, hence λ_{\max}^p is not convex.

This covers all cases. \square

B.3 Eigenvalue clipping is a Bures-Wasserstein contraction

Convex sets play an important role in Euclidean optimization because projection onto a convex set is a contraction (c.f. [Bub15, Lemma 3.1]), and hence projected GD can be used to solve constrained optimization. Unfortunately, as the Bures-Wasserstein space is positively curved, we cannot automatically conclude that projection onto a geodesically convex set is a projection. Nevertheless, we can verify by hand the following result. In what follows, define for $0 < \beta < \infty$ the operator $\text{clip}^\beta : \mathbb{S}_{++}^d \rightarrow \mathbb{S}_{++}^d$ in the following way: if $\Sigma = \sum_{i=1}^d \lambda_i u_i u_i^\top$ is an eigenvalue decomposition of Σ , then

$$\text{clip}^\beta \Sigma := \sum_{i=1}^d (\lambda_i \wedge \beta) u_i u_i^\top.$$

Proposition 3. *The operator clip^β is a contraction w.r.t. the Bures-Wasserstein metric, i.e., $W_2(\text{clip}^\beta \Sigma, \text{clip}^\beta \Sigma') \leq W_2(\Sigma, \Sigma')$.*

To prove this proposition, we first extend the clipping operation to an operator $\mathbb{R}^{d \times d} \rightarrow \mathbb{R}^{d \times d}$ via the singular values; namely, given a singular value decomposition $A = \sum_{i=1}^d s_i u_i v_i^\top$, we let $\text{clip}^\beta A := \sum_{i=1}^d (s_i \wedge \beta) u_i v_i^\top$.

Proof of Proposition 3. Fix $X, Y \in \mathbb{S}_{++}^d$. It is known (see e.g. [BJL19]) that

$$W_2(X, Y) = \min_{\substack{A, B \in \mathbb{R}^{d \times d} \\ AA^\top = X \\ BB^\top = Y}} \|A - B\|_F.$$

Let (\bar{A}, \bar{B}) be a minimizing pair in the above expression. We aim to show

$$W_2(\text{clip}^\beta X, \text{clip}^\beta Y) \leq \|\text{clip}^{\sqrt{\beta}} \bar{A} - \text{clip}^{\sqrt{\beta}} \bar{B}\|_F \stackrel{?}{\leq} \|\bar{A} - \bar{B}\|_F = W_2(X, Y).$$

We only have to show the second inequality, and we do so by showing that the operator $\text{clip}^M : \mathbb{R}^{d \times d} \rightarrow \mathbb{R}^{d \times d}$ satisfies

$$\text{clip}^M A = \arg \min_{\tilde{A} \in \mathbb{R}^{d \times d}, \|\tilde{A}\| \leq M} \|A - \tilde{A}\|_F, \quad A \in \mathbb{R}^{d \times d}. \quad (14)$$

This will prove that clip^M is the Euclidean *projection* onto the closed convex set $\{\|\cdot\| \leq M\}$, and such a projection is automatically 1-Lipschitz.

Indeed, showing (14) is standard. Write $A = U\Sigma V^\top$ for its singular value decomposition.

$$\begin{aligned} \arg \min_{\tilde{A} \in \mathbb{R}^{d \times d}, \|\tilde{A}\| \leq M} \|\tilde{A} - A\|_F^2 &= \arg \min_{\tilde{A} \in \mathbb{R}^{d \times d}, \|\tilde{A}\| \leq M} \|\tilde{A} - U\Sigma V^\top\|_F^2 = \arg \min_{\tilde{A} \in \mathbb{R}^{d \times d}, \|\tilde{A}\| \leq M} \|U^\top \tilde{A} V - \Sigma\|_F^2 \\ &= \arg \min_{\tilde{A} \in \mathbb{R}^{d \times d}, \|\tilde{A}\| \leq M} \left\{ \sum_{i=1}^d \{\Sigma[i, i] - (U^\top \tilde{A} V)[i, i]\}^2 + \sum_{\substack{i, j \in [d] \\ i \neq j}} (U^\top \tilde{A} V)[i, j]^2 \right\}. \end{aligned}$$

On the other hand,

$$\begin{aligned} \min_{\tilde{A} \in \mathbb{R}^{d \times d}, \|\tilde{A}\| \leq M} \left\{ \sum_{i=1}^d \{\Sigma[i, i] - (U^\top \tilde{A} V)[i, i]\}^2 + \sum_{\substack{i, j \in [d] \\ i \neq j}} (U^\top \tilde{A} V)[i, j]^2 \right\} \\ \geq \sum_{i=1}^d \{(\Sigma[i, i] - M)_+\}^2, \end{aligned}$$

with equality attained at the unique minimizer \tilde{A} satisfying $U^\top \tilde{A} V = \text{clip}^M \Sigma$, i.e., $\tilde{A} = \text{clip}^M A$. \square

C Proofs for barycenters

C.1 Riemannian gradient descent

In this section, we detail the obstacles faced by previous analyses and then show how our geodesic convexity result, Theorem 1, enables us to overcome the prior exponential dependence on dimension and obtain the dimension-free rates in Theorems 2 and 3.

We begin by recalling the proof strategy of [Che+20]. Let F denote the barycenter functional,

$$F(\Sigma) := \frac{1}{2} \int W_2^2(\Sigma, \cdot) \, dP. \quad (15)$$

Standard optimization guarantees are often proven under the assumption that the objective function F is smooth and convex. Since we are considering Riemannian descent, this should be interpreted as convex and smooth along geodesics, as in [ZS16]. Unfortunately, the functional F is not geodesically convex (see [Che+20, Appendix B.2]), and so we must look for weaker conditions which still imply convergence of GD/SGD. A gradient domination condition known as the *Polyak-Łojasiewicz inequality* (henceforth *PL inequality*) was introduced in the non-convex optimization literature as an appropriate substitute for strong convexity [Pol64; Loj63], see also e.g., [KNS16; Bol+17]). Establishing a PL inequality in the present setting plays a key role in the analysis.

The following properties of the barycenter functional were proven in [Che+20].

Theorem 7. *Let $0 < \lambda_{\min} \leq \lambda_{\max} < \infty$ and write $\kappa := \lambda_{\max}/\lambda_{\min}$.*

1. ([Che+20, Theorem 7]) *The barycenter functional F is 1-geodesically smooth.*
2. ([Che+20, Theorem 17]) *Assume that the covariance matrices in the support of P have eigenvalues in the range $[\lambda_{\min}, \lambda_{\max}]$. Then, F satisfies a variance inequality,*

$$F(\Sigma) - F(\Sigma^*) \geq \frac{1}{2\kappa} W_2^2(\Sigma, \Sigma^*), \quad \text{for all } \Sigma \in \mathbb{S}_{++}^d.$$

3. ([Che+20, Theorem 19]) *Assume that the covariance matrices in the support of P , as well as Σ itself, have eigenvalues in the range $[\lambda_{\min}, \lambda_{\max}]$. Then, F satisfies a PL inequality at the matrix Σ :*

$$F(\Sigma) - F(\Sigma^*) \leq 2\kappa^2 \|\nabla F(\Sigma)\|_\Sigma^2.$$

Geodesic smoothness together with a PL inequality at every iterate are enough to obtain convergence guarantees for GD/SGD in objective value (i.e., the quantity $F(\Sigma) - F(\Sigma^*)$), c.f. [Che+20, Theorems 4-5]. The variance inequality is then used to deduce convergence of the iterate to Σ^* .

The main difficulty when applying these results is the assumption required for the third point: it requires *a priori* control over the eigenvalues of the iterates of GD/SGD.

This difficulty is addressed in [Che+20] via the following strategy: identify a geodesically convex subset \mathcal{S} of the Bures-Wasserstein manifold for which we can prove uniform bounds on the eigenvalues of matrices in \mathcal{S} . Since the iterates of SGD travel along geodesics, if P is supported in \mathcal{S} and the algorithm is initialized in \mathcal{S} , it follows that all iterates of SGD will remain in \mathcal{S} . The situation is similar for GD, except that “geodesics” must be replaced by “generalized geodesics”.

We can now describe the source of the exponential dependence on dimension in the result of [Che+20]: if the covariance matrices in the support of P have eigenvalues in $[\lambda_{\min}, \lambda_{\max}]$, then the subset \mathcal{S} used in the analysis of Chewi et al. is substantially larger than the support of P , and in particular the eigenvalues of matrices in \mathcal{S} can only be proven to lie in the range $[\lambda_{\min}/\kappa^{d-1}, \lambda_{\max}]$. The main improvement in the present analysis is to use our geodesic convexity result (Theorem 1) to prove the following result.

Lemma 1. *Suppose that the covariance matrices in the support of P have eigenvalues in the range $[\lambda_{\min}, \lambda_{\max}]$, and that we initialize GD (respectively SGD) at a point in $\text{supp } P$. Then, the iterates of GD (respectively SGD) also have eigenvalues in the range $[\lambda_{\min}, \lambda_{\max}]$.*

Proof. From Theorem 1, the set of matrices with eigenvalues in $[\lambda_{\min}, \lambda_{\max}]$ is closed under generalized geodesics. Since the update of GD (respectively SGD) moves along generalized geodesics (respectively geodesics), the result follows. \square

This combined with the arguments below is enough to alleviate the exponential dimension dependence. However, before continuing to the main argument, we prove sharper bounds for the last two statements of Theorem 7. This allows us to also improve our convergence rates’ dependence on the conditioning.

This improved version of Theorem 7 rests on the following observation. [Che+20, Lemma 16] shows that if Σ, Σ' have eigenvalues which lie in the range $[\lambda_{\min}, \lambda_{\max}]$, then the eigenvalues of the transport map $T_{\Sigma \rightarrow \Sigma'}$ lie in $[\kappa^{-1}, \kappa]$. However, these bounds are loose, as following lemma shows.

Lemma 2. *Suppose that $\Sigma, \Sigma' \in \mathbb{S}_{++}^d$ have eigenvalues which lie in the range $[\lambda_{\min}, \lambda_{\max}]$, and let $\kappa := \lambda_{\max}/\lambda_{\min}$ denote the condition number. Then, the eigenvalues of the transport map $T_{\Sigma \rightarrow \Sigma'}$ lie in the range $[1/\sqrt{\kappa}, \sqrt{\kappa}]$.*

Proof. The transport map $T_{\Sigma \rightarrow \Sigma'}$ is explicitly given in (6), and it can be recognized as the matrix geometric mean of Σ^{-1} and Σ' . Applying a norm bound for the matrix geometric mean [BG12, Theorem 3], we deduce that

$$\lambda_{\max}(T_{\Sigma \rightarrow \Sigma'}) \leq \lambda_{\max}(\Sigma'^{1/4} \Sigma^{-1/2} \Sigma'^{1/4}) \leq \sqrt{\kappa}.$$

The symmetry of Σ and Σ' together with Fact 3 in Appendix A.5 yields $\lambda_{\min}(T_{\Sigma \rightarrow \Sigma'}) \geq 1/\sqrt{\kappa}$. \square

Using this lemma, we now state and prove the refinement of Theorem 7.

Theorem 8. *Let $0 < \lambda_{\min} \leq \lambda_{\max} < \infty$ and write $\kappa := \lambda_{\max}/\lambda_{\min}$.*

1. ([Che+20, Theorem 7]) *The barycenter functional F is 1-geodesically smooth.*
2. *Assume that the covariance matrices in the support of P have eigenvalues in the range $[\lambda_{\min}, \lambda_{\max}]$. Then, F satisfies a variance inequality,*

$$F(\Sigma) - F(\Sigma^*) \geq \frac{1}{2\sqrt{\kappa}} W_2^2(\Sigma, \Sigma^*), \quad \text{for all } \Sigma \in \mathbb{S}_{++}^d.$$

3. *Assume that the covariance matrices in the support of P , as well as Σ itself, have eigenvalues in the range $[\lambda_{\min}, \lambda_{\max}]$. Then, F satisfies a PL inequality at the matrix Σ :*

$$F(\Sigma) - F(\Sigma^*) \leq 2\kappa^{3/2} \|\nabla F(\Sigma)\|_{\Sigma}^2.$$

Proof. The second statement follows from the general variance inequality ([Che+20, Theorem 6]) together with Lemma 2. Similarly, the third statement follows from the proof of [Che+20, Theorem 19] using the improved variance inequality. \square

The proof of Theorem 2 now follows from the analysis of [Che+20, §4.2] by adapting their Theorems 4 and 5. We now sketch the modifications required to prove Theorem 3.

Proof of Theorem 3. We prove the stronger result that this holds for $\bar{\kappa}$, as this implies the result for κ^* because

$$\bar{\kappa} = \frac{\overline{\lambda_{\max}}}{\overline{\lambda_{\min}}} = \left(\frac{\int \lambda_{\max}(\Sigma)^{1/2} dP(\Sigma)}{\int \lambda_{\min}(\Sigma)^{1/2} dP(\Sigma)} \right)^2 \leq \sup_{\Sigma \in \text{supp}(P)} \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)} = \kappa^*.$$

Above, the inequality follows from rearranging $\int \sqrt{\lambda_{\max}(\Sigma)} dP(\Sigma) \leq \sqrt{\kappa^*} \int \sqrt{\lambda_{\min}(\Sigma)} dP(\Sigma)$.

We begin by checking that the variance inequality and PL inequality from Theorem 7 continue to hold under these assumptions.

Variance inequality. From the geodesic convexity of $-\sqrt{\lambda_{\min}}$ and $\sqrt{\lambda_{\max}}$, the barycenter Σ^* of P has eigenvalues in $[\overline{\lambda_{\min}}, \overline{\lambda_{\max}}]$. By modifying the proof of Lemma 2 and using Fact 3 in Appendix A.5, the transport map $T_{\Sigma^* \rightarrow \Sigma}$ has eigenvalues bounded below as

$$\lambda_{\min}(T_{\Sigma^* \rightarrow \Sigma}) = \frac{1}{\lambda_{\max}(T_{\Sigma \rightarrow \Sigma^*})} \geq \frac{1}{\lambda_{\max}(\Sigma^{*1/4} \Sigma^{-1/2} \Sigma^{*1/4})} \geq \frac{\lambda_{\min}(\Sigma)^{1/2}}{\overline{\lambda_{\max}}^{1/2}}.$$

From [Che+20, Theorem 6], we can deduce that the variance inequality holds for P with constant

$$\int \lambda_{\min}(T_{\Sigma^* \rightarrow \Sigma}) dP(\Sigma) \geq \frac{\overline{\lambda_{\min}}^{1/2}}{\overline{\lambda_{\max}}^{1/2}} = \frac{1}{\bar{\kappa}^{1/2}}.$$

PL inequality. Similarly, a modification of the proof of [Che+20, Theorem 19] using the improved variance inequality shows that a PL inequality holds at Σ :

$$F(\Sigma) - F(\Sigma^*) \leq 2\bar{\kappa}^{1/2} \frac{\overline{\lambda_{\max}}}{\overline{\lambda_{\min}}} \|\nabla F(\Sigma)\|_{\Sigma}^2.$$

Putting it together. From Corollary 1, the iterates of either GD or SGD all have eigenvalues in the range $[\overline{\lambda_{\min}}, \overline{\lambda_{\max}}]$. Hence, the PL inequality in (3) of Theorem 8 holds at every iterate with $\bar{\kappa}$ replacing κ . The convergence rates now follow as before. \square

C.2 Euclidean gradient descent approach

We now present our results for the Euclidean geometry. [BJL19] prove that the barycenter functional is strictly convex on the positive semidefinite cone (w.r.t. the standard Euclidean geometry). We extend their results by showing that it is in fact strongly convex and smooth (again w.r.t. the standard Euclidean geometry). Besides yielding an analysis of Euclidean projected GD and SGD, these results also aid our analysis of the regularized barycenter problem in the sequel.

Fix $0 < \alpha \leq \beta$ and denote by $\mathcal{K}_{\alpha,\beta}$ the subset of covariance matrices whose spectrum lies within $[\alpha, \beta]$. Let F denote the barycenter functional, defined in (15).

Lemma 3. For all $\Sigma \in \mathcal{K}_{\alpha,\beta}$ and non-zero $Y \in \mathbb{S}^d$,

$$\frac{\alpha^3}{4\beta^4} \leq \frac{\langle Y, D^2 F(\Sigma)[Y] \rangle_F}{\|Y\|_F^2} \leq \frac{\beta^3}{4\alpha^4}. \quad (16)$$

Proof. It suffices to consider the case where $P = \frac{1}{N} \sum_{i=1}^N \delta_{\Sigma_i}$ for some $\Sigma_i \in \mathcal{K}_{\alpha,\beta}$, $i \in [N]$, as the case of general P supported on $\mathcal{K}_{\alpha,\beta}$ follows by compactness. Fix $\Sigma \in \mathcal{K}_{\alpha,\beta}$. Standard calculations as in [BJL19] show that the first derivative satisfies

$$2D F(\Sigma) = I_d - \frac{1}{N} \sum_{i=1}^N GM(\Sigma_i, \Sigma^{-1}).$$

We now compute the second derivative. Define the functions

$$\begin{aligned}\text{inv}(\Sigma) &:= \Sigma^{-1}, \\ \text{conj}_A(\Sigma) &:= A\Sigma A, \\ \text{sqrt}(\Sigma) &:= \Sigma^{1/2}.\end{aligned}$$

For $Y \in \mathbb{S}^d$, the above maps have derivatives

$$\begin{aligned}\text{D inv}(\Sigma)[Y] &= -\Sigma^{-1}Y\Sigma^{-1}, \\ \text{D conj}_A(\Sigma)[Y] &= AY A, \\ \text{D sqrt}(\Sigma)[Y] &= \int_0^\infty e^{-t\Sigma^{1/2}} Y e^{-t\Sigma^{1/2}} dt.\end{aligned}$$

With these definitions in hand, we can write

$$2 \text{D } F(\Sigma) = I_d - \frac{1}{N} \sum_{i=1}^N \text{conj}_{\Sigma_i^{1/2}} \circ \text{sqrt} \circ \text{conj}_{\Sigma_i^{-1/2}} \circ \text{inv}(\Sigma).$$

Taking the derivative in a symmetric direction $Y \in \mathbb{S}^d$ and applying the chain rule repeatedly,

$$\begin{aligned}2 \text{D}^2 F(\Sigma)[Y] \\ = \frac{1}{N} \sum_{i=1}^N \int_0^\infty \Sigma_i^{1/2} e^{-t(\Sigma_i^{1/2} \Sigma \Sigma_i^{1/2})^{-1/2}} \Sigma_i^{-1/2} \Sigma^{-1} Y \Sigma^{-1} \Sigma_i^{-1/2} e^{-t(\Sigma_i^{1/2} \Sigma \Sigma_i^{1/2})^{-1/2}} \Sigma_i^{1/2} dt.\end{aligned}$$

Let $g(t, x) = \exp(-t/\sqrt{x}) x^{-1}$ on $(t, x) \in (0, \infty) \times (0, \infty)$ and $Z_i = \Sigma_i^{1/2} \Sigma \Sigma_i^{1/2}$. Since $g(t, \cdot)$ is analytic on its domain, the Riesz-Dunford calculus (see [DS88]) applies and we may write

$$2 \langle Y, \text{D}^2 F(\Sigma)[Y] \rangle_F = \frac{1}{N} \sum_{i=1}^N \int_0^\infty \text{tr}(g(t, Z_i) \Sigma_i^{1/2} Y \Sigma_i^{1/2} g(t, Z_i) \Sigma_i^{1/2} Y \Sigma_i^{1/2}) dt.$$

Using the spectral mapping theorem and Lemma 5 below we further write

$$\leq \frac{\|Y\|_F^2}{N} \sum_{i=1}^N \lambda_{\max}(\Sigma_i)^2 \int_0^\infty \max_{\lambda \in \text{spec}(Z_i)} g(t, \lambda)^2 dt.$$

An analogous argument gives the lower bound

$$2 \langle Y, \text{D}^2 F(\Sigma)[Y] \rangle_F \geq \frac{\|Y\|_F^2}{N} \sum_{i=1}^N \lambda_{\min}(\Sigma_i)^2 \int_0^\infty \min_{\lambda \in \text{spec}(Z_i)} g(t, \lambda)^2 dt.$$

To bound the integral, we note that

$$e^{-t/\sqrt{\lambda_{\min}(Z_i)}} \lambda_{\max}(Z_i)^{-1} \leq g(t, \lambda) \leq e^{-t/\sqrt{\lambda_{\max}(Z_i)}} \lambda_{\min}(Z_i)^{-1}$$

for all $\lambda \in \text{spec}(Z_i)$. Since we assume $\alpha I_d \preceq \Sigma_i, \Sigma \preceq \beta I_d$, then $\alpha^2 I_d \preceq Z_i \preceq \beta^2 I_d$, so

$$2 \langle Y, \text{D}^2 F(\Sigma)[Y] \rangle_F \leq \beta^2 \int_0^\infty \exp\left(-\frac{2t}{\beta}\right) \frac{1}{\alpha^4} dt = \frac{\beta^3}{2\alpha^4}.$$

An analogous calculation for the lower bound finishes the proof. \square

Remark 4. Similar to Theorem 3, one can obtain improved strong convexity and smoothness parameters for F based on non-uniform notions of conditioning.

We can now describe the projected GD and projected SGD updates. Let $\Pi_{\alpha, \beta} : \mathbb{S}^d \rightarrow \mathcal{K}_{\alpha, \beta}$ denote the Euclidean projection onto $\mathcal{K}_{\alpha, \beta}$ and let $\eta = 4\lambda_{\min}^4 / \lambda_{\max}^3$. Given a starting matrix Σ_0 , the projected GD scheme to minimize the barycenter functional of a measure P supported on $\mathcal{K}_{\lambda_{\min}, \lambda_{\max}}$ is given by

$$\Sigma_{t+1}^{\text{EGD}} := \Pi_{\lambda_{\min}, \lambda_{\max}}(\Sigma_t - \eta \text{D } F(\Sigma_t^{\text{EGD}})), \quad t \geq 0. \quad (17)$$

Also, suppose that $\Sigma_1, \dots, \Sigma_t$ are i.i.d. samples from P . Then, the projected stochastic gradient scheme is

$$\Sigma_{t+1}^{\text{ESGD}} := \Pi_{\lambda_{\min}, \lambda_{\max}}(\Sigma_t^{\text{ESGD}} - \eta_{t+1} \{I_d - \Sigma_{t+1} \# (\Sigma_t^{\text{ESGD}})^{-1}\}), \quad t \geq 0, \quad (18)$$

where following [LSB12] we take the step size to be $\eta_t = 8\lambda_{\max}^4 / (\lambda_{\min}^3(t+1))$.

We now state the convergence guarantees for these two algorithms.

Theorem 9 (guarantees for Euclidean GD/SGD). *Assume that P is supported on covariance matrices whose eigenvalues lie in the range $[\lambda_{\min}, \lambda_{\max}]$, $0 < \lambda_{\min} \leq \lambda_{\max} < \infty$. Let $\kappa := \lambda_{\max}/\lambda_{\min}$ denote the condition number. Assume that we initialize at $\Sigma_0 \in \text{supp } P$.*

1. (EGD) *Let Σ_T^{EGD} denote the T -th iterate of projected Euclidean GD (17). Then,*

$$\|\Sigma_T^{\text{EGD}} - \Sigma^*\|_{\text{F}}^2 \leq \exp\left(-\frac{T}{\kappa^7}\right) \|\Sigma_0 - \Sigma^*\|_{\text{F}}^2. \quad (19)$$

2. (ESGD) *Let Σ_T^{ESGD} denote the T -th iterate of Euclidean projected SGD (18). Then,*

$$\mathbb{E}[\|\Sigma_T^{\text{ESGD}} - \Sigma^*\|_{\text{F}}^2] \leq \frac{64d\lambda_{\max}^2\kappa^{6.5}}{T}.$$

Proof. (1) The preceding lemma shows that the barycenter functional F is strongly convex and smooth with condition number κ^7 . By [Bub15, Theorem 3.10], projected GD (17) converges at the stated rate.

(2) For ESGD, we must compute a bound on the Euclidean variance of the stochastic gradient. Using Lemma 2, we get the two-sided control

$$\frac{1}{\sqrt{\kappa}} I_d \preceq \Sigma_{t+1} \#(\Sigma_t^{\text{ESGD}})^{-1} \preceq \sqrt{\kappa} I_d$$

and thus

$$\|I_d - \Sigma_{t+1} \#(\Sigma_t^{\text{ESGD}})^{-1}\|_{\text{F}}^2 \leq d(\sqrt{\kappa} - 1) \leq d\sqrt{\kappa}.$$

The result now follows from the preceding lemma and [LSB12]. \square

Remark 5. To compare the guarantees of Theorems 2 and 9, first we have

$$\frac{1}{2} \|\Sigma_T^{1/2} - \Sigma^{*1/2}\|_{\text{F}}^2 \leq W_2^2(\Sigma_T, \Sigma^*) \leq \|\Sigma_T^{1/2} - \Sigma^{*1/2}\|_{\text{F}}^2$$

as a consequence of [CV21, Lemma 3.5]. Moreover, under our assumptions,

$$\frac{1}{4\lambda_{\max}} \|\Sigma_T - \Sigma^*\|_{\text{F}}^2 \leq \|\Sigma_T^{1/2} - \Sigma^{*1/2}\|_{\text{F}}^2 \leq \frac{1}{4\lambda_{\min}} \|\Sigma_T - \Sigma^*\|_{\text{F}}^2,$$

where the first inequality is elementary and follows from

$$A - B = A^{1/2}(A^{1/2} - B^{1/2}) + (A^{1/2} - B^{1/2})B^{1/2},$$

whereas the second inequality uses [Bha97, (X.46)].

For the iterations given by (17) and (18) to be practical, we need the projection step to be implementable. The following lemma takes care of this.

Lemma 4. *Let $\Pi_{\alpha,\beta} : \mathbb{S}^d \rightarrow \mathcal{K}_{\alpha,\beta}$ be the projection with respect to the Frobenius norm. Then*

$$\Pi_{\alpha,\beta}(Y) = \sum_{i=1}^d [(\lambda_i \wedge \beta) \vee \alpha] v_i v_i^{\top}$$

where $Y = \sum_{i=1}^d \lambda_i v_i v_i^{\top}$ is an orthogonal eigendecomposition of Y .

Proof. Let $Y = Q\Lambda Q^{\top}$ be an orthogonal eigendecomposition of Y . Since the Frobenius norm is unitarily invariant, we have

$$\Pi_{\alpha,\beta}(Y) = \arg \min_{X \in \mathcal{K}_{\alpha,\beta}} \|X - Q\Lambda Q^{\top}\|_{\text{F}}^2 = \arg \min_{X \in \mathcal{K}_{\alpha,\beta}} \|Q^{\top} X Q - \Lambda\|_{\text{F}}^2 = Q \left(\arg \min_{X \in \mathcal{K}_{\alpha,\beta}} \|X - \Lambda\|_{\text{F}}^2 \right) Q^{\top}$$

and the result follows. \square

Finally, we state and prove the elementary lemma we used in the proof of Lemma 3.

Lemma 5. *Let $A, B \in \mathbb{S}_{++}^d$ and $Y \in \mathbb{S}^d$. Then*

$$\lambda_{\min}(A) \lambda_{\min}(B) \|Y\|_{\text{F}}^2 \leq \text{tr}(AYBY) \leq \lambda_{\max}(A) \lambda_{\max}(B) \|Y\|_{\text{F}}^2.$$

Proof. The result follows immediately from $\text{tr}(AYBY) = \|A^{1/2}YB^{1/2}\|_{\text{F}}^2$ and $\lambda_{\min}(A^{1/2}) = \lambda_{\min}(A)^{1/2}$ (similarly for B). \square

C.3 SDP formulation

The SDP formulation of the Bures-Wasserstein barycenter is as follows. Suppose that P is a discrete distribution, $P = \sum_{i=1}^k p_i \delta_{\Sigma_i}$. The Wasserstein distance between $\Sigma_0, \Sigma_1 \in \mathbb{S}_{++}^d$ can be expressed as

$$W_2^2(\Sigma_0, \Sigma_1) = \min_{S \in \mathbb{R}^{d \times d}} \left\{ \text{tr}(\Sigma_0 + \Sigma_1 - 2S) \quad \text{such that} \quad \begin{bmatrix} \Sigma_0 & S \\ S^\top & \Sigma_1 \end{bmatrix} \succeq 0 \right\}.$$

It follows that the barycenter Σ^* of P solves the optimization problem

$$\min_{\substack{\Sigma^* \in \mathbb{S}_{++}^d \\ S_1, \dots, S_k \in \mathbb{R}^{d \times d}}} \left\{ \text{tr}\left(\Sigma^* - 2 \sum_{i=1}^k p_i S_i\right) \quad \text{such that} \quad \begin{bmatrix} \Sigma_i & S_i \\ S_i^\top & \Sigma^* \end{bmatrix} \succeq 0, \forall i \in [k] \right\}.$$

D Proofs for entropically-regularized barycenters

We begin by remarking how the non-centered case can be reduced to the centered case.

Remark 6. For a probability measure μ , let m_μ denote its mean and let $\bar{\mu}$ denote the centered version of μ . Using Fact 5 in Appendix A.5, one can verify that

$$\begin{aligned} & \frac{1}{2} \int W_2^2(b, \mu) dP(\mu) + \gamma \text{KL}(b \parallel \mathcal{N}(0, I_d)) \\ &= \frac{1}{2} \int \|m_b - m_\mu\|^2 dP(\mu) + \frac{\gamma}{2} \|m_b\|^2 + \frac{1}{2} \int W_2^2(\bar{b}, \bar{\mu}) dP(\mu) + \gamma \text{KL}(\bar{b} \parallel \mathcal{N}(0, I_d)). \end{aligned}$$

This shows that the objective of the entropically-regularized barycenter decouples into two parts, one involving the mean of b and the other involving the centered version of b . Explicitly, we can compute

$$m^* := \frac{1}{1+\gamma} \int m_\mu dP(\mu)$$

and the entropically-regularized barycenter \bar{b}^* of the centered versions of the distributions in P . Then, if $\tau : \mathbb{R}^d \rightarrow \mathbb{R}^d$ denotes the translation $x \mapsto x + m^*$, the solution to the original entropically-regularized barycenter problem is $\tau_\# \bar{b}^*$.

We now overview the proof strategy; proofs are then provided in the subsequent subsections. Throughout this section let P be supported on $\mathcal{K}_{1/\sqrt{\kappa}, \sqrt{\kappa}}$, the subset of matrices in \mathbb{S}_{++}^d with eigenvalues in the range $[1/\sqrt{\kappa}, \sqrt{\kappa}]$.

An important observation driving our analysis is that the gradient of the KL divergence at Σ has the following form:

$$\nabla \text{KL}(\cdot \parallel I_d)(\Sigma) = I_d - \Sigma^{-1} = I_d - T_{\Sigma \rightarrow \Sigma^{-1}} = -\log_\Sigma(\Sigma^{-1}). \quad (20)$$

This can be shown by observing that

$$\text{KL}(\Sigma \parallel I_d) = \frac{1}{2} \text{tr} \Sigma - \frac{1}{2} \ln \det \Sigma - \frac{d}{2},$$

computing the Euclidean gradient, and appealing to Fact 1 in Appendix A.5. This gradient identity is convenient for applying our generalized geodesic convexity results and allows us to prove the following Lemma in Subsection D.1. Put $\Sigma^+ := \exp_\Sigma(-\eta \nabla F_\gamma(\Sigma))$.

Lemma 6. *If $\Sigma \in \mathcal{K}_{1/\sqrt{\kappa}, \sqrt{\kappa}}$, then so is Σ^+ .*

We also establish a couple of properties of our objective function in Subsection D.2.

Proposition 4. *Define $G : \mathcal{K}_{1/\sqrt{\kappa}, \sqrt{\kappa}} \rightarrow \mathbb{R}$ to take $\Sigma \mapsto \text{KL}(\Sigma \parallel I_d)$. Then, the following hold:*

1. *G is $2\sqrt{\kappa}$ -smooth with respect to Wasserstein geodesics.*
2. *F_γ is $1/(4\kappa^{7/2})$ -strongly convex with respect to Euclidean geodesics on $\mathcal{K}_{1/\sqrt{\kappa}, \sqrt{\kappa}}$.*
3. *F_γ is strictly convex on all of \mathbb{S}_{++}^d .*

With these facts, we can establish existence and uniqueness of Σ^* and prove Proposition 1 in Subsection D.3.

Next we prove smoothness and PL inequalities in Subsection D.4.

Lemma 7 (Smoothness). *If $\Sigma \in \mathcal{K}_{1/\sqrt{\kappa}, \sqrt{\kappa}}$ and we take the step size $\eta = 1/(1 + 2\gamma\sqrt{\kappa})$, then*

$$F_\gamma(\Sigma^+) - F_\gamma(\Sigma) \leq -\frac{1}{2(1 + 2\gamma\sqrt{\kappa})} \|\nabla F_\gamma(\Sigma)\|_\Sigma^2.$$

Lemma 8 (PL inequality). *If $\Sigma \in \mathcal{K}_{1/\sqrt{\kappa}, \sqrt{\kappa}}$, then*

$$F_\gamma(\Sigma) - F_\gamma(\Sigma^*) \leq \frac{\kappa^4}{2} \|\nabla F_\gamma(\Sigma)\|_\Sigma^2.$$

When the regularization parameter γ is large, we can instead use a different argument to improve the PL constant.

Lemma 9 (PL inequality, large regularization). *If $\Sigma \in \mathcal{K}_{1/\sqrt{\kappa}, \sqrt{\kappa}}$ and $\gamma \geq 14\kappa^4$, then*

$$F_\gamma(\Sigma) - F_\gamma(\Sigma^*) \leq \frac{1}{\gamma} \|\nabla F_\gamma(\Sigma)\|_\Sigma^2.$$

The main theorem now follows by combining these lemmas. \square

Proof of Theorem 4. By Lemma 6, the Lemmas 7, 8, and 9 hold throughout the optimization trajectory. Let $C = 2/\kappa^4$ if we apply Lemma 8, and let $C = \gamma$ if we apply Lemma 9. Then,

$$\begin{aligned} F_\gamma(\Sigma_{t+1}) - F_\gamma(\Sigma^*) &= F_\gamma(\Sigma_{t+1}) - F_\gamma(\Sigma_t) + F_\gamma(\Sigma_t) - F_\gamma(\Sigma^*) \\ &\leq -\frac{1}{2(1 + 2\gamma\sqrt{\kappa})} \|\nabla F_\gamma(\Sigma_t)\|_\Sigma^2 + F_\gamma(\Sigma_t) - F_\gamma(\Sigma^*) \\ &\leq \left(1 - \frac{C}{2(1 + 2\gamma\sqrt{\kappa})}\right) \{F_\gamma(\Sigma_t) - F_\gamma(\Sigma^*)\}. \end{aligned}$$

Iterating yields the result. \square

D.1 Trapping the iterates

Proof of Lemma 6. Combining (20) with the formula for the gradient of the squared Bures-Wasserstein distance (Fact 2 in Appendix A.5), we see that in fact

$$-\nabla F_\gamma(\Sigma) = \int \log_\Sigma(\Sigma') dP(\Sigma') + \gamma \log_\Sigma(\Sigma^{-1}).$$

We then apply Theorem 1 (see also the discussion in Appendix A.2) and the fact that that $\eta(1 + \gamma) \leq 1$ to yield

$$\begin{aligned} \sqrt{\lambda_{\min}}(\Sigma^+) &= \sqrt{\lambda_{\min}} \circ \exp_\Sigma \left(\eta \int \log_\Sigma(\Sigma') dP(\Sigma') + \eta\gamma \log_\Sigma(\Sigma^{-1}) + (1 - \eta - \eta\gamma) \underbrace{\log_\Sigma(\Sigma)}_{=0} \right) \\ &\geq \eta \int \sqrt{\lambda_{\min}}(\Sigma') dP(\Sigma') + \eta\gamma \sqrt{\lambda_{\min}}(\Sigma^{-1}) + [1 - \eta(1 + \gamma)] \sqrt{\lambda_{\min}}(\Sigma) \\ &\geq \frac{1}{\kappa^{1/4}}. \end{aligned}$$

The analogous argument shows that $\sqrt{\lambda_{\max}}(\Sigma^+) \leq \kappa^{1/4}$, hence the result. \square

D.2 Properties of the KL divergence

Proof of Proposition 4. For the first claim, fix $\Sigma_0, \Sigma_1 \in \mathcal{K}_{1/\sqrt{\kappa}, \sqrt{\kappa}}$ and let T denote the transport map from Σ_0 to Σ_1 . Then put

$$\Sigma_s := ((1-s)I_d + sT)\Sigma_0((1-s)I_d + sT) = (1-s)^2 \Sigma_0 + s^2 \Sigma_1 + s(1-s)(\Sigma_0 T + T\Sigma_0).$$

In other words, $(\Sigma_s)_{s \in [0,1]}$ is the Bures-Wasserstein geodesic between Σ_0 and Σ_1 (see (10)). It suffices to show (see Appendix A.3) that

$$\partial_s^2 \text{KL}(\Sigma_s \| I_d)|_{s=0} \leq 2\sqrt{\kappa} W_2^2(\Sigma_0, \Sigma_1).$$

Since $\text{KL}(\Sigma_s \| I_d) = \frac{1}{2}(\text{tr}(\Sigma_s) - \ln \det \Sigma_s + \text{constant})$, we analyze the first two terms separately. First, we note that

$$\partial_s^2 \text{tr}(\Sigma_s)|_{s=0} = 2 \text{tr}(\Sigma_0 + \Sigma_1 - 2\Sigma_0 T) = 2W_2^2(\Sigma_0, \Sigma_1),$$

where the equality follows from (9). For the second term we start by observing that

$$-\ln \det \Sigma_s = -\ln \det \Sigma_0 - 2 \ln \det((1-s)I_d + sT).$$

Using this identity we see that

$$-\partial_s^2 \left(\frac{1}{2} \ln \det(\Sigma_s) \right)|_{s=0} = \text{tr}((T - I_d)^2) \leq \sqrt{\kappa} \|T - I_d\|_{\Sigma_0}^2 = \sqrt{\kappa} W_2^2(\Sigma_0, \Sigma_1).$$

Putting these bounds together yields the result.

For the second claim, using the convexity of $-\ln \det$, it follows that the Euclidean Hessian of F_γ satisfies $D^2 F_\gamma \succeq D^2 F$, where F is the barycenter functional. It follows from Lemma 3 that

$$D^2 F_\gamma \succeq \frac{1}{4\kappa^{7/2}} I_d$$

on the set $\mathcal{K}_{1/\sqrt{\kappa}, \sqrt{\kappa}}$.

Finally, for the third claim we can use the convexity of F and the strict convexity of $-\ln \det$ (together with $\gamma > 0$) to argue that $D^2 F_\gamma \succ 0$ on \mathbb{S}_{++}^d . \square

D.3 Existence and uniqueness of the minimizer

Proof of Proposition 1. First, we prove that when restricted to \mathbb{S}_{++}^d , the functional F_γ has a unique minimizer. Let $H: \mathcal{K}_{1/\sqrt{\kappa}, \sqrt{\kappa}} \rightarrow \mathbb{S}_{++}^d$ take

$$\Sigma \mapsto \exp_\Sigma \left(-\frac{1}{1+\gamma} \nabla F_\gamma(\Sigma) \right).$$

Then by an analogous calculation to the proof of Lemma 6, H must map into $\mathcal{K}_{1/\sqrt{\kappa}, \sqrt{\kappa}}$. We may thus apply Brouwer's fixed point theorem to guarantee a fixed point of H in $\mathcal{K}_{1/\sqrt{\kappa}, \sqrt{\kappa}}$, call it Σ^* . Note that this means precisely that $\nabla F_\gamma(\Sigma^*) = 0$. By the equivalence of Euclidean and Bures-Wasserstein gradients (Fact 1 in Appendix A.5), we conclude that $D F_\gamma(\Sigma^*) = 0$ as well. By the strict convexity of F_γ (the third claim of Proposition 4), we deduce that Σ^* is the unique minimizer of F_γ on \mathbb{S}_{++}^d (actually, on all of \mathbb{S}_+^d , since $-\ln \det$ blows up if the determinant approaches 0).

Next, let b be a probability measure on \mathbb{R}^d which has mean m and covariance matrix Σ . Let \bar{b} denote the centered version of b . We now claim that

$$F_\gamma(b) \geq F_\gamma(\bar{b}) \geq F_\gamma(\mathcal{N}(0, \Sigma)) \geq F_\gamma(\mathcal{N}(0, \Sigma^*)).$$

The first inequality is due to Remark 6 and it is strict unless $b = \bar{b}$. The second inequality follows from Fact 6 in Appendix A.5, together with the classical fact that the Gaussian maximizes entropy among all centered distributions with the same covariance matrix; this latter fact is proven in [CT06, Theorem 8.6.5], and it also shows that the inequality is strict unless $\bar{b} = \mathcal{N}(0, \Sigma)$. Finally, the last inequality is what we have shown above, and it is also strict unless $\Sigma = \Sigma^*$. \square

D.4 PL and smoothness inequalities

Proof of Lemma 7. By Proposition 4 and the 1-geodesic smoothness of the barycenter functional [Che+20, Theorem 7] we deduce that $F_\gamma = F + \gamma G$ is $(1 + 2\gamma\sqrt{\kappa})$ -smooth, i.e.

$$F_\gamma(\Sigma^+) - F_\gamma(\Sigma) \leq \langle \nabla F_\gamma(\Sigma), \log_\Sigma(\Sigma^+) \rangle_\Sigma + \frac{1 + 2\gamma\sqrt{\kappa}}{2} W_2^2(\Sigma, \Sigma^+).$$

Substituting in $\log_\Sigma(\Sigma^+) = -\eta \nabla F_\gamma(\Sigma)$ and the step size $\eta = 1/(1 + 2\gamma\sqrt{\kappa})$ yields the result. \square

Proof of Lemma 8. From the second claim in Proposition 4, and since $\mathcal{K}_{1/\sqrt{\kappa}, \sqrt{\kappa}}$ is convex with respect to Euclidean geodesics, we see that for $\Sigma \in \mathcal{K}_{1/\sqrt{\kappa}, \sqrt{\kappa}}$

$$\begin{aligned} F_\gamma(\Sigma) - F_\gamma(\Sigma^*) &\leq \langle D F_\gamma(\Sigma), \Sigma - \Sigma^* \rangle - \frac{1}{8\kappa^{7/2}} \|\Sigma - \Sigma^*\|_F^2 \\ &= \frac{1}{2} \langle \nabla F_\gamma(\Sigma), \Sigma - \Sigma^* \rangle - \frac{1}{8\kappa^{7/2}} \|\Sigma - \Sigma^*\|_F^2, \end{aligned}$$

where the last line uses Fact 1 in Appendix A.5. Next we observe that by combining Cauchy-Schwarz with Young's inequality we get that for all $r > 0$,

$$\begin{aligned} \frac{1}{2} \langle \nabla F_\gamma(\Sigma), \Sigma - \Sigma^* \rangle &\leq \frac{1}{2} \|\nabla F_\gamma(\Sigma)\|_\Sigma \|\Sigma - \Sigma^*\|_{\Sigma^{-1}} \leq \frac{r}{16} \|\nabla F_\gamma(\Sigma)\|_\Sigma^2 + \frac{1}{r} \|\Sigma - \Sigma^*\|_{\Sigma^{-1}}^2 \\ &\leq \frac{r}{16} \|\nabla F_\gamma(\Sigma)\|_\Sigma^2 + \frac{\sqrt{\kappa}}{r} \|\Sigma - \Sigma^*\|_F^2. \end{aligned}$$

Putting $r = 8\kappa^4$ yields the result. \square

Proof of Lemma 9. From the 1-smoothness of the barycenter functional F [Che+20, Theorem 7],

$$\begin{aligned} F(\Sigma) - F(\Sigma^*) &\leq \langle \nabla F(\Sigma^*), T_{\Sigma^* \rightarrow \Sigma} - I_d \rangle_{\Sigma^*} + \frac{1}{2} W_2^2(\Sigma, \Sigma^*) \\ &= -\langle \nabla F(\Sigma^*) T_{\Sigma \rightarrow \Sigma^*}, T_{\Sigma \rightarrow \Sigma^*} - I_d \rangle_\Sigma + \frac{1}{2} W_2^2(\Sigma, \Sigma^*). \end{aligned}$$

On the other hand, it is a celebrated fact that G is 1-strongly convex w.r.t. the Wasserstein geometry (see [Vil03, §5]; on the Bures-Wasserstein space, it can also be read off from the proof of Proposition 4). It implies

$$G(\Sigma) - G(\Sigma^*) \leq -\langle \nabla G(\Sigma), T_{\Sigma \rightarrow \Sigma^*} - I_d \rangle_\Sigma - \frac{1}{2} W_2^2(\Sigma, \Sigma^*).$$

Combining these inequalities yields

$$\begin{aligned} F_\gamma(\Sigma) - F_\gamma(\Sigma^*) &\leq -\langle \nabla F_\gamma(\Sigma), T_{\Sigma \rightarrow \Sigma^*} - I_d \rangle_\Sigma \\ &\quad + \langle \nabla F(\Sigma) - \nabla F(\Sigma^*) T_{\Sigma \rightarrow \Sigma^*}, T_{\Sigma \rightarrow \Sigma^*} - I_d \rangle_\Sigma - \frac{\gamma - 1}{2} W_2^2(\Sigma, \Sigma^*). \end{aligned}$$

We next bound $\|\nabla F(\Sigma) - \nabla F(\Sigma^*) T_{\Sigma \rightarrow \Sigma^*}\|_\Sigma$. Write

$$\begin{aligned} \|\nabla F(\Sigma) - \nabla F(\Sigma^*) T_{\Sigma \rightarrow \Sigma^*}\|_\Sigma &= \left\| \int \{T_{\Sigma \rightarrow \Sigma'} - T_{\Sigma^* \rightarrow \Sigma'} T_{\Sigma \rightarrow \Sigma^*} + T_{\Sigma \rightarrow \Sigma^*} - I_d\} dP(\Sigma') \right\|_\Sigma \\ &\leq \left\| \int \{T_{\Sigma \rightarrow \Sigma'} - T_{\Sigma^* \rightarrow \Sigma'} T_{\Sigma \rightarrow \Sigma^*}\} dP(\Sigma') \right\|_\Sigma + W_2(\Sigma, \Sigma^*). \end{aligned}$$

For the first term, start with the triangle inequality,

$$\begin{aligned} \|T_{\Sigma \rightarrow \Sigma'} - T_{\Sigma^* \rightarrow \Sigma'} T_{\Sigma \rightarrow \Sigma^*}\|_\Sigma &\leq \|T_{\Sigma \rightarrow \Sigma'} - T_{\Sigma^* \rightarrow \Sigma'} + T_{\Sigma^* \rightarrow \Sigma'} - T_{\Sigma^* \rightarrow \Sigma'} T_{\Sigma \rightarrow \Sigma^*}\|_\Sigma \\ &\leq \|T_{\Sigma \rightarrow \Sigma'} - T_{\Sigma^* \rightarrow \Sigma'}\|_\Sigma + \|T_{\Sigma^* \rightarrow \Sigma'} - T_{\Sigma^* \rightarrow \Sigma'} T_{\Sigma \rightarrow \Sigma^*}\|_\Sigma \\ &\leq \|T_{\Sigma \rightarrow \Sigma'} - T_{\Sigma^* \rightarrow \Sigma'}\|_\Sigma + \sqrt{\kappa} \|I_d - T_{\Sigma \rightarrow \Sigma^*}\|_\Sigma \\ &= \|T_{\Sigma \rightarrow \Sigma'} - T_{\Sigma^* \rightarrow \Sigma'}\|_\Sigma + \sqrt{\kappa} W_2(\Sigma, \Sigma^*), \end{aligned}$$

where we use the fact that the eigenvalues of the transport map $T_{\Sigma^* \rightarrow \Sigma'}$ are bounded in magnitude by $\sqrt{\kappa}$ (Lemma 2). Next, consider the distribution $\delta_{\Sigma'}$ (a point mass on Σ') and let $F^{\Sigma'}$ denote the barycenter functional for the distribution $\delta_{\Sigma'}$. Then, the Euclidean smoothness bound for $F^{\Sigma'}$ (Lemma 3) yields

$$\|D F^{\Sigma'}(\Sigma) - D F^{\Sigma'}(\Sigma^*)\|_F \leq \frac{\kappa^{7/2}}{4} \|\Sigma - \Sigma^*\|_F.$$

Hence,

$$\begin{aligned} \|T_{\Sigma \rightarrow \Sigma'} - T_{\Sigma^* \rightarrow \Sigma'}\|_\Sigma^2 &= \|D F^{\Sigma'}(\Sigma) - D F^{\Sigma'}(\Sigma^*)\|_\Sigma^2 \leq \sqrt{\kappa} \|D F^{\Sigma'}(\Sigma) - D F^{\Sigma'}(\Sigma^*)\|_F^2 \\ &\leq \frac{\kappa^{15/2}}{16} \|\Sigma - \Sigma^*\|_F^2 \leq \frac{\kappa^8}{2} W_2^2(\Sigma, \Sigma^*), \end{aligned}$$

where the last inequality follows from Remark 5. Putting these inequalities together,

$$\|\nabla F(\Sigma) - \nabla F(\Sigma^*) T_{\Sigma \rightarrow \Sigma^*}\|_\Sigma \leq 3\kappa^4 W_2(\Sigma, \Sigma^*).$$

Continuing from before, we obtain

$$\begin{aligned} F_\gamma(\Sigma) - F_\gamma(\Sigma^*) &\leq \|\nabla F_\gamma(\Sigma)\|_\Sigma W_2(\Sigma, \Sigma^*) \\ &\quad + \|\nabla F(\Sigma) - \nabla F(\Sigma^*) T_{\Sigma \rightarrow \Sigma^*}\|_\Sigma W_2(\Sigma, \Sigma^*) - \frac{\gamma-1}{2} W_2^2(\Sigma, \Sigma^*) \\ &\leq \|\nabla F_\gamma(\Sigma)\|_\Sigma W_2(\Sigma, \Sigma^*) + 3\kappa^4 W_2^2(\Sigma, \Sigma^*) - \frac{\gamma-1}{2} W_2^2(\Sigma, \Sigma^*) \\ &\leq \|\nabla F_\gamma(\Sigma)\|_\Sigma W_2(\Sigma, \Sigma^*) - \frac{\gamma}{4} W_2^2(\Sigma, \Sigma^*) \end{aligned}$$

provided that γ is sufficiently large, $\gamma \geq 14\kappa^4$. For this large regularization, we can then prove

$$\begin{aligned} F_\gamma(\Sigma) - F_\gamma(\Sigma^*) &\leq \frac{1}{\gamma} \|\nabla F_\gamma(\Sigma)\|_\Sigma^2 + \frac{\gamma}{4} W_2^2(\Sigma, \Sigma^*) - \frac{\gamma}{4} W_2^2(\Sigma, \Sigma^*) \\ &\leq \frac{1}{\gamma} \|\nabla F_\gamma(\Sigma)\|_\Sigma^2. \end{aligned}$$

This completes the proof. \square

E Proofs for geometric medians

E.1 Convergence guarantee for smoothed Riemannian gradient descent

We begin with the proof of Proposition 2.

Proof of Proposition 2. Let $F : \mathcal{P}_2(\mathbb{R}^d) \rightarrow \mathbb{R}$ be the geometric median functional, $F(b) := \int W_2(b, \cdot) dP$. If we regard F as a functional over the Bures-Wasserstein space, then by continuity of F and compactness of the set $\{\|\cdot\| \leq \lambda_{\max}\} \subseteq \mathbb{S}_+^d$, there exists a minimizer Σ_{median}^* of F on this set. We will show that the Gaussian b_{median}^* with covariance Σ_{median}^* minimizes F over all of Wasserstein space.

First, recall the map $\text{clip}^{\lambda_{\max}}$ in Proposition 3, which is a contraction w.r.t. the Bures-Wasserstein metric. Then, for any $\Sigma \in \mathbb{S}_+^d$, it holds that

$$\begin{aligned} F(\Sigma) &= \int W_2(\Sigma, \Sigma') dP(\Sigma') \geq \int W_2(\text{clip}^{\lambda_{\max}} \Sigma, \Sigma') dP(\Sigma') \\ &\geq \int W_2(\Sigma_{\text{median}}^*, \Sigma') dP(\Sigma') = F(\Sigma_{\text{median}}^*), \end{aligned}$$

so that Σ_{median}^* minimizes F over \mathbb{S}_+^d .

Next, using Fact 6 in Appendix A.5, if $b \in \mathcal{P}_2(\mathbb{R}^d)$ has covariance matrix Σ , then

$$F(b) = \int W_2(b, \cdot) dP \geq \int W_2(\gamma_{0,\Sigma}, \cdot) dP \geq \int W_2(\Sigma_{\text{median}}^*, \cdot) dP = F(b_{\text{median}}^*),$$

so that b_{median}^* minimizes F over $\mathcal{P}_2(\mathbb{R}^d)$.

By definition, Σ_{median}^* has eigenvalues upper bounded by λ_{\max} . To finish the proof, we must show that the eigenvalues are also lower bounded by λ_{\min} ; we defer this part of the proof until the end of this section. \square

As the main difficulty in the analysis of the geometric median is the lack of both convexity and smoothness, we now pause to justify these remarks.

Remark 7. We claim that the unsquared Wasserstein distance $W_2(\cdot, \Sigma')$ is neither geodesically convex nor geodesically smooth. For the former statement, note that the geodesic convexity of $W_2(\cdot, \Sigma')$ would imply the geodesic convexity of $W_2^2(\cdot, \Sigma')$, but the squared Wasserstein distance is known

to not be geodesically convex (in general, it is not even *semi-convex*, see [AGS08, Example 9.1.5]; for a Gaussian example, see [Che+20, Appendix B.2]). In fact, unsquared metrics are almost never geodesically smooth; if this were the case, then there would exist a constant $\beta < \infty$ for which $W_2(\Sigma, \Sigma') \leq \frac{\beta}{2} W_2^2(\Sigma, \Sigma')$, which is manifestly false.

Moreover, the function $W_2(\cdot, \Sigma')$ is neither Euclidean convex nor Euclidean smooth. To see this, observe that in one dimension we have $W_2(\Sigma, \Sigma') = |\sqrt{\Sigma} - \sqrt{\Sigma'}|$ (see Fact 4 in Appendix A.5), which is neither convex nor smooth. It is notable that for this one-dimensional example, $W_2(\Sigma, \Sigma')$ is convex with respect to the variable $\sqrt{\Sigma}$, but it appears that reparameterization does not help in general; numerics indicate that the function $A \mapsto W_2(A^2, \Sigma')$ is not Euclidean convex on \mathbb{S}_{++}^d .

We now proceed with the analysis of the smoothed Riemannian GD algorithm given as Algorithm 4. Recall that F_ε denotes the smoothed geometric median functional. The first step is to show that the smoothing does not affect the objective significantly.

Lemma 10. *For any $\Sigma \in \mathbb{S}_{++}^d$, we have $|F(\Sigma) - F_\varepsilon(\Sigma)| \leq \varepsilon$.*

Proof. This follows from

$$|W_2(\Sigma, \Sigma') - \sqrt{W_2^2(\Sigma, \Sigma') + \varepsilon^2}| = |\sqrt{W_2^2(\Sigma, \Sigma')} - \sqrt{W_2^2(\Sigma, \Sigma') + \varepsilon^2}| \leq \varepsilon$$

and integrating. \square

Hence, if we can find a point $\hat{\Sigma}$ with $F_\varepsilon(\hat{\Sigma}) - \inf F_\varepsilon \leq \varepsilon$, it will then follow that $F(\hat{\Sigma}) - \inf F \leq 3\varepsilon$.

We next show that replacing W_2 by $W_{2,\varepsilon}$ indeed yields smoothness.

Lemma 11. *The functional F_ε is $1/\varepsilon$ -geodesically smooth.*

Proof. Recall from Theorem 7 that one-half of the squared Wasserstein distance is 1-smooth. This means that for any W_2 geodesic $(\Sigma_t)_{t \in \mathbb{R}}$, the following Hessian bound holds:

$$\frac{1}{2} \partial_t^2|_{t=0} W_2^2(\Sigma_t, \Sigma') \leq \|\dot{\Sigma}_0\|_{\Sigma_0}^2.$$

Here, $\dot{\Sigma}_0$ denotes the Bures-Wasserstein tangent vector, see the end of Appendix A.1. We use this to compute the smoothness of F_ε . Riemannian calculus yields

$$\begin{aligned} \partial_t F_\varepsilon(\Sigma_t) &= \int \frac{\partial_t W_2^2(\Sigma_t, \Sigma')}{2W_{2,\varepsilon}(\Sigma_t, \Sigma')} dP(\Sigma'), \\ \partial_t^2|_{t=0} F_\varepsilon(\Sigma_t) &= \int \left[\frac{\partial_t^2|_{t=0} W_2^2(\Sigma_t, \Sigma')}{2W_{2,\varepsilon}(\Sigma_0, \Sigma')} - \frac{\{\partial_t|_{t=0} W_2^2(\Sigma_t, \Sigma')\}^2}{4W_{2,\varepsilon}^3(\Sigma_0, \Sigma')} \right] dP(\Sigma'). \end{aligned}$$

The second term is non-positive. For the first term,

$$\int \frac{\partial_t^2|_{t=0} W_2^2(\Sigma_t, \Sigma')}{2W_{2,\varepsilon}(\Sigma_0, \Sigma')} dP(\Sigma') = \int \underbrace{\frac{1}{W_{2,\varepsilon}(\Sigma_0, \Sigma')}}_{\leq 1/\varepsilon} \underbrace{\frac{1}{2} \partial_t^2|_{t=0} W_2^2(\Sigma_t, \Sigma')}_{\leq \|\dot{\Sigma}_0\|_{\Sigma_0}^2} dP(\Sigma') \leq \frac{1}{\varepsilon} \|\dot{\Sigma}_0\|_{\Sigma_0}^2.$$

Hence, F_ε is $1/\varepsilon$ -smooth. \square

The next step is to prove a gradient domination condition for the functional F_ε . Let Σ_ε^* denote a minimizer of F_ε .

Lemma 12. *Suppose that the covariance matrices in the support of P , as well as Σ itself, have eigenvalues which lie in $[\lambda_{\min}, \lambda_{\max}]$. Let $\kappa := \lambda_{\max}/\lambda_{\min}$ denote the condition number. Then,*

$$F_\varepsilon(\Sigma) - F_\varepsilon(\Sigma_\varepsilon^*) \leq 2\kappa^{1/4} F_\varepsilon(\Sigma) \|\nabla F_\varepsilon(\Sigma)\|_\Sigma^{1/2}.$$

Proof. As the first part of the proof relies on general optimal transport arguments, we use the notation of general Wasserstein space. Given two measures μ, ν , let $\varphi_{\mu \rightarrow \nu}$ denote the Kantorovich potential

from μ to ν , and also denote $\psi_{\mu \rightarrow \nu} := \|\cdot\|^2/2 - \varphi_{\mu \rightarrow \nu}$. Also, let b_ε^* denote the minimizer of F_ε . Then, from Kantorovich duality,

$$W_2(b, \mu) = \sqrt{2 \int \psi_{\mu \rightarrow b} d\mu + 2 \int \psi_{b \rightarrow \mu} db},$$

$$W_2(b_\varepsilon^*, \mu) \geq \sqrt{\max \left\{ 2 \int \psi_{\mu \rightarrow b} d\mu + 2 \int \psi_{b \rightarrow \mu} db_\varepsilon^*, 0 \right\}}.$$

It follows that

$$W_{2,\varepsilon}(b, \mu) - W_{2,\varepsilon}(b_\varepsilon^*, \mu) \leq \sqrt{2 \max \left\{ \int \psi_{b \rightarrow \mu} d(b - b_\varepsilon^*), 0 \right\}}.$$

Integrating,

$$\begin{aligned} F_\varepsilon(b) - F_\varepsilon(b_\varepsilon^*) &\leq \int \sqrt{2 \left| \int \psi_{b \rightarrow \mu} d(b - b_\varepsilon^*) \right|} dP(\mu) \\ &= \int \sqrt{2 W_{2,\varepsilon}(b, \mu) \left| \int \frac{\psi_{b \rightarrow \mu}}{W_{2,\varepsilon}(b, \mu)} d(b - b_\varepsilon^*) \right|} dP(\mu) \\ &\leq \sqrt{2 \int W_{2,\varepsilon}(b, \mu) dP(\mu) \left| \iint \frac{\psi_{b \rightarrow \mu}}{W_{2,\varepsilon}(b, \mu)} d(b - b_\varepsilon^*) dP(\mu) \right|}. \end{aligned}$$

Following [Che+20], we introduce the constant-speed W_2 geodesic $(b_s)_{s \in [0,1]}$ from b to b_ε^* , and applying [Che+20, Lemma 13] we obtain

$$F_\varepsilon(b) - F_\varepsilon(b_\varepsilon^*) \leq \sqrt{2 F_\varepsilon(b) W_2(b, b_\varepsilon^*) \int_0^1 \|\nabla F_\varepsilon(b)\|_{L^2(b_s)} ds}.$$

This can be simplified via

$$F_\varepsilon(b) = \int W_{2,\varepsilon}(b, \mu) dP(\mu) \geq W_{2,\varepsilon}(b, b_\varepsilon^*) - \int W_{2,\varepsilon}(b_\varepsilon^*, \mu) dP(\mu),$$

and since b_ε^* is assumed to be a minimizer of F_ε , then

$$W_2(b, b_\varepsilon^*) \leq 2 F_\varepsilon(b).$$

Hence,

$$F_\varepsilon(b) - F_\varepsilon(b_\varepsilon^*) \leq 2 F_\varepsilon(b) \sqrt{\int_0^1 \|\nabla F_\varepsilon(b)\|_{L^2(b_s)} ds}.$$

Next, we specialize the result to the Bures-Wasserstein space. In this case, using the assumptions on P and Σ , we can argue as in [Che+20, Theorem 19] that $\|\nabla F_\varepsilon(\Sigma)\|_{\Sigma}^2 \leq \kappa \|\nabla F_\varepsilon(\Sigma)\|_{\Sigma}^2$, which completes the proof. \square

In order to proceed with the analysis, we must study the dynamics of the smoothed Riemannian GD algorithm to see if we can satisfy the hypotheses of Lemma 12. To study these dynamics, it is helpful to again adopt the notation and calculus of general Wasserstein space. The Wasserstein gradient of F_ε is

$$\nabla F_\varepsilon(b) = - \int \frac{T_{b \rightarrow \mu} - \text{id}}{W_{2,\varepsilon}(b, \mu)} dP(\mu),$$

and one step of the Riemannian GD iteration with step size ε (which is motivated by Lemma 11) is

$$b^+ := \exp(-\varepsilon \nabla F_\varepsilon(b)) = \left[\text{id} + \varepsilon \int \frac{T_{b \rightarrow \mu} - \text{id}}{W_{2,\varepsilon}(b, \mu)} dP(\mu) \right] \# b.$$

We will rewrite this in the following way. Define the weight

$$\rho(\mu) := \frac{W_{2,\varepsilon}(b, \mu)^{-1}}{\int W_{2,\varepsilon}(b, \cdot)^{-1} dP}.$$

Then,

$$\begin{aligned} & \text{id} + \varepsilon \int \frac{T_{b \rightarrow \mu} - \text{id}}{W_{2,\varepsilon}(b, \mu)} dP(\mu) \\ &= \left(1 - \varepsilon \int W_{2,\varepsilon}(b, \cdot)^{-1} dP\right) \text{id} + \left(\varepsilon \int W_{2,\varepsilon}(b, \cdot)^{-1} dP\right) \int T_{b \rightarrow \mu} \rho(\mu) dP(\mu). \end{aligned}$$

Since $\int W_{2,\varepsilon}(b, \cdot)^{-1} dP \leq 1/\varepsilon$, this is a convex combination of two terms. Let us call the weights $1 - \lambda$ and λ respectively. If we define the probability measure $\tilde{P} := (1 - \lambda)\delta_b + \lambda\rho P$, then this can also be written as

$$b^+ = \left(\int T_{b \rightarrow \mu} d\tilde{P}(\mu) \right)_\# b.$$

This expression proves the following fact (see also Appendix A.2).

Lemma 13. *The next iterate b^+ of smoothed Riemannian GD starting at b (with step size ε) is a generalized barycenter of the distribution \tilde{P} with base b .*

Combined with our geodesic convexity result (Theorem 1), we can now conclude the following important facts about the dynamics of smoothed GD.

Corollary 1. *Assume that all of the covariance matrices in the support of P have eigenvalues which lie in the range $[\lambda_{\min}, \lambda_{\max}]$, and that we initialize Algorithm 4 at an element of $\text{supp } P$. Then, all of the iterates of Algorithm 4 satisfy the same eigenvalue bounds.*

We can now prove Theorem 5.

Proof of Theorem 5. As discussed after Lemma 10, it suffices to find the number of iterations until we find an ε -approximate minimizer of F_ε . Combining the smoothness (Lemma 11) and gradient domination (Lemma 12), as well as Corollary 1, we obtain

$$F_\varepsilon(\Sigma_{t+1}) - F_\varepsilon(\Sigma_t) \leq -\frac{\varepsilon}{2} \|\nabla F_\varepsilon(\Sigma_t)\|_{\Sigma_t}^2 \leq -\frac{\varepsilon}{32\kappa F_\varepsilon(\Sigma_0)^4} \{F_\varepsilon(b_t) - F_\varepsilon(b^\star)\}^4.$$

Let us write

$$\delta_t := F_\varepsilon(\Sigma_t) - F_\varepsilon(\Sigma_\varepsilon^\star), \quad \zeta := \frac{\varepsilon}{32\kappa F_\varepsilon(\Sigma_0)^4}.$$

Then, we can rewrite the recursion as

$$\delta_{t+1} \leq \delta_t - \zeta \delta_t^4.$$

We solve the recursion via induction; we claim that

$$\delta_t \leq \frac{1}{\zeta^{1/3} (1+t)^{1/3}}.$$

This holds when $t = 0$ because $\zeta^{-1/3} \geq \delta_0$. Assuming that the inequality holds at some iteration t , we proceed to verify this for iteration $t + 1$. If $\delta_t \leq 1/[\zeta^{1/3} (2+t)^{1/3}]$, then this is immediate because $\delta_{t+1} \leq \delta_t$. Otherwise, $\delta_t \geq 1/[\zeta^{1/3} (2+t)^{1/3}]$, and we obtain

$$1 - \zeta \delta_t^3 \leq 1 - \frac{1}{2+t} = \frac{1+t}{2+t} \leq \left(\frac{1+t}{2+t}\right)^{1/3}$$

and hence

$$\delta_{t+1} \leq \delta_t (1 - \zeta \delta_t^3) \leq \frac{1}{\zeta^{1/3} (1+t)^{1/3}} \left(\frac{1+t}{2+t}\right)^{1/3} = \frac{1}{\zeta^{1/3} (2+t)^{1/3}}.$$

The theorem now follows. \square

Finally, we finish the last part of Proposition 2.

Proof of Proposition 2 (Continued). Let $(\varepsilon_k)_{k \in \mathbb{N}}$ be a sequence of positive numbers tending to zero. The guarantee for smoothed GD in Theorem 5, along with the control over the iterates in Corollary 1, allows us to assert that for each $k \in \mathbb{N}$, there is a point $\Sigma_k \in \mathbb{S}_{++}^d$ with eigenvalues in $[\lambda_{\min}, \lambda_{\max}]$ (the output of smoothed GD) with suboptimality gap $F(\Sigma_k) - \inf F \leq \varepsilon_k$. By compactness, we can extract a convergent subsequence of $(\Sigma_k)_{k \in \mathbb{N}}$, which must therefore converge to a minimizer of F (since F is continuous). The limit point must have eigenvalues in the range $[\lambda_{\min}, \lambda_{\max}]$, which completes the proof. \square

E.2 Reduction for non-zero means

In this section, we suppose that P is supported on non-degenerate, not necessarily centered Gaussians, whose covariance matrices have eigenvalues in the range $[\lambda_{\min}, \lambda_{\max}]$. We begin with the observation that if b_{median}^* denotes a Gaussian minimizer of the median functional for P , then the mean of b_{median}^* is not necessarily the Euclidean geometric median of the means of distributions in $\text{supp } P$. To see this, consider the case when the Gaussians are one-dimensional. Then, if we identify each Gaussian $\mu \in \text{supp } P$ with its mean and *standard deviation* (the square root of the variance) (m_μ, σ_μ) , then the W_2 distance between Gaussians is isometric to the standard Euclidean metric on the pairs (m, σ) in \mathbb{R}^2 (see Facts 4 and 5 in Appendix A.5). Therefore, the Wasserstein geometric median of P is equivalent to the Euclidean geometric median of the pairs (m, σ) , and the statement whose validity is being investigated is tantamount to asking: is the first coordinate of the Euclidean geometric median in \mathbb{R}^2 equal to the median of the first coordinates? This statement is manifestly false.

Next, we describe the reduction. Let $(m, \Sigma), (m', \Sigma')$ denote two pairs of means and covariance matrices in the support of P . Then,

$$W_2^2((m, \Sigma), (m', \Sigma')) = \|m - m'\|^2 + W_2^2(\Sigma, \Sigma') ,$$

where the LHS denotes the squared Wasserstein distance between Gaussians with parameters (m, Σ) and (m', Σ') respectively. The idea behind the reduction is that since the Wasserstein metric on diagonal matrices is the same as the Euclidean metric between the *square roots* of the matrices (Fact 4 in Appendix A.5), we can embed the mean vectors as diagonal matrices, and take the direct sum of these diagonal matrices with the covariance matrices to form augmented matrices; then, we can apply the geometric median algorithm (Algorithm 4) to the augmented matrices. In this reduction, however, we must take care that when we embed the mean vectors, we embed them into *positive definite* diagonal matrices.

Hence, define the augmented matrices

$$\Sigma := \begin{bmatrix} \text{diag}((m + C)^2) & \\ & \Sigma \end{bmatrix} , \quad \Sigma' := \begin{bmatrix} \text{diag}((m' + C)^2) & \\ & \Sigma' \end{bmatrix} ,$$

where the constant $C \geq \sqrt{\lambda_{\min}} + \max\{\|m\|_\infty, \|m'\|_\infty\}$ is chosen to ensure that $\Sigma, \Sigma' \succeq \lambda_{\min} I_d$ (and that $m + C, m' + C \geq 0$). The Wasserstein distance between the augmented matrices is

$$\begin{aligned} W_2^2(\Sigma, \Sigma') &= \|(m + C) - (m' + C)\|_{\text{F}}^2 + W_2^2(\Sigma, \Sigma') = \|m - m'\|^2 + W_2^2(\Sigma, \Sigma') \\ &= W_2^2((m, \Sigma), (m', \Sigma')) . \end{aligned}$$

Hence, after preprocessing the mean vectors and covariance matrices to form these augmented matrices, we may apply Algorithm 4 to the augmented matrices in a black box manner. It is easy to check that the set of such diagonal block matrices (where the upper block is itself diagonal) is convex under generalized geodesics. Hence, as long as the Algorithm 4 is initialized at such a matrix every iterate will remain in that form, and therefore the iterates will, when transformed back through the augmentation operation described above, indeed approach optimality for the original median problem. Note that the new value of the condition number will be

$$\kappa = \max \left\{ \kappa, \left(1 + \frac{2B}{\sqrt{\lambda_{\min}}} \right)^2 \right\} , \quad B := \sup_{\mu \in \text{supp } P} \|m_\mu\|_\infty .$$

The convergence guarantee of Theorem 5 applies, with κ replacing κ .

Of course, it is likely that analyzing smoothed Riemannian GD directly for the non-centered case could produce sharper results (in particular, with a dependence on κ rather than κ), but this simple approach already gives dimension-free convergence rates for the Bures-Wasserstein geometric median.

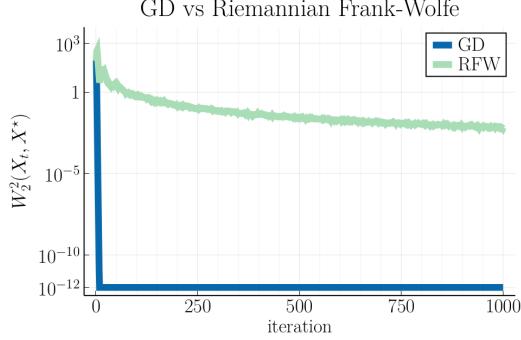


Figure 7: Riemannian GD vs Riemannian Frank-Wolfe, for computing Bures-Wasserstein barycenters.

F Further experiments and details

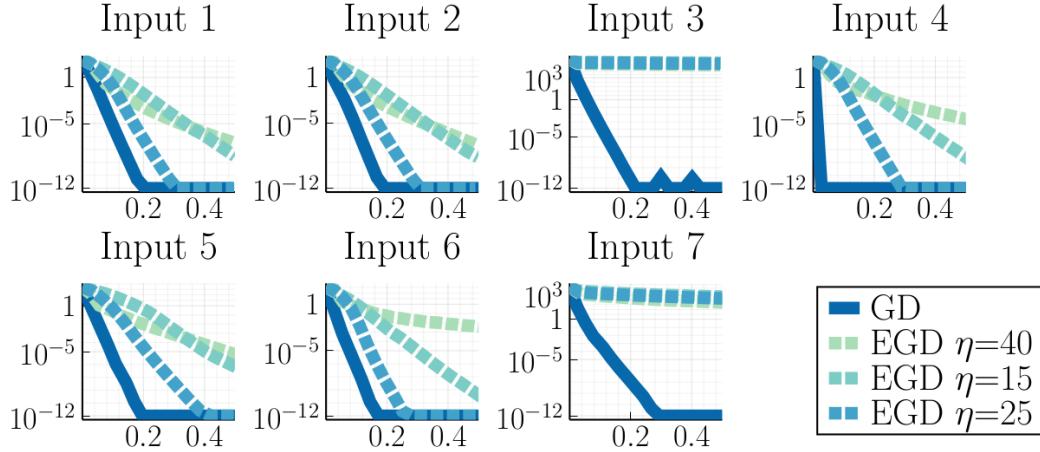
Reproducibility details. Input generation details for Figures 1, 2, 5, and 6 are provided in the main text. For Figures 3 and 4, recall that we generated matrices from a distribution whose barycenter is known to be the identity. By [ZP19, Theorem 2], if the mean of the distribution $(\log_{I_d})_# P$ is 0, then I_d is the barycenter of P . In particular, if Q is a mean zero distribution supported on symmetric matrices that lie in the domain of the exponential map, then $P = (\exp_{I_d})_# Q$ has I_d as its barycenter. In our experiments, we defined Q to be the law of a random matrix with Haar eigenbasis and uniform eigenvalues from the interval $[-(1 - \delta), 1 - \delta]$ for a parameter $\delta \in (0, 1)$. At the identity, the exponential map takes the simple form $\exp_{I_d} S = (I_d + S)^2$ and we see that P is then supported on covariance matrices with spectrum in $[\delta^2, (2 - \delta)^2]$. Both figures were generated with $\delta = 0.1$. All experiments were performed using Julia 1.5.1 on a desktop computer running Ubuntu 18.04 with an Intel i7-10700 CPU.

Riemannian Frank-Wolfe. Here we provide an empirical comparison with the Riemannian Frank-Wolfe algorithm for computing Bures-Wasserstein barycenters, as described in [WS17, Algorithm 3]. Figure 7 demonstrates the superior practical performance of Riemannian GD, the algorithm studied in this paper. In this experiment, the input is as in Figure 2, and X^* denotes the best iterate.

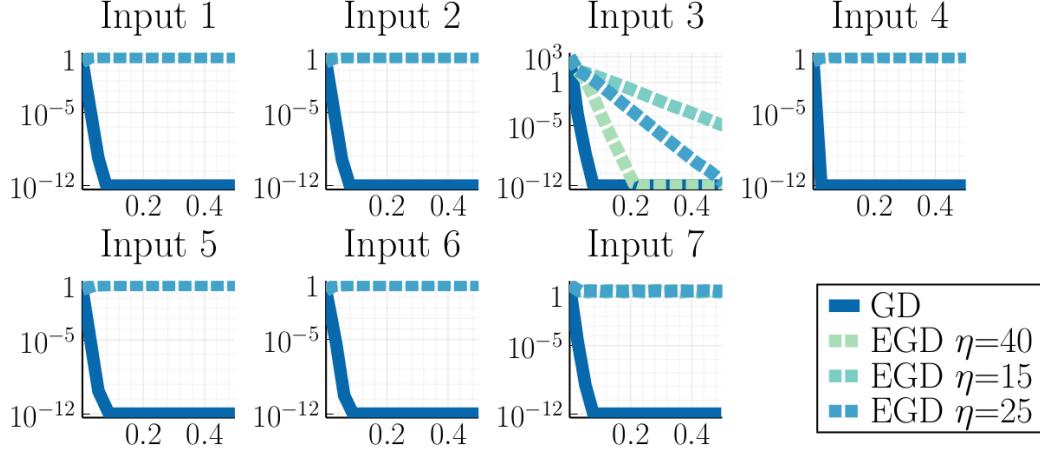
Further empirical comparisons. Here we further investigate the comparison of Riemannian and Euclidean GD done in Figure 2 by demonstrating qualitatively similar results for a variety of synthetic datasets. For each dataset, the measure P is the empirical measure of n matrices of dimension $d \times d$ that are drawn randomly as follows.

1. Haar eigenbasis and linearly spaced eigenvalues in $[\alpha, \beta]$.
2. Haar eigenbasis and i.i.d. $\text{Unif}[\alpha, \beta]$ eigenvalues.
3. First split the matrices into 3 groups. Each matrix has Haar eigenbasis and i.i.d. $\text{Unif}[\alpha, \beta]$ eigenvalues where $[\alpha, \beta] = 10^i \times [1, \kappa]$ for $i \in \{-2, 0, 2\}$ depending on its group.
4. Same as method 2 above, except all matrices have the same eigenbasis. (Note that GD converges in 1 step here since the matrices commute.)
5. Haar eigenbasis and eigenvalues uniform on a set of size $m \leq d$, whose elements are i.i.d. $\text{Unif}[\alpha, \beta]$.
6. Same as method 5 above, except all matrices use the same eigenvalues.
7. Mix of all methods above.

Figures 8a and 8b compare Euclidean and Riemannian GD on the barycenter problem as in Figure 2, but now with these 7 different input families. We average well-conditioned matrices in Figure 8a, and ill-conditioned matrices in Figure 8b. The plots are generated using $n = d = 50$ and $m = d/4$. For Method 7, the 50 matrices are divided into 6 groups of roughly equal size. The y -axis measures the W_2^2 distance to the best iterate; and the x -axis measures time in seconds.



(a) Here, the matrices are poorly conditioned, namely $[\alpha, \beta] = [0.03, 30]$ whereby $\kappa = 1000$.



(b) Here, the matrices are well-conditioned, namely $[\alpha, \beta] = [1, 2]$ whereby $\kappa = 2$.

Figure 8: Comparison of high-precision barycenter algorithms for various types of synthetic data.

In these figures we had to hand-tune the stepsize for Euclidean GD since the stepsize indicated by Theorem 9 performs quite poorly. We used the same range of stepsizes ($\eta \in \{15, 25, 40\}$) in all plots to demonstrate that the performance of Euclidean GD is quite sensitive to its stepsize. In contrast, GD performs well on all inputs with its (untuned) stepsize of 1.