On the Convergence of Projected Bures-Wasserstein Gradient Descent under Euclidean Strong Convexity

Junyi Fan [*1] **Yuxuan Han** [*1] **Zijian Liu** [2] **Jian-Feng Cai** [1] **Yang Wang** [1 3] **Zhengyuan Zhou** [2 4]

Abstract

The Bures-Wasserstein (BW) gradient descent method has gained considerable attention in various domains, including Gaussian barycenter, matrix recovery and variational inference problems, due to its alignment with the Wasserstein geometry of normal distributions. Despite its popularity, existing convergence analysis are often contingent upon specific loss functions, and the exploration of constrained settings within this framework remains limited. In this work, we make an attempt to bridge this gap by providing a general convergence rate guarantee for BW gradient descent when the Euclidean strong convexity of the loss and the constraints is assumed. In an effort to advance practical implementations, we also derive a closed-form solution for the projection onto BW distance-constrained sets, which enables the fast implementation of projected BW gradient descent for problems that arise in the constrained barycenter and distributionally robust optimization literature. Experimental results demonstrate significant improvements in computational efficiency and convergence speed, underscoring the efficacy of our method in practical scenarios.

1. Introduction

In this work, we consider constrained optimization problems of the form

$$\min_{\Sigma} f(\Sigma) \quad \text{subject to } \Sigma \in C \tag{1}$$

for some function $f : \mathbb{R}^{n\times n} \to \mathbb{R}$ and closed set C defined over the space of $n \times n$ positive-definite matrices:

$$\mathbb{S}_+^n := \{\Sigma \in \mathbb{R}^{n\times n} : \Sigma = \Sigma^\top, \lambda_{\min}(\Sigma) > 0\},$$

which have appeared in various applications, including kernel and metric learning (Han et al., 2021; Guillaumin et al., 2009; Suárez et al., 2021; Tsuda et al., 2005), variational inference (Blei et al., 2017; Lambert et al., 2022; Diao et al., 2023), image processing (Lenglet et al., 2006; Pennec, 2020), and distributionally robust optimization (Shafieezadeh Abadeh et al., 2018; Nguyen et al., 2023; Taşkesen et al., 2023) among others.

When endowed with a particular metric, S_+^n forms a Riemannian manifold; various metrics and the corresponding Riemannian optimization algorithms over S_+^n have been extensively studied in previous works (Pennec et al., 2006; Sra, 2012; Ha Quang et al., 2014; Lin, 2019; Han et al., 2021).

The Bures-Wasserstein (BW) distance has garnered increasing attention among various metrics for its connections to optimal transport theory and the Wasserstein metric. Within BW geometry, the corresponding Riemannian gradient descent algorithm is widely adopted for its straightforward implementation and superior performance in handling BW distance-related objective functions, including Gaussian barycenter (Chewi et al., 2020; Altschuler et al., 2021), low-rank matrix recovery (Luo & Trillos, 2022; Maunu et al., 2023), and variational inference problems (Lambert et al., 2022; Diao et al., 2023).

In barycenter-related problems, the global convergence of BW gradient descent has been demonstrated by Chewi et al. (2020); Altschuler et al. (2021). Remarkably, the barycenter objective function, while strongly convex and smooth in Euclidean geometry, lacks geodesic convexity in BW geometry–the usual prerequisite for Riemannian gradient convergence (Zhang & Sra, 2016; 2018; Han et al., 2021)–making the analysis and result notably valuable.

The success of Chewi et al. (2020); Altschuler et al. (2021) motivates us to explore whether a more general global convergence result for BW gradient descent is possible under the sole assumption of Euclidean convexity and smoothness. In this paper, we address an even broader scenario as for-

[*]Equal contribution [1]Department of Mathematics, Hong Kong University of Science and Technology [2]Stern School of Business, New York University [3]Department of Industrial Engineering and Decision Analytics, Hong Kong University of Science and Technology [4]Arena Technologies. Correspondence to: Zhengyuan Zhou <zhengyuanzhou24@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

mulated in (1), where we assume only the Euclidean strong convexity and smoothness of the function f, together with the Euclidean convexity of the constraint set C. This formulation can further encompass both constrained barycenter problems and WDRO related problems considered in (Nguyen et al., 2023). We summarize our contributions in Section 1.1.

1.1. Our Contributions

Global convergence of BW gradient descent for Euclidean strongly convex & smooth functions By exploring the connection between BW geometry and Euclidean geometry, we establish the global convergence of BW gradient descent for optimizing Euclidean strongly convex and smooth objective functions, with a convergence rate that depends on the eigenvalues of the resulting iterative sequence. We also show the necessity of this eigenvalue dependency by providing a lower bound result of the BW gradient method. Furthermore, we provide a sufficient condition for the linear convergence of BW gradient descent and, by applying this condition to the barycenter problem, demonstrate that our results imply a linear convergence rate for the barycenter problem as shown in Chewi et al. (2020); Altschuler et al. (2021).

Global convergence of Projected BW gradient descent and Closed form solution for projection to BW ball We further extend the global convergence result to the setting with a Euclidean convex set constraint, resulting in problems with both geodesically non-convex objectives and constraints. Besides the general convergence analysis, for the specific BW ball constraint set, we establish the existence of a closed-form solution for the projection under the BW distance—a contrast to scenarios under Euclidean distance. This closed-form solution enables efficient implementation of the projected BW gradient method for BW ball-constrained problems, including constrained BW barycenter problems and least-favorable distribution seeking problems that arise in Wasserstein distributionally robust optimization literature (Nguyen et al., 2023). We further provide experimental results on several applications of the projected BW gradient descent to show its superiority in both computational complexity and convergence rate.

1.2. Related Works

Optimization over Riemannian Manifolds Our problem setting lies in the more general Riemannian optimization context, where non-Euclidean convexity and smoothness are defined, and optimization algorithms are analyzed under such conditions. Non-asymptotic convergence of (stochastic) projected Riemannian gradient descent and other first-order methods has been well-studied when both the objective function and the constraint set are geodesically convex (Udriste, 2013; Zhang & Sra, 2016; 2018; Kim & Yang, 2022; Weber & Sra, 2023). In this paper, our main focus is on the setting where both the objective function and the constraint set are not geodesically convex, and geodesic convexity under BW distance is difficult to verify. Although there are results for non-convex Riemannian optimization, the general theory primarily concerns convergence to stationary points (Han & Gao, 2021; Criscitiello & Boumal, 2023). To the best of our knowledge, a general global convergence result under geodesic non-convexity is only considered in Boumal et al. (2019), whose result is not applicable in BW geometry[1].

Convergence Analysis of the BW Gradient Descent Algorithms Non-asymptotic convergence of the BWGD algorithm for unconstrained problems has been studied for general geodesically convex functions in (Han et al., 2021) and in specific problems, including Bures-Wasserstein barycenters (Chewi et al., 2020; Altschuler et al., 2021), Bures–Wasserstein geometric median (Altschuler et al., 2021), variational inference (Lambert et al., 2022; Diao et al., 2023), and low-rank matrix sensing (Luo & Trillos, 2022; Maunu et al., 2023). To the best of our knowledge, the convergence of projected BWGD for constrained problems has not been studied in previous literature.

Optimization with the BW Ball Constraint The BW ball-constrained optimization problems have arisen in Wasserstein distributionally robust optimization literature (Kuhn et al., 2019; Gao & Kleywegt, 2023), especially in finding the least-favorable distributions (Shafieezadeh Abadeh et al., 2018; Nguyen et al., 2023; Taşkesen et al., 2023). Previous works attempt to solve such problems based on Euclidean optimization algorithms. Since the projection operator to the BW ball under the Euclidean norm is computationally expensive, the Frank-Wolfe algorithm is adopted in (Shafieezadeh Abadeh et al., 2018; Nguyen et al., 2023; Taşkesen et al., 2023). It is also worth noting that there are several works studying general constrained optimization problems and the optimality conditions for optimizing over Wasserstein spaces (Lanzetti et al., 2022; Yue et al., 2021).

2. Backgrounds

Notations We use $A \succ 0$ to denote A is positive definite. And $A \succ B$ to denote $A - B \succ 0$. For matrices A, B and $\Sigma \succ 0$, we denote $\langle A, B\rangle = \mathrm{tr}(A^\top B)$ and $\langle A, B\rangle_\Sigma = \mathrm{tr}(A^\top \Sigma B)$. We use $a \lesssim b$ to denote $a \leq cb$ for some absolute constant c.

Moreover, throughout the paper, we assume that the considered function $f : \mathbb{R}^{n\times n} \to \mathbb{R}$ is symmetric in the sense

[1] See also the remark under Proposition 3.7.

Concept	Euclidean Geometry	BW Geometry
Squared distance $d^2(\Sigma, \Sigma')$	$\|\Sigma - \Sigma'\|_F^2$	$\mathrm{tr}\big(\Sigma + \Sigma' - 2(\Sigma^{1/2}\Sigma'\Sigma^{1/2})^{1/2}\big)$
Tangent space T_Σ and $\langle\cdot,\cdot\rangle_\Sigma$	$\mathbb{S}^n, \langle A, B\rangle_\Sigma = \mathrm{tr}(AB)$	$\mathbb{S}^n, \langle A, B\rangle_\Sigma = \mathrm{tr}(A\Sigma B)$
Geodesic joining Σ, Σ'	$t\Sigma' + (1-t)\Sigma, t \in [0,1]$	$\big((1-t)I + tT_{\Sigma\Sigma'}\big)\Sigma\big((1-t)I + tT_{\Sigma\Sigma'}\big), t \in [0,1]$
Exp map and its inverse $\exp_\Sigma(\cdot), \log_\Sigma(\cdot)$	$\exp_\Sigma(S) = \Sigma + S, \log_\Sigma(\Sigma') = \Sigma' - \Sigma$	$\exp_\Sigma(S) = (I+S)\Sigma(I+S), \log_\Sigma(\Sigma') = T_{\Sigma\Sigma'} - I$
β-smoothness	$f(\Sigma') \le f(\Sigma) + \langle Df(\Sigma), \Sigma' - \Sigma\rangle + \frac{\beta}{2}\|\Sigma - \Sigma'\|_F^2$	$f(\Sigma') \le f(\Sigma) + 2\langle Df(\Sigma), T_{\Sigma\Sigma'} - I\rangle_\Sigma + \frac{\beta}{2}d^2(\Sigma, \Sigma')$
α-strong convexity	$f(\Sigma') \ge f(\Sigma) + \langle Df(\Sigma), \Sigma' - \Sigma\rangle + \frac{\alpha}{2}\|\Sigma - \Sigma'\|_F^2$	$f(\Sigma') \ge f(\Sigma) + 2\langle Df(\Sigma), T_{\Sigma\Sigma'} - I\rangle_\Sigma + \frac{\alpha}{2}d^2(\Sigma, \Sigma')$
gradient update in (6)	$\Sigma^+ \leftarrow \Sigma - \eta Df(\Sigma)$	$\Sigma^+ \leftarrow \big(I - 2\eta Df(\Sigma)\big)\Sigma\big(I - 2\eta Df(\Sigma)\big)$

Table 1. Formulas of geometric concepts for BW and Euclidean Geometry of $\mathbb{S}_+^n$, where $T_{\Sigma\Sigma'}$ is defined in (5) and Df is the Euclidean gradient of f. A part of results in this table is reproduced from those in table 2.1 of Diao (2023).

that $f(M) = f(M^\top)$ for all $M \in \mathbb{R}^{n\times n}$. For a general function f, we can adapt it to our setting by defining $\tilde{f}(M) := \frac{1}{2}(f(M) + f(M^\top))$ for all M. Such a replacement will not alter its minimum point $\Sigma^\star$ when $\Sigma^\star \succ 0$, see also Luo & Trillos (2022).

2.1. Riemannian Manifold and Geodesic Convexity

In this section, we introduce the notation in general Riemannian manifold and geodesic convexity & smoothness. To keep the readability and fluency of the article, we just provide the minimal concepts that we will use in the main text, and we refer interested readers for (Absil et al., 2008; Bacak, 2014; Zhang & Sra, 2016; Boumal, 2023) for more detailed treatment.

On a general manifold $\mathcal{M}$, a Riemannian metric is a smooth, bilinear, symmetric positive definite function defined on the tangent space T_x at any point $x \in \mathcal{M}$, denoted $\langle\cdot,\cdot\rangle_x$. A geodesic on the manifold $\gamma : [0,1] \to \mathcal{M}$ is a curve that is locally the shortest path and has zero acceleration. For $x \in \mathcal{M}$ and $u \in T_x$, the value of the exponential map $\exp_x(\cdot) : T_x \to \mathcal{M}$ at u is given by $\gamma(1)$, where γ is the geodesic satisfying $\gamma(0) = x$ and $\gamma'(0) = u$. The logarithmic map $\log_x(\cdot) : \mathcal{M} \to T_x$ is defined as the inverse of the exponential map.

And for a differentiable function f over $\mathcal{M}$, we define the following concept of geodesic strong convexity and smoothness:

Definition 2.1. A function $f : \mathcal{M} \to \mathbb{R}$ is said to be *geodesically α-strongly convex* if for any $x, x' \in \mathcal{M}$, it holds that

$$f(x') \ge f(x) + \langle \nabla f(x), \log_x(x')\rangle_x + \frac{\alpha}{2} d_{\mathcal{M}}(x, x')^2, \quad (2)$$

with $\nabla f(x) \in T_x$ the Riemannian gradient of f at x.

Definition 2.2. A function $f : \mathcal{M} \to \mathbb{R}$ is said to be *geodesically β-smooth* if for any $x, x' \in \mathcal{M}$, it holds that

$$f(x') \le f(x) + \langle \nabla f(x), \log_x(x')\rangle_x + \frac{\beta}{2} d_{\mathcal{M}}(x, x')^2, \quad (3)$$

with $\nabla f(x) \in T_x$ the Riemannian gradient of f at x.

Definition 2.3. A function $f : \mathcal{M} \to \mathbb{R}$ is said to be *Euclidean L-Lipschitz* if for any $x \in \mathcal{M}$, its Euclidean gradient $Df(x)$ satisfies

$$\|Df(x)\| \le L. \quad (4)$$

2.2. BW and Euclidean Geometry over Positive Definite Matrices

In this section, we introduce the Riemannian geometry of $\mathbb{S}_+^n$ with the BW distance and the Euclidean distance. As in the previous section, we just provide the minimal concepts that we will use in the main text, and we refer interested readers to Appendix A of Altschuler et al. (2021) and section 2 of Diao (2023) for more detailed treatment.

The BW distance defined over $\mathbb{S}_+^n$ is given by

$$d_{\mathrm{BW}}(\Sigma, \Sigma') = \sqrt{\mathrm{tr}\big(\Sigma + \Sigma' - 2(\Sigma^{1/2}\Sigma'\Sigma^{1/2})^{1/2}\big)}.$$

The BW distance is intimately connected to the concept of optimal transport, in that the Wasserstein-2 distance between Gaussian distributions $\mathcal{N}(0, \Sigma)$ and $\mathcal{N}(0, \Sigma')$ is precisely the BW distance between the covariance matrices Σ and Σ'.

On the other hand, the Euclidean distance defined over $\mathbb{S}_+^n$ is straightforwardly determined by the Frobenius norm of the matrix difference:

$$d_{\|\cdot\|_F}(\Sigma, \Sigma') = \|\Sigma - \Sigma'\|_F.$$

This means that there are Riemannian manifold structures defined over $\mathbb{S}_+^n$ for which the geodesic distances correspond to d_{BW} and $d_{\|\cdot\|_F}$, respectively. We present the related geometric quantities and the conditions for smoothness and convexity under these structures in Table 1, where we introduce the following transformation:

$$T_{\Sigma\Sigma'} := \Sigma^{-1/2}(\Sigma^{1/2}\Sigma'\Sigma^{1/2})^{1/2}\Sigma^{-1/2}, \quad (5)$$

which can be interpreted as the optimal transport map from $\mathcal{N}(0, \Sigma)$ to $\mathcal{N}(0, \Sigma')$.

Additionally, within the context of the smooth and convex conditions presented in Table 1, we employ the identity

$\nabla_{\mathrm{BW}} f(\Sigma) = 2Df(\Sigma)$ between Euclidean gradient Df and the Riemannian gradient $\nabla_{\mathrm{BW}} f$ under BW geometry.

2.3. Projected BW Gradient Descent Update

In this paper, we aim to study the convergence of projected Riemannian gradient descent iteration described in Algorithm 1 for solving (1).

Algorithm 1 Projected BW Gradient Descent

Input: Objective function f, constraint set C, Initial point Σ_0, step size η, time-horizon T.

for $t = 0, 1, 2, \ldots, T$ **do**

$$\Sigma_t^+ \leftarrow \exp_{\Sigma_t}(-\eta \nabla f(\Sigma_t)), \tag{6}$$

$$\Sigma_{t+1} \leftarrow P_C(\Sigma_t^+). \tag{7}$$

The update rule in (6) can be viewed as a generalization of the gradient descent update to a Riemannian manifold and simplifies to the standard gradient descent update in the case of flat geometry, as indicated in the last row of Table 1. The projection step in (7) solves the problem

$$\min_{\Sigma} d(\Sigma, \Sigma_t^+) \quad \text{subject to } \Sigma \in C,$$

where d is the geodesic distance induced by the manifold's geometry. The explicit form of the general update equation (6) within the Bures-Wasserstein geometry is presented in the last row of Table 1.

3. Analysis of BW Gradient Descent for Unconstrained Problems

We would first consider the unconstrained case, whose analysis is relatively straightforward, as a warm-up, we will then extend the idea of analysis to the constrained setting in the next section. From now on, we set $\nabla f = \nabla_{\mathrm{BW}} f$ and $d = d_{\mathrm{BW}}$ for simplicity of notation.

For a Euclidean strongly convex and smooth f, we would show the convergence of BWGD based on the following general fact, as shown in (Chewi et al., 2020; Altschuler et al., 2021)

Definition 3.1 (Polyak-Łojasiewicz inequality)**.** We say a function f satisfies the μ-Polyak-Łojasiewicz (PL) inequality at Σ, if it holds that

$$\mu\big(f(\Sigma) - f(\Sigma^\star)\big) \le \|\nabla f(\Sigma)\|_\Sigma^2. \tag{8}$$

Definition 3.2 (Descent condition)**.** We say a function f satisfies the ζ-descent condition at Σ under the BWGD stepsize η, if it holds that for Σ and $\Sigma^+ = \exp_\Sigma\big(-\eta \nabla f(\Sigma)\big)$,

$$f(\Sigma^+) - f(\Sigma) \le -\zeta \|\nabla f(\Sigma)\|_\Sigma^2 \tag{9}$$

for some $\zeta > 0$.

In particular, a straightforward fact, by the update formula of Riemannian gradient, is that the descent condition holds with $\eta \le \frac{2}{\beta}$ and $\zeta = \eta - \frac{\beta\eta^2}{2}$ when f is β-geodesically smooth.

It is easy to verify that the PL inequality and the descent condition together can lead to a contraction guarantee:

Proposition 3.3. *Suppose function f satisfies μ-PL inequality and the ζ-descent condition at Σ, then it holds that*

$$f(\Sigma^+) - f(\Sigma^\star) \le (1 - \zeta\mu)\,(f(\Sigma) - f(\Sigma^\star))$$

Based on Proposition 3.3, we would divide the proof contraction result into verifying the PL inequality (8) and the descent condition for f.

Verifying the PL inequality

Verifying the PL inequality is rather straightforward based on the tangent space structure of the BW manifold: Actually, as long as the Euclidean PL inequality

$$\tilde{\mu}\big(f(\Sigma) - f(\Sigma^\star)\big) \le \|Df(\Sigma)\|^2. \tag{10}$$

holds, we can verify by definition that (8) holds with $\mu_{\mathrm{PL}} = 4\lambda_{\min}(\Sigma)\tilde{\mu}$. On the other hand, we would recall the following well-known fact in convex optimization:

Lemma 3.4. *If f is Euclidean α-strongly convex, then* (10) *holds with $\tilde{\mu} = 2\alpha$.*

Thus the α-Euclidean convexity together with Lemma 3.4 implies the $8\alpha\lambda_{\min}(\Sigma)$-PL inequality of f at every Σ.

Verifying the Descent Condition

To verify the descent condition at Σ, we aim to relate the expansion in (3) at Σ under Euclidean geometry to its counterpart in BW geometry.

For the squared distance term, the following equivalence between the BW distance and the Frobenius norm has been established by Altschuler et al. (2021)[2]:

Lemma 3.5 ((Altschuler et al., 2021), Remark 5)**.** *For every $\Sigma, \Sigma' \in \mathbb{S}_+^n$, $\lambda_{\min} I \preceq \Sigma, \Sigma' \preceq \lambda_{\max} I$, we have*

$$4\lambda_{\min} d^2(\Sigma, \Sigma') \le \|\Sigma - \Sigma'\|_F^2 \le 5\lambda_{\max} d^2(\Sigma, \Sigma')$$

It remains to relate the first-order term $\langle Df(\Sigma), \Sigma' - \Sigma\rangle$ to $\langle \nabla f(\Sigma), \log_\Sigma(\Sigma')\rangle_\Sigma$. To this end, we introduce a lemma that captures the difference in geometry between the traditional Euclidean space and the BW geometry, which is

[2]We have also provided an alternative proof of the second inequality of Lemma 3.5 in Appendix F, with a minor improvement on the absolute constant.

essential for understanding the behavior of optimization algorithms when transitioning from Euclidean to BW geometry:

Lemma 3.6. *Given any two symmetric positive definite matrices $\Sigma, \Sigma' \in \mathbb{S}_+^n$ and a symmetric matrix $g \in \mathbb{R}^{n\times n}$, we have*

$$|\langle g, \Sigma - \Sigma'\rangle - \langle 2g, \log_\Sigma(\Sigma')\rangle_\Sigma| \le \|g\| \cdot d^2(\Sigma, \Sigma').$$

Given the identity $2Df(\Sigma) = \nabla_{\mathrm{BW}} f(\Sigma)$, Lemma 3.6 suggests that

$$\begin{aligned}&\big|\langle Df(\Sigma), \Sigma - \Sigma'\rangle - \langle \nabla f(\Sigma), \log_\Sigma(\Sigma')\rangle_\Sigma\big| \\ &\le \|Df\| \cdot d^2(\Sigma, \Sigma').\end{aligned}$$

This lemma indicates that the difference between the directional derivatives in the Euclidean and BW geometries is bounded by a term proportional to the squared BW distance. By combining this result with Lemma 3.5, we can provide a geodesic smoothness guarantee for f:

Proposition 3.7. *If function f is Euclidean β-smooth and Euclidean L-Lipschitz, then it holds that for any Σ, Σ'*

$$f(\Sigma') \le f(\Sigma) + \langle \nabla f(\Sigma), \log_\Sigma(\Sigma')\rangle_\Sigma + \frac{\beta'}{2} d^2(\Sigma, \Sigma')$$

with $\beta' = 5\beta \max\{\|\Sigma\|, \|\Sigma'\|\} + 2L$.

Noticing that when $\eta \le \frac{1}{20\beta\|\Sigma\|+16L}$, we have

$$\|\Sigma^+\| \le \|(I - \eta\nabla f)\|^2 \|\Sigma\| \le 4\|\Sigma\|,$$

Then we have by Proposition 3.7,

$$\begin{aligned}f(\Sigma^+) &\le f(\Sigma) - \eta\|\nabla f\|_\Sigma^2 + \frac{20\beta\|\Sigma\| + 2L}{2}\eta^2\|\nabla f\|_\Sigma^2 \\ &\le f(\Sigma) - \frac{\eta}{2}\|\nabla f\|_\Sigma^2.\end{aligned} \tag{11}$$

Remark 3.8. It is worth noting that the result in Lemma 3.6 and Proposition 3.7 can be interpreted as a special case that verifies condition A3 in (Boumal et al., 2019) within the BW geometry. Similarly, Lemma 4 in (Boumal et al., 2019), which ensures geodesic smoothness from Euclidean smoothness, serves a more general analogous to Proposition 3.7. However, Lemma 4 in (Boumal et al., 2019) assumes that the considered manifold is a sub-manifold of Euclidean space that inherits the Euclidean inner product, which is not the case in our setting. Therefore, new arguments utilizing the properties of the BW distance are provided in the proofs of Lemma 3.6 and Proposition 3.7.

Now combining Proposition 3.1, Lemma 3.4 and 11 we have the following contraction guarantee of BW gradient descent:

Theorem 3.9. *At every $\Sigma \in \mathbb{S}_+^n$, if we select $\eta \le \frac{1}{20\beta\|\Sigma\|+16L}$, then it holds that*

$$f(\Sigma^+) - f(\Sigma^\star) \le \big(1 - \frac{\alpha\lambda_{\min}(\Sigma)}{16}\eta\big)\big(f(\Sigma) - f(\Sigma^\star)\big)$$

In contrast to the classical linear convergence rates achieved by gradient descent for strongly convex and smooth functions, our contraction result introduces an additional dependence on $\lambda_{\min}(\Sigma)$, which may slow down the convergence rate as $\lambda_{\min}(\Sigma) \to 0$. It is therefore worth discussing whether this dependency can be removed. In fact, we have the following claim:

Proposition 3.10. *There exists a Euclidean strongly convex and smooth objective function f and an absolute constant $c > 0$ so that for every step-size $\eta > 0$ there exists a initial point Σ_0 so that the BWGD iteration starting from Σ_0 satisfies $\max_k \lambda_{\max}(\Sigma_k)/\lambda_{\min}(\Sigma_k) = 1$ and*

$$\liminf_k \lambda_{\min}^{-1}(\Sigma_k) \log\left(\frac{f(\Sigma_k) - f(\Sigma^\star)}{f(\Sigma_{k+1}) - f(\Sigma^\star)}\right) \le c.$$

This demonstrates that even if every matrix in the sequence is well-conditioned, the convergence rate may still depend on the minimal eigenvalue.

To explain Proposition 3.10, if we omit the $\liminf$ on the left-hand side for convenience and take the exponential of both sides of the inequality in Proposition 3.8, we obtain the following expression:

$$\begin{aligned}f(\Sigma_{k+1}) - f(\Sigma^\star) &\ge \exp(-c\lambda_{\min}(\Sigma_k))(f(\Sigma_k) - f(\Sigma^\star)) \\ &\ge (1 - c\lambda_{\min}(\Sigma_k))(f(\Sigma_k) - f(\Sigma^\star)),\end{aligned}$$

where we have used $\exp(-t) \ge 1 - t$ in the second inequality. This inequality implies that the contraction factor at the k-th step is lower bounded by $(1 - c\lambda_{\min}(\Sigma_k))$, which depends on $\lambda_{\min}(\Sigma_k)$ in a manner similar to our convergence rate result in Theorem 3.9. This indicates that our dependency on $\lambda_{\min}(\Sigma)$ is indeed tight.

Finally, it is worth mentioning that, based on Theorem 3.9, we can obtain the following linear convergence result by imposing a boundedness assumption on the eigenvalues of the iteration sequence generated by the BW gradient descent update:

Corollary 3.11. *Suppose there exists constants $0 < \lambda_{\min} < \lambda_{\max} < +\infty$ so that for the sequence $\{\Sigma_k\}_{k=1}^\infty$ generated by the BW gradient update satisfies $\lambda_{\min} I \preceq \Sigma_k \preceq \lambda_{\max} I$ for all k, then it holds that when selecting $\eta \le \frac{2}{20\beta\lambda_{\max}+16L}$,*

$$f(\Sigma_k) - f(\Sigma^\star) \le (1 - \frac{\alpha\lambda_{\min}}{16})^k f(\Sigma_0).$$

In particular, for the Gaussian barycenter problem, Lemma 1 in Altschuler et al. (2021) demonstrates that the eigenvalue upper and lower bounds required in Corollary 3.11 exist under mild assumptions (see also Section A for more detailed discussions). Thus, our result also implies a linear convergence rate for unconstrained Gaussian barycenter problems.

4. Analysis of BW Gradient Descent for Constrained Problems

4.1. Closed Form Projection to the BW Ball

Considering the computational feasibility of the projection operator is crucial when applying projected gradient methods. In the BW ball constrained problems, we need to consider the projection of any Σ onto the BW ball $\mathcal{W}(\Sigma_0, \rho)$, which is given by the following problem:

$$\begin{aligned} &\min_{\tilde{\Sigma} \in \mathcal{W}(\Sigma_0, \rho)} \quad d^2(\Sigma, \tilde{\Sigma}) \\ &\text{with} \quad \mathcal{W}(\Sigma_0, \rho) := \{\Sigma' : d^2(\Sigma', \Sigma_0) \leq \rho^2\}. \end{aligned} \tag{12}$$

Although (12) is a convex problem in Euclidean geometry due to the convexity of the squared BW distance, making it computationally feasible, solving it directly results in an n-dimensional semidefinite program (Appendix I.3), which can be relatively slow compared to projection-free methods. Our first observation is that problem (12) can be solved with a closed-form solution.

Proposition 4.1 (Closed form projection to BW ball)**.** *The problem* (12) *attains its minimizer* $P_{\mathcal{W}}(\Sigma)$ *at*

$$\gamma^2 \Sigma + (1-\gamma)^2 \Sigma_0 + (1-\gamma)\gamma\big(\Sigma_0 T_{\Sigma_0 \Sigma} + \Sigma T_{\Sigma \Sigma_0}\big) \tag{13}$$

where $\gamma = 1$ *when* $d^2(\Sigma, \Sigma_0) \leq \rho^2$*, and* $\gamma = \frac{\rho}{d(\Sigma, \Sigma_0)}$ *when* $d^2(\Sigma, \Sigma_0) > \rho^2$.

In the following text, we denote $\mathcal{W}(\Sigma_0, \rho)$ by $\mathcal{W}$, when there is no ambiguity. We have two remarks about the Proposition 4.1:

Computational Cost of the Projection We aim to evaluate the computational expense of the projection in comparison with the linear oracle employed in projection-free optimization methods. Specifically, when the Frank-Wolfe method is applied to minimizing some objective function $-f$ over $\mathcal{W}$, its each iteration necessitates solving the following sub-problem within the BW ball:

$$\max_{\Sigma \in \mathcal{W}} \langle \Sigma, \nabla f(\Sigma) \rangle \tag{14}$$

When $\nabla f(\Sigma) \preceq 0$—a condition holds in the least-favorable distribution seeking problems as in Shafieezadeh Abadeh et al. (2018); Nguyen et al. (2023); Taşkesen et al. (2023)—is satisfied, Taşkesen et al. (2023) has shown that (14) has a quasi closed-form solution, which can be obtained by taking matrix inversion and solving a one-dimensional convex problem, which can be solved highly efficiently via bisection. By contrast, the projection described in (4.1) includes a singular value decomposition operation when computing $T_{\Sigma\Sigma_0}$, which has a cost on par with matrix inversion, but no longer needs the one-dimensional optimization. Consequently, in this case, the computational cost of each projected gradient update is always not slower than that of the projection-free methods.

On the other hand, in the general scenario where $\nabla f(\Sigma)$ is not necessarily positive definite (for example, in the constrained barycenter problem), although Proposition A.2 of (Taşkesen et al., 2023) still shows that Equation (14) has a quasi-closed form solution, its associated one-dimensional problem is strictly more challenging than in the case where $\nabla f(\Sigma) \preceq 0$. Consequently, the bisection algorithm guarantee, developed by the authors, no longer holds, which may lead to a less efficient method for solving Equation (14). In comparison, our closed-form projection approach is still valid in such scenarios, thus offering superior computational efficiency.

Inconsistency between BW Projection and Euclidean Gradient Update Since our focus is on a Euclidean convex objective function, it is natural to inquire if the BW projection (12) can be smoothly incorporated into the Euclidean gradient update process. Specifically, whether the update step $\Sigma_{t+1} \leftarrow P_{\mathcal{W}}(\Sigma_t - \eta \nabla F(\Sigma_t))$ can have the desired global convergence guarantee. However, we demonstrate numerically that applying classical projected gradient descent (GD) convergence results to this setting is not straightforward. In section 5, our simulations using Euclidean gradient descent (EGD) update with projection (12) fail to converge to the global minimum, despite the objective function being strongly convex and smooth in the Euclidean sense.

4.2. Properties of the BW Projection onto Euclidean Convex Sets

In this section, we present several properties of the BW projection onto a general Euclidean convex set C. It is important to emphasize that while solving the projection problem is convex in the context of Euclidean geometry, the projected set C generally exhibits non-convexity in terms of geodesic distance within BW geometry (Bhatia et al., 2017). Our first result is the variational characterization of the BW projection

Proposition 4.2. *If* C *is a Euclidean convex closed subset of* $\mathbb{S}^n_+$*, then for any* $\Sigma \succ 0$*, we have* $\tilde{\Sigma} \in P_C(\Sigma)$ *if and only if it holds*

$$\begin{aligned} &\langle \log_{\tilde{\Sigma}}(\Sigma), \log_{\tilde{\Sigma}}(\Sigma') \rangle_{\tilde{\Sigma}} \\ &\leq -\frac{1}{2} \langle \log_{\tilde{\Sigma}}(\Sigma), \Sigma' + \tilde{\Sigma} - 2\tilde{\Sigma} T_{\tilde{\Sigma}, \Sigma'} \rangle \end{aligned} \tag{15}$$

for every $\Sigma' \in C$. In particular, we have

$$\begin{aligned}&\langle \log_{P_C(\Sigma)}(\Sigma), \log_{P_C(\Sigma)}(\Sigma')\rangle_{P_C(\Sigma)} \\ &\leq \frac{1}{2}\|\log_\Sigma P_C(\Sigma))\|d^2(\Sigma', P_C(\Sigma))\end{aligned} \tag{16}$$

Remark 4.3. When compared with the projection property for geodesically convex sets, as stated in the following lemma:

Lemma 4.4 ((Walter, 1974), Lemma 2). *If C is a geodesically convex and closed set, then for any $\Sigma' \in C$,*

$$\langle \log_{P_C(\Sigma)}(\Sigma), \log_{P_C(\Sigma)}(\Sigma')\rangle_{P_C(\Sigma)} \leq 0. \tag{17}$$

Unlike in (17), the right-hand side term in (15) might be positive due to the geodesic non-convexity of C. The magnitude of the right-hand side term can be interpreted as an indicator of how closely C approaches geodesic convexity. Furthermore, the inequality 16, which controls this magnitude, implies that it will converge to zero at a cubic rate as both $d(\Sigma, P_C(\Sigma))$ and $d(P_C(\Sigma), \Sigma')$ approach zero.

4.3. Convergence Analysis of Projected BW Gradient Descent

In this section, we provide our main results on the convergence of the projected BW gradient descent method for Euclidean strongly convex and smooth objective functions with the Euclidean convex constraint sets.

While it is also possible to derive the one-step contraction result as in Theorem 3.9, for the sake of analytical simplicity, we assume that the following bounded eigenvalue condition holds to derive the linear convergence guarantee as in Corollary 3.11. As will be demonstrated in section A, this condition can indeed be verified in both the constrained barycenter problem and the least-favorable distribution seeking problem, which are the foundational examples motivating our study.

Assumption 4.5. There exists constants $0 < \lambda_{\min} < \lambda_{\max} < +\infty$ so that for the sequence $\{\Sigma_k\}_{k=1}^\infty$ generated by the projected BW gradient update satisfies $\lambda_{\min} I \preceq \Sigma_k \preceq \lambda_{\max} I$ for all k. We denote $\kappa = \frac{\lambda_{\max}}{\lambda_{\min}}$.

One implication of Assumption 4.5, when combined with Proposition 3.7, is that if the function f is Euclidean β-smooth and L-Lipschitz, then it is geodesically $(5\beta\lambda_{\max} + 2L)$-smooth under the BW geometry. For simplicity, we denote $\beta' = 5\beta\lambda_{\max} + 2L$ in this section.

In the presence of the constraint set, the PL inequality and geodesic smoothness no longer guarantee global convergence of projected BW gradient descent due to the projection operator. Motivated by the classical approach as seen in Nesterov (2013); Bubeck et al. (2015), we define the following alternative to the gradient to measure the increment produced by each projected BW gradient descent update:

$$g_C(\Sigma) := \frac{1}{\eta}\log_\Sigma(\Sigma^+), \Sigma^+ := P_C\left(\exp_\Sigma(-\eta\nabla_{\mathrm{BW}} f(\Sigma))\right).$$

The quantity $g_C(\Sigma)$ describes the direction of the increment of the BW gradient descent (BWGD) update after projection. It can be seen that when the projection operator is not invoked, we have $g_C(\Sigma) = -\nabla_{\mathrm{BW}} f(\Sigma)$. We will show the convergence result by establishing the following variant of the PL inequality and the descent condition for $g_C(\Sigma)$:

$$\mu\big(f(\Sigma) - f(\Sigma^\star)\big) \leq \|g_C(\Sigma)\|_\Sigma^2, \tag{18}$$

$$f(\Sigma^+) - f(\Sigma) \leq -\zeta\|g_C(\Sigma)\|_\Sigma^2. \tag{19}$$

By firstly applying (19) to $f(\Sigma^+) - f(\Sigma)$, then invoking (18), it is straightforward to verify the following one-step contraction result holds:

Lemma 4.6. *Suppose* (18) *and* (19) *holds at some Σ with η, μ, ζ, then we have*

$$f(\Sigma^+) - f(\Sigma^\star) \leq (1 - \mu\zeta)(f(\Sigma^+) - f(\Sigma^\star))$$

Thus it is sufficient to verify (18) and (19).

4.3.1. Verifying the PL inequality

For the condition (18), we have the following result:

Lemma 4.7. *As long as Assumption 4.5 is satisfied, we have the condition* (18) *holds with*

$$\tilde{\mu} \gtrsim \left(\frac{(\beta'+1)\kappa^5}{\alpha} L\rho^2\kappa \max\{\lambda_{\min}^{-1}, \lambda_{\min}^{-1/2}\}(\eta^{-1}+\eta^3)\right)^{-1}$$

The proof of Lemma 4.7 is technical and lengthy; therefore, it is deferred to the appendix to maintain the readability and flow of the text. It is also noteworthy that, compared to the PL coefficient in the unconstrained case, the result in Lemma 4.7 exhibits a worse dependency on problem parameters. This is due to the increased difficulty in dealing with the projection and the constraint set when analyzing projected BW gradient descent.

4.3.2. Verifying the descent condition

To verify the descent condition, firstly noticing that by Proposition 3.7, we have

$$f(\Sigma^+) - f(\Sigma) \leq \langle \nabla f(\Sigma), \log_\Sigma(\Sigma^+)\rangle_\Sigma + \frac{\beta'\eta^2}{2}\|g(\Sigma)\|_\Sigma^2,$$

thus it is sufficient to show that there exists some large enough c so that $\langle \nabla f(\Sigma), \log_\Sigma(\Sigma^+)\rangle_\Sigma \leq -c\|g(\Sigma)\|_\Sigma^2$. If we denote $\tilde{\Sigma} := \exp_\Sigma(-\eta\nabla f(\Sigma))$, then we can check the decomposition

$$\eta\big(\langle \nabla f(\Sigma), \log_\Sigma(\Sigma^+)\rangle_\Sigma + \eta\|g_C(\Sigma)\|_\Sigma^2\big)$$

$$= \underbrace{\langle T_{\Sigma^+,\tilde{\Sigma}} - T_{\Sigma^+,\Sigma}T_{\Sigma,\tilde{\Sigma}}, \log_{\Sigma_+}(\Sigma)\rangle_{\Sigma^+}}_{:=I_1} + \underbrace{\langle \log_{\Sigma_+}(\tilde{\Sigma}), \log_{\Sigma_+}(\Sigma)\rangle_{\Sigma^+}}_{:=I_2}.$$

For I_1, we have

$$|I_1| \le \eta \|T_{\Sigma^+,\tilde{\Sigma}} - T_{\Sigma^+,\Sigma}T_{\Sigma,\tilde{\Sigma}}\|_{\Sigma^+}\|g_C(\Sigma)\|_\Sigma$$

For I_2, we have by Proposition 4.2,

$$\begin{aligned}|I_2| &\le \frac{1}{2}|\langle \log_{\Sigma_+}(\tilde{\Sigma}), \Sigma_+ + \Sigma - 2\Sigma_+ T_{\Sigma^+,\Sigma}\rangle| \\ &\le \sqrt{\kappa}L\eta^3\|g_C(\Sigma)\|_\Sigma^2,\end{aligned}$$

where the second line is by (16) and the following inequalities:

$$\lambda_{\min}^{1/2}\|\log_{\Sigma_+}(\tilde{\Sigma})\| \le d(\Sigma_+,\tilde{\Sigma}) \le d(\Sigma,\tilde{\Sigma}) \le \lambda_{\max}^{1/2} L\eta$$

Now combining our results for I_1, I_2 and the decomposition, we have

$$\begin{aligned}f(\Sigma^+) - f(\Sigma) \le (\frac{\beta'\eta^2}{2} + \sqrt{\kappa}L\eta^2 - \eta)\|g_C(\Sigma)\|_\Sigma^2 \\ + \|T_{\Sigma^+,\tilde{\Sigma}} - T_{\Sigma^+,\Sigma}T_{\Sigma,\tilde{\Sigma}}\|_{\Sigma^+}\|g_C(\Sigma)\|_\Sigma.\end{aligned}$$

While the first term is guaranteed to be less than or equal to $-\frac{\eta}{2}\|g_C(\Sigma)\|_\Sigma^2$ for $\eta \le \frac{1}{4}\min\{(\sqrt{\kappa}L)^{-1}, \beta'^{-1/2}\}$, we need to demonstrate that the second term can be dominated by the first for a sufficiently small η.

To this end, we establish the following general second-order bound for the optimal transport maps. This bound is derived through perturbation analysis of the Lyapunov equation (Hewer & Kenney, 1987) and may be of independent interest:

Lemma 4.8. *For any matrices A, B, and C satisfying $\lambda_{\min} I \preceq A, B, C \preceq \lambda_{\max} I$, the following inequality holds:*

$$\|T_{AB}T_{BC} - T_{AC}\|_A \le 8\kappa\lambda_{\min}^{-1/2} d(A,B)d(B,C).$$

This lemma provides a quantitative measure of the deviation between the direct transport map T_{AC} and the composition of two successive transport maps T_{AB} and T_{BC}, with respect to the metric induced by matrix A. Specifically, applying it with $A = \Sigma^+, B = \Sigma, C = \tilde{\Sigma}$ leads to

$$\begin{aligned}&\|T_{\Sigma^+,\tilde{\Sigma}} - T_{\Sigma^+,\Sigma}T_{\Sigma,\tilde{\Sigma}}\|_{\Sigma^+}\|g_C(\Sigma)\|_\Sigma \\ \le& 8\kappa\lambda_{\min}^{-1/2} d(\Sigma^+,\Sigma)d(\tilde{\Sigma},\Sigma)\|g_C(\Sigma)\|_\Sigma\end{aligned}$$

Now noticing that

$$d(\Sigma^+,\Sigma)d(\tilde{\Sigma},\Sigma) \le \lambda_{\max}^{1/2} L\eta^2 \|g_C(\Sigma)\|_\Sigma.$$

Thus the second term is bounded by $8\kappa^{3/2}L\eta^2\|g_C(\Sigma)\|_\Sigma^2$, in particular letting $\eta \le \frac{1}{32\kappa^{3/2}L}$, we have then

$$f(\Sigma^+) - f(\Sigma) \le -\frac{\eta}{4}\|g_C(\Sigma)\|_\Sigma^2.$$

We can summarize the above result as the following:

Lemma 4.9. *As long as Assumption 4.5 is satisfied and*

$$\eta \le \frac{1}{4}\min\left\{(\sqrt{\kappa}L)^{-1}, \beta'^{-1/2}, (4\kappa^{3/2}L)^{-1}\right\},$$

we have then the descent condition (19) holds with $\zeta = \eta/4$.

4.3.3. Linear Convergence Result

Combining Lemma 4.7 and Lemma 4.9, we have the following linear convergence guarantee of projected BW gradient descent:

Theorem 4.10. *As long as Assumption 4.5 is satisfied and*

$$\eta \le \frac{1}{4}\min\left\{(\sqrt{\kappa}L)^{-1}, \beta'^{-1/2}, (4\kappa^{3/2}L)^{-1}\right\},$$

thus with $\tilde{\mu}$ specified in Lemma 4.7, we have

$$f(\Sigma_k) - f(\Sigma^\star) \lesssim (1 - \eta\tilde{\mu}/16)^k \big(f(\Sigma_0) - f(\Sigma^\star)\big).$$

5. Numerical Results

In this section, we provide the convergence behavior of our projected BWGD algorithm under two different scenarios: WDRO-MMSE and constrained Gaussian barycenter. The detailed introduction and assumption verification of these two problems are in Appendix A.

For comparison, we also experiment with frank-wolfe (FW), fully adaptive frank-wolfe (FAFW), projected BWGD with heuristic Armijo search (Armijo BWGD), and Euclidean gradient descent (EGD) with projection (12).

For the WDRO-MMSE problem, we aim to solve the following problem:

$$\begin{aligned}\min_{\Sigma_x,\Sigma_w} \; &-\mathrm{tr}\big(\Sigma_x - \Sigma_x H^\top (H\Sigma_x H^\top + \Sigma_w)^{-1} H\Sigma_x\big) \\ \text{s.t. } &\Sigma_x \in \mathcal{W}(\widehat{\Sigma}_x, \rho_x), \Sigma_w \in \mathcal{W}(\widehat{\Sigma}_w, \rho_w).\end{aligned}$$

where dimension $n = 200$, Wasserstein radii $\rho_x = \rho_w = \sqrt{n}$, $\widehat{\Sigma}_x = U_x\Lambda_x U_x^T, \widehat{\Sigma}_w = U_w\Lambda_w U_w^T$ with U_x, U_w the orthonormal eigenvector of $Q_x + Q_x^T, Q_w + Q_w^T$ and Q_x, Q_w are sampled from standard normal distribution on $\mathbb{R}^{n\times n}$, Λ_x and Λ_w are diagonal matrices with elements uniformly sampled form $[1,5]$ and $[1,2]$.

For constrained barycenter, we aim to solve:

$$\min_{\Sigma\in\mathcal{W}(\widehat{\Sigma},\rho)} \sum_{i=1}^N \beta_i d^2(\Sigma,\Sigma_i).$$

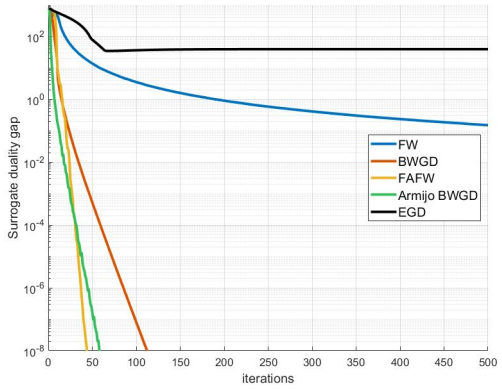

Figure 1. WDRO-MMSE

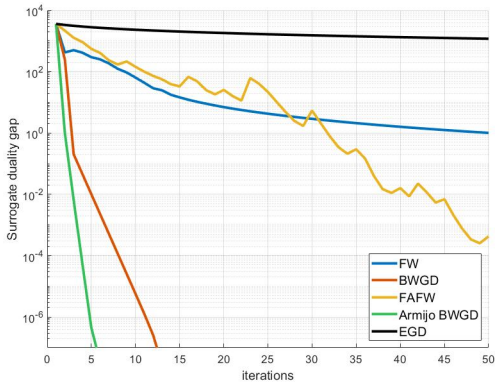

Figure 2. Constrained Barycenter

where $n = 30, \rho = 4\sqrt{n}, N = 50, \sum_{i=1}^{N} \beta_i = 1$. And $\beta_i = \frac{\alpha_i}{\sum_{i=1}^{N} \alpha_i}$ with α_i generated from $\alpha_i \sim \mathcal{U}(0,1)$, $\widehat{\Sigma} = U\Lambda_x U^T, \widehat{\Sigma}_i = U_i \Lambda_i U_i^T$ with U, U_i generated the same way as in WDRO-MMSE and elements of Λ_x and Λ_w are from $[1,2]$ and $[0.1, 100]$.

We run 5 experiments for WDRO-MMSE and 1 experiment for constrained barycenter and plot the surrogate duality gap (Appendix I.4) as the convergence criteria.

Figure 1 shows the result of WDRO-MMSE, and Figure 2 shows the result of the constrained barycenter. When choosing the vanilla step size, the projected BWGD is much faster than Frank-Wolfe. By incorporating the Armijo search technique Iusem (2003) into projected BWGD, we can achieve comparable results in WDRO-MMSE and perform significantly superior to FAFW in constrained barycenter, where FAFW shows fluctuation due to its dependency on three hyper-parameters. We leave the theoretical proof of convergence of Armijo BWGD to potential future directions.

Moreover, we comment that EGD exhibits poor convergence behavior and numerically shows the inconsistency between our BW projection and the Euclidean gradient descent as we mentioned in section 4.

Regarding the potential applications of the BWGD algorithm beyond the BW ball constraint, we added a new experiment involving the Wasserstein barycenter problem with the matrix interval constraint, as discussed in Weber & Sra (2022; 2023). A detailed introduction and the results of this experiment are presented in Appendix J.

We have provided the Matlab code to reproduce our numerical results[3] in https://github.com/Junyifannnn/ProjBWGD.

[3]require MOSEK as the SDP solver

6. Conclusion

In this work, we have established the convergence criteria of the projected BWGD under Euclidean strong convexity and smoothness. The key idea of our proof lies in drawing the connection between the Euclidean PL inequality and smoothness to their geodesic analogs, which may be of independent interest. Moreover, an analytical formula for the BW ball projection under the BW distance is provided and used to implement our algorithm in several numerical examples. We hope our study and the provided tools can lead to more insight into optimization under the BW distance and its connection to Euclidean geometry. Several questions are left for future study, including similar convergence guarantees when replacing Euclidean strong convexity and smoothness with weaker convexity and Lipschitz continuity, and the stochastic optimization analogue of our current results.

Acknowledgements

We would like to thank the anonymous reviewers for valuable suggestions and bringing important related works Boumal et al. (2019); Lanzetti et al. (2022); Weber & Sra (2022; 2023) to our attention. This work is generously supported by RGC CRF 8730063, RGC GRFs 16306821 and 16307023, and NSF grant CCF-2312205.

Impact Statement

Our theoretical work, which has applications in several specific problems, is designed to advance the field without any known harmful broader impacts. The insights and methodologies developed in this study are intended to contribute positively to optimization and related areas, providing valuable tools and frameworks for future research and practical applications.

References

Absil, P.-A., Mahony, R., and Sepulchre, R. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, 2008. ISBN 9781400830244.

Altschuler, J., Chewi, S., Gerber, P. R., and Stromme, A. Averaging on the bures-wasserstein manifold: dimension-free convergence of gradient descent. *Advances in Neural Information Processing Systems*, 34:22132–22145, 2021.

Bacak, M. *Convex Analysis and Optimization in Hadamard Spaces*. De Gruyter, Berlin, München, Boston, 2014. ISBN 9783110361629.

Bhatia, R., Jain, T., and Lim, Y. On the bures-wasserstein distance between positive definite matrices. *arXiv preprint arXiv:1712.01504*, 2017.

Bhatia, R., Jain, T., and Lim, Y. Strong convexity of sandwiched entropies and related optimization problems. *Reviews in Mathematical Physics*, 30(09):1850014, 2018.

Bhatia, R., Jain, T., and Lim, Y. On the bures–wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Boumal, N. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.

Boumal, N., Absil, P.-A., and Cartis, C. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1):1–33, 2019.

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

Chewi, S., Maunu, T., Rigollet, P., and Stromme, A. J. Gradient descent algorithms for bures-wasserstein barycenters. In *Conference on Learning Theory*, pp. 1276–1304. PMLR, 2020.

Criscitiello, C. and Boumal, N. An accelerated first-order method for non-convex optimization on manifolds. *Foundations of Computational Mathematics*, 23(4):1433–1509, 2023.

Diao, M. Z. *Proximal Gradient Algorithms for Gaussian Variational Inference: Optimization in the Bures–Wasserstein Space*. PhD thesis, Massachusetts Institute of Technology, 2023.

Diao, M. Z., Balasubramanian, K., Chewi, S., and Salim, A. Forward-backward gaussian variational inference via jko in the bures-wasserstein space. In *International Conference on Machine Learning*, pp. 7960–7991. PMLR, 2023.

Gao, R. and Kleywegt, A. Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655, 2023.

Guillaumin, M., Verbeek, J., and Schmid, C. Is that you? metric learning approaches for face identification. In *2009 IEEE 12th international conference on computer vision*, pp. 498–505. IEEE, 2009.

Ha Quang, M., San Biagio, M., and Murino, V. Log-hilbert-schmidt metric between positive definite operators on hilbert spaces. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

Han, A. and Gao, J. Improved variance reduction methods for riemannian non-convex optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44 (11):7610–7623, 2021.

Han, A., Mishra, B., Jawanpuria, P. K., and Gao, J. On riemannian optimization over positive definite matrices with the bures-wasserstein geometry. *Advances in Neural Information Processing Systems*, 34:8940–8953, 2021.

Hewer, G. and Kenney, C. The sensitivity of the stable lyapunov equation. In *26th IEEE Conference on Decision and Control*, volume 26, pp. 2122–2122, 1987.

Iusem, A. On the convergence properties of the projected gradient method for convex optimization. *Computational and Applied Mathematics*, 22:37–52, 01 2003.

Kim, J. and Yang, I. Accelerated gradient methods for geodesically convex optimization: Tractable algorithms and convergence analysis. In *International Conference on Machine Learning*, pp. 11255–11282. PMLR, 2022.

Kuhn, D., Esfahani, P. M., Nguyen, V. A., and Shafieezadeh-Abadeh, S. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations research & management science in the age of analytics*, pp. 130–166. Informs, 2019.

Lambert, M., Chewi, S., Bach, F., Bonnabel, S., and Rigollet, P. Variational inference via wasserstein gradient flows. *Advances in Neural Information Processing Systems*, 35: 14434–14447, 2022.

Lanzetti, N., Bolognani, S., and Dörfler, F. First-order conditions for optimization in the wasserstein space. *arXiv preprint arXiv:2209.12197*, 2022.

Lenglet, C., Rousson, M., Deriche, R., and Faugeras, O. Statistics on the manifold of multivariate normal distributions: Theory and application to diffusion tensor mri processing. *J. Math. Imaging Vis.*, 25(3):423–444, oct 2006. ISSN 0924-9907.

Lin, Z. Riemannian geometry of symmetric positive definite matrices via cholesky decomposition. *SIAM Journal on Matrix Analysis and Applications*, 40(4):1353–1370, 2019.

Luo, Y. and Trillos, N. G. Nonconvex matrix factorization is geodesically convex: Global landscape analysis for fixed-rank matrix optimization from a riemannian perspective. *arXiv preprint arXiv:2209.15130*, 2022.

Mathias, R. Perturbation bounds for the polar decomposition. *SIAM Journal on Matrix Analysis and Applications*, 14(2):588–597, 1993. doi: 10.1137/0614041.

Maunu, T., Le Gouic, T., and Rigollet, P. Bures-wasserstein barycenters and low-rank matrix recovery. In *International Conference on Artificial Intelligence and Statistics*, pp. 8183–8210. PMLR, 2023.

Moral, P. D. and Niclas, A. A taylor expansion of the square root matrix functional. *arXiv preprint arXiv:1705.08561*, 2018.

Nesterov, Y. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Nguyen, V. A., Shafieezadeh-Abadeh, S., Kuhn, D., and Mohajerin Esfahani, P. Bridging bayesian and minimax mean square error estimation via wasserstein distributionally robust optimization. *Mathematics of Operations Research*, 48(1):1–37, 2023.

Pennec, X. Manifold-valued image processing with spd matrices. In Pennec, X., Sommer, S., and Fletcher, T. (eds.), *Riemannian Geometric Statistics in Medical Image Analysis*, pp. 75–134. Academic Press, 2020. ISBN 978-0-12-814725-2.

Pennec, X., Fillard, P., and Ayache, N. A riemannian framework for tensor computing. *International Journal of computer vision*, 66:41–66, 2006.

Shafieezadeh Abadeh, S., Nguyen, V. A., Kuhn, D., and Mohajerin Esfahani, P. M. Wasserstein distributionally robust kalman filtering. *Advances in Neural Information Processing Systems*, 31, 2018.

Sra, S. A new metric on the manifold of kernel matrices with application to matrix geometric means. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

Suárez, J. L., García, S., and Herrera, F. A tutorial on distance metric learning: Mathematical foundations, algorithms, experimental analysis, prospects and challenges. *Neurocomputing*, 425:300–322, 2021.

Taşkesen, B., Iancu, D. A., Koçyiğit, Ç., and Kuhn, D. Distributionally robust linear quadratic control. *arXiv preprint arXiv:2305.17037*, 2023.

Tsuda, K., Rätsch, G., and Warmuth, M. K. Matrix exponentiated gradient updates for on-line learning and bregman projection. *Journal of Machine Learning Research*, 6 (Jun):995–1018, 2005.

Udriste, C. *Convex functions and optimization methods on Riemannian manifolds*, volume 297. Springer Science & Business Media, 2013.

Walter, R. On the metric projection onto convex sets in riemannian spaces. *Archiv der Mathematik*, 25(1):91–98, 1974.

Weber, M. and Sra, S. Projection-free nonconvex stochastic optimization on riemannian manifolds. *IMA Journal of Numerical Analysis*, 42(4):3241–3271, 2022.

Weber, M. and Sra, S. Riemannian optimization via frank-wolfe methods. *Mathematical Programming*, 199(1-2): 525–556, 2023.

Yue, M., Kuhn, D., and Wiesemann, W. On linear optimization over wasserstein balls. *Mathematical Programming*, June 2021. ISSN 0025-5610. doi: 10.1007/s10107-021-01673-8. Funding Information: The authors gratefully acknowledge funding from the Swiss National Science Foundation under Grant BSCGI0157733, the UK's Engineering and Physical Sciences Research Council under Grant EP/R045518/1 and the Hong Kong Research Grants Council under the Grant 25302420. Publisher Copyright: © 2021, Springer-Verlag GmbH Germany, part of Springer Nature and Mathematical Optimization Society.

Zhang, H. and Sra, S. First-order methods for geodesically convex optimization. In *Conference on Learning Theory*, pp. 1617–1638. PMLR, 2016.

Zhang, H. and Sra, S. An estimate sequence for geodesically convex optimization. In *Conference On Learning Theory*, pp. 1703–1723. PMLR, 2018.

A. Background Examples

In this section, we provide two problems that fit our general framework and have their own importance in their area.

(Constrained) Gaussian Barycenter Problem In the BW barycenter problem, we are given a mixture distribution P over mean-zero Gaussian distributions, and our objective is to find the barycenter distribution

$$\mu^\star \in \text{argmin}_\mu \int W^2(\mu, \mu_0) dP(\mu_0). \tag{20}$$

Proposition 15 in Chewi et al. (2020) demonstrates that if the covariance matrices $\Sigma(\mu_0)$ of all $\mu_0 \in \text{supp}(P)$ are uniformly bounded above and below by constants $0 < \kappa_l \leq \kappa_u < +\infty : \kappa_l I \preceq \Sigma(\mu_0) \preceq \kappa_u I$, then $\mu^\star$ must also be Gaussian. Consequently, problem (20) simplifies to finding the optimal covariance matrix

$$\Sigma^\star \in \text{argmin}_{\Sigma \in \mathbb{S}^n_+} \int W^2(\Sigma, \Sigma_0) dP(\Sigma_0). \tag{21}$$

When it is crucial to maintain a predefined proximity to a certain distribution(for instance, if we are working with a series of medical images where the barycenter should not deviate significantly from the standard human anatomy), a hard constraint on the Wasserstein distance is introduced, leading to the following constrained barycenter problem:

$$\Sigma^\star \in \text{argmin}_{\Sigma \in \text{BW}(\tilde{\Sigma}; \rho)} \int \text{BW}^2(\Sigma, \Sigma_0) dP(\Sigma_0), \tag{22}$$

with $\text{BW}(\tilde{\Sigma}; \rho)$ denoting the set of covariance matrices within the Wasserstein distance ρ from a specified matrix $\tilde{\Sigma}$.

WDRO-MMSE Problem In the MMSE problem, one needs to estimate unknown parameter $x \in \mathbb{R}^n$ based on its noisy measurement

$$y = Hx + w \in \mathbb{R}^m$$

where w is the random noise and the observation matrix $H \in \mathbb{R}^{m \times n}$ is assumed to be known. The quality of an estimator $\psi : \mathbb{R}^m \to \mathbb{R}^n$ is measured by the mean squared error risk:

$$R(\psi, x) = \mathbb{E}_{\mathbb{P}_{y|x}} \left[\|x - \psi(y)\|^2 \right],$$

and when x is given from some prior distribution P_x, the MMSE problem aims to find

$$\psi^\star = \min_\psi \mathbb{E}_{P_x}[R(\psi, x)]. \tag{23}$$

The WDRO MMSE aims to find the robust version of (23), where both $\mathbb{P}_{y|x}$ and $\mathbb{P}_x$ lies in the Wasserstein ambiguity sets $\mathcal{P}$ and the learner wants to optimize the worst-case performance

$$\psi^\star = \min_\psi \max_{P_x, P_{y|x} \in \mathcal{P}} \mathbb{E}_{P_x}[R(\psi, x)]. \tag{24}$$

In particular, if we consider the case that both P_x and $P_{y|x}$ are Gaussian and $\mathcal{P}$ is given by the Wasserstein ball, then it has been shown in Nguyen et al. (2023)(24) can be solved by first finding the least-favorable distribution $P_x^\star, P_{y|x}^\star$, then solving (23) with $P_x^\star, P_{y|x}^\star$. And Theorem 3.5 of Nguyen et al. (2023) shows the least-favorable distribution seeking problem can be formulated as the following BW ball constrained problem

$$\min_{\Sigma_x, \Sigma_w} -\text{tr}\big(\Sigma_x - \Sigma_x H^\top (H\Sigma_x H^\top + \Sigma_w)^{-1} H \Sigma_x\big) \text{ s.t. } \Sigma_x \in \mathcal{W}(\widehat{\Sigma}_x, \rho_x), \Sigma_w \in \mathcal{W}(\widehat{\Sigma}_w, \rho_w) \tag{25}$$

While this formulation involves an optimization problem over two positive-definite matrices, each subject to an independent BW ball constraint, our theory and algorithm have been developed for a single positive-definite matrix. However, it is relatively straightforward to consider the product geometry and extend our framework, provided that the objective function is jointly strongly convex and smooth with respect to both Σ_x and Σ_w.

We would show that both examples satisfies Assumption 4.5 (see Appendix I)

Proposition A.1. *The constrained barycenter problem satisfies Assumption 4.5*

Proposition A.2. *The WDRO MMSE problem satisfies Assumption 4.5*

B. Proof of Proposition 4.1

Consider the following BW-ball projection problem:

$$g(\hat{\Sigma}, X) = \min_{S \in \mathcal{S}_+^p} \quad d(S, X),$$
$$\text{s.t.} \quad d^2(S, \hat{\Sigma}) \leq \rho^2.$$

which is equivalent to

$$g(\hat{\Sigma}, X) = \min_{S \in \mathcal{S}_+^p} \quad \text{tr}(S) + \text{tr}(X) - 2\text{tr}\big(\sqrt{X^{\frac{1}{2}} S X^{\frac{1}{2}}}\big),$$
$$\text{s.t.} \quad \text{tr}(S) + \text{tr}(\hat{\Sigma}) - 2\text{tr}\big(\sqrt{\hat{\Sigma}^{\frac{1}{2}} S \hat{\Sigma}^{\frac{1}{2}}}\big) \leq \rho^2.$$

With Lagrangian function and Slater's condition, we have

$$\begin{aligned} g(\hat{\Sigma}, X) &= \min_{S \in \mathcal{S}_+^p} \max_{\gamma \geq 0} \text{tr}(S) + \text{tr}(X) - 2\text{tr}\big(\sqrt{X^{\frac{1}{2}} S X^{\frac{1}{2}}}\big) + \gamma(\text{tr}(S) + \text{tr}(\hat{\Sigma}) - 2\text{tr}\big(\sqrt{\hat{\Sigma}^{\frac{1}{2}} S \hat{\Sigma}^{\frac{1}{2}}}\big) - \rho^2) \\ &= \max_{\gamma \geq 0} \min_{S \in \mathcal{S}_+^p} \text{tr}(S) + \text{tr}(X) - 2\text{tr}\big(\sqrt{X^{\frac{1}{2}} S X^{\frac{1}{2}}}\big) + \gamma(\text{tr}(S) + \text{tr}(\hat{\Sigma}) - 2\text{tr}\big(\sqrt{\hat{\Sigma}^{\frac{1}{2}} S \hat{\Sigma}^{\frac{1}{2}}}\big) - \rho^2) \\ &= \max_{\gamma \geq 0} \gamma(\text{tr}(\hat{\Sigma}) - \rho^2) + \min_{S \in \mathcal{S}_+^p} \langle S, (1+\gamma) I \rangle - 2\text{tr}\big(\sqrt{X^{\frac{1}{2}} S X^{\frac{1}{2}}}\big) - 2\gamma \text{tr}\big(\sqrt{\hat{\Sigma}^{\frac{1}{2}} S \hat{\Sigma}^{\frac{1}{2}}}\big) \end{aligned}$$

We first solve the subproblem

$$\min_{S \in \mathcal{S}_+^p} \langle S, (1+\gamma) I \rangle - 2\text{tr}\big(\sqrt{X^{\frac{1}{2}} S X^{\frac{1}{2}}}\big) - 2\gamma \text{tr}\big(\sqrt{\hat{\Sigma}^{\frac{1}{2}} S \hat{\Sigma}^{\frac{1}{2}}}\big). \tag{26}$$

Taking derivative with respect to S, the optimal S satisfies

$$(1+\gamma) I = X^{\frac{1}{2}} \big(\sqrt{X^{\frac{1}{2}} S X^{\frac{1}{2}}}\big)^{-\frac{1}{2}} X^{\frac{1}{2}} + \gamma \hat{\Sigma}^{\frac{1}{2}} \big(\sqrt{\hat{\Sigma}^{\frac{1}{2}} S \hat{\Sigma}^{\frac{1}{2}}}\big)^{-\frac{1}{2}} \hat{\Sigma}^{\frac{1}{2}}.$$

which is equivalent to

$$S = \frac{1}{1+\gamma} \big(S^{\frac{1}{2}} X S^{\frac{1}{2}}\big)^{\frac{1}{2}} + \frac{\gamma}{1+\gamma} \big(S^{\frac{1}{2}} \hat{\Sigma} S^{\frac{1}{2}}\big)^{\frac{1}{2}}.$$

This equation is a special Wasserstein barycenter problem with explicit solution Bhatia et al. (2017)

$$S = \frac{1}{(1+\gamma)^2} X + \frac{\gamma^2}{(1+\gamma)^2} \hat{\Sigma} + \frac{\gamma}{(1+\gamma)^2} \big(X T_{X\hat{\Sigma}} + \hat{\Sigma} T_{\hat{\Sigma} X}\big). \tag{27}$$

Combine (26), (27),

$$\begin{aligned} &\max_{\gamma \geq 0} \gamma(\text{tr}(\hat{\Sigma}) - \rho^2) + \min_{S \in \mathcal{S}_+^p} \langle S, (1+\gamma) I \rangle - 2\text{tr}\big(\sqrt{X^{\frac{1}{2}} S X^{\frac{1}{2}}}\big) - 2\gamma \text{tr}\big(\sqrt{\hat{\Sigma}^{\frac{1}{2}} S \hat{\Sigma}^{\frac{1}{2}}}\big) \\ =& \max_{\gamma \geq 0} \gamma(\text{tr}(\hat{\Sigma}) - \rho^2) - \frac{1}{1+\gamma} \text{tr}(X) - \frac{\gamma^2}{1+\gamma} \text{tr}(\hat{\Sigma}) - \frac{\gamma}{1+\gamma} \text{tr}\big(X T_{X\hat{\Sigma}} + \hat{\Sigma} T_{\hat{\Sigma} X}\big). \end{aligned}$$

Taking derivative with respect to γ, we have

$$(1+\gamma)^2 = \frac{d^2(X, \hat{\Sigma})}{\rho^2}.$$

If $d(X, \hat{\Sigma}) > \rho$, $\gamma = \frac{d(X, \hat{\Sigma})}{\rho} - 1$, otherwise $\gamma = 0$.

Hence the solution of Bures-Wasserstein projection $\hat{S}$

$$\hat{S} = \arg\min_{S \in \mathcal{S}_+^p} \quad d(S, X),$$

$$\text{s.t.} \quad d^2(S, \hat{\Sigma}) \leq \rho^2.$$

has closed-form solution

$$\hat{S} = \begin{cases} \frac{\rho^2}{d^2(X,\hat{\Sigma})} X + (1 - \frac{\rho}{d(X,\hat{\Sigma})})^2 \hat{\Sigma} + (1 - \frac{\rho}{d(X,\hat{\Sigma})}) \frac{\rho}{d(X,\hat{\Sigma})} \big(X T_{X\hat{\Sigma}} + \hat{\Sigma} T_{\hat{\Sigma}X} \big), & d(X, \hat{\Sigma}) > \rho, \\ X, & d(X, \hat{\Sigma}) \leq \rho. \end{cases}$$

C. Proof of Proposition 4.2

Noticing that for given $\hat{\Sigma}, \Sigma_0$, the projection problem

$$\min_{\Sigma \in \mathbb{S}^d_{++}} d^2(\Sigma, \Sigma_0) \quad \text{s.t. } d^2(\Sigma, \hat{\Sigma}) \leq \rho$$

is a convex optimization problem in Euclidean geometry. If we denote $f(\Sigma) := d^2(\Sigma, \Sigma_0)$ and $\mathcal{C}$ the $\hat{\Sigma}$-centered BW ball with radius ρ, we have the minimizer $P_C(\Sigma_0)$ of above problem satisfies the characterization

$$\langle Df(P_C(\Sigma_0)), P_C(\Sigma_0) - \Sigma \rangle \leq 0, \quad \forall \Sigma \in C.$$

Noticing that

$$Df(\Sigma) = D_\Sigma d^2(\Sigma, \Sigma_0) = -\log_\Sigma(\Sigma_0),$$

we have by (29)

$$\begin{aligned} \langle Df(P_C(\Sigma_0)), P_C(\Sigma_0) - \Sigma \rangle &= \langle \log_{P_C(\Sigma_0)}(\Sigma_0), \Sigma - P_C(\Sigma_0) \rangle \\ &= 2 \langle \log_{P_C(\Sigma_0)}(\Sigma_0), \log_{P_C(\Sigma_0)}(\Sigma) \rangle_{P_C(\Sigma_0)} \\ &\quad + \langle \log_{P_C(\Sigma_0)}(\Sigma_0), \Sigma + P_C(\Sigma_0) - 2P_C(\Sigma_0) T_{P_C(\Sigma_0),\Sigma} \rangle. \end{aligned}$$

That leads to the following characterization equation of the BW ball projection:

$$\langle \log_{P_C(\Sigma_0)}(\Sigma_0), \log_{P_C(\Sigma_0)}(\Sigma) \rangle_{P_C(\Sigma_0)} \leq \underbrace{-\frac{1}{2} \langle \log_{P_C(\Sigma_0)}(\Sigma_0), \Sigma + P_C(\Sigma_0) - 2P_C(\Sigma_0) T_{P_C(\Sigma_0),\Sigma} \rangle}_{:= \kappa(\Sigma_0, \Sigma)} \tag{28}$$

D. Proof of Lemma 3.6

For any symmetric G, we have

$$\begin{aligned} \langle G, \Sigma' - \Sigma \rangle &= \langle G, \Sigma(\Sigma^{-1}\Sigma' - I) \rangle \\ &= \langle G, 2\Sigma(T_{\Sigma,\Sigma'} - I) \rangle + \langle \nabla G, \Sigma\big(\Sigma^{-1}\Sigma' + I - 2T_{\Sigma,\Sigma'}\big) \rangle \\ &= \langle 2G, \log_\Sigma(\Sigma') \rangle_\Sigma + \langle G, \Sigma' - \Sigma + 2\Sigma - 2\Sigma T_{\Sigma,\Sigma'} \rangle, \end{aligned} \tag{29}$$

that shows the identity, to show the inequality, just noticing that for any A symmetric, $B \succeq 0$ we have $\big|\text{tr}(AB)\big| \leq \|A\| \text{tr}(B)$ and the fact

$$\big(\Sigma' - \Sigma + 2\Sigma - 2\Sigma T_{\Sigma,\Sigma'}\big) + \big(\Sigma' - \Sigma + 2\Sigma - 2\Sigma T_{\Sigma,\Sigma'}\big)^\top \succeq 0$$

are shown by

$$\begin{aligned} &\Sigma + \Sigma' - \Sigma T_{\Sigma,\Sigma'} - T_{\Sigma,\Sigma'}\Sigma \\ =&\Sigma^{-1/2}(\Sigma^2 + \Sigma^{1/2}\Sigma'\Sigma^{1/2} - \Sigma^{3/2} T_{\Sigma,\Sigma'} \Sigma^{1/2} - \Sigma^{1/2} T_{\Sigma,\Sigma'} \Sigma^{3/2}) \Sigma^{-1/2} \\ =&\Sigma^{-1/2}(\Sigma - \Sigma^{1/2} T_{\Sigma,\Sigma'} \Sigma^{1/2})(\Sigma - \Sigma^{1/2} T_{\Sigma,\Sigma'} \Sigma^{1/2}) \Sigma^{-1/2} \succeq 0. \end{aligned}$$

E. Verifying The PL Inequality

To verify the PL inequality, we would first establish a dominant result of $\|g_C(\Sigma)\|_\Sigma^2$ with respect to the distance-to-optimal $d(\Sigma, \Sigma^\star)$:

Proposition E.1. *Suppose f is Euclidean L-Lipschitz,α-strongly convex,and β-smooth, C is Euclidean convex and Assumption 4.5 is satisfied, then it holds that*

$$d^2(\Sigma, \Sigma^\star) \leq \big((2 + \frac{10\beta\kappa}{2\alpha})\eta^2 + \frac{25\kappa^5}{4\alpha} \cdot \lambda_{\min}^{-2}\big)\|g_C(\Sigma)\|_\Sigma^2$$

On the other hand, we have the following lemma resulted in the non-negative curvature of the BW geometry (Chewi et al., 2020; Altschuler et al., 2021):

Lemma E.2. *For any $A, B, C \succ 0$, it holds that*

$$d_{BW}(A, B) \leq \|\log_C(A) - \log_C(B)\|_C \tag{30}$$

If we denote $\theta_1 = 2 + \frac{10\beta\kappa}{2\alpha}, \theta_2 = \frac{25\kappa^5}{4\alpha} \cdot \lambda_{\min}^{-2}$, then applying the non-negative curvature property in (30) with $A = \Sigma^\star, B = \Sigma^+, C = \Sigma$ implies

$$2\langle \log_\Sigma(\Sigma^+), \log_\Sigma(\Sigma^\star)\rangle_\Sigma \leq d^2(\Sigma, \Sigma^\star) + d^2(\Sigma, \Sigma^+) \lesssim (\theta_1\eta^2 + \theta_2)\|g_C(\Sigma)\|_\Sigma^2, \tag{31}$$

where in the second line we have used Proposition E.1 and the fact $\|g_C(\Sigma)\|_\Sigma^2 = \eta^2 d^2(\Sigma, \Sigma^+)$.

On the other hand, noticing that by Euclidean convexity of f, we have

$$\begin{aligned} f(\Sigma) - f(\Sigma^\star) &\leq \langle Df(\Sigma), \Sigma^\star - \Sigma\rangle \\ &\leq \langle \nabla f(\Sigma), \log_\Sigma(\Sigma^\star)\rangle_\Sigma \\ &\quad + \|Df(\Sigma)\| d^2(\Sigma, \Sigma^\star), \end{aligned}$$

where in the second inequality we have used Proposition 3.6. Since the second term in above inequality can be bounded by $O(L(\theta_1\eta^2 + \theta_2)\|g_C(\Sigma)\|_\Sigma^2$ by the Lipschtiz property of f and the Proposition E.1, it remains to bound $\langle \nabla f(\Sigma), \log_\Sigma(\Sigma^\star)\rangle_\Sigma$ to show the PL condition, actually, we can bound the gap between $\langle \nabla f(\Sigma), \log_\Sigma(\Sigma^\star)\rangle_\Sigma$ and $\langle \log_\Sigma(\Sigma^+), \log_\Sigma(\Sigma^\star)\rangle_\Sigma$ in the following statement:

Lemma E.3. *Under the same condition as in Proposition E.1, we have*

$$\begin{aligned} &|\langle \nabla f(\Sigma), \log_\Sigma(\Sigma^\star)\rangle_\Sigma + \frac{1}{\eta}\langle \log_\Sigma(\Sigma^+), \log_\Sigma(\Sigma^\star)\rangle_\Sigma| \\ &\leq 256(\eta + 1)L \max\{\lambda_{\min}^{-1}, \lambda_{\min}^{-1/2}\}\rho^2\kappa(2\theta_1\eta^2 + 2\theta_2)\|g_C(\Sigma)\|_\Sigma^2. \end{aligned}$$

Now applying (31) and Lemma E.3, we have there exists some absolute constant c_0 so that (18) holds with

$$\mu = c_0\bigg(\theta_2\eta^{-1} + (\eta + 1)L \max\{\lambda_{\min}^{-1}, \lambda_{\min}^{-1/2}\}\rho^2\kappa(\theta_1\eta^2 + \theta_2)\bigg)^{-1}\|g_C(\Sigma)\|_\Sigma^2.$$

Moreover, noticing that

$$(\eta + 1)(\theta_1\eta^2 + \theta_2) \leq 4(\theta_1 + \theta_2)(\eta^3 + 1) \leq 8(\theta_1 + \theta_2)(\eta^3 + \eta^{-1})$$

and $(\theta_1 + \theta_2) \leq 256\frac{(\beta' + 1)\kappa^5}{\alpha}$ we get there exists some absolute constant c_0' so that

$$\mu = c_0'\frac{(\beta' + 1)\kappa^5}{\alpha}L\rho^2\kappa \max\{\lambda_{\min}^{-1}, \lambda_{\min}^{-1/2}\}(\eta^{-1} + \eta^3)$$

E.1. Proof of Proposition E.1

Proof. Noticing that by Euclidean convexity, we have

$$\langle Df(\Sigma^\star), \Sigma^\star - \Sigma\rangle \leq 0, \forall \Sigma \in C.$$

Thus it holds that

$$\begin{aligned} f(\Sigma^+) &\geq f(\Sigma^\star) + \langle Df(\Sigma^\star), \Sigma^+ - \Sigma^\star\rangle + \frac{\alpha}{2}\|\Sigma^+ - \Sigma^\star\|_F^2 \\ &\geq f(\Sigma^\star) + 2\alpha\lambda_{\min} d^2(\Sigma^+, \Sigma^\star), \end{aligned}$$

Thus it holds that $d^2(\Sigma^+, \Sigma^\star) \leq \frac{\Delta^+}{2\alpha\lambda_{\min}}$ for $\Delta^+ := f(\Sigma^+) - f(\Sigma^\star)$.

On the other hand, we have

$$\Delta^+ \leq \alpha\lambda_{\min} d^2(\Sigma^+, Y) + \big(\frac{5\beta\lambda_{\max}}{2} + \frac{25\kappa^5}{8\alpha\lambda_{\min}} \cdot \frac{1}{\eta^2} d^2(\Sigma^+, \Sigma)\big) \tag{32}$$

That leads to

$$d^2(\Sigma^+, \Sigma^\star) \leq \big(\frac{5\beta\kappa}{2\alpha} + \frac{25\kappa^5}{8\alpha} \cdot \lambda_{\min}^{-2} \cdot \frac{1}{\eta^2}\big) d^2(\Sigma^+, \Sigma).$$

which then implies the desired result by

$$\begin{aligned} d^2(\Sigma, \Sigma^\star) &\leq 2d^2(\Sigma, \Sigma^+) + 2d^2(\Sigma^+, \Sigma^\star) \\ &\leq \big(2 + \frac{10\beta\kappa}{2\alpha} + \frac{25\kappa^5}{4\alpha} \cdot \lambda_{\min}^{-2} \cdot \frac{1}{\eta^2}\big) d^2(\Sigma^+, \Sigma) \\ &\leq \big((2 + \frac{10\beta\kappa}{2\alpha})\eta^2 + \frac{25\kappa^5}{4\alpha} \cdot \lambda_{\min}^{-2}\big) \|g_C(\Sigma)\|_\Sigma^2 \end{aligned}$$

Proof of (32): For any $Y \in C$, the characterization of projection implies that

$$-\langle \log_{\Sigma^+}(\tilde{\Sigma}), \Sigma^+ - Y\rangle \leq 0.$$

on the other hand, the Euclidean convexity implies

$$\begin{aligned} f(\Sigma^+) - f(Y) &= f(\Sigma^+) - f(\Sigma) + f(\Sigma) - f(Y) \\ &\leq \langle Df(\Sigma), \Sigma^+ - \Sigma\rangle + \frac{\beta}{2}\|\Sigma^+ - \Sigma\|_F^2 - \langle Df(\Sigma), Y - \Sigma\rangle - \frac{\alpha}{2}\|\Sigma - Y\|_F^2 \\ &\leq \langle Df(\Sigma), \Sigma^+ - Y\rangle + \frac{\beta}{2}\|\Sigma^+ - \Sigma\|_F^2 \\ &= -\frac{1}{\eta}\langle \log_\Sigma(\tilde{\Sigma}), \Sigma^+ - Y\rangle + \frac{\beta}{2}\|\Sigma^+ - \Sigma\|_F^2 \\ &\leq -\frac{1}{\eta}\langle \log_\Sigma(\tilde{\Sigma}) - \log_{\Sigma^+}(\tilde{\Sigma}), \Sigma^+ - Y\rangle + \frac{\beta}{2}\|\Sigma^+ - \Sigma\|_F^2 \\ &\leq \frac{1}{\eta}\|\log_\Sigma(\tilde{\Sigma}) - \log_{\Sigma^+}(\tilde{\Sigma})\|_F \|\Sigma^+ - Y\|_F + \frac{5\beta\lambda_{\max}}{2} d^2(\Sigma^+, \Sigma) \\ &\leq \frac{5\sqrt{2}}{2\eta}\kappa^{5/2} d(\Sigma^+, \Sigma) d(\Sigma^+, Y) + \frac{5\beta\lambda_{\max}}{2} d^2(\Sigma^+, \Sigma) \\ &\leq \mu\lambda_{\min} d^2(\Sigma^+, Y) + \big(\frac{5\beta\lambda_{\max}}{2} + \frac{25\kappa^5}{8\alpha\lambda_{\min}} \cdot \frac{1}{\eta^2} d^2(\Sigma^+, \Sigma)\big) \end{aligned}$$

Where we use Cauchy-Schwarz inequality in the last line. Setting $Y = \Sigma^\star$ leads to the result. □

E.2. Proof the Proposition E.3

Denoting $\tilde{\Sigma} := \exp_{\Sigma}(-\eta \nabla f(\Sigma))$, then

In fact, we have by (29),

$$\begin{aligned}-\eta\langle \nabla f(\Sigma), \log_{\Sigma}(\Sigma^\star)\rangle_{\Sigma} &= \langle \log_{\Sigma}(\tilde{\Sigma}), \log_{\Sigma}(\Sigma^\star)\rangle_{\Sigma} \\ &= \langle \log_{\Sigma}(\Sigma^+), \log_{\Sigma}(\Sigma^\star)\rangle_{\Sigma} + \underbrace{\langle \log_{\Sigma}(\tilde{\Sigma}) - \log_{\Sigma}(\Sigma^+), \log_{\Sigma}(\Sigma^\star)\rangle_{\Sigma}}_{:=R},\end{aligned}$$

and the residual term R can be further written as

$$\begin{aligned}R &= \langle T_{\Sigma,\tilde{\Sigma}} - T_{\Sigma,\Sigma^+}, \log_{\Sigma}(\Sigma^\star)\rangle_{\Sigma} \\ &= \langle T_{\Sigma^+,\Sigma}T_{\Sigma,\tilde{\Sigma}} - I, T_{\Sigma^+,\Sigma}\big(T_{\Sigma,\Sigma^\star} - I\big)\rangle_{\Sigma^+} \\ &= \langle T_{\Sigma^+,\Sigma}T_{\Sigma,\tilde{\Sigma}} - T_{\Sigma^+,\tilde{\Sigma}} + \log_{\Sigma^+}(\tilde{\Sigma}), T_{\Sigma^+,\Sigma}T_{\Sigma,\Sigma^\star} - T_{\Sigma^+,\Sigma} + I - I\rangle_{\Sigma^+} \\ &= -\langle \log_{\Sigma^+}(\tilde{\Sigma}), \log_{\Sigma^+}(\Sigma)\rangle_{\Sigma^+} + \underbrace{\langle T_{\Sigma^+,\Sigma}T_{\Sigma,\tilde{\Sigma}} - T_{\Sigma^+,\tilde{\Sigma}}, T_{\Sigma^+,\Sigma}T_{\Sigma,\Sigma^\star} - T_{\Sigma^+,\Sigma}\rangle_{\Sigma^+}}_{:=R_1} \\ &\quad + \langle \log_{\Sigma^+}(\tilde{\Sigma}), \log_{\Sigma^+}(\Sigma^\star)\rangle_{\Sigma^+} + \underbrace{\langle \log_{\Sigma^+}(\tilde{\Sigma}), T_{\Sigma^+,\Sigma}T_{\Sigma,\Sigma^\star} - T_{\Sigma^+,\Sigma^\star}\rangle_{\Sigma^+}}_{:=R_2}\end{aligned}$$

That leads to

$$\begin{aligned}&|\langle \nabla f(\Sigma), \log_{\Sigma}(\Sigma^\star)\rangle_{\Sigma} + \frac{1}{\eta}\langle \log_{\Sigma}(\Sigma^+), \log_{\Sigma}(\Sigma^\star)\rangle_{\Sigma}| \\ &\leq \frac{1}{\eta}\big(R_1 + R_2\big) + \frac{1}{\eta}\langle \log_{\Sigma^+}(\tilde{\Sigma}), \log_{\Sigma^+}(\Sigma^\star) - \log_{\Sigma^+}(\Sigma)\rangle_{\Sigma^+}\end{aligned}$$

Now we would provide bounds for each term above case by case:

Bounding $\langle \log_{\Sigma^+}(\tilde{\Sigma}), \log_{\Sigma^+}(\Sigma))\rangle_{\Sigma^+}$:

By the projection characterization (28), we have

$$\begin{aligned}\frac{1}{\eta}\langle \log_{\Sigma^+}(\tilde{\Sigma}), \log_{\Sigma^+}(\Sigma))\rangle_{\Sigma^+} &\leq -\frac{1}{2\eta}\langle \log_{\Sigma^+}(\tilde{\Sigma}), \Sigma^+ + \Sigma - 2\Sigma_+ T_{\Sigma^+,\Sigma}\rangle \\ &\leq \frac{1}{2\eta}\lambda_{\min}^{-1/2}(\Sigma^+)\underbrace{\|\log_{\Sigma^+}(\tilde{\Sigma})\|_{\Sigma^+}}_{=d(\Sigma^+,\tilde{\Sigma})} d^2(\Sigma^+, \Sigma) \\ &\leq \frac{1}{2\eta}\lambda_{\min}^{-1/2} d(\Sigma, \tilde{\Sigma}) d^2(\Sigma^+, \Sigma) \\ &\leq \frac{L}{2}\lambda_{\min}^{-1/2}(\Sigma^+)\|\log_{\Sigma}(\Sigma^+)\|_{\Sigma}^2.\end{aligned}$$

i.e.

$$\frac{1}{\eta}\langle \log_{\Sigma^+}(\tilde{\Sigma}), \log_{\Sigma^+}(\Sigma))\rangle_{\Sigma^+} \leq \frac{L\eta^2}{2}\lambda_{\min}^{-1/2}(\Sigma^+)\|g_C(\Sigma)\|_{\Sigma}^2. \tag{33}$$

Bounding $\langle \log_{\Sigma^+}(\tilde{\Sigma}), \log_{\Sigma^+}(\Sigma^\star)\rangle_{\Sigma^+}$:

By the projection characterization (28),we have

$$\begin{aligned}\langle \log_{\Sigma^+}(\tilde{\Sigma}), \log_{\Sigma^+}(\Sigma^\star))\rangle_{\Sigma^+} &\leq -\frac{1}{2}\langle \log_{\Sigma^+}(\tilde{\Sigma}), \Sigma^+ + \Sigma^\star - 2\Sigma_+ T_{\Sigma^+,\Sigma^\star}\rangle \\ &\leq \frac{1}{2}\lambda_{\min}^{-1/2}(\Sigma^+)\|\log_{\Sigma^+}(\Sigma)\|_{\Sigma_+} d^2(\Sigma^+, \Sigma^\star)\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{L\eta}{2}\lambda_{\min}^{-1/2} d^2(\Sigma^+, \Sigma^\star) \\
&\leq L\eta\lambda_{\min}^{-1/2}\big(d^2(\Sigma^+, \Sigma) + d^2(\Sigma, \Sigma^\star)\big) \\
&\leq L\eta\lambda_{\min}^{-1/2}\big(\eta^2\|g_C(\Sigma)\|_\Sigma^2 + (\theta_1\eta^2 + \theta_2)\|g_C(\Sigma)\|_\Sigma^2\big)
\end{aligned}
$$

i.e.

$$
\frac{1}{\eta}\langle \log_{\Sigma^+}(\tilde{\Sigma}), \log_{\Sigma^+}(\Sigma^\star))\rangle_{\Sigma^+} \leq L\lambda_{\min}^{-1/2}(2\theta_1\eta^2 + \theta_2)\|g_C(\Sigma)\|_\Sigma^2 \tag{34}
$$

Bounding R_1: We have

$$
\begin{aligned}
R_1 &\leq \|T_{\Sigma^+,\Sigma}T_{\Sigma,\tilde{\Sigma}} - T_{\Sigma^+,\tilde{\Sigma}}\|_{\Sigma^+} \cdot \|T_{\Sigma^+,\Sigma}T_{\Sigma,\Sigma^\star} - T_{\Sigma^+,\Sigma}\|_{\Sigma^+} \\
&\leq 64\eta\kappa^2\lambda_{\min}^{-1} d(\Sigma^+, \Sigma)d(\Sigma, \tilde{\Sigma}) \cdot d(\Sigma^+, \Sigma)d(\Sigma, \Sigma^\star) \\
&\leq 64\eta L\kappa^2\lambda_{\min}^{-1} d^2(\Sigma^+, \Sigma)d(\Sigma, \Sigma^\star)
\end{aligned}
$$

thus with $d(\Sigma^\star, \Sigma) \leq d(\Sigma^\star, \hat{\Sigma}) + d(\Sigma, \hat{\Sigma}) \leq 2\rho$,

$$
\begin{aligned}
\frac{1}{\eta}R_1 &\leq 64L\rho\kappa^2\lambda_{\min}^{-1} d^2(\Sigma^+, \Sigma)d(\Sigma, \Sigma^\star) \\
&\leq 128L\rho^2\eta^2\kappa^2\lambda_{\min}^{-1}\|g_C(\Sigma)\|_\Sigma^2
\end{aligned} \tag{35}
$$

Bounding R_2: With the property of projection, we have

$$
\begin{aligned}
R_2 = \langle \log_{\Sigma^+}(\tilde{\Sigma}), T_{\Sigma^+,\Sigma}T_{\Sigma,\Sigma^\star} - T_{\Sigma^+,\Sigma^\star}\rangle_{\Sigma^+} &\leq 8\kappa\lambda_{\min}^{-1/2} d(\Sigma^+, \tilde{\Sigma})d(\Sigma, \Sigma^\star)d(\Sigma, \Sigma^+) \\
&\leq 8\kappa\lambda_{\min}^{-1/2} d(\Sigma, \tilde{\Sigma})d(\Sigma, \Sigma^\star)d(\Sigma, \Sigma^+) \\
&\leq 8\eta L\kappa\lambda_{\min}^{-1/2} d(\Sigma, \Sigma^\star)d(\Sigma, \Sigma^+)
\end{aligned}
$$

thus

$$
\begin{aligned}
\frac{1}{\eta}R_2 &\leq 8L\kappa\lambda_{\min}^{-1/2} d(\Sigma, \Sigma^\star)d(\Sigma, \Sigma^+) \\
&\leq 8\eta\sqrt{\theta_1\eta^2 + \theta_2}L\kappa\lambda_{\min}^{-1/2}\|g_C(\Sigma)\|_\Sigma^2
\end{aligned} \tag{36}
$$

Combining equations (33)–(36), we have

$$
\begin{aligned}
&|\langle \nabla f(\Sigma), \log_\Sigma(\Sigma^\star)\rangle_\Sigma + \frac{1}{\eta}\langle \log_\Sigma(\Sigma^+), \log_\Sigma(\Sigma^\star)\rangle_\Sigma| \\
\leq &\frac{1}{\eta}\big(R_1 + R_2\big) + \frac{1}{\eta}\langle \log_{\Sigma^+}(\tilde{\Sigma}), \log_{\Sigma^+}(\Sigma^\star) - \log_{\Sigma^+}(\Sigma)\rangle_{\Sigma^+} \\
\leq &\bigg(8\eta\sqrt{\theta_1\eta^2 + \theta_2}L\kappa\lambda_{\min}^{-1/2} + 128L\rho^2\eta^2\kappa^2\lambda_{\min}^{-1} + L\lambda_{\min}^{-1/2}(2\theta_1\eta^2 + \theta_2) + \frac{L\eta^2}{2}\lambda_{\min}^{-1/2}\bigg)\|g_C(\Sigma)\|_\Sigma^2 \\
\leq &\bigg(8\eta\sqrt{\theta_1\eta^2 + \theta_2}L\kappa\lambda_{\min}^{-1/2} + 128L\rho^2\eta^2\kappa^2\lambda_{\min}^{-1} + L\lambda_{\min}^{-1/2}(2\theta_1\eta^2 + 2\theta_2)\bigg)\|g_C(\Sigma)\|_\Sigma^2 \\
\leq &\bigg(8\eta\sqrt{\theta_1\eta^2 + \theta_2}L\kappa\lambda_{\min}^{-1/2} + 256L\max\{\lambda_{\min}^{-1}, \lambda_{\min}^{-1/2}\}\rho^2\kappa(2\theta_1\eta^2 + 2\theta_2)\bigg)\|g_C(\Sigma)\|_\Sigma^2 \\
\leq &256\bigg((\eta+1)L\max\{\lambda_{\min}^{-1}, \lambda_{\min}^{-1/2}\}\rho^2\kappa(2\theta_1\eta^2 + 2\theta_2)\bigg)\|g_C(\Sigma)\|_\Sigma^2,
\end{aligned}
$$

that leads to the desired result.

F. Proof of Lemma 3.5

The first inequality has been shown in Altschuler et al. (2021), and we will provide a new proof of the second inequality with sharper constants:

By $T_{AB}AT_{AB} = B$,

$$\begin{aligned}
&\mathrm{tr}\big((A+B+AT_{A,B}+T_{A,B}A)(A+B-AT_{A,B}-T_{A,B}A)\big)\\
=&\mathrm{tr}(A^2+B^2-2A^2T_{A,B}^2)\\
=&\|A-B\|_F^2+2\mathrm{tr}(AT_{A,B}AT_{A,B})-2\mathrm{tr}(A^2T_{A,B}^2).
\end{aligned}\tag{37}$$

Since

$$\begin{aligned}
&\mathrm{tr}(AT_{A,B}AT_{A,B})-\mathrm{tr}(A^2T_{A,B}^2)\\
=&\mathrm{tr}\big(A(T_{A,B}A-AT_{A,B})T_{A,B}\big)\\
=&\mathrm{tr}\big(A(T_{A,B}A-AT_{A,B})(T_{A,B}-I)\big)\\
=&\mathrm{tr}\big((T_{A,B}-I)A(T_{A,B}-I)A\big)+\mathrm{tr}\big((T_{A,B}-I)A^2(I-T_{A,B})\big)\\
\geq&(\lambda_{\min}-\lambda_{\max})\mathrm{tr}((T_{A,B}-I)A(T_{A,B}-I))\\
=&(\lambda_{\min}-\lambda_{\max})d^2(A,B).
\end{aligned}\tag{38}$$

where in the second equation we have used $\mathrm{tr}(AT_{A,B}A-A^2T_{A,B})=0$. Combining (37), (38), we have

$$\begin{aligned}
&4\lambda_{\max}d^2(A,B)\\
=&4\lambda_{\max}\mathrm{tr}(A+B-AT_{A,B}-T_{A,B}A)\\
\geq&\mathrm{tr}\big((A+B+AT_{A,B}+T_{A,B}A)(A+B-AT_{A,B}-T_{A,B}A)\big)\\
=&\|A-B\|_F^2+2\mathrm{tr}(AT_{A,B}AT_{A,B})-2\mathrm{tr}(A^2T_{A,B}^2)\\
\geq&\|A-B\|_F^2+(\lambda_{\min}-\lambda_{\max})d^2(A,B)
\end{aligned}\tag{39}$$

Hence

$$\|A-B\|_F^2\leq(5\lambda_{\max}-\lambda_{\min})d^2(A,B).$$

In third line of (39), we use that for matrix C, D, $\lambda_{\min}(C)I\preceq C\preceq\lambda_{\max}(C)I$ and $D\succeq O$,

$$\lambda_{\min}(C)\mathrm{tr}(D)\leq\mathrm{tr}(CD)\leq\lambda_{\max}(C)\mathrm{tr}(D).$$

Obviously $A+B+AT_{A,B}+T_{A,B}A\preceq 4\lambda_{\max}I$. And we also require matrix $A+B-AT_{A,B}-T_{A,B}A$ to be positive semi-definite, which can be shown via

$$\begin{aligned}
&A+B-AT_{A,B}-T_{A,B}A\\
=&A^{-1/2}(A^2+A^{1/2}BA^{1/2}-A^{3/2}T_{A,B}A^{1/2}-A^{1/2}T_{A,B}A^{3/2})A^{-1/2}\\
=&A^{-1/2}(A-A^{1/2}T_{A,B}A^{1/2})(A-A^{1/2}T_{A,B}A^{1/2})A^{-1/2}\succeq O.
\end{aligned}$$

F.1. Proof of $\|T_{A,C}-T_{B,C}\|_F\lesssim d(A,B)$.

Notice that

$$\begin{aligned}
\|T_{A,C}-T_{B,C}\|_F&=\|T_{A,C}(T_{C,B}-T_{C,A})T_{B,C}\|_F\\
&\leq\|T_{A,C}\|_2\|T_{B,C}\|_2\|(T_{C,B}-T_{C,A})\|_F\\
&\leq\kappa\|(T_{C,B}-T_{C,A})\|_F\\
&\leq\kappa\|C^{-1/2}\big((C^{1/2}BC^{1/2})^{1/2}-(C^{1/2}AC^{1/2})^{1/2}\big)C^{-1/2}\|_F\\
&\leq\kappa\lambda_{\min}^{-1}\|(C^{1/2}BC^{1/2})^{1/2}-(C^{1/2}AC^{1/2})^{1/2}\|_F.
\end{aligned}\tag{40}$$

Where we use $\|T_{A,C}\|_2 \leq \sqrt{\kappa}, \|T_{A,B}\|_2 \leq \sqrt{\kappa}$. (Appendix C, Lemma 2 (Altschuler et al., 2021)).

Consider polar decomposition:

$$B^{1/2}C^{1/2} = U_1P_1, A^{1/2}C^{1/2} = U_2P_2.$$

where U_1, U_2 are unitary matrices and P_1, P_2 are positive semi-definite matrices.

Then

$$\|(C^{1/2}BC^{1/2})^{1/2} - (C^{1/2}AC^{1/2})^{1/2}\|_F = \|P_1 - P_2\|_F. \tag{41}$$

From Theorem 5.1 in Mathias (1993),

$$\begin{aligned}\|P_1 - P_2\|_F &\leq \sqrt{2}\|B^{1/2}C^{1/2} - A^{1/2}C^{1/2}\|_F \\ &\leq \sqrt{2}\|C^{1/2}\|_2\|B^{1/2} - A^{1/2}\|_F \\ &\leq \frac{\sqrt{2}}{2}\kappa^{1/2}\|B - A\|_F. \end{aligned} \tag{42}$$

And hence by combining (40), (41), (42)

$$\|T_{A,C} - T_{B,C}\|_F \leq \frac{\sqrt{2}}{2}\kappa^{3/2}\lambda_{\min}^{-1}\|B - A\|_F.$$

G. Proof of Lemma 4.8

Noticing that,

$$(T_{AB} - I)(T_{BC} - I) - (T_{AB}T_{BC} - T_{AC}) = (T_{BB} - T_{BC}) - (T_{AB} - T_{AC}) = f(B) - f(A)$$

with $f(X)$ defined as

$$f(X) := T_{XB} - T_{XC}.$$

Thus we have

$$\begin{aligned} d^2(A,B)d^2(B,C) &= \text{tr}\big((T_{AB} - I)A(T_{AB} - I)\big)\text{tr}\big((T_{BC} - I)B(T_{BC} - I)\big) \\ &\geq \text{tr}\big((T_{BC} - I)(T_{AB} - I)A(T_{AB} - I)(T_{BC} - I)B\big) \\ &\geq \lambda_{\min}\text{tr}((T_{BC} - I)(T_{AB} - I)A(T_{AB} - I)(T_{BC} - I)). \end{aligned} \tag{43}$$

And

$$\begin{aligned} &\text{tr}((T_{BC} - I)(T_{AB} - I)A(T_{AB} - I)(T_{BC} - I)) \\ =&\text{tr}((T_{BC} - I)(T_{AB} - I)A(T_{AB}T_{BC} - T_{AC} + f(B) - f(A))) \\ \geq&\text{tr}\big((T_{AB}T_{BC} - T_{AC} + f(B) - f(A))A(T_{AB}T_{BC} - T_{AC})\big) - \|f(B) - f(A)\|_A^2 - \|(T_{BC} - I)(T_{AB} - I)\|_A^2 \\ \geq&(1-\alpha)\text{tr}\big((T_{AB}T_{BC} - T_{AC})A(T_{AB}T_{BC} - T_{AC})\big) - \alpha\|f(B) - f(A)\|_A^2 - \|(T_{BC} - I)(T_{AB} - I)\|_A^2 \end{aligned} \tag{44}$$

where α is a constant, and we will specify its value in the following proof. Denoting $g(X) = Xf(X)$, then

$$\begin{aligned} \|f(B) - f(A)\|_A^2 &= \text{tr}\big((f(B) - f(A)^\top A(f(B) - f(A)\big) \\ &\leq \lambda_{\min}^{-1}\text{tr}\big([(A - B)f(B) + Bf(B) - Af(A)]^\top[(A - B)f(B) + Bf(B) - Af(A)]\big) \\ &= \lambda_{\min}^{-1}\text{tr}\big([\underbrace{(A - B)f(B)}_{:=D_1} + \underbrace{g(B) - g(A)}_{:=D_2}]^\top[(A - B)f(B) + g(B) - g(A)]\big) \\ &= \lambda_{\min}^{-1}\big(\text{tr}(D_1^\top D_1) + 2\text{tr}(D_2^\top D_1) + \text{tr}(D_2^\top D_2)\big). \end{aligned} \tag{45}$$

Since it holds directly that

$$\begin{aligned}\|D_1\|_F^2 &\leq \|f(B)\|_2^2\|A-B\|_F^2\\ &\leq \|T_{B,C}-I\|_F^2\|A-B\|_F^2\\ &\leq (5\lambda_{\max}-\lambda_{\min})\lambda_{\min}^{-1}d^2(B,C)d^2(A,B)\\ &\leq 5\kappa d^2(B,C)d^2(A,B)\end{aligned} \tag{46}$$

we need only show that $\|D_2\|_F \lesssim d(A,B)d(B,C)$.

By the mean-value inequality, we have for any matrix $M_1, M_2 \in \mathbb{R}^{n\times n}$ and smooth $\psi : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$

$$\|\psi(M_1)-\psi(M_2)\|_F \leq \sup_{M\in\gamma}\|D\psi(M)\|_{op}\|\|M_1-M_2\|_F. \tag{47}$$

where $\gamma := \{t \in [0,1] : tM_1+(1-t)M_2\}$ and $\|D\psi\|_{op} := \max_{\|E\|_F=1}\|D\psi(E)\|_F$. Thus we need only show that $\|Dg(M)\|_{op} \lesssim d(B,C)$.

Noticing that

$$g(X) = XT_{XB} - XT_{XC} = T_{BX}B - T_{CX}C$$

and

$$T_{BX}B = B^{-1/2}(B^{1/2}XB^{1/2})^{1/2}B^{1/2}, T_{CX}C = C^{-1/2}(C^{1/2}XC^{1/2})^{1/2}C^{1/2}$$

Denote $Z_B = B^{1/2}XB^{1/2}, Z_C = C^{1/2}XC^{1/2}$ we have

$$\begin{aligned}\frac{dg(X)}{dX} =&\frac{d}{dX}T_{BX}B - \frac{d}{dX}T_{CX}C\\ =&B^{-1/2}\frac{d}{dX}(B^{1/2}XB^{1/2})^{1/2}B^{1/2} - C^{-1/2}\frac{d}{dX}(C^{1/2}XC^{1/2})^{1/2}B^{1/2}\\ =&\frac{d(Z_B)^{1/2}}{dZ_B}B - \frac{d(Z_C)^{1/2}}{dZ_C}C\\ =&\left(\frac{d(Z_B)^{1/2}}{dZ_B} - \frac{d(Z_C)^{1/2}}{dZ_C}\right)B + \frac{d(Z_C)^{1/2}}{dZ_C}(B-C).\end{aligned}$$

From Moral & Niclas (2018), we have

$$\|\frac{d(Z_C)^{1/2}}{dZ_C}\|_{op} \leq \frac{1}{2}\lambda_{\min}^{-1/2}(Z_C) \leq \frac{1}{2}\lambda_{\min}^{-1}.$$

Hence

$$\|\frac{d(Z_C)^{1/2}}{dZ_C}(B-C)\|_F \leq \|\frac{d(Z_C)^{1/2}}{dZ_C}\|_{op}\|B-C\|_F \leq \frac{\sqrt{5}}{2}\kappa^{1/2}\lambda_{\min}^{-1/2}d(B,C).$$

And for any matrix E, according to Moral & Niclas (2018), we have

$$\begin{aligned}&\left(\frac{d(Z_B)^{1/2}}{dZ_B} - \frac{d(Z_C)^{1/2}}{dZ_C}\right)E\\ =&\int_0^\infty e^{-t(B^{1/2}XB^{1/2})^{1/2}}Ee^{-t(B^{1/2}XB^{1/2})^{1/2}}dt - \int_0^\infty e^{-t(C^{1/2}XC^{1/2})^{1/2}}Ee^{-t(C^{1/2}XC^{1/2})^{1/2}}dt\\ =&D_B - D_C,\end{aligned}$$

where $D_B = \int_0^\infty e^{-t(B^{1/2}XB^{1/2})^{1/2}}Ee^{-t(B^{1/2}XB^{1/2})^{1/2}}dt$, $D_C = \int_0^\infty e^{-t(C^{1/2}XC^{1/2})^{1/2}}Ee^{-t(C^{1/2}XC^{1/2})^{1/2}}dt$, by definition, D_B, D_C satisfy the Lyapunov equations:

$$(B^{1/2}XB^{1/2})^{1/2}D_B + D_B(B^{1/2}XB^{1/2})^{1/2} = -E$$

$$(C^{1/2}XC^{1/2})^{1/2}D_C + D_C(C^{1/2}XC^{1/2})^{1/2} = -E$$

Hence

$$\begin{aligned}
&\big(-(B^{\frac{1}{2}}XB^{\frac{1}{2}})^{\frac{1}{2}}\big)(D_B - D_C) + (D_B - D_C)\big(-(B^{\frac{1}{2}}XB^{\frac{1}{2}})^{\frac{1}{2}}\big)\\
&= E + (B^{1/2}XB^{1/2})^{1/2}D_C + D_C(B^{1/2}XB^{1/2})^{1/2}\\
&= \big((B^{1/2}XB^{1/2})^{1/2} - (C^{1/2}XC^{1/2})^{1/2}\big)D_C + D_C\big((B^{1/2}XB^{1/2})^{1/2} - (C^{1/2}XC^{1/2})^{1/2}\big).
\end{aligned}$$

Then $D_B - D_C$ satisfies a Lyapunov equation, and

$$\begin{aligned}
D_B - D_C = \int_0^\infty e^{-t(B^{1/2}XB^{1/2})^{1/2}}\Big(&\big((B^{1/2}XB^{1/2})^{1/2} - (C^{1/2}XC^{1/2})^{1/2}\big)D_C\\
&+ D_C\big((B^{1/2}XB^{1/2})^{1/2} - (C^{1/2}XC^{1/2})^{1/2}\big)\Big)e^{-t(B^{1/2}XB^{1/2})^{1/2}}dt
\end{aligned}$$

Hence

$$\begin{aligned}
&\|D_B - D_C\|_F\\
&\le 2\|D_C\|_F\|(B^{1/2}XB^{1/2})^{1/2} - (C^{1/2}XC^{1/2})^{1/2}\|_F\int_0^\infty \|e^{-t(B^{1/2}XB^{1/2})^{1/2}}\|_2^2 dt\\
&\le 2(\frac{1}{2\lambda_{\max}})^2\|(B^{1/2}XB^{1/2})^{1/2} - (C^{1/2}XC^{1/2})^{1/2}\|_F\\
&\le \frac{\sqrt{2}}{4}\kappa^{1/2}\lambda_{\max}^{-2}\|B - C\|_F\\
&\le \frac{\sqrt{10}}{4}\kappa^{1/2}\lambda_{\max}^{-3/2}d(B,C).
\end{aligned}$$

Therefore

$$\begin{aligned}
\|\frac{dg(X)}{dX}\|_F &= \|\Big(\frac{d(Z_B)^{1/2}}{dZ_B} - \frac{d(Z_C)^{1/2}}{dZ_C}\Big)B + \frac{d(Z_C)^{1/2}}{dZ_C}(B - C)\|_F\\
&\le \|B\|_2\|\Big(\frac{d(Z_B)^{1/2}}{dZ_B} - \frac{d(Z_C)^{1/2}}{dZ_C}\Big)\|_{op} + \|\frac{d(Z_C)^{1/2}}{dZ_C}(B - C)\|_F\\
&\le (\frac{\sqrt{10}}{4}\kappa^{1/2}\lambda_{\max}^{-1/2} + \frac{\sqrt{5}}{2}\kappa^{1/2}\lambda_{\min}^{-1/2})d(B,C)\\
&\le \sqrt{5}\kappa^{1/2}\lambda_{\min}^{-1/2}d(B,C).
\end{aligned}$$

Then by mean-value theorem

$$\begin{aligned}
\|D_2\|_F = \|g(A) - g(B)\|_F &\le \sqrt{5}\kappa^{1/2}\lambda_{\min}^{-1/2}d(B,C)\|A - B\|_F\\
&\le 5\kappa d(A,B)d(B,C) \qquad (48)
\end{aligned}$$

With (45), (46), (48),

$$\begin{aligned}
\lambda_{\min}\|f(B) - f(A)\|_A^2 &\le 2(\|D_1\|_F^2 + \|D_2\|_F^2)\\
&\le 60\kappa^2 d^2(A,B)d^2(B,C). \qquad (49)
\end{aligned}$$

Meanwhile, we have

$$\begin{aligned}
\|(T_{BC} - I)(T_{AB} - I)\|_A^2 &\le \|(T_{BC} - I)\|_F^2 d^2(A,B)\\
&\le \lambda_{\min}^{-1}d^2(A,B)d^2(B,C) \qquad (50)
\end{aligned}$$

With (49), (50), (44), we have

$$\lambda_{\min}(1 - \alpha)\|T_{AB}T_{BC} - T_{AC}\|_A^2 \le (2 + 60\alpha\kappa^2)d^2(A,B)d^2(B,C).$$

Let $\alpha = \frac{1}{2}$, we have

$$\begin{aligned}\|T_{AB}T_{BC} - T_{AC}\|_A &\leq 2\sqrt{1+15\kappa^2}\lambda_{\min}^{-1/2} d(A,B)d(B,C)\\ &\leq 8\kappa\lambda_{\min}^{-1/2} d(A,B)d(B,C).\end{aligned}$$

H. Proof of Proposition 3.10

Consider the following minimization problem

$$\min_{A\in\mathbb{S}_+^n} \frac{1}{2}\|A\|_F^2$$

Its corresponding BWGD update is $\Sigma_{k+1} = (I-\eta\Sigma_k)\Sigma_k(I-\eta\Sigma_k)$. Consider starting with $\Sigma_0 = \frac{1}{\eta}I$, then each Σ_k is also a diagonal matrix with element $(\Sigma_k)_{ii}$. Noted that $(\Sigma_k)_{ii} = \lambda_{\min}(\Sigma_k)$, hence for each $(\Sigma_k)_{ii}$, we have update $(\Sigma_{k+1})_{ii} = (1-\eta(\Sigma_k)_{ii})^2\cdot(\Sigma_k)_{ii} \geq (1-\eta\lambda_{\min}(\Sigma_k))^2\cdot(\Sigma_k)_{ii}$. Hence

$$\begin{aligned}\lambda_{\min}^{-1}(\Sigma_k)\log\left(\frac{f(\Sigma_k)-f(\Sigma^\star)}{f(\Sigma_{k+1})-f(\Sigma^\star)}\right) &= \lambda_{\min}^{-1}(\Sigma_k)\log\big(\frac{\|\Sigma_k\|_F^2}{\|\Sigma_{k+1}\|_F^2}\big)\\ &\leq 2\lambda_{\min}^{-1}(\Sigma_k)\log(\frac{1}{1-\eta\lambda_{\min}(\Sigma_k)})\end{aligned}$$

Hence

$$\liminf_k \lambda_{\min}^{-1}(\Sigma_k)\log\left(\frac{f(\Sigma_k)-f(\Sigma^\star)}{f(\Sigma_{k+1})-f(\Sigma^\star)}\right) \leq 2.$$

I. Details of Examples

I.1. Proof of Proposition A.1

Our proof relies on the following result on the projection operator:

Lemma I.1. *Suppose some Σ is projected to the BW ball $\mathcal{W}(\widehat{\Sigma},\rho)$, and if there exists some $\lambda_{\min},\lambda_{\max}$ so that $\lambda_{\min}I \preceq \Sigma,\widehat{\Sigma} \preceq \lambda_{\max}I$, then it holds that $\lambda_{\max}I \preceq P_{\mathcal{W}}(\Sigma) \preceq \lambda_{\min}I$.*

Proof of Lemma I.1. From the explicit form of projection (13), we only need to prove $2\lambda_{\max}I \preceq \Sigma T_{\Sigma,\widehat{\Sigma}} + \widehat{\Sigma}T_{\widehat{\Sigma},\Sigma} \preceq 2\lambda_{\min}I$, which can be shown by

$$\begin{aligned}\lambda(\Sigma T_{\Sigma,\widehat{\Sigma}} + \widehat{\Sigma}T_{\widehat{\Sigma},\Sigma}) &= \lambda(\Sigma^{\frac{1}{2}}(\Sigma^{\frac{1}{2}}\widehat{\Sigma}\Sigma^{\frac{1}{2}})^{\frac{1}{2}}\Sigma^{-\frac{1}{2}} + \Sigma^{-\frac{1}{2}}(\Sigma^{\frac{1}{2}}\widehat{\Sigma}\Sigma^{\frac{1}{2}})^{\frac{1}{2}}\Sigma^{\frac{1}{2}})\\ &= \lambda((\Sigma^{\frac{1}{2}}\widehat{\Sigma}\Sigma^{\frac{1}{2}})^{\frac{1}{2}} + \Sigma^{-1}(\Sigma^{\frac{1}{2}}\widehat{\Sigma}\Sigma^{\frac{1}{2}})^{\frac{1}{2}}\Sigma)\end{aligned}$$

And $\lambda(\Sigma^{-1}(\Sigma^{\frac{1}{2}}\widehat{\Sigma}\Sigma^{\frac{1}{2}})^{\frac{1}{2}}\Sigma) = \lambda((\Sigma^{\frac{1}{2}}\widehat{\Sigma}\Sigma^{\frac{1}{2}})^{\frac{1}{2}}) \in [\lambda_{\min},\lambda_{\max}]$. □

Proof. In Lemma 1 of Altschuler et al. (2021), it has been shown that when the mixture measure P is supported over covariance with eigenvalue at range $[\lambda_{\min},\lambda_{\max}]$, then at every time-step t, if $\Sigma_t \succeq \lambda_{\min}/4$, then $\tilde{\Sigma}_t := \exp_{\Sigma_t}(-\nabla f(\Sigma_t))$ with $\eta \lesssim \kappa^{-1}$ will also satisfy $\tilde{\Sigma}_t \succeq \lambda_{\min}/4$.

In our case, we can set

$$\lambda_{\min} := \min\{\lambda_{\min}(P), \frac{1}{4}\sigma_{\min}(\widehat{\Sigma})\}, \lambda'_{\max} := \max\{\lambda_{\max}(P), \frac{1}{4}\sigma_{\max}(\widehat{\Sigma})\}, \kappa = \lambda'_{\max}/\lambda'_{\min}$$

Then applying Lemma I.1 and the above result leads to the desired result. □

I.2. Proof of Proposition A.2

Proof. Firstly, noticing that the upper bound $\lambda_{\max}$ exists directly by the boundedness of the BW ball, thus it remains to show the existence of $\lambda_{\min}$: By Lemma A.7 in Nguyen et al. (2023), we have the gradient for both Σ_x and Σ_w are negative definite, thus as long as the initialization $\Sigma_{x,k} \succeq \widehat{\Sigma}_x$, we have $\Sigma_{x,k+1} \succeq \widehat{\Sigma}_x$ also holds by

$$\big(I - \eta \nabla_x F(\Sigma_{x,k})\big)\Sigma_{x,k}\big(I - \eta \nabla_x F(\Sigma_{x,k})\big) \succeq \widehat{\Sigma}_x$$

and Lemma I.1. The proof for Σ_w is nearly the same. □

I.3. SDP formulation of the BW Projection

By Proposition 2.3 in Nguyen et al. (2023), the BW distance between two matrices Σ, Σ' can be equivalently formulated as the solution of the SDP problem:

$$\begin{aligned} &\min_{D \in \mathbb{R}^{n\times n}} \operatorname{tr}\big(\Sigma + \Sigma' - 2D\big), \\ &\text{subject to } \begin{bmatrix} \Sigma & D \\ D^\top & \Sigma' \end{bmatrix} \succeq 0. \end{aligned} \tag{51}$$

On the other hand, the Lemma D.1 of Taşkesen et al. (2023) shows that the constrained set $\mathcal{W} := \{\Sigma : d(\Sigma, \widehat{\Sigma}) \le \rho\}$ can be written as

$$\left\{ Z \in \mathbb{S}_+^n, \exists E_z \in \mathbb{S}_+^n \text{ with } \operatorname{tr}(Z + \widehat{\Sigma} - 2E_z) \le \rho^2, \begin{bmatrix} \widehat{\Sigma}^{1/2} Z \widehat{\Sigma}^{1/2} & E_z \\ E_z & I \end{bmatrix} \succeq 0 \right\} \tag{52}$$

Combining (51) and (52), we have for any $\Sigma, \widehat{\Sigma}, \rho$, let

$$\begin{aligned} D^\star, E_z^\star, Z^\star = &\operatorname{argmin}_{D \in \mathbb{R}^{n\times n}, E \in \mathbb{S}_+^n, Z \in \mathbb{S}_+^n} \operatorname{tr}\big(\Sigma + Z - 2D\big), \\ &\text{subject to } \begin{bmatrix} \Sigma & D \\ D^\top & Z \end{bmatrix} \succeq 0, \quad \begin{bmatrix} \widehat{\Sigma}^{1/2} Z \widehat{\Sigma}^{1/2} & E_z \\ E_z & I \end{bmatrix} \succeq 0, \operatorname{tr}(Z + \widehat{\Sigma} - 2E_z) \le \rho^2, \end{aligned} \tag{53}$$

Then $Z^\star$ is the desired projection of Σ to $\widehat{\Sigma}$ centered $\mathcal{W}$.

I.4. Surrogate dual gap

The following materials are basically from section 6.1 in Nguyen et al. (2023).

We consider constrained optimization problems of the following form (1)

$$\min_{M \in C} f(M) \tag{54}$$

and we assume access to an inexact oracle $F : C \to C$, such that for any $\Sigma \in C$

$$\langle F(\Sigma) - \Sigma, Df(\Sigma)\rangle \le \delta \min_{z \in C} \langle z - \Sigma, Df(\Sigma)\rangle.$$

where $\delta \in [0, 1]$ is a precision parameter. Obviously, if $\delta > 0$, when Σ solve the (54), $-\langle F(\Sigma) - \Sigma, Df(\Sigma)\rangle = 0$. Hence in k-th iteration, the *surrogate duality gap* $g_k = -\langle F(\Sigma_k) - \Sigma_k, Df(\Sigma_k)\rangle$ can be regarded as the criteria of convergence.

Especially, when C is the BW ball $C := \{\Sigma \in \mathbb{S}_+^n \mid d(\Sigma, \hat{\Sigma}) \le \rho\}$ and the Euclidean gradient $Df(\Sigma) \preceq 0$, Nguyen et al. (2023) finds that the dual of the linear oracle subproblem is equivalent to solving a univariate algebraic equation, which can be solved efficiently via a bisection algorithm with given δ.

I.5. Armijo search strategy

As discussed in Iusem (2003), there are two types of Armijo search strategies for determining the step sizes in projected gradient descent (PGD): the Armijo search along the feasible direction and the Armijo search along the boundary of the

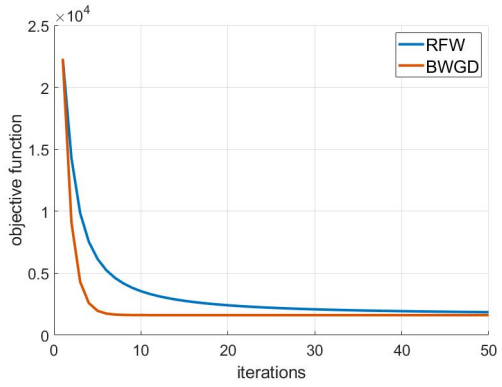

Figure 3. Case 1: Unconstrained minimizer lies inside C

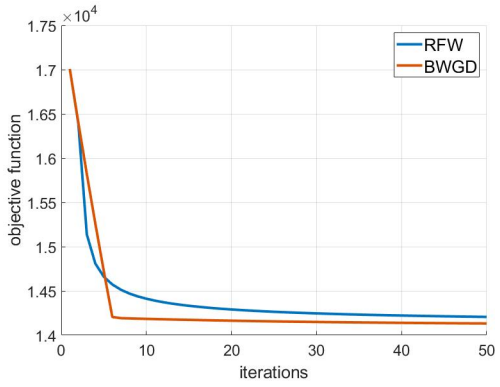

Figure 4. Case 2: Unconstrained minimizer lies outside C

constrained set C. Here, we adopt the latter strategy, which we will refer to as the Armijo search for simplicity. Specifically, we consider PGD with iteration described as $x^{k+1} = P_C(x^k - \beta_k Df(x^k))$. Given $\bar{\beta} > 0, \sigma \in (0,1)$, in k-th iteration, the Armijo search strategy finds $\beta_k = \bar{\beta}2^{-l(k)}$, where

$$l(k) = \left\{j \in \mathbb{N} \mid f(z^{k,j}) \leq f(x^k) - \sigma Df(x^k)^\intercal (x^k - z^{k,j})\right\} \text{ with } z^{k,j} = P_C(x^k - \bar{\beta}2^{-j} Df(x^k)). \tag{55}$$

For projected BWGD algorithm with iteration described as $\Sigma^{k+1} = P_C(\exp_{\Sigma^k}(-\beta_k \nabla f(\Sigma^k)))$, we replace those under Euclidean geometry with BW space counterparts. Then projected BWGD with Armijo search finds $\beta_k = \bar{\beta}2^{-l(k)}$, where

$$l(k) = \left\{j \in \mathbb{N} \mid f(Z^{k,j}) \leq f(\Sigma^k) + \sigma \langle \nabla f(\Sigma^k), \log_{\Sigma^k}(Z^{k,j}) \rangle_{\Sigma^k}\right\} \text{ with } Z^{k,j} = P_C(\exp_{\Sigma^k}(-\bar{\beta}2^{-j} \nabla f(\Sigma^k))). \tag{56}$$

J. Supplementary Experiment

In this section, we introduce the Wasserstein barycenter problem with the matrix interval constraint, as discussed in Weber & Sra (2022; 2023). Additionally, we compare the convergence speeds of the projected BWGD algorithm and the Riemannian Frank-Wolfe (RFW) algorithm, proposed in Weber & Sra (2023), for this problem.

J.1. Problem Formulation and BW Projection

In this problem, we aim to minimize the objective function

$$\min_{\Sigma \in C} \sum_{i=1}^{N} \beta_i d^2(\Sigma, \Sigma_i)$$

with the matrix interval constraint set $C := \{\Sigma \mid A \preceq \Sigma \preceq B\}$ for two pre-determined matrices $A \preceq B$ and $\sum_{i=1}^{N} \beta_i = 1$. For such a constraint set, as far as we know, an analytical projection of a matrix Σ to C

$$\min_{\tilde{\Sigma} \in C} d^2(\Sigma, \tilde{\Sigma}),$$

does not exist. Thus we propose to compute the projection by solving for $Z^\star$ in the following SDP:

$$\begin{aligned} Z^\star, D^\star = &\arg\min_{Z \in \mathbb{S}^n_+, D \in \mathbb{R}^{n \times n}} \operatorname{tr}\big(\Sigma + Z - 2D\big), \\ &\text{subject to } \begin{bmatrix} \Sigma & D \\ D^\top & Z \end{bmatrix} \succeq 0, \quad B - Z \succeq 0, \quad Z - A \succeq 0. \end{aligned} \tag{57}$$

J.2. Experiment Setting

Within the proposed projection operation, we compare the projected BWGD algorithm and the RFW algorithm in the following two different cases:

- **Case 1:** The unconstrained minimum lies within the interior of the interval (i.e., the constrained optimum also lies in the interior of the interval).
- **Case 2:** The unconstrained minimum lies outside the interval (i.e., the constrained optimum lies on the boundary of the interval).

More precisely, denote $\mathrm{AM}(\beta; \{\Sigma_i\})$ as the arithmetic mean of all N matrices, defined as

$$\mathrm{AM}(\beta; \{\Sigma_i\}) = \sum_{i=1}^{N} \beta_i \Sigma_i.$$

We consider the matrix interval constraints in two cases as follows:

In Case 1, we consider the same setting as discussed in Weber & Sra (2022; 2023), where $A = \alpha_l I$ with $\alpha_l = \min_{1 \le i \le N} \lambda_{\min}(\Sigma_i)$ and $B = \mathrm{AM}(\beta; \{\Sigma_i\})$. It has been shown in Bhatia et al. (2018; 2019) that these A and B serve as natural lower and upper bounds for the Wasserstein barycenter, indicating that the unconstrained minimum lies within the interval's interior.

In Case 2, we set $A = \alpha_l I$ with $\alpha_l = \lambda_{\min}(\mathrm{AM}(\beta; \{\Sigma_i\}))$ and $B = \alpha_u I$ with $\alpha_u = \frac{1}{N} \sum_{i=1}^{N} \lambda_i(\mathrm{AM}(\beta; \{\Sigma_i\}))$. Numerical results indicate that the constrained optimum lies on the boundary of the interval.

In both cases, we set $n = 60$, $N = 20$, $\beta_i = \frac{1}{N}, \forall i = 1, \cdots, N$ and generate the matrices $\{\Sigma_i\}$ following the same method as described in Weber & Sra (2022). For more details, please refer to Weber & Sra (2022) and the code. It should be noted that solving the SDP in equation (57) is required in each iteration of BWGD, which can be time-consuming. Figure 3 and Figure 4 show the result of Case 1 and Case 2 respectively, demonstrating that the BWGD algorithm outperforms the RFW algorithm.