Bures-Wasserstein Barycenters and Low-Rank Matrix Recovery

Tyler Maunu
Brandeis University

Thibaut Le Gouic
École Centrale de Marseille

Philippe Rigollet
Massachusetts Institute of Technology

Abstract

We revisit the problem of recovering a low-rank positive semidefinite matrix from rank-one projections using tools from optimal transport. More specifically, we show that a variational formulation of this problem is equivalent to computing a Wasserstein barycenter. In turn, this new perspective enables the development of new geometric first-order methods with strong convergence guarantees in Bures-Wasserstein distance. Experiments on simulated data demonstrate the advantages of our new methodology over existing methods.

1 INTRODUCTION

Recovering a low-rank matrix is a fundamental primitive across many settings, such as matrix completion [Fazel, 2002, Candès and Recht, 2009, Candès and Tao, 2010], phase retrieval [Candès et al., 2015], principal component analysis [Pearson, 1901, Hotelling, 1933], robust subspace recovery [Lerman and Maunu, 2018], and robust principal component analysis [Chandrasekaran et al., 2009, Candès et al., 2011, Xu et al., 2010]. This line of work can be understood as a generalization of the classical compressed sensing question [Donoho, 2006, Candès et al., 2006], where the goal is the recovery of a sparse vector. Indeed, the sparse vector recovery problem can be cast as a low-rank recovery problem over diagonal matrices. In all of these settings, the assumption of a low-rank structure is essential for efficient estimation and optimization in high-dimensional settings.

While the above applications all aim at recovering a low-rank matrix $\boldsymbol{S}$, the observational—a.k.a sensing—mechanism that governs access to $\boldsymbol{S}$ comes in many declinations. For the purpose of applications, it is often sufficient to focus on linear measurements of the form $\langle \boldsymbol{S}, \boldsymbol{A} \rangle$ for some given sensing matrix $\boldsymbol{A}$. This setup covers a wide variety of applications ranging from covariance sketching [Chen et al., 2015] and low-rank matrix completion [Candès and Plan, 2010, Recht et al., 2010] to phase retrieval [Fienup, 1978, Candès et al., 2013, 2015] and quantum state tomography [Gross et al., 2010]. Due to the flexibility of this formulation, new solutions to this problem can have many practical implications.

In this paper, we focus on a specific instantiation of this problem, where the measurement matrix $\boldsymbol{A} = \boldsymbol{x}\boldsymbol{x}^\top$ is rank-one and positive semidefinite (PSD), so that $\langle \boldsymbol{x}\boldsymbol{x}^\top, \boldsymbol{S} \rangle = \boldsymbol{x}^\top \boldsymbol{S} \boldsymbol{x}$. This important case of the low-rank matrix recovery problem has received significant attention over the past few decade [Cai and Zhang, 2015, Chen et al., 2015, Sanghavi et al., 2017, Li et al., 2019].

Assume that we observe a sample

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{S} \boldsymbol{x}_i \,, i = 1, \dots, n \tag{1.1}$$

where $\boldsymbol{S} \in \mathbb{R}^{d \times d}$ is an unknown rank r PSD matrix and $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ are i.i.d from some distribution. Our goal is to recover or estimate $\boldsymbol{S}$ from the pairs $(y_i, \boldsymbol{x}_i)$, $i = 1, \dots, n$. Throughout, we denote by $\mathbb{S}_+^d$ the set of $d \times d$ PSD matrices and $\mathbb{S}_{++}^d$ is the set of positive definite (PD) matrices.

Finding a low-rank matrix $\boldsymbol{S}$ subject to constraints (1.1) is a semidefinite program (SDP) that can be implemented in polynomial time using general-purpose solvers. Furthermore, the specific structure of this SDP may be leveraged to derive faster algorithms. Such solutions include the Burer-Monteiro approach to solving semidefinite programs [Burer and Monteiro, 2003] or nonconvex gradient descent methods for low-rank programs [Sanghavi et al., 2017]. Often, these approaches result in *nonconvex* optimization programs for which theoretical results are limited.

In this work, we take a principled approach to solving this problem by eliciting convexity using a specific geometry on the the space of PSD matrices. More precisely, we employ the Bures-Wasserstein (hereafter BW) geometry, which comes independently from optimal transport and quantum information theory [Bures, 1969, Bhatia et al., 2019]. This geometry allows us to solve the original problem by computing a *BW barycenter* [Agueh and Carlier, 2011, Álvarez-Esteban et al., 2016, Chewi et al., 2020,

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

Altschuler et al., 2021]. In turn, we employ geodesic gradient descent on the BW manifold to compute said barycenter. We propose both full gradient and stochastic gradient based methods that are guaranteed to efficiently recover a low-rank matrix. These methods have low computational cost (per iteration complexity of $O(ndr)$ for gradient descent and $O(dr)$ for stochastic gradient descent), have minimal parameter tuning, are easily implemented, and show excellent practical performance. We demonstrate an example application of phase retrieval in Figure 1. In this set-up, BWgradient descent (BWGD) recovers the image faster than Wirtinger Flow (WF) [Candès et al., 2015], and BWGD needs no parameter tuning.

Figure 1: Recovered images after 140 iterations of Wirtinger Flow [Candès et al., 2015] and BWGD (our method). BWGD recovers a much sharper image within the same number of iterations; note that the iteration complexity of both methods is the same.

Main contributions. The main results of this paper are:

1. We prove that a barycenter of a certain distribution of rank-one Gaussians exactly recovers the target low-rank matrix $\boldsymbol{S}$.
2. With this connection, we give novel geodesic gradient descent and stochastic geodesic gradient descent algorithms for solving the low-rank PSD matrix recovery problem using existing first-order algorithms for computing BW barycenters.
3. Existing first-order algorithms for computing BW barycenters are only guaranteed to work for full rank distributions. Since our method considers barycenters of rank-one PSD matrices, we develop new theory and give a guarantee of local linear convergence in BW distance for the gradient descent method. We also discuss initialization of our method.
4. We demonstrate the competitive edge of our algorithms in a few experimental settings.

Related Work. Many methods have been proposed to solve variants of the matrix recovery problem. Original ideas for this problem trace back to linear systems theory, low-rank matrix completion, low-dimensional Euclidean embeddings, and image compression [Recht et al., 2010].

We focus here on rank-one projections of positive semidefinite matrices as in (1.1). This setup is either specifically considered or a special case of a large number of works including [Candès et al., 2015, Cai and Zhang, 2015, Chen et al., 2015, Zhong et al., 2015, Wang and Giannakis, 2016, Wang et al., 2017, Sanghavi et al., 2017, Li et al., 2019]. These methods can be clustered into two families. The first one aims at minimizing convex relaxations of an energy functional that often based on the nuclear norm [Cai and Zhang, 2015, Chen et al., 2015]. Such convex programs can be solved via standard solvers. Another family of methods directly implement the low-rank constraint into a nonconvex constraint [Li et al., 2019] thus manipulating candidate matrices with smaller representations and thereby boosting computational efficiency; see Chi et al. [2019] for an overview of such algorithms. The special case where $\boldsymbol{S}$ has rank one corresponds to the classical phase retrieval problem and has received much attention with dedicated algorithms [Fienup, 1978, Candès et al., 2015, Wang and Giannakis, 2016, Wang et al., 2017, Chi et al., 2019].

Note that the methods introduced in the present paper can be readily extended to the *covariance recovery* problem framed in Cai and Zhang [2015], and that is similar to estimation in a random effects model. Under this model, rather than (1.1), we observe $y_i = \langle \boldsymbol{x}_i, \boldsymbol{w}_i \rangle + \epsilon_i$, where $\boldsymbol{w}_i \sim N(\boldsymbol{0}, \boldsymbol{S})$, where $N(\boldsymbol{0}, \boldsymbol{S})$ is the centered Gaussian distribution on $\mathbb{R}^d$ with PSD covariance matrix $\boldsymbol{S}$. The goal here is to recover the covariance matrix $\boldsymbol{S}$ of the weight vectors from observations of $(y_i, \boldsymbol{x}_i)$, $i = 1, \dots, n$. This problem has natural connections to mixture of regressions models as well [De Veaux, 1989, Yi et al., 2014, Zhong et al., 2016, Sedghi et al., 2016].

In terms of the complexity of various methods, solving the semidefinite program using off-the-shelf solvers takes $O(nd^2 + d^3)$ complexity. Gradient descent on PSD matrices (see the initialization phase in [Tu et al., 2016]) can solve this with per-iteration complexity $O(nd^2)$, and one can prove linear convergence under certain assumptions. The most directly comparable methods are nonconvex gradient descent [Li et al., 2019], which have per-iteration complexity $O(ndr)$ and utilize a Burer-Monteiro factorization [Burer and Monteiro, 2003]. As we will see, our method also has per-iteration complexity of $O(ndr)$.

Finally, we mention several works dedicated to the computation of the nonconvex Wasserstein barycenter problem [Agueh and Carlier, 2011, Álvarez-Esteban et al., 2016, Zemel and Panaretos, 2019, Chewi et al., 2020,

Altschuler et al., 2021]. No theoretical study in these works allows for rank deficiency.

Notation. Bold capital letters denote matrices while bold lower-case letters denote vectors. The Hilbert-Schmidt inner product is $\langle \cdot, \cdot \rangle$, and $\langle \cdot, \cdot \rangle_{\gamma_{\boldsymbol{\Sigma}}} = \langle \cdot, \cdot \rangle_{\boldsymbol{\Sigma}}$ is the Riemannian metric associated to the (fixed-rank) BW manifold at $\boldsymbol{\Sigma}$. Their corresponding norms are written as $\| \cdot \|$ and $\| \cdot \|_{\boldsymbol{\Sigma}}$, respectively. The orthogonal projection onto the column span of $\boldsymbol{\Sigma}$ is $\boldsymbol{P}_{\boldsymbol{\Sigma}} = \boldsymbol{P}_{\mathrm{Sp}(\boldsymbol{\Sigma})}$. Similarly, $\boldsymbol{P}_{\boldsymbol{\Sigma}}^{\perp}$ is the projection onto null-space of $\boldsymbol{\Sigma}$ (orthogonal complement of $\mathrm{Sp}(\boldsymbol{\Sigma})$). The Dirac distribution at a point x denoted by δ_x.

Outline. We begin by outlining our approach to matrix recovery in Section 2. We then give the main theoretical results for our approach in Section 3. After this, we give experiments demonstrating these advantages in Section 4, and discuss the limitations of our work in Section 5.

2 THE BURES-WASSERSTEIN BARYCENTER APPROACH

Recall that we aim at find the rank r matrix $\boldsymbol{S} \in \mathbb{S}_+^d$ given the observations (1.1). We will use the notation $y_i = \mathcal{A}(\boldsymbol{S})_i = \langle \boldsymbol{x}_i \boldsymbol{x}_i^\top, \boldsymbol{S} \rangle$, $i = 1, \dots, n$. To recover a low-rank matrix $\boldsymbol{S}$ from these measurements, most past work has focused on some form of energy minimization. For example, some works have looked at convex nuclear norm minimization methods. Cai and Zhang [2015] and Chen et al. [2015] concurrently developed a nuclear norm minimization procedure that solves

$$\min_{\boldsymbol{\Sigma} \in \mathbb{S}_+^d,\ \mathcal{A}(\boldsymbol{\Sigma}) = \boldsymbol{y}} \mathrm{Tr}(\boldsymbol{\Sigma}), \tag{2.1}$$

where $\boldsymbol{y} = [y_1, \dots, y_n]^\top$ and $\mathcal{A}(\boldsymbol{\Sigma}) = [\mathcal{A}(\boldsymbol{\Sigma})_1, \dots, \mathcal{A}(\boldsymbol{\Sigma})_n]^\top$. One can directly solve the semidefinite program using standard convex optimization packages. A Lagrangian formulation of (2.1) yields the energy $\frac{1}{2}\|\mathcal{A}(\boldsymbol{\Sigma}) - \boldsymbol{y}\|_2^2 + \lambda \mathrm{Tr}(\boldsymbol{\Sigma})$, which can also be minimized using a variety of methods.

In the following, we lay out our approach to the low-rank matrix recovery problem, which focuses on a new energy minimization procedure. In Section 2.1 we discuss the common nonconvex approaches to matrix recovery and outline our novel optimization program. Then, in Section 2.2, we discuss the BW barycenter problem, and show how it recovers solutions to the energy minimization we propose. After this, in Section 2.3 we outline the first-order algorithms for computing BW barycenters. We finish in Section 2.4 by discussing a regularization procedure that allows one to estimate higher rank proxies, from which it is possible to recover $\boldsymbol{S}$.

2.1 Nonconvex Approaches for Matrix Recovery

Suppose that we know an upper bound for the rank of the underlying matrix $\boldsymbol{S}$. We could utilize this information in a nonconvex optimization program such as

$$\min_{\boldsymbol{\Sigma} \in \mathbb{S}_+^d,\ \mathrm{rank}(\boldsymbol{\Sigma}) \leq r} \frac{1}{2n} \|\mathcal{A}(\boldsymbol{\Sigma}) - y\|_2^2. \tag{2.2}$$

Without the rank restriction (i.e., $r = d$) this problem is in fact convex. For any fixed $r \leq d$, we can parameterize the rank r matrices in $\mathbb{S}_+^d$ by $\boldsymbol{U}\boldsymbol{U}^\top$, for $\boldsymbol{U} \in \mathbb{R}^{d \times r}$, which is now commonly referred to as Burer-Monteiro factorization [Burer and Monteiro, 2003]. We thus define the set of PSD matrices of rank at most r using this factorization: $\mathbb{S}_+^{d,r} := \{\boldsymbol{\Sigma} \in \mathbb{S}_+^d : \boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{U}^\top,\ \boldsymbol{U} \in \mathbb{R}^{d \times r}\}$. With this parametrization, the matrix recovery problem in (2.2) is equivalent to

$$\min_{\boldsymbol{U} \in \mathbb{R}^{d \times r}} \frac{1}{2} \|\mathcal{A}(\boldsymbol{U}\boldsymbol{U}^\top) - \boldsymbol{y}\|_2^2. \tag{2.3}$$

While past work has focused on these least squares formulations, there have not been many modifications of this energy. We propose the following modifications to the energies (2.2) and (2.3):

$$\min_{\boldsymbol{\Sigma} \in \mathbb{S}_+^d,\ \mathrm{rank}(\boldsymbol{\Sigma}) \leq r} \frac{1}{2n} \|\sqrt{\mathcal{A}(\boldsymbol{\Sigma})} - \sqrt{\boldsymbol{y}}\|_2^2 \tag{2.4}$$

$$= \min_{\boldsymbol{U} \in \mathbb{R}^{d \times r}} \frac{1}{2n} \|\sqrt{\mathcal{A}(\boldsymbol{U}\boldsymbol{U}^\top)} - \sqrt{\boldsymbol{y}}\|_2^2, \tag{2.5}$$

where the square root is taken componentwise. As we demonstrate in the following sections, this problem has a natural solution as a BW barycenter.

2.2 The Bures-Wasserstein Barycenter Problem

To explain the connection of (2.4) to BW barycenters, we will first explain how BW space arises from the perspective of optimal transport [Villani, 2009]. Let $\mathcal{P}_2(\mathbb{R}^d)$ be the set of all measures on $\mathbb{R}^d$ with finite second moment. The 2-Wasserstein distance between measures μ and $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ is defined by

$$W_2^2(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \mathbb{E}_{(\boldsymbol{x}, \boldsymbol{y}) \sim \pi} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2, \tag{2.6}$$

where $\Pi(\mu, \nu)$ denotes the set of all couplings between μ and ν (i.e., the set of all joint distributions on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and ν). The 2-Wasserstein distance defines a metric over $\mathcal{P}_2(\mathbb{R}^d)$, and the resulting geodesic metric space is referred to as 2-Wasserstein space.

Let $\mathcal{N}(\mathbb{R}^d)$ denote the set of Gaussian distributions on $\mathbb{R}^d$, and $\mathcal{N}_0(\mathbb{R}^d)$ be the set of centered Gaussian distributions. Both are geodesically weakly convex subsets of 2-Wasserstein space, meaning there always exist 2-Wasserstein geodesics between points in these sets that are

contained within these sets. Letting $N(\mathbf{0}, \mathbf{\Sigma}) \in \mathcal{N}_0(\mathbb{R}^d)$ denote the Gaussian distribution on $\mathbb{R}^d$ with mean zero and covariance matrix $\mathbf{\Sigma} \in \mathbb{S}^d_+$, the 2-Wasserstein distance between $N(\mathbf{0}, \mathbf{\Sigma}_0)$ and $N(\mathbf{0}, \mathbf{\Sigma}_1) \in \mathcal{N}_0(\mathbb{R}^d)$ has the explicit form

$$W_2^2(N(\mathbf{0}, \mathbf{\Sigma}_0), N(\mathbf{0}, \mathbf{\Sigma}_1)) \\ = \operatorname{Tr}\left[\mathbf{\Sigma}_0 + \mathbf{\Sigma}_1 - 2(\mathbf{\Sigma}_0^{1/2}\mathbf{\Sigma}_1\mathbf{\Sigma}_0^{1/2})^{1/2}\right]. \tag{2.7}$$

Notice that this is purely a function of the covariance matrices, and so the Wasserstein distance induces a distance metric on PSD matrices called the *Bures-Wasserstein distance* [Bhatia et al., 2019]. To refer to this distance over PSD matrices rather than the Gaussian distributions, we will write

$$d_{\mathsf{BW}}(\mathbf{\Sigma}_0, \mathbf{\Sigma}_1) = W_2(N(\mathbf{0}, \mathbf{\Sigma}_0), N(\mathbf{0}, \mathbf{\Sigma}_1)). \tag{2.8}$$

More than just giving the set of PSD matrices a distance metric, this identification endows $\mathbb{S}^d_+$ with a natural Riemannian structure that it inherits from $(\mathcal{N}_0(\mathbb{R}^d), W_2)$.

The barycenter problem seeks to generalize the notion of averages to non-Euclidean spaces. In the 2-Wasserstein barycenter problem, one seeks a solution to

$$\min_{b \in \mathcal{P}_2(\mathbb{R}^d)} \frac{1}{2}\mathbb{E}_{\mu \sim Q} W_2^2(\mu, b), \tag{2.9}$$

where Q is a distribution over $\mathcal{P}_2(\mathbb{R}^d)$ with finite second moment, which we write as $Q \in \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$. When Q is supported on Gaussians, the minimum is achieved on Gaussians [Knott and Smith, 1994, Agueh and Carlier, 2011, Álvarez-Esteban et al., 2016]. For $\mathcal{N}_0(\mathbb{R}^d)$, due to the identification in (2.7), this is equivalent to the Fréchet mean of PSD matrices on the BW manifold. Without loss of generality, if Q is supported on mean zero Gaussians, we will therefore think of Q as a distribution over PSD matrices.

We finally arrive at the connection between low-rank PSD matrix recovery and BW barycenters. The following proposition connects the barycenter problem (2.9) when $Q = Q_n = \frac{1}{n}\sum_{i=1}^n \delta_{N(\mathbf{0}, y_i \boldsymbol{x}_i \boldsymbol{x}_i^\top)}$ to our new low-rank matrix recovery program (2.4), provided that $\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\top = \boldsymbol{I}$. This proposition indicates that we can recover the matrix $\boldsymbol{S}$ by solving a Wasserstein barycenter problem.

Proposition 1. *If $\frac{1}{n}\sum_i \boldsymbol{x}_i \boldsymbol{x}_i^\top = \boldsymbol{I}$, then*

$$\operatorname*{argmin}_{\mathbf{\Sigma} \in \mathbb{S}^d_+} \frac{1}{2n} \|\sqrt{\mathcal{A}(\mathbf{\Sigma})} - \sqrt{\boldsymbol{y}}\|_2^2 \\ = \operatorname*{argmin}_{\mathbf{\Sigma} \in \mathbb{S}^d_+} \frac{1}{2n}\sum_{i=1}^n d^2_{\mathsf{BW}}(\mathbf{\Sigma}, y_i \boldsymbol{x}_i \boldsymbol{x}_i^\top). \tag{2.10}$$

To make this result practical, we cannot assume in general that $\frac{1}{n}\sum_i \boldsymbol{x}_i \boldsymbol{x}_i = \boldsymbol{I}$. If we instead encounter a case where $\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\top = \boldsymbol{C}_n$, where $\boldsymbol{C}_n$ is the PD sample covariance matrix of the vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, then the transformation $\boldsymbol{x}_i \mapsto \boldsymbol{C}_n^{-1/2}\boldsymbol{x}_i$ outputs vectors with identity covariance (i.e., $\frac{1}{n}\sum_{i=1}^n \boldsymbol{C}_n^{-1/2}\boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{C}_n^{-1/2} = \boldsymbol{I}$). We are able use this fact to recover the matrix $\boldsymbol{S}$, as we show in the following proposition.

Proposition 2. *Let $\boldsymbol{C}_n = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\top \in \mathbb{S}^d_{++}$. Then*

$$\operatorname*{argmin}_{\mathbf{\Sigma} \in \mathbb{S}^d_+} \frac{1}{2n} \|\sqrt{\mathcal{A}(\boldsymbol{C}_n^{-1/2}\mathbf{\Sigma}\boldsymbol{C}_n^{-1/2})} - \sqrt{\boldsymbol{y}}\|_2^2 \\ = \operatorname*{argmin}_{\mathbf{\Sigma} \in \mathbb{S}^d_+} \frac{1}{2n}\sum_{i=1}^n d^2_{\mathsf{BW}}(\mathbf{\Sigma}, y_i \boldsymbol{C}_n^{-1/2}\boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{C}_n^{-1/2}), \tag{2.11}$$

and $\boldsymbol{C}_n^{1/2}\boldsymbol{S}\boldsymbol{C}_n^{1/2}$ is a solution to both problems.

Notice that one can recover $\boldsymbol{S}$ from $\boldsymbol{C}_n^{1/2}\boldsymbol{S}\boldsymbol{C}_n^{1/2}$ as $\boldsymbol{S} = \boldsymbol{C}_n^{-1/2}\boldsymbol{C}_n^{1/2}\boldsymbol{S}\boldsymbol{C}_n^{1/2}\boldsymbol{C}_n^{-1/2}$. Thus, in the sample setting where $\boldsymbol{C}_n$ is not exactly the identity and assuming we can solve the barycenter problem, we envision a two stage procedure: 1) recover the barycenter $\mathbf{\Sigma}_n$ of $y_1\boldsymbol{C}_n^{-1/2}\boldsymbol{x}_1\boldsymbol{x}_1^\top\boldsymbol{C}_n^{-1/2}, \ldots, y_n\boldsymbol{C}_n^{-1/2}\boldsymbol{x}_n\boldsymbol{x}_n^\top\boldsymbol{C}_n^{-1/2}$, and 2) transform the barycenter by $\boldsymbol{C}_n^{-1/2}\mathbf{\Sigma}_n\boldsymbol{C}_n^{-1/2}$ to find $\boldsymbol{S}$.

The whitening step can be efficiently computed since we can use any linear transformation $\boldsymbol{L}^{-1}$ such that $\boldsymbol{L}^{-1}\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{L}^{-1\top} = \boldsymbol{I}$. For example, with the Cholesky factorization $\boldsymbol{C}_n = \boldsymbol{L}\boldsymbol{L}^\top$, we can solve the equations $\boldsymbol{L}\boldsymbol{z}_i = \boldsymbol{x}_i$ for $\boldsymbol{z}_i$, and these satisfy $\frac{1}{n}\sum_i \boldsymbol{z}_i \boldsymbol{z}_i^\top = \boldsymbol{I}$.

Remark 3. *The assumption that $\boldsymbol{C}_n$ is full rank can be relaxed by using the pseudoinverse of $\boldsymbol{C}_n$ rather than the inverse. However, for our later theorems we assume $n = O(dr)$ i.i.d. samples from a $N(\mathbf{0}, \boldsymbol{I})$ distribution in order to obtain a restricted isometry property. In this case, the matrix $\boldsymbol{C}_n$ will be full rank with probability 1.*

The connections established by Propositions 1 and 2 enables the development of novel methods for the matrix recovery problem, since we can solve a specific Wasserstein barycenter problem rather than the original matrix recovery problem (2.4). In other words, any methods that solve this Wasserstein barycenter problem could be used to solve the matrix recovery problem. Since barycenters are geometric notions of averages, this naturally leads to the development of novel geometric methods for matrix recovery.

2.3 Algorithms for Barycenters

The primary way to compute BW barycenters involves Riemannian gradient descent [Álvarez-Esteban et al., 2016, Chewi et al., 2020, Altschuler et al., 2021]. Following the results in the last section, we wish to find the BW barycenter of the matrices $\boldsymbol{X}_i = y_i\boldsymbol{C}_n^{-1/2}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{C}_n^{-1/2}$, $i = 1, \ldots, n$. In other words, we seek to minimize the energy function $F : \mathbb{S}^d_+ \to \mathbb{R}$ given by

$$F(\mathbf{\Sigma}) = \frac{1}{2n}\sum_{i=1}^n d^2_{\mathsf{BW}}(\boldsymbol{X}_i, \mathbf{\Sigma}). \tag{2.12}$$

For ease of notation, we will write F as an expectation over $y_i \boldsymbol{C}_n^{-1/2} \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{C}_n^{-1/2} \sim Q$, and whether or not Q is a discrete distribution will be made clear from context.

The gradient of F at full rank $\boldsymbol{\Sigma}_0$ is

$$\nabla F(\boldsymbol{\Sigma}_0) = \boldsymbol{I} - \mathbb{E}_{\boldsymbol{X}\sim Q} \frac{\boldsymbol{X}}{\sqrt{\mathrm{Tr}(\boldsymbol{X}\boldsymbol{\Sigma}_0)}} =: \boldsymbol{I} - \tilde{T}(\boldsymbol{\Sigma}_0),$$

where for convenience we have defined the quantity $\tilde{T}(\boldsymbol{\Sigma}_0)$. At low-rank $\boldsymbol{\Sigma}_0$, this is a subgradient [Clarke, 1990]. We note that this holds for general measures Q where we can differentiate under the integral.

Wasserstein gradient descent uses the gradient to determine a geodesic along which to move. In Wasserstein space, this is the "pushforward" direction in a base measure is transported. For a complete description of Wasserstein gradient descent, see [Álvarez-Esteban et al., 2016, Zemel and Panaretos, 2019, Chewi et al., 2020, Altschuler et al., 2021], and for a more complete discussion of BW geometry, the reader should consult Bhatia et al. [2019]. We have included a discussion of the geodesic structure of BW space in the appendix. For our purposes, we consider BWGD (BWGD) with step size η_k:

$$\boldsymbol{\Sigma}_{k+1} = (\boldsymbol{I} - \eta_k \nabla F(\boldsymbol{\Sigma}_k))\boldsymbol{\Sigma}_k(\boldsymbol{I} - \eta_k \nabla F(\boldsymbol{\Sigma}_k)). \quad (2.13)$$

When $\eta_k = 1$, this corresponds to the fixed point iteration of Álvarez-Esteban et al. [2016] (see also Appendix B). Note that this is easy to extend to the stochastic setting: if we observe a stochastic gradient $\boldsymbol{G}_k$ rather than $\nabla F(\boldsymbol{\Sigma}_k)$, then the BW stochastic gradient descent (BWSGD) iteration would use $\boldsymbol{G}_k$ in place of $\nabla F(\boldsymbol{\Sigma}_k)$. For example, one common variant of such a stochastic gradient method uses

$$\boldsymbol{G}_k = \boldsymbol{I} - \frac{\boldsymbol{X}_k}{\mathrm{Tr}(\boldsymbol{X}_k \boldsymbol{\Sigma}_k)^{1/2}}, \quad (2.14)$$

where $k = 1, \dots, n$. In other words, this variant of stochastic gradient descent passes over each sample one at a time. At each point in time, we take a gradient with respect to that sample alone and move in the minus gradient direction.

To save computational time and to allow efficient computation in the low-rank case, we modify the BWGD iteration (2.13) and the BWSGD iteration to instead operate on factorized matrices (i.e., on the formulation seen in (2.5)). This means that instead of storing the sequence $\boldsymbol{\Sigma}_k$ for $k \in \mathbb{N}$, we instead store the sequence

$$\boldsymbol{U}_{k+1} = (\boldsymbol{I} - \eta_k \nabla F(\boldsymbol{U}_k \boldsymbol{U}_k^\top))\boldsymbol{U}_k. \quad (2.15)$$

We note that if $\boldsymbol{\Sigma}_k$ is low-rank in (2.13), then the update in (2.15) is equivalent to (2.13). In particular, it is not hard to show that if $\boldsymbol{U}_k \boldsymbol{U}_k^\top = \boldsymbol{\Sigma}_k$, then $\boldsymbol{U}_{k+1}\boldsymbol{U}_{k+1}^\top = \boldsymbol{\Sigma}_{k+1}$. In the same way as before, stochastic gradient methods naturally extend to the low-rank setting. This update corresponds to a geodesic gradient descent update over the fixed rank BW manifold [Massart and Absil, 2020].

When $Q = Q_n$ is a discrete distribution, which is the main focus of this paper, the BWGD updates can be computed in $O(ndr)$ time. This follows from the fact that we can rewrite the update in (2.15) as

$$\boldsymbol{U}_{k+1} = (1 - \eta_k)\boldsymbol{U}_k + \frac{\eta_k}{n} \sum_{i=1}^n \frac{\boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{U}_k}{\|\boldsymbol{U}_k^\top \boldsymbol{x}_i\|_2}. \quad (2.16)$$

In the case of single-sample streaming BWSGD, which uses the gradient in (2.14), the updates take $O(dr)$ time. We also note that both the BWGD and BWSGD iterations maintain the rank of the updated matrix for all $\eta_k \in [0, 1)$. For $\eta_k = 1$, the rank is maintained for BWGD as long as $\mathrm{rank}(\sum_i \boldsymbol{x}_i \boldsymbol{x}_i^\top) = d$.

2.4 Regularization through Perturbed Gradient Descent

As we demonstrate later, our current theorem only applies for $r \geq 3$, despite the fact that we observe good performance in practice for the iteration (2.15) when $r = 1$ and $r = 2$. We comment that a modified version of our method can yield results in the cases of $r = 1$ or $r = 2$.

To demonstrate how this works, consider the observation model $y_i = \langle \boldsymbol{x}_i \boldsymbol{x}_i^\top, \boldsymbol{S} \rangle$ for a rank r matrix $\boldsymbol{S}$ and $i = 1, \dots, n$. If we choose an arbitrary rank r' matrix $\boldsymbol{\Delta}$, we could instead try to recover the rank at most $r + r'$ matrix $\boldsymbol{S} + \boldsymbol{\Delta}$ from the observations $\tilde{y}_i = \langle \boldsymbol{x}_i \boldsymbol{x}_i^\top, \boldsymbol{S} + \boldsymbol{\Delta} \rangle = \langle \boldsymbol{x}_i \boldsymbol{x}_i^\top, \boldsymbol{S} \rangle + \langle \boldsymbol{x}_i \boldsymbol{x}_i^\top, \boldsymbol{\Delta} \rangle$, which can be computed by computing and adding the factors $\langle \boldsymbol{x}_i \boldsymbol{x}_i^\top, \boldsymbol{\Delta} \rangle$ to the observations y_i. If we then find the barycenter of $\tilde{y}_i \boldsymbol{x}_i \boldsymbol{x}_i^\top$ (provided that it is unique, which we prove later under conditions on the vectors $\boldsymbol{x}_i$), then this would recover $\boldsymbol{S} + \boldsymbol{\Delta}$! Since this is true for any such $\boldsymbol{\Delta}$, the user can pick $\boldsymbol{\Delta}$ beforehand and then recover $\boldsymbol{S}$ from simple subtraction: $\boldsymbol{S} = (\boldsymbol{S} + \boldsymbol{\Delta}) - \boldsymbol{\Delta}$. In brief, every rank r recovery problem can be solved by instead first solving the rank $r + r'$ problem to find $\boldsymbol{S} + \boldsymbol{\Delta}$ and then subtracting off the perturbation factor $\boldsymbol{\Delta}$.

3 THEORETICAL RESULTS

We now discuss the main theoretical results of this paper. We make the following assumption on our measurement model in (1.1).

Assumption 1. *We observe data from the model (1.1) with* $\boldsymbol{x}_i \overset{i.i.d.}{\sim} N(\boldsymbol{0}, \boldsymbol{I})$. *The underlying matrix* $\boldsymbol{S}$ *is rank* r *and satisfies* $m \leq \lambda_r(\boldsymbol{S}) \leq \lambda_1(\boldsymbol{S}) \leq M$.

The assumption of Gaussianity can be weakened to sub-Gaussianity, but we state the result for Gaussian vectors since it simplifies some of the later proofs (for example, the proof of Euclidean smoothness in Appendix C.2.3). The sub-Gaussianity is essential for an ℓ_2/ℓ_1 restricted isometry property [Chen et al., 2015, Cai and Zhang, 2015] that we need to hold (see Appendix C.2.1).

Our main result on matrix recovery with gradient descent for BW barycenters is given in the following theorem.

Theorem 4. *Suppose that we observe $y_i = \langle \boldsymbol{x}_i \boldsymbol{x}_i^\top, \boldsymbol{S} \rangle$, $i = 1, \dots, n$, where $\boldsymbol{x}_i$ and $\boldsymbol{S}$ satisfy Assumption 1. Suppose further that $r = \operatorname{rank}(\boldsymbol{S}) \geq 5$, and let $\boldsymbol{S}_n = \boldsymbol{C}_n^{1/2} \boldsymbol{S} \boldsymbol{C}_n^{1/2}$. Then, for constants c_1, c_2, if $n \gtrsim dr$, with probability at least $1 - \exp(-c_2 n) - 1/n$,*

1. *$\boldsymbol{S}_n$ is the unique global minimizer of F over $\mathbb{S}_+^d$.*
2. *Let $\boldsymbol{U}_0$ be the initial iterate of BWGD. If $\lambda_1(\boldsymbol{U}_0 \boldsymbol{U}_0^\top) \leq M$, and $F(\boldsymbol{U}_0 \boldsymbol{U}_0^\top) - F(\boldsymbol{S}_n) \leq \frac{c_1^5 m^8 (r-2)^2 \beta^{10}}{6^5 M^{15/2} d^3 r^2}$ for an $\boldsymbol{S}$ dependent constant β, then BWGD with step size 1 satisfies the following bound for $\mathcal{C} = O\big(c_1 m / M^{3/2}\big)$.*

$$d_{\mathsf{BW}}^2(\boldsymbol{U}_k \boldsymbol{U}_k^\top, \boldsymbol{C}_n^{1/2} \boldsymbol{S} \boldsymbol{C}_n^{1/2}) \leq (1 - \mathcal{C})^k (F(\boldsymbol{U}_0 \boldsymbol{U}_0^\top) - F(\boldsymbol{S})). \tag{3.1}$$

This is the first result for convergence in Bures-Wassersten distance in the literature. The constant $\mathcal{C}$ is the local strong geodesic convexity constant seen in (C.50). We note that this amounts to showing that our new energy given by (2.4) has a positive definite Hessian in a neighborhood around $\boldsymbol{S}$. A couple of remarks are in order to discuss two issues that arise with the analysis of our method: initialization and the rank constraint on $\boldsymbol{S}$. In practice, we observe convergence to $\boldsymbol{S}$ from random initialization, but do not currently have a proof of this fact.

Remark 5 (Initialization)**.** *We note that the convergence bound in Theorem 4 is local. A few procedures can be used to initialize in the correct neighborhood. First, under the Gaussian assumption, $\mathbb{E} y(\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{I}) = 2\boldsymbol{S}$, and so one could use a rank r approximation of $\frac{1}{2n} \sum_{i=1}^n y_i (\boldsymbol{x}_i \boldsymbol{x}_i^\top - \boldsymbol{I})$ to initialize the gradient descent. This approximation can be computed in $O(ndr)$ time with the power method. A finite sample approximation result follows from concentration of the fourth moment tensor, see [Diakonikolas et al., 2019, Theorem 4.13] for details. Another path to initialization would be to use gradient descent on full rank PSD matrices in the first stage since, as we show in the Appendix, the function (2.10) is convex over PD matrices. This would allow one to recover a good approximation to $\boldsymbol{S}$, and take a rank r approximation of it, and then run BWGD from there. The downside of this method is that it takes complexity $O(nd^2)$ to compute, but it would not need concentration of the fourth moment tensor. Also, the gradient descent procedure with overspecified rank converges sublinearly.*

Remark 6 (Rank Constraint)**.** *The result holds for $\operatorname{rank}(\boldsymbol{S}) \geq 5$. This is due to a smoothness bound that requires bounds on the expectation and variance of $\mathbb{E}\sqrt{1 + (x_{r+1}^2 + \dots + x_{2r}^2)/(x_1^2 + \dots + x_r)^2}$ for $x_i \overset{i.i.d.}{\sim} N(0,1)$ random variables. In order for $\mathbb{E} 1/(x_1^2 + \dots + x_r)^2$ to exist, we need $r \geq 3$, and in order for $1/(x_1^2 + \dots + x_r)^2$ to have finite variance, we need $r \geq 5$. In practice, and as we show in the experiments, the method still succeeds much of the time for $r = 1, \dots, 5$. To have a theoretically guaranteed method in these settings, one can use the regularized BWGD method of Section 2.4.*

We give a sketch of the proof of Theorem 4.

- We first show that the energy F in (2.12) is Euclidean strongly convex over $\mathbb{S}_+^d$ and that $\boldsymbol{S}_n$ is the unique minimizer with high probability. This result uses the ℓ_2/ℓ_1-RIP condition of Chen et al. [2015] (or restricted uniform boundedness condition of Cai and Zhang [2015]).
- We then prove that the gradient is locally $1/2$-Hölder continuous: $\|\nabla F(\boldsymbol{\Sigma}) - \nabla F(\boldsymbol{S}_n)\|_F \lesssim \|\boldsymbol{\Sigma} - \boldsymbol{S}_n\|_F^{1/2}$ for rank r matrices $\boldsymbol{\Sigma}$.
- Using this smoothness result, we are able so show that F is *locally-BW-geodesically convex* around $\boldsymbol{S}_n$ in $\mathbb{S}_+^{d,r}$.
- Local strong geodesic convexity along with a geodesic smoothness result yields the local linear convergence result via standard optimization arguments (see, for example, Bubeck [2015], Zhang and Sra [2016]).

We also include the following theorem on the convergence of BWSGD. This guarantees a slow rate of convergence for BWSGD with gradient given by (2.14).

Theorem 7. *Suppose that we observe $y_i = \langle \boldsymbol{x}_i \boldsymbol{x}_i^\top, \boldsymbol{S} \rangle$, $i = 1, \dots, n$, where $\boldsymbol{x}_i$ and $\boldsymbol{S}$ satisfy Assumption 1. Suppose that we run single sample streaming BWSGD, which uses gradient (2.14), for n iterations with step size $1/\sqrt{n}$. Then, for a constant c_1, if $n \gtrsim dr$, with probability at least $1 - \exp(-c_1 n) - 1/n$,*

$$\min_{k=1,\dots,n} \mathbb{E} \|\nabla F(\boldsymbol{\Sigma}_k)\|_{\boldsymbol{\Sigma}_k}^2 = O\big(n^{-1/2}\big), \tag{3.2}$$

where $\|\cdot\|_{\boldsymbol{\Sigma}}^2 = \mathbb{E}_{\boldsymbol{z} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})} \|\cdot \boldsymbol{z}\|_2^2$ is the norm induced by the BW Riemannian metric.

This states that the best iterate of the BWSGD sequence outputs an approximate stationary point with respect to the norm induced by the BW Riemannian metric. However, we cannot guarantee that this stationary point is the global minimum. We comment that we do not tend to run into spurious local minima in practice, and future work should go into studying this fact. While we do not have a justification for this, we give a theorem on the $r = 1$ case, where we show that the energy function has no local minima in the asymptotic limit. The proof of this theorem is left to the supplementary material.

Theorem 8. *Consider the observation model with the rank one matrix $\boldsymbol{S} = \boldsymbol{v}\boldsymbol{v}^\top$ and $y = \langle \boldsymbol{x}\boldsymbol{x}^\top, \boldsymbol{S} \rangle$. Then, the only*

fixed points of the population version of (2.15) *(which corresponds to gradient descent on* (2.12) *with the sum replaced by an integral) are* $\boldsymbol{v}$ *or orthogonal to* $\boldsymbol{v}$*. In particular, this implies that population* BWGD *from any initialization such that* $\boldsymbol{u}_0 \not\perp \boldsymbol{v}$ *converges to* $\boldsymbol{v}$*.*

For the case of general r, we believe that a similar result holds, although we have not yet been able to show it. Furthermore, we also believe that these results can be extended to high probability results in the finite sample case.

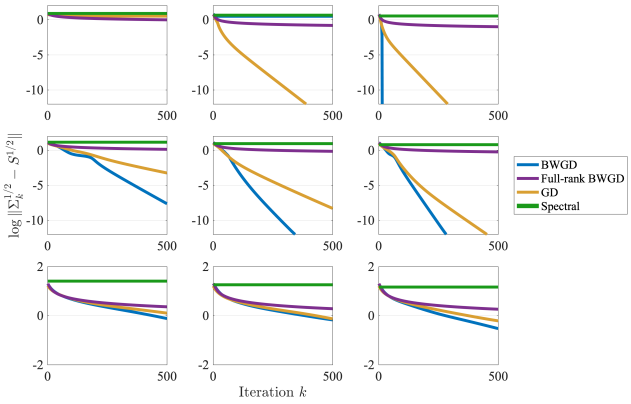

Figure 2: Convergence of BWGD (2.15) compared with full rank BWGD as well as the Euclidean GD [Li et al., 2019]. Here, $d = 32$, and down the columns we use $r = 1, 4, 16$, respectively. Across the rows we vary the number of points, with $n = 3dr$, $n = 10dr$, and $n = 20dr$, respectively. As we can see, for low to moderate ranks, the BWGD method converges much faster than the standard GD method.

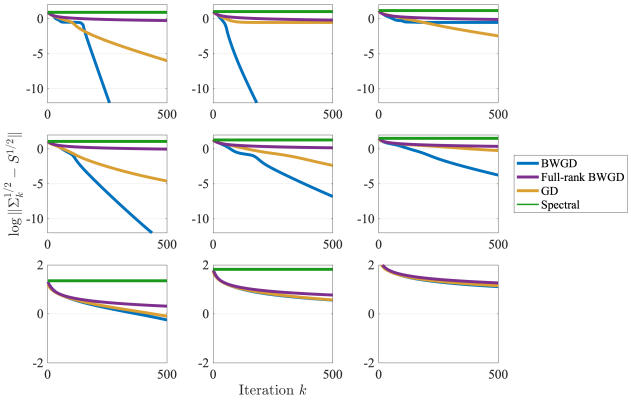

Figure 3: Convergence of BWGD (2.15) compared with full rank BWGD as well as the Euclidean GD [Li et al., 2019]. Here, $d = 32$, and down the columns we use $r = 2, 4, 16$, respectively. Across the rows we vary scale factor α, with $\alpha = 0, 1, 2$, respectively. As we can see, the BWGD method converges much faster than the standard GD method.

4 EXPERIMENTS

Here we present some numerical simulations that demonstrate the effectiveness of our method. All experiments were run in MATLAB on a 2020 Macbook Pro with a 2Ghz Quad-Core Intel Core i5 CPU and 16 GB of RAM.

4.1 Synthetic Experiments

We begin with some experiments on generated datasets to better understand the performance of our method. Since our main results focus on the performance of gradient descent rather than stochastic gradient descent, we focus on its performance here.

The methods we compare are rank r BWGD, full rank BWGD, the nonconvex Euclidean Gradient Descent (GD) of Li et al. [2019], and a spectral method, which takes a low-rank approximation to $\boldsymbol{S}_n = \frac{1}{2n}\sum_{i=1}^n y_i(\boldsymbol{x}_i\boldsymbol{x}_i^\top - \boldsymbol{I})$ since $\frac{1}{2}\mathbb{E}y^2(\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{I}) = \boldsymbol{S}$ [Sedghi et al., 2016]. As an error metric, we compute $\|\boldsymbol{\Sigma}_k^{1/2} - \boldsymbol{S}^{1/2}\|_F = \Theta(d_{\mathsf{BW}}(\boldsymbol{\Sigma}_k, \boldsymbol{S}))$. We do not compare with other matrix recovery algorithms since they are not guaranteed to work in the symmetric rank one projection setting. Also, we find the comparison with GD to be the most relevant, since BWGD and GD have comparable complexity and are both first-order methods.

In our first experiment, we test the accuracy of the methods over time as we vary the rank of $\boldsymbol{S}$ and the sample size. Since the spectral method is not iterative, we include it as a horizontal line. Figure 2 displays the results of this experiment. Across the rows we vary the number of points and down the columns we vary the rank of $\boldsymbol{S}$. The fixed dimension is $d = 32$, the ranks from top to bottom are $r = 1, 4, 16$, and across the rows the number number points are $3dr$, $10dr$, and $20dr$, respectively. For each frame, we generate 20 datasets and run the four methods on them. All methods are run with random initialization, where the entries of $\boldsymbol{U}_0$ are i.i.d. $N(0, 1)$. For $r = 1$, we see that rank r BWGD succeeds once the number of points is sufficiently large. Furthermore, the convergence when it is successful is extremely fast. For moderate ranks, rank r BWGD converges faster than the previous GD method of Li et al. [2019]. We note that in this figure, there are some conflating factors that affect the convergence. Most notably, the conditioning in this problem (m versus M) gets worse as rank increases since $\boldsymbol{S} = \boldsymbol{V}\boldsymbol{V}^T$ and the entries of $\boldsymbol{V}$ are i.i.d. $N(0, 1)$. Better conditioning leads to faster for higher ranks, see Figure 7 in the appendix, where the columns of $\boldsymbol{V}$ are orthogonal and length $\sqrt{d}$. Also, there is the additional factor of c_1 in the strong convexity constant in Theorem 3, which is the RIP constant. While in theory the constant is universal once the number of points is sufficiently large, one expects some scaling in terms of rank for finite n. .

In Figure 3, we examine the performance of the methods under varying conditioning of $\boldsymbol{S}$. We set $d = 32$ and $n = 5dr$ and vary r as well as the conditioning of the matrix $\boldsymbol{S}$. Here, $\boldsymbol{S} = \boldsymbol{V}\mathrm{diag}(r^\alpha, (r-1)^\alpha, \ldots, 1^\alpha)\boldsymbol{V}^\top$, where the entries of $\boldsymbol{V}$ are i.i.d. $N(0,1)$ and α is a scale factor. Figure 3 displays the results on 20 randomly generated datasets per frame. The rows correspond to $r = 2, 4$, and 16 respectively. The columns correspond to $\alpha = 0, 1$ and 2 respectively. Rank r BWGD performs uniformly well throughout.

In Figure 4 we show the dependence on sample size. Here, $d = 64$ and r is varied from 1 to 20. The error of BWGD and GD after 200 iterations is shown for sample sizes of d, $2d$, $\ldots$, $20d$ for each value of r. For each r, n pair, we generate 20 datasets and compute the average error across them. The color indicates the average error value across these datasets. As we see, BWGD performs the best out of these methods. GD with linesearch is also competitive, but is more time consuming.

For the final synthetic experiment in Figure 5, we demonstrate the scalability of BWGD to higher dimensions. We note that it scales much better in terms of actual computational time when compared with the Euclidean GD method of Li et al. [2019].

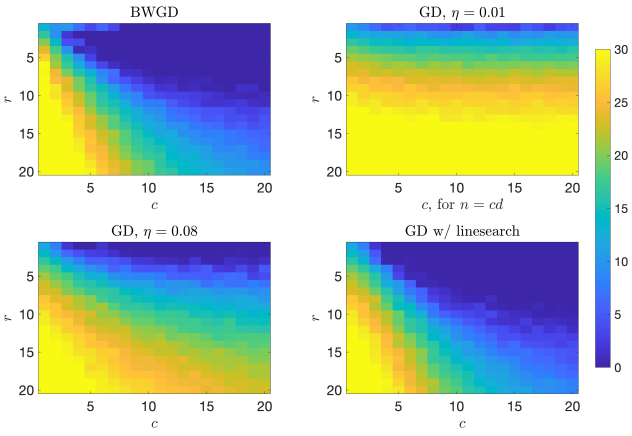

Figure 4: Convergence of BWGD (2.15) with the GD algorithm of [Li et al., 2019]. Here, $d = 64$, and down the columns of each inset we use $r = 1, \ldots, 20$, respectively. Across the rows we vary the scale factor c, where $n = cd$. As we can see, the BWGD method converges much faster than the standard GD method with fixed step size, and performs on par with GD with linesearch. We note that the GD method with line search is much more computationally intensive than BWGD, as is displayed in Figure 5.

4.2 Real Data Experiment: Phase Retrieval

Here we give the details of the experiment displayed in Figure 1. We replicate the phase retrieval experiments in Candès et al. [2015]. In this paper, the authors study the nonconvex Wirtinger Flow algorithm. Since phase retrieval is equivalent to the recovery problem in (1.1) with a rank one complex $\boldsymbol{S}$, we can apply our algorithm in this setting.

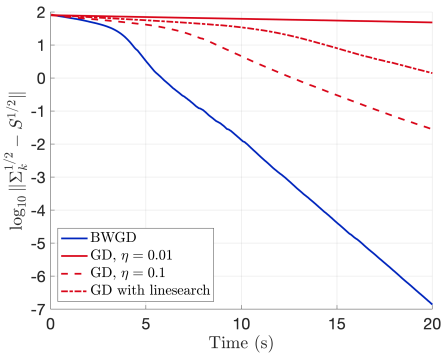

Figure 5: Convergence of BWGD (2.15) and GD [Li et al., 2019]. Here, $d = 512$, $r = 8$, and $n = 3rd$. As we can see, the BWGD method converges much faster than the standard GD method with any choice of step sizes. The per-iteration convergence rate for BWGD is comparable to GD with linesearch, but doing linesearch increases computational burden.

Here, we use $i = \sqrt{-1}$. We use an image of Denali National Park, which is denoted by the 2-dimensional array $\boldsymbol{J}$ for each color band. In our simulated acquisition model, for $m = (u, v, \ell)$, we acquire data of the form

$$y_m = \left| \sum_{j=1,k=1}^{j=d_1,k=d_2} \boldsymbol{J}_{jk} \bar{d}_\ell(j,k) e^{-i2\pi(ju+kv)} \right|^2. \qquad (4.1)$$

Here, $d_\ell(j,k) \sim b_1 b_2$, where b_1 is uniform on $\{1, -1, i, -i\}$, and b_2 takes values $\sqrt{2}/2$ with probability $4/5$ and $\sqrt{3}$ with probability $1/5$. The goal is to recover the image $\boldsymbol{J}$ from these measurements. We note that these measurements can be equivalently written as $y_m = \boldsymbol{F}_{u,v,\ell}^{\flat\top} \boldsymbol{J}^\flat \boldsymbol{J}^{\flat\top} \boldsymbol{F}_{u,v,\ell}^\flat$, where $\cdot^\flat$ denotes the vectorization operation and $\boldsymbol{F}$ is a matrix with entries $\bar{d}_\ell(j,k)e^{-i2\pi(ju+kv)}$. This notation makes clear the connection with the original matrix recovery problem in (1.1). We display the errors versus runtimes in Figure 6. Despite the fact that this example has $r = 1$, BWGD still efficiently recovers the underlying image, even though our current theory only works for $r \geq 5$.

We also give additional phase retrieval experiments on more images in Appendix D.2.

5 LIMITATIONS

There are a few notable limitations for the current work. First of all, the theory does not directly extend to the cases of $r = 1, \ldots, 4$, and we must resort instead to the regularized methods discussed in Section 2.4. It is unclear whether

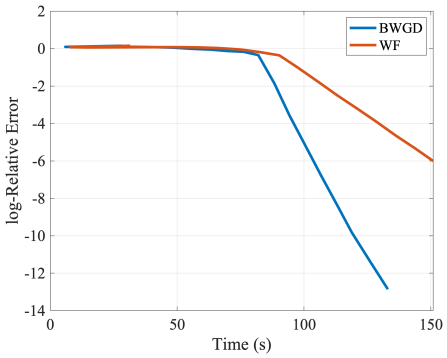

Figure 6: Error vs Runtime of BWGD (our method) and Wirtinger Flow Candès et al. [2015] on the phase retrieval experiment in Figure 1. As we can see, BWGD recovers the underlying image faster. Both methods are initialized with the power method as is done in [Candès et al., 2015].

or not this is a limitation of the methods or the analysis, although experiments indicate that the method works well for small ranks in practice. Second, while our experiments indicate well-behaved energy landscapes, we are only able to prove local convergence results. Third, while we give the first optimization rates in terms of BW distance for this problem, our constants are not optimized. Fourth, we assume that one knows the rank of $\boldsymbol{S}$ in advance. This is not a large issue since our framework allows for one to pick any $r' \geq r$ and still recover r, albeit at a slower rate.

6 CONCLUSION

We have presented a novel connection between the BW barycenter problem and low-rank matrix recovery from rank one measurements. We show that a novel energy minimization problem coincides with the barycenter minimization. This connection allows us to extend algorithms from the barycenter problem to the matrix recovery problem, giving new algorithms for recovering low-rank PSD matrices. Our methods are guaranteed to have local linear convergence in BW distance, which is a stronger guarantee than existing methods.

This work leaves open many unexplored directions. For example, it would be interesting to show that the energy landscape does not exhibit spurious local minima. Beyond this, it would also be interesting to know when and how one can go beyond the RIP assumption, which currently relies on sub-Gaussianity of the sensing vectors. Finally, it would be interesting to connect these ideas to optimization landscapes for neural networks [Zhong et al., 2017, Du et al., 2019, Li et al., 2019].

7 ACKNOWLEDGEMENTS

P. Rigollet is supported by NSF grants IIS-1838071, DMS-2022448, and CCF-2106377.

References

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

Jason Altschuler, Sinho Chewi, Patrik R Gerber, and Austin Stromme. Averaging on the Bures-Wasserstein manifold: dimension-free convergence of gradient descent. *Advances in Neural Information Processing Systems*, 34: 22132–22145, 2021.

Pedro C Álvarez-Esteban, E Del Barrio, JA Cuesta-Albertos, and C Matrán. A fixed-point approach to barycenters in Wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016.

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

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

C. L. Blake and C. J. Merz. UCI repository of machine learning databases, 1998. URL https://archive.ics.uci.edu/ml/datasets/abalone.

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

Samuel Burer and Renato DC Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

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

T Tony Cai and Anru Zhang. ROP: Matrix recovery via rank-one projections. *The Annals of Statistics*, 43(1): 102–138, 2015.

Emmanuel J Candès and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE*

Transactions on Information Theory, 56(5):2053–2080, 2010.

Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

Emmanuel J Candès, Thomas Strohmer, and Vladislav Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

Emmanuel J Candès, Yonina C Eldar, Thomas Strohmer, and Vladislav Voroninski. Phase retrieval via matrix completion. *SIAM review*, 57(2):225–251, 2015.

Venkat Chandrasekaran, Sujay Sanghavi, Pablo A Parrilo, and Alan S Willsky. Sparse and low-rank matrix decompositions. *IFAC Proceedings Volumes*, 42(10):1493–1498, 2009.

Yuxin Chen, Yuejie Chi, and Andrea J Goldsmith. Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Transactions on Information Theory*, 61(7):4034–4059, 2015.

Sinho Chewi, Tyler Maunu, Philippe Rigollet, and Austin J Stromme. Gradient descent algorithms for Bures-Wasserstein barycenters. In *Conference on Learning Theory*, pages 1276–1304. PMLR, 2020.

Yuejie Chi, Yue M Lu, and Yuxin Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67 (20):5239–5269, 2019.

Frank H Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

Richard D De Veaux. Mixtures of linear regressions. *Computational Statistics & Data Analysis*, 8(3):227–245, 1989.

Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.

David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pages 1675–1685. PMLR, 2019.

Maryam Fazel. *Matrix rank minimization with applications*. PhD thesis, PhD thesis, Stanford University, 2002.

James R Fienup. Reconstruction of an object from the modulus of its fourier transform. *Optics letters*, 3(1):27–29, 1978.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

David Gross, Yi-Kai Liu, Steven T Flammia, Stephen Becker, and Jens Eisert. Quantum state tomography via compressed sensing. *Physical review letters*, 105(15): 150401, 2010.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

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

Gilad Lerman and Tyler Maunu. An overview of robust subspace recovery. *Proceedings of the IEEE*, 106(8): 1380–1410, 2018.

Yuanxin Li, Cong Ma, Yuxin Chen, and Yuejie Chi. Nonconvex matrix factorization from rank-one measurements. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1496–1505. PMLR, 2019.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

Estelle Massart and P-A Absil. Quotient geometry with simple geodesics for the manifold of fixed-rank positive-semidefinite matrices. *SIAM Journal on Matrix Analysis and Applications*, 41(1):171–198, 2020.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2004.

Karl Pearson. LIII. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2 (11):559–572, 1901.

Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

Sujay Sanghavi, Rachel Ward, and Chris D White. The local convexity of solving systems of quadratic equations. *Results in Mathematics*, 71(3):569–608, 2017.

Hanie Sedghi, Majid Janzamin, and Anima Anandkumar. Provable tensor methods for learning mixtures of generalized linear models. In *Artificial Intelligence and Statistics*, pages 1223–1231. PMLR, 2016.

Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Ben Recht. Low-rank solutions of linear matrix equations via Procrustes flow. In *International Conference on Machine Learning*, pages 964–973. PMLR, 2016.

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

Gang Wang and Georgios Giannakis. Solving random systems of quadratic equations via truncated generalized gradient flow. *Advances in Neural Information Processing Systems*, 29, 2016.

Gang Wang, Georgios Giannakis, Yousef Saad, and Jie Chen. Solving most systems of random quadratic equations. *Advances in Neural Information Processing Systems*, 30, 2017.

Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust pca via outlier pursuit. *Advances in neural information processing systems*, 23, 2010.

Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi. Alternating minimization for mixed linear regression. In *International Conference on Machine Learning*, pages 613–621. PMLR, 2014.

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

Hongyi Zhang and Suvrit Sra. First-order methods for geodesically convex optimization. In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1617–1638, Columbia University, New York, New York, USA, 2016.

Kai Zhong, Prateek Jain, and Inderjit S Dhillon. Efficient matrix sensing using rank-1 Gaussian measurements. In *International conference on algorithmic learning theory*, pages 3–18. Springer, 2015.

Kai Zhong, Prateek Jain, and Inderjit S Dhillon. Mixed linear regression with multiple components. In *NIPS*, pages 2190–2198, 2016.

Kai Zhong, Zhao Song, Prateek Jain, Peter L Bartlett, and Inderjit S Dhillon. Recovery guarantees for one-hidden-layer neural networks. In *International conference on machine learning*, pages 4140–4149. PMLR, 2017.

A GEOMETRIES OF $\mathbb{S}_+^d$

Since the BW barycenter problem is inherently a geometric minimization problem over $\mathbb{S}_+^d$, we will quickly comment on how the choice of geometry effects optimization algorithms over this space.

There are many ways to define geometries over $\mathbb{S}_+^d$. For sake of comparison with previous methods for matrix recovery, we will compare BW space to the Euclidean geometry over PSD matrices. Among other things, the choice of geometry gives us a way of defining defining distance minimizing paths, or *geodesics*, over $\mathbb{S}_+^d$.

Consider two matrices $\mathbf{\Sigma}_0, \mathbf{\Sigma}_1 \in \mathbb{S}_+^d$ such that $\operatorname{rank}(\mathbf{\Sigma}_0\mathbf{\Sigma}_1) = \operatorname{rank}(\mathbf{\Sigma}_0) = \operatorname{rank}(\mathbf{\Sigma}_1)$. This is a sufficient but not necessary condition to ensure that our following definition of BW geodesic is well defined since the same map can work for transporting higher rank PSD matrices to lower rank matrices. In any case, under this assumption, there exists a transport map from $\mathbf{\Sigma}_0$ to $\mathbf{\Sigma}_1$ given by

$$\boldsymbol{T} = \mathbf{\Sigma}_1^{1/2}(\mathbf{\Sigma}_1^{1/2}\mathbf{\Sigma}_0\mathbf{\Sigma}_1^{1/2})^{-1/2}\mathbf{\Sigma}_1^{1/2}, \tag{A.1}$$

where the inverse is actually a pseudoinverse and one can check that $T\mathbf{\Sigma}_0 T = \mathbf{\Sigma}_1$. In this case, the Euclidean and BW geodesics $\mathbf{\Sigma}_t : [0,1] \to \mathbb{S}_+^d$ are given by

$$\mathbf{\Sigma}_t = (1-t)\mathbf{\Sigma}_0 + t\mathbf{\Sigma}_1, \tag{EG}$$
$$\mathbf{\Sigma}_t = (\boldsymbol{I} + t(\boldsymbol{T} - \boldsymbol{I}))\mathbf{\Sigma}_0(I + t(\boldsymbol{T} - \boldsymbol{I})). \tag{BWG}$$

The first choice of geodesic, the Euclidean Geodesic (EG), corresponds to the distance functional $\|\mathbf{\Sigma}_0 - \mathbf{\Sigma}_1\|_F$. The second choice of geodesic, the BW Geodesic (BWG), corresponds to the BW distance functional $d_{\mathsf{BW}}(\mathbf{\Sigma}_0, \mathbf{\Sigma}_1)$.

We note that (BWG) is equivalent to

$$\mathbf{\Sigma}_t = (I + t(\boldsymbol{T} - \boldsymbol{I}))\mathbf{\Sigma}_0(I + t(\boldsymbol{T} - \boldsymbol{I})) \tag{A.2}$$
$$= (\mathbf{\Sigma}_0^{1/2} + t\mathbf{\Delta})(\mathbf{\Sigma}_0^{1/2} + t\mathbf{\Delta})^\top. \tag{A.3}$$

where $\mathbf{\Delta} = (\boldsymbol{T} - \boldsymbol{I})\mathbf{\Sigma}_0^{1/2}$.

One of the big differences in these two paths is that while the Euclidean geodesics are linear in t, which is a result of the underlying flatness of the space, the BW geodesic contains terms that are quadratic in t. This points to the fact that this choice of geometry adds *curvature* to the space $\mathbb{S}_+^d$.

If we restrict ourselves to rank r PSD matrices, $\mathbb{S}_+^{d,r}$, the geodesic (BWG) becomes

$$\begin{aligned}\mathbf{\Sigma}_t &= (I + t(T - I)\boldsymbol{U}_0\boldsymbol{U}_0^\top(I + t(T - I))^\top \\ &= ((1-t)\boldsymbol{U}_0 + t\boldsymbol{U}_1)((1-t)\boldsymbol{U}_0 + t\boldsymbol{U}_1)^\top,\end{aligned} \tag{A.4}$$

where we use the fact that $\boldsymbol{T}\mathbf{\Sigma}_0\boldsymbol{T} = \boldsymbol{T}\boldsymbol{U}_0\boldsymbol{U}_0^\top\boldsymbol{T} = \mathbf{\Sigma}_1 = \boldsymbol{U}_1\boldsymbol{U}_1^\top$. Note that there is an inherent rotational symmetry in the problem, since for any $R \in \mathbb{R}^{r\times r}$ such that $\boldsymbol{R}\boldsymbol{R}^\top = \boldsymbol{I}$ and $\boldsymbol{U}\boldsymbol{U}^\top \in \mathbb{S}_+^{d,r}$, $\boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{U}\boldsymbol{R}\boldsymbol{R}^\top U^\top$.

B FIXED POINTS AND BWGD

Here we briefly show how one can interpret the fixed point equation of Álvarez-Esteban et al. [2016] as gradient descent with step size 1. Consider the case where $\mathbf{\Sigma}_k$ is full rank. The fixed point iteration is given by

$$\mathbf{\Sigma}_{k+1} = \mathbf{\Sigma}_k^{-1/2}\Big(\frac{1}{n}\sum_{i=1}^n(\mathbf{\Sigma}_k^{1/2}\boldsymbol{X}_i\mathbf{\Sigma}_k^{1/2})^{1/2}\Big)^2\mathbf{\Sigma}_k^{-1/2} \tag{B.1}$$

We can write

$$\begin{aligned}&\mathbf{\Sigma}_k^{-1/2}\Big(\frac{1}{n}\sum_{i=1}^n(\mathbf{\Sigma}_k^{1/2}\boldsymbol{X}_i\mathbf{\Sigma}_k^{1/2})^{1/2}\Big)^2\mathbf{\Sigma}_k^{-1/2} \\ &= \mathbf{\Sigma}_k^{-1/2}\Big(\frac{1}{n}\sum_{i=1}^n(\mathbf{\Sigma}_k^{1/2}\boldsymbol{X}_i\mathbf{\Sigma}_k^{1/2})^{1/2}\Big)\mathbf{\Sigma}_k^{-1/2}\mathbf{\Sigma}_k\mathbf{\Sigma}_k^{-1/2}.\end{aligned} \tag{B.2}$$

$$\Big(\frac{1}{n}\sum_{i=1}^n(\boldsymbol{\Sigma}_k^{1/2}\boldsymbol{X}_i\boldsymbol{\Sigma}_k^{1/2})^{1/2}\Big)\boldsymbol{\Sigma}_k^{-1/2}$$
$$=(\boldsymbol{I}-\nabla F(\boldsymbol{\Sigma}_k))\boldsymbol{\Sigma}_k(\boldsymbol{I}-\nabla F(\boldsymbol{\Sigma}_k)).$$

On the other hand, we note that the fixed point equation in Álvarez-Esteban et al. [2016] does not work for low-rank matrices, since the span in (B.1) is not allowed to change between iterations. Instead, one must use the transport map $\boldsymbol{X}_i^{1/2}(\boldsymbol{X}_i^{1/2}\boldsymbol{\Sigma}_k\boldsymbol{X}_i^{1/2})^{+1/2}\boldsymbol{X}_i^{1/2}$, which results in the fixed point iteration

$$\boldsymbol{\Sigma}_{k+1}=\Big(\frac{1}{n}\sum_{i=1}^n\boldsymbol{X}_i^{1/2}(\boldsymbol{X}_i^{1/2}\boldsymbol{\Sigma}_k\boldsymbol{X}_i^{1/2})^{+1/2}\boldsymbol{X}_i^{1/2}\Big)\boldsymbol{\Sigma}_k\cdot \tag{B.3}$$
$$\Big(\frac{1}{n}\sum_{i=1}^n\boldsymbol{X}_i^{1/2}(\boldsymbol{X}_i^{1/2}\boldsymbol{\Sigma}_k\boldsymbol{X}_i^{1/2})^{+1/2}\boldsymbol{X}_i^{1/2}\Big)$$

and fixed point equation

$$\boldsymbol{\Sigma}=\Big(\frac{1}{n}\sum_{i=1}^n\boldsymbol{X}_i^{1/2}(\boldsymbol{X}_i^{1/2}\boldsymbol{\Sigma}\boldsymbol{X}_i^{1/2})^{+1/2}\boldsymbol{X}_i^{1/2}\Big)\boldsymbol{\Sigma}\cdot \tag{B.4}$$
$$\Big(\frac{1}{n}\sum_{i=1}^n\boldsymbol{X}_i^{1/2}(\boldsymbol{X}_i^{1/2}\boldsymbol{\Sigma}\boldsymbol{X}_i^{1/2})^{+1/2}\boldsymbol{X}_i^{1/2}\Big).$$

In essence, this says that $\boldsymbol{\Sigma}$ is an exponential barycenter.

C SUPPLEMENTARY PROOFS

C.1 Proof of Propositions 1 and 2

Proposition 1 *If* $\frac{1}{n}\sum_i \boldsymbol{x}_i\boldsymbol{x}_i^\top=\boldsymbol{I}$, *then*

$$\operatorname*{argmin}_{\boldsymbol{\Sigma}\in\mathbb{S}_+^d}\frac{1}{2n}\|\sqrt{\mathcal{A}(\boldsymbol{\Sigma})}-\sqrt{\boldsymbol{y}}\|_2^2$$
$$=\operatorname*{argmin}_{\boldsymbol{\Sigma}\in\mathbb{S}_+^d}\frac{1}{2n}\sum_{i=1}^n d_{\mathsf{BW}}^2(\boldsymbol{\Sigma},y_i\boldsymbol{x}_i\boldsymbol{x}_i^\top).$$

Proof of Proposition 1. First, if $\frac{1}{n}\sum_i \boldsymbol{x}_i\boldsymbol{x}_i^\top=\boldsymbol{I}$, then

$$\frac{1}{n}\sum_i\mathcal{A}(\boldsymbol{\Sigma})_i=\operatorname{Tr}(\boldsymbol{\Sigma}).$$

Indeed, this follows from the fact that

$$\frac{1}{n}\sum_i\mathcal{A}(\boldsymbol{\Sigma})_i=\frac{1}{n}\sum_i\boldsymbol{x}_i^\top\boldsymbol{\Sigma}\boldsymbol{x}_i$$
$$=\operatorname{Tr}\left(\boldsymbol{\Sigma}\frac{1}{n}\sum_i\boldsymbol{x}_i\boldsymbol{x}_i^\top\right)$$
$$=\operatorname{Tr}(\boldsymbol{\Sigma}).$$

With this in mind, we expand the square in (2.4)

$$\frac{1}{2n}\|\sqrt{\mathcal{A}(\boldsymbol{\Sigma})}-\sqrt{y}\|_2^2= \tag{C.1}$$

$$\frac{1}{2n}\sum_i \mathcal{A}(\boldsymbol{\Sigma})_i + \frac{1}{2n}\sum_i y_{\boldsymbol{I}} - \frac{1}{n}\sum_i \sqrt{y_i}\sqrt{\boldsymbol{x}_i^\top \boldsymbol{\Sigma}\boldsymbol{x}_i}$$
$$= \frac{1}{2}\operatorname{Tr}(\boldsymbol{\Sigma}) - \frac{1}{n}\sum_i \sqrt{y_i}\sqrt{\boldsymbol{x}_i^\top \boldsymbol{\Sigma}\boldsymbol{x}_i} + \frac{1}{2n}\sum_i y_i.$$

Thus, the minimization in (2.4) is equivalent to the program

$$\min_{\boldsymbol{\Sigma}\in\mathbb{S}_+^d} \operatorname{Tr}(\boldsymbol{\Sigma}) - \frac{2}{n}\sum_{i=1}^n \sqrt{y_i}\sqrt{\boldsymbol{x}_i^\top \boldsymbol{\Sigma}\boldsymbol{x}_i} \tag{C.2}$$

On the other hand, it is not hard to show that the BW distance between a matrix $\boldsymbol{\Sigma} \in \mathbb{S}_+^d$ and a rank one matrix $\boldsymbol{w}\boldsymbol{w}^\top$ is

$$d_{\mathsf{BW}}^2(\boldsymbol{\Sigma}, \boldsymbol{w}\boldsymbol{w}^\top) = \Big[\operatorname{Tr}(\boldsymbol{\Sigma}) + \operatorname{Tr}(\boldsymbol{w}\boldsymbol{w}^\top) - 2\sqrt{\operatorname{Tr}(\boldsymbol{w}\boldsymbol{w}^\top\boldsymbol{\Sigma})}\Big]. \tag{C.3}$$

Letting $\boldsymbol{w}\boldsymbol{w}^\top = y_i\boldsymbol{x}_i\boldsymbol{x}_i^\top$ and summing over i, we see that (C.2) is equivalent to minimizing (2.9) when

$$Q = \frac{1}{n}\sum_{i=1}^n \delta_{y_i\boldsymbol{x}_i\boldsymbol{x}_i^\top}.$$

□

Proposition 2 *Let* $\boldsymbol{C}_n = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top \in \mathbb{S}_{++}^d$. *Then*

$$\operatorname*{argmin}_{\boldsymbol{\Sigma}\in\mathbb{S}_+^d} \frac{1}{2n}\|\sqrt{\mathcal{A}(\boldsymbol{C}_n^{-1/2}\boldsymbol{\Sigma}\boldsymbol{C}_n^{-1/2})} - \sqrt{\boldsymbol{y}}\|^2$$
$$= \operatorname*{argmin}_{\boldsymbol{\Sigma}\in\mathbb{S}_+^d} \frac{1}{2n}\sum_{i=1}^n d_{\mathsf{BW}}^2(\boldsymbol{\Sigma}, y_i\boldsymbol{C}_n^{-1/2}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{C}_n^{-1/2}),$$

and $\boldsymbol{C}_n^{1/2}\boldsymbol{S}\boldsymbol{C}_n^{1/2}$ *is a solution to both problems.*

Proof of Proposition 2. Consider the first order condition for the rank one matrices $y_1^2\boldsymbol{C}_n^{-1/2}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{C}_n^{-1/2}$ given by

$$\frac{1}{n}\sum_{i=1}^n \frac{\operatorname{Tr}(\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{S})^{1/2}\boldsymbol{C}_n^{-1/2}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{C}_n^{-1/2}}{\operatorname{Tr}(\boldsymbol{C}_n^{-1/2}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{C}_n^{-1/2}\boldsymbol{\Sigma}_n)^{1/2}} = \boldsymbol{I}. \tag{C.4}$$

Since $\boldsymbol{C}_n^{-1/2}\boldsymbol{C}_n\boldsymbol{C}_n^{-1/2} = \boldsymbol{I}$, we see that $\boldsymbol{C}_n^{-1/2}\boldsymbol{\Sigma}_n\boldsymbol{C}_n^{-1/2} = \boldsymbol{S}$ is a sufficient condition for $\boldsymbol{\Sigma}_n$ to be a barycenter of

$$Q = \frac{1}{n}\sum_{i=1}^n \delta_{y_i\boldsymbol{C}_n^{-1/2}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{C}_n^{-1/2}} \tag{C.5}$$

□

C.2 Proof of Theorem 4

The proof of Theorem 4 proceeds in the following sections. In Section C.2.1, we establish some restricted isometry properties (RIP) that will be used in our proof. Then, Section C.2.2 proves that the function F is Euclidean strongly convex over $\mathbb{S}_+^d$ with high probability. After this, Section C.2.3 shows that F also satisfies a certain Euclidean smoothness over rank r matrices. Section C.2.4 discusses first order optimality conditions for the barycenter, and in particular gives sufficient conditions for a point $\overline{\boldsymbol{\Sigma}}$ to be the minimizer of F. Section C.2.5 gives a descent lemma for the fixed-rank gradient descent method. After this, Section C.2.6 proves local strong convexity of F over fixed-rank PSD matrices in $\mathbb{S}_+^{d,r}$ that are close to $\boldsymbol{S}$. We finish in Section C.2.7 by putting all of these facts together.

C.2.1 RIP Conditions

We discuss here a case an RIP condition that becomes essential for our later proof. As discussed in Cai and Zhang [2015], issues arise in trying to prove a full ℓ_2 RIP for this problem, due to the fact that the fourth moments of $\boldsymbol{x}$ that show up. Instead, both Cai and Zhang [2015], Chen et al. [2015] prove the following ℓ_2/ℓ_1 RIP condition (also referred to as "Restricted Uniform Boundedness").

Theorem 9 (Chen et al. [2015] Proposition 1)**.** *Suppose that $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ are a sample from a sub-Gaussian distribution with $\mathbb{E}\boldsymbol{x}_i = \boldsymbol{0}$, $\mathbb{E}x_{ij}^2 = 1$, and $\mathbb{E}x_{ij}^4 > 1$. Then, there are constants $c_1, c_2, c_3 > 0$ such that with probability exceeding $1 - \exp(-c_3 n)$*

$$c_1\|\boldsymbol{\Delta}\|_F \leq \frac{1}{n}\left\|\begin{bmatrix} \boldsymbol{x}_1^\top \boldsymbol{\Delta}\boldsymbol{x}_1 - \boldsymbol{x}_2^\top \boldsymbol{\Delta}\boldsymbol{x}_2 \\ \boldsymbol{x}_3^\top \boldsymbol{\Delta}\boldsymbol{x}_3 - \boldsymbol{x}_4^\top \boldsymbol{\Delta}\boldsymbol{x}_4 \\ \vdots \\ \boldsymbol{x}_{n-1}^\top \boldsymbol{\Delta}\boldsymbol{x}_{n-1} - \boldsymbol{x}_n^\top \boldsymbol{\Delta}\boldsymbol{x}_n \end{bmatrix}\right\|_1 \leq c_2\|\boldsymbol{\Delta}\|_F. \tag{C.6}$$

hold simultaneously for all rank r matrices $\boldsymbol{\Delta}$ provided that $n \gtrsim dr$.

We have the following corollary to this theorem.

Corollary 10. *Suppose that a random vector $\boldsymbol{w} = (w_1, \dots, w_d)^\top$ is sub-Gaussian with $\mathbb{E}\boldsymbol{w} = \boldsymbol{0}$, $\mathbb{E}w_j^2 = 1$, $\mathbb{E}w_j^4 > 1$. Then, the following population RIP holds:*

$$\mathbb{E}_{\boldsymbol{w}} \operatorname{Tr}(\boldsymbol{w}\boldsymbol{w}^\top \boldsymbol{A})^2 \geq c_1\|\boldsymbol{A}\|_F^2. \tag{C.7}$$

Furthermore, for a sample of n i.i.d. copies of W, $\boldsymbol{w}_1, \dots, \boldsymbol{w}_n$, there are constants $c_2, c_3 > 0$ such that with probability exceeding $1 - \exp(-c_3 n)$

$$\frac{1}{n}\sum_{i=1}^n \operatorname{Tr}(\boldsymbol{w}_i\boldsymbol{w}_i^\top \boldsymbol{A})^2 \geq c_2\|\boldsymbol{A}\|_F^2 \tag{C.8}$$

hold simultaneously for all rank r matrices $\boldsymbol{\Delta}$ provided that $n \gtrsim dr$.

Proof. The first part holds from the argument in Appendix A of Chen et al. [2015].

For the second part, we compute

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \operatorname{Tr}(\boldsymbol{w}_i\boldsymbol{w}_i^\top \boldsymbol{A})^2 = \frac{1}{n}\sum_{i=1}^n \operatorname{Tr}(\boldsymbol{w}_i\boldsymbol{w}_i^\top \boldsymbol{A})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n/2} \operatorname{Tr}(\boldsymbol{w}_{2i}\boldsymbol{w}_{2i}^\top \boldsymbol{A})^2 + \operatorname{Tr}(\boldsymbol{w}_{2i-1}\boldsymbol{w}_{2i-1}^\top \boldsymbol{A})^2 \\
&\geq \frac{1}{2n}\sum_{i=1}^{n/2} [\operatorname{Tr}(\boldsymbol{w}_{2i}\boldsymbol{w}_{2i}^\top \boldsymbol{A}) - \operatorname{Tr}(\boldsymbol{w}_{2i-1}\boldsymbol{w}_{2i-1}^\top \boldsymbol{A})]^2 \\
&= \frac{1}{n^2}\frac{n}{2}\sum_{i=1}^{n/2} [\operatorname{Tr}(\boldsymbol{w}_{2i}\boldsymbol{w}_{2i}^\top \boldsymbol{A}) - \operatorname{Tr}(\boldsymbol{w}_{2i-1}\boldsymbol{w}_{2i-1}^\top \boldsymbol{A})]^2 \\
&\geq \frac{1}{n^2}\Big[\sum_{i=1}^{n/2} |\operatorname{Tr}(\boldsymbol{w}_{2i}\boldsymbol{w}_{2i}^\top \boldsymbol{A}) - \operatorname{Tr}(\boldsymbol{w}_{2i-1}\boldsymbol{w}_{2i-1}^\top \boldsymbol{A})|\Big]^2 \\
&\geq c_2^2\|\boldsymbol{A}\|_F^2.
\end{aligned} \tag{C.9}$$

□

C.2.2 Euclidean Strong Convexity

We begin by proving strong convexity along Euclidean geodesics. In particular, this guarantees that $\boldsymbol{S}$ is the unique minimizer of F over $\mathbb{S}_+^d$ in the population setting and $\boldsymbol{S}_n$ is the unique rank r global minimizer over $\mathbb{S}_+^{d,r}$.

Population Setting: First, it is easy to see that $\boldsymbol{S}$ is stationary because

$$\nabla F(\boldsymbol{S}) = \boldsymbol{I} - \mathbb{E}_Q \sqrt{\frac{\boldsymbol{x}^\top \boldsymbol{S}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{S}\boldsymbol{x}}} \boldsymbol{x}\boldsymbol{x}^\top = \boldsymbol{I} - \mathbb{E}_Q \boldsymbol{x}\boldsymbol{x}^\top = 0. \tag{C.10}$$

Define the set

$$\mathcal{S}(m, M) = \{\boldsymbol{\Sigma} \in \mathbb{S}^d_+ : m \leq \lambda_r(\boldsymbol{\Sigma}) \leq \lambda_1(\boldsymbol{\Sigma}) \leq M\}. \tag{C.11}$$

We have the following lemma over this set.

Lemma 11. *Let $y = \langle \boldsymbol{x}\boldsymbol{x}^\top, \boldsymbol{S}\rangle$, where $\boldsymbol{x} \sim N(\boldsymbol{0}, \boldsymbol{I})$, $\boldsymbol{S}$ is rank r, and m, M be such that $\boldsymbol{S} \in \mathcal{S}(m, M)$. Then, the population obective F is Euclidean strongly convex over $\mathcal{S}(0, M)$ with constant $\frac{c_1 \beta^2 d}{6M^{3/2}}$.*

In other words, the function $F(\boldsymbol{\Sigma}_t)$ is strongly convex when $\boldsymbol{\Sigma}_t$ is defined by (EG).

Proof. Let $\boldsymbol{\Sigma}_t$ denote the geodesic between $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ given by $(1-t)\boldsymbol{\Sigma}_0 + t\boldsymbol{\Sigma}_1$. We compute the derivatives of the BW distance as

$$\partial_t d_{\mathsf{BW}}(\boldsymbol{\Sigma}_t, \boldsymbol{x}\boldsymbol{x}^\top)^2 = \operatorname{Tr}(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0) \tag{C.12}$$

$$- \frac{\operatorname{Tr}((\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0)\boldsymbol{x}\boldsymbol{x}^\top)}{\sqrt{\boldsymbol{x}^\top \boldsymbol{\Sigma}_t \boldsymbol{x}}}, \tag{C.13}$$

$$\partial_t^2 d_{\mathsf{BW}}(\boldsymbol{\Sigma}_t, \boldsymbol{x}\boldsymbol{x}^\top)^2 = \frac{1}{2} \frac{\operatorname{Tr}(\boldsymbol{x}\boldsymbol{x}^\top(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0))^2}{\operatorname{Tr}(\boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{\Sigma}_t)^{3/2}} \tag{C.14}$$

By the definition of M, we have the uniform bound

$$\partial_t^2 F(\boldsymbol{\Sigma}_t)|_{t=s} = \mathbb{E}_{\boldsymbol{x}\boldsymbol{x}^\top \sim Q} \frac{1}{2} \frac{\operatorname{Tr}(\boldsymbol{x}\boldsymbol{x}^\top(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0))^2}{(\operatorname{Tr}(\boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{\Sigma}_s))^{3/2}} \tag{C.15}$$

$$= \mathbb{E}_{\boldsymbol{x}\boldsymbol{x}^\top \sim Q} \frac{1}{2\|\boldsymbol{x}\|^3} \frac{\operatorname{Tr}(\boldsymbol{x}\boldsymbol{x}^\top(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0))^2}{(\operatorname{Tr}(\boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{\Sigma}_s/\|\boldsymbol{x}\|^2))^{3/2}}$$

$$\geq \frac{1}{2(2M)^{3/2}} \mathbb{E}_{\boldsymbol{x}\boldsymbol{x}^\top \sim Q} \frac{1}{\|\boldsymbol{x}\|^3} \operatorname{Tr}(\boldsymbol{x}\boldsymbol{x}^\top(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0))^2.$$

This then implies that

$$\partial_t^2 F(\boldsymbol{\Sigma}_t)|_{t=s} \tag{C.16}$$

$$\geq \frac{1}{6M^{3/2}} \mathbb{E}_{\boldsymbol{x} \sim N(\boldsymbol{0}, \boldsymbol{I})} \frac{\sqrt{\boldsymbol{x}^\top \boldsymbol{S}\boldsymbol{x}}}{\|\boldsymbol{x}\|^3} \operatorname{Tr}(\boldsymbol{x}\boldsymbol{x}^\top(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0))^2$$

$$\geq \frac{1}{6M^{3/2}} \mathbb{E}_{\boldsymbol{x} \sim N(\boldsymbol{0}, \boldsymbol{I})} \|\boldsymbol{x}\|^2 \sqrt{\frac{\boldsymbol{x}^\top \boldsymbol{S}\boldsymbol{x}}{\|\boldsymbol{x}\|^2}} \operatorname{Tr}(\frac{\boldsymbol{x}\boldsymbol{x}^\top}{\|\boldsymbol{x}\|^2}(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0))^2$$

$$= \frac{d}{6M^{3/2}} \mathbb{E}_{\boldsymbol{x} \sim N(\boldsymbol{0}, \boldsymbol{I})} \sqrt{\frac{\boldsymbol{x}^\top \boldsymbol{S}\boldsymbol{x}}{\|\boldsymbol{x}\|^2}} \operatorname{Tr}(\frac{\boldsymbol{x}\boldsymbol{x}^\top}{\|\boldsymbol{x}\|^2}(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0))^2.$$

Define the random vector

$$\boldsymbol{w} := \left(\frac{\boldsymbol{x}^\top \boldsymbol{S}\boldsymbol{x}}{\|\boldsymbol{x}\|^2}\right)^{1/8} \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}. \tag{C.17}$$

By symmetry, $\boldsymbol{w}$ is mean zero. Furthermore, $\boldsymbol{w}$ is bounded and thus sub-Gaussian. Furthermore, it is a simple exercise to show that

$$\boldsymbol{C}_W := \mathbb{E}\boldsymbol{w}\boldsymbol{w}^\top \succeq \beta \boldsymbol{I}, \tag{C.18}$$

for some dimension and $\boldsymbol{S}$-dependent constant β. We can thus bound

$$\mathbb{E}_{\boldsymbol{x}} \operatorname{Tr}\left(\boldsymbol{w}\boldsymbol{w}^\top(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0)\right)^2 \tag{C.19}$$

$$= \mathbb{E}_W \operatorname{Tr}\left(\boldsymbol{C}_W^{1/2} \boldsymbol{C}_W^{-1/2} \boldsymbol{w}\boldsymbol{w}^\top \boldsymbol{C}_W^{-1/2} \boldsymbol{C}_W^{1/2}(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0)\right)^2$$

$$\geq \beta^2 \mathbb{E}_W \operatorname{Tr}(\boldsymbol{C}_W^{-1/2}\boldsymbol{w}\boldsymbol{w}^\top \boldsymbol{C}_W^{-1/2}(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0))^2.$$

Using the fact that $\boldsymbol{C}_W^{-1/2}\boldsymbol{w}$ is mean zero sub-Gaussian with identity covariance (and furthermore the 4th moment condition is trivially satisfied), we can now use the RIP condition of Corollary 10 to bound

$$\begin{aligned}\mathbb{E}_{\boldsymbol{w}\boldsymbol{w}^\top \sim Q} \operatorname{Tr}(\boldsymbol{C}_W^{-1/2}\boldsymbol{w}\boldsymbol{w}^\top \boldsymbol{C}_W^{-1/2}(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0))^2 \\ \geq c_1 \|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0\|_F^2.\end{aligned} \tag{C.20}$$

Putting this all together,

$$\begin{aligned}\partial_t^2 F(\boldsymbol{\Sigma}_t)|_{t=s} &\geq \frac{c_1\beta^2 d}{2(2M)^{3/2}}\|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0\|_F^2 \\ &\geq \frac{c_1\beta^2 d}{6M^{3/2}}\|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0\|_F^2.\end{aligned} \tag{C.21}$$

□

Sample Setting: The sample setting requires a bit more work. In this case, we observe y_i and $\boldsymbol{x}_i$, $i = 1, \dots, n$, and to recover the matrix $\boldsymbol{S}$ we first compute the barycenter of $y_i \boldsymbol{C}_n^{-1/2}\boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{C}_n^{-1/2}$, where $\boldsymbol{C}_n$ is the sample covariance. We denote this whitened discrete distribution by $\tilde{Q}$. In this case, by Proposition 2, the matrix we hope to recover is $\boldsymbol{S}_n = \boldsymbol{C}_n^{-1/2}\boldsymbol{S}\boldsymbol{C}_n^{-1/2}$

Lemma 12. *If $n \gtrsim d^2$, then with probability at least $1 - \exp(-c_2 n)$ for some constant c_2, F is Euclidean strongly convex over $\mathcal{S}(0, M)$ with constant $\frac{c_3\beta^2 d}{6M^{3/2}}$.*

Proof. In this case, the second statement of Corollary 10 holds with rank d,

$$\begin{aligned}&\partial_t^2 F(\boldsymbol{\Sigma}_t)|_{t=s} \geq \\ &\frac{1}{6M^{3/2}}\frac{1}{n}\sum_{i=1}^n \frac{\sqrt{\boldsymbol{x}_i^\top \boldsymbol{S}\boldsymbol{x}_i}}{\|\boldsymbol{C}_n^{-1/2}\boldsymbol{x}_i\|^3}\operatorname{Tr}(\boldsymbol{C}_n^{-1/2}\boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{C}_n^{-1/2}(\boldsymbol{S}_n - \boldsymbol{\Sigma}))^2 \\ &\geq \frac{1}{6M^{3/2}}\frac{\lambda_{\min}(\boldsymbol{C}_n)^{3/2}}{\lambda_{\max}(\boldsymbol{C}_n)^2}\frac{1}{n}\sum_{i=1}^n \frac{\sqrt{\boldsymbol{x}_i^\top \boldsymbol{S}\boldsymbol{x}_i}}{\|\boldsymbol{x}_i\|^3}\operatorname{Tr}(\boldsymbol{x}_i\boldsymbol{x}_i^\top(\boldsymbol{S}_n - \boldsymbol{\Sigma}))^2 \\ &= \frac{d}{6M^{3/2}}\frac{\lambda_{\min}(\boldsymbol{C}_n)^{3/2}}{\lambda_{\max}(\boldsymbol{C}_n)^2}\cdot \\ &\quad \frac{1}{n}\sum_{i=1}^n \operatorname{Tr}\Big(\frac{\|\boldsymbol{x}_i\|}{d}\Big(\frac{\boldsymbol{x}_i^\top \boldsymbol{S}\boldsymbol{x}_i}{\|\boldsymbol{x}_i\|^2}\Big)^{1/4}\frac{\boldsymbol{x}_i\boldsymbol{x}_i^\top}{\|\boldsymbol{x}_i\|^2}(\boldsymbol{S}_n - \boldsymbol{\Sigma})\Big)^2.\end{aligned} \tag{C.22}$$

Then, following the same line of reasoning as in the previous proof to obtain strong convexity the sub-Gaussianity of the random vector

$$\boldsymbol{w}_i = \frac{\|\boldsymbol{x}_i\|^{1/2}}{\sqrt{d}}\Big(\frac{\boldsymbol{x}_i^\top \boldsymbol{S}\boldsymbol{x}_i}{\|\boldsymbol{x}_i\|^2}\Big)^{1/8}\frac{\boldsymbol{x}_i}{\|\boldsymbol{x}_i\|}, \tag{C.23}$$

which again a simple exercise shows $\mathbb{E}\boldsymbol{w}_i\boldsymbol{w}_i^\top \succeq \beta \boldsymbol{I}$ and $\mathbb{E}\boldsymbol{w}_i = 0$. Then, with probability at least $1 - \exp(-c_2 n)$, the discrete distribution is strongly convex with constant $\frac{c_3\beta^2 d}{6M^{3/2}}$. □

Notice that in both the population and the sample setting with high probability we get a unique barycenter $\boldsymbol{S}$ (and $\boldsymbol{S}_n$, respectively).

Remark 13. *When $n \gtrsim dr$, we can repeat the argument in Lemma 12 by applying Corollary 10 with the rank set to $2r$ to yield that $\boldsymbol{S}_n$ is the unique rank r minimizer of F over $\mathbb{S}_+^{d,r}$.*

C.2.3 Local Generalized Euclidean Smoothness

We remind ourselves that

$$\nabla F(\boldsymbol{\Sigma}) = \boldsymbol{I} - \mathbb{E}_Q \sqrt{\frac{\boldsymbol{x}^\top \boldsymbol{S}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{\Sigma}\boldsymbol{x}}} \boldsymbol{x}\boldsymbol{x}^\top, \tag{C.24}$$

In gradient-based optimization, a function is typically called *smooth* if it has a Lipschitz continuous gradient in the Euclidean sense. In our case, this would correspond to proving a bound of the form

$$\|\nabla F(\boldsymbol{S}) - \nabla F(\boldsymbol{\Sigma})\|_F \lesssim \|\boldsymbol{\Sigma} - \boldsymbol{S}\|_F. \tag{C.25}$$

While we are not able to prove a bound of this form, locally we are able to prove that the gradient is $1/2$-Hölder continuous.

Using this, we have the following lemma.

Lemma 14. *Let $y = \langle \boldsymbol{x}\boldsymbol{x}^\top, \boldsymbol{S}$, $X \sim N(\boldsymbol{0}, \boldsymbol{I})$, $\operatorname{rank}(\boldsymbol{\Sigma}) \geq 3$, and $\lambda_r(\boldsymbol{\Sigma}) \geq m$. Then,*

$$\begin{aligned} \|\nabla F(\boldsymbol{S}) - \nabla F(\boldsymbol{\Sigma})\|_F &= \|\nabla F(\boldsymbol{\Sigma})\|_F \\ &\leq \frac{d\sqrt{r}}{m\sqrt{r-2}} \|\boldsymbol{\Sigma} - \boldsymbol{S}\|_F^{1/2} \end{aligned} \tag{C.26}$$

Proof. We have that

$$\begin{aligned} &\left\| \mathbb{E}_Q \sqrt{\frac{\boldsymbol{x}^\top \boldsymbol{S}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{S}\boldsymbol{x}}} \boldsymbol{x}\boldsymbol{x}^\top - \mathbb{E}_Q \sqrt{\frac{\boldsymbol{x}^\top \boldsymbol{S}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{\Sigma}\boldsymbol{x}}} \boldsymbol{x}\boldsymbol{x}^\top \right\| \leq \\ &\mathbb{E}_Q \left| \frac{\sqrt{\boldsymbol{x}^\top \boldsymbol{\Sigma}\boldsymbol{x}} - \sqrt{\boldsymbol{x}^\top \boldsymbol{S}\boldsymbol{x}}}{\sqrt{\boldsymbol{x}^\top \boldsymbol{\Sigma}\boldsymbol{x}}} \right| \|\boldsymbol{x}\|^2 \\ &= d\mathbb{E}_Q \frac{1}{\sqrt{\boldsymbol{x}^\top \boldsymbol{\Sigma}\boldsymbol{x}}} \sqrt{|\boldsymbol{x}^\top (\boldsymbol{\Sigma} - \boldsymbol{S})\boldsymbol{x}|} \\ &\leq d\|\boldsymbol{\Sigma} - \boldsymbol{S}\|_F^{1/2} \mathbb{E}_Q \frac{\|\boldsymbol{P}_{\boldsymbol{\Sigma}-\boldsymbol{S}}\boldsymbol{x}\|}{\sqrt{\boldsymbol{x}^\top \boldsymbol{\Sigma}\boldsymbol{x}}} \end{aligned} \tag{C.27}$$

As long as $r \geq 3$ and $\lambda_r(\boldsymbol{\Sigma}) \geq m$, we have that

$$\mathbb{E}_Q \frac{\|\boldsymbol{P}_{\boldsymbol{\Sigma}-\boldsymbol{S}}\boldsymbol{x}\|}{\sqrt{\boldsymbol{x}^\top \boldsymbol{\Sigma}\boldsymbol{x}}} \leq \frac{1}{m} \sqrt{1 + \frac{r}{r-2}} = \frac{\sqrt{r}}{m\sqrt{r-2}}. \tag{C.28}$$

□

This demonstrates that as $\boldsymbol{\Sigma} \to S$ for $\operatorname{rank}(\boldsymbol{\Sigma}) \geq 3$ $\|\nabla F(\boldsymbol{\Sigma})\|_F \to 0$.

This can be extended to the finite sample case as well using concentration of the quantity

$$\frac{1}{n} \sum_{i=1}^n \frac{\|\boldsymbol{P}_{\boldsymbol{\Sigma}-\boldsymbol{S}}\boldsymbol{x}_i\|_2}{\sqrt{\boldsymbol{x}_i^T \boldsymbol{\Sigma}\boldsymbol{x}_i}}. \tag{C.29}$$

In particular, we have that

$$\frac{1}{n} \sum_{i=1}^n \frac{\|\boldsymbol{P}_{\boldsymbol{\Sigma}-\boldsymbol{S}}\boldsymbol{x}_i\|_2}{\sqrt{\boldsymbol{x}_i^T \boldsymbol{\Sigma}\boldsymbol{x}_i}} \leq \sqrt{\frac{1}{n} \sum_{i=1}^n \frac{\|\boldsymbol{P}_{\boldsymbol{\Sigma}-\boldsymbol{S}}\boldsymbol{x}_i\|_2^2}{\boldsymbol{x}_i^T \boldsymbol{\Sigma}\boldsymbol{x}_i}} \tag{C.30}$$

and

$$\frac{1}{n} \sum_{i=1}^n \frac{\|\boldsymbol{P}_{\boldsymbol{\Sigma}-\boldsymbol{S}}\boldsymbol{x}_i\|_2^2}{\boldsymbol{x}_i^T \boldsymbol{\Sigma}\boldsymbol{x}_i} \leq \frac{1}{m} \left(1 + \frac{1}{n} \sum_{i=1}^n \frac{Y_i}{W_i} \right), \tag{C.31}$$

where Y_i is a χ^2_r random variable and W_i is a χ^2_r random variable. By Chebyshevs inequality, denoting $\mathrm{var}(Y_i/W_i) = \sigma^2$, we have

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{W_i} - \frac{r}{r-2}\right| \geq \frac{\epsilon\sigma^2}{n}\right) \leq \frac{1}{\epsilon^2}. \tag{C.32}$$

Letting $\epsilon = \sqrt{n}$, we arrive at the following result.

Lemma 15. *Let n be such that $\sigma^2/\sqrt{n} \leq r/(2(r-2))$. Then, with probability at least $1 - 1/n$*

$$\|\nabla F(\boldsymbol{S}_n) - \nabla F(\boldsymbol{\Sigma})\|_F \leq \frac{d\sqrt{r}}{m\sqrt{r-2}}\|\boldsymbol{\Sigma} - \boldsymbol{S}_n\|_F^{1/2} \tag{C.33}$$

C.2.4 First Order Optimality of the Low-Rank Barycenter

Moving on to the BW geometry, we first show that the low-rank barycenter is a first-order stationary point with respect to BW distance. While this follows from the previous results due to the fact that it is a global minimum over $\mathbb{S}^d_+$, we take a different approach here based on the fixed point iteration of Agueh and Carlier [2011]. This fixed point iteration forms the basis of our efficient low-rank algorithm.

By Agueh and Carlier [2011], a sufficient condition for $\gamma_{\boldsymbol{\Sigma}}$ to solve (2.9), is

$$\mathbb{E}_{\boldsymbol{x}\boldsymbol{x}^\top \sim Q}\frac{\boldsymbol{x}\boldsymbol{x}^\top}{\mathrm{Tr}(\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\Sigma})^{1/2}} = \boldsymbol{I}, \tag{C.34}$$

where $\boldsymbol{I}$ is the identity matrix in $\mathbb{R}^d$. Notice that this corresponds to the gradient ∇F being equal to $\mathbf{0}$. In the following, we let

$$\tilde{T}(\boldsymbol{\Sigma}) := \mathbb{E}_Q\frac{\boldsymbol{x}\boldsymbol{x}^\top}{\mathrm{Tr}(\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\Sigma})^{1/2}} \tag{C.35}$$

Consider the map corresponding to gradient descent with step size 1,

$$\boldsymbol{\Sigma}_{k+1} = \left(\tilde{T}(\boldsymbol{\Sigma}_k)\right)\boldsymbol{\Sigma}_k\left(\tilde{T}(\boldsymbol{\Sigma}_k)\right). \tag{C.36}$$

The corresponding fixed point equation is

$$\boldsymbol{\Sigma} = \left(\tilde{T}(\boldsymbol{\Sigma})\right)\boldsymbol{\Sigma}\left(\tilde{T}(\boldsymbol{\Sigma})\right). \tag{C.37}$$

Chewi et al. [2020] prove that the operator norm $\|\cdot\|_2$ is convex along generalized geodesics. Therefore,(C.37) maps a compact subset of $\mathbb{S}_+$ to itself, and one can apply the Brouwer fixed point theorem to guarantee a solution. The fixed point satisfies a restricted first order condition given by

$$\boldsymbol{P}_{\mathrm{Sp}(\boldsymbol{\Sigma})}\tilde{T}(\boldsymbol{\Sigma})\boldsymbol{P}_{\mathrm{Sp}(\boldsymbol{\Sigma})} = \boldsymbol{P}_{\mathrm{Sp}(\boldsymbol{\Sigma})} = \mathrm{id}_{\mathrm{Sp}(\boldsymbol{\Sigma})} = \boldsymbol{P}_{\mathrm{Sp}(\boldsymbol{\Sigma})}. \tag{C.38}$$

Notice that if $\boldsymbol{\Sigma}$ is full rank, then this implies (C.34). More generally, we need an extra condition on top of first order optimaltiy to guarantee that $\boldsymbol{\Sigma}$ is a barycenter.

Proposition 16 (Sufficient Condition for Barycenter)**.** *If $\boldsymbol{\Sigma}$ satisfies the first order conditions*

$$\boldsymbol{P}_{\boldsymbol{\Sigma}}\tilde{T}(\boldsymbol{\Sigma})\boldsymbol{P}_{\boldsymbol{\Sigma}} = \boldsymbol{P}_{\boldsymbol{\Sigma}}, \tag{C.39}$$

$$\tilde{T}(\boldsymbol{\Sigma}) \preceq \boldsymbol{I}. \tag{C.40}$$

then $\boldsymbol{\Sigma}$ is a barycenter of Q. Furthermore, in our observation model where Q is the law of $\sqrt{\boldsymbol{x}^\top\boldsymbol{S}\boldsymbol{x}}\boldsymbol{x}\boldsymbol{x}^\top$, for $\boldsymbol{x} \sim N(\mathbf{0}, \boldsymbol{I})$, $\boldsymbol{S}$ satisfies (C.39) and (C.40).

Proof. The first equation, (C.39), guarantees that $\boldsymbol{\Sigma}$ is a fixed point satisfying (C.37). On the other hand, (C.40) guarantees that all directional derivatives are positive (and that $\tilde{T}(\boldsymbol{\Sigma}) \preceq \boldsymbol{I}$), and thus $\boldsymbol{\Sigma}$ is a local minimum. To see this, suppose that $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$ is a fixed point and $\boldsymbol{\Sigma}_1$ is another PSD matrix. The directional derivatives are

$$\partial_t F(\boldsymbol{\Sigma}_t)|_{t=0} = \mathrm{Tr}\left((\boldsymbol{I} - \tilde{T}(\boldsymbol{\Sigma}))(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0)\right) \tag{C.41}$$

$$= \operatorname{Tr}\Big((\boldsymbol{I} - \tilde{T}(\boldsymbol{\Sigma}))\boldsymbol{\Sigma}_1\Big).$$

If $\tilde{T}(\boldsymbol{\Sigma}) \not\succeq \boldsymbol{I}$, then there is a $\boldsymbol{\Sigma}_1$ such that this is less than zero, and therefore $\boldsymbol{\Sigma}_0$ cannot be a minimum. On the other hand, if $\tilde{T}(\boldsymbol{\Sigma}) \preceq \boldsymbol{I}$, then all directional derivatives are positive. Combined with the Euclidean convexity result in the paper, this proves that $\boldsymbol{\Sigma}$ would be the global minimum and thus the barycenter.

Finally, in our observation model, we have

$$\tilde{T}(\boldsymbol{S}) = \boldsymbol{I}, \tag{C.42}$$

and so both (C.39) and (C.40) hold. □

Notice alternatively that this also implies that the barycenter is the only stationary point in the set where $\tilde{T}(\boldsymbol{\Sigma}) \succeq \boldsymbol{I}$. In particular, this is because at all points where $\tilde{T}(\boldsymbol{\Sigma}) \neq \boldsymbol{I}$, one can find a direction of decrease.

C.2.5 Smoothness and a Descent Lemma

The barycenter functional is *geodesically smooth*, as is shown in Theorem 7 of Chewi et al. [2020]. This result extends to non absolutely continuous measures by noting that 1) nonnegative curvature extends to measures that are not absolutely continuous with respect to Lebesgue, (see Ambrosio et al. [2008] Theorem 7.3.2), and 2) the characterization of the derivative of the Wasserstein distance extends to cases where the measures are not absolutely continuous (see Ambrosio et al. [2008] Lemma 7.3.6).

With smoothness, we have the following descent lemma over fixed rank BW space. Note that such a descent lemma is standard in the analysis of gradient descent methods (see Nesterov [2004, Theorem 2.1.5]). Here, $\boldsymbol{\Sigma}^+$ is the update after one gradient step.

Lemma 17. *For any $\boldsymbol{\Sigma} \in \mathbb{S}_+^{d,r}$ and $\boldsymbol{\Sigma}^+ = \tilde{T}(\boldsymbol{\Sigma})\boldsymbol{\Sigma}\tilde{T}(\boldsymbol{\Sigma})$, it holds that*

$$F(\boldsymbol{\Sigma}^+) - F(\boldsymbol{\Sigma}) \leq -\frac{1}{2}\left\|\boldsymbol{I} - \tilde{T}(\boldsymbol{\Sigma})\right\|_{\boldsymbol{\gamma}_{\boldsymbol{\Sigma}}}^2. \tag{C.43}$$

C.2.6 Local Geodesic Strong Convexity

We now prove local strong convexity in the population setting. By the previous section, we know that $\boldsymbol{S}$ is the unique barycenter of Q since the variance inequality holds for arbitrarily small m and ρ and arbitrarily large M.

Proposition 18. *Let $y = \langle \boldsymbol{x}\boldsymbol{x}^\top, \boldsymbol{S}\rangle$, $X \sim N(\boldsymbol{0}, \boldsymbol{I})$, $\operatorname{rank}(\boldsymbol{S}) \geq 3$, and m, M be such that $m < \lambda_r(\boldsymbol{S}) \leq \lambda_1(\boldsymbol{S}) < M$. Then, if*

$$\rho \leq \frac{m^2(r-2)}{d^2 r}\frac{c_1^2\beta^4 m^2}{36M^3} \tag{C.44}$$

F is locally geodesically strongly convex over the the largest BW ball inside the set

$$\mathcal{S}(m, M) \cap \{\boldsymbol{\Sigma} \in \mathbb{S}_+^{d,r} : \|\boldsymbol{\Sigma} - \boldsymbol{S}\|_F \leq \rho\}. \tag{C.45}$$

Proof. Fix $\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1 \in \mathcal{S}(m, M) \cap \{\boldsymbol{\Sigma} \in \mathbb{S}_+^{d,r} : \|\boldsymbol{\Sigma} - \boldsymbol{S}\|_F \leq \rho\}$.

Suppose there exists a transport map T between $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$, which we are guaranteed for ρ sufficiently small. Then for the BW geodesic $\boldsymbol{\Sigma}_t = ((1-t)\boldsymbol{I} + t\boldsymbol{T})\boldsymbol{\Sigma}_0((1-t)\boldsymbol{I} + t\boldsymbol{T})$, we can compute

$$\begin{aligned}\partial_t^2 F(\boldsymbol{\Sigma}_t)|_{t=s} &= \mathbb{E}_{\boldsymbol{x}\boldsymbol{x}^\top \sim Q} \operatorname{Tr}[(\boldsymbol{T} - \boldsymbol{I})\boldsymbol{\Sigma}_0(\boldsymbol{T} - \boldsymbol{I})] \\ &+ 2\frac{\operatorname{Tr}(\boldsymbol{x}\boldsymbol{x}^\top(\boldsymbol{T} - \boldsymbol{I})\boldsymbol{\Sigma}_0)^2}{(\operatorname{Tr}(\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\Sigma}_s))^{3/2}} \\ &- \operatorname{Tr}\left[\frac{\boldsymbol{x}\boldsymbol{x}^\top(\boldsymbol{T} - \boldsymbol{I})\boldsymbol{\Sigma}_0(\boldsymbol{T} - \boldsymbol{I})}{\sqrt{\operatorname{Tr}(\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\Sigma}_s)}}\right].\end{aligned} \tag{C.46}$$

Some manipulation when $s = 0$ yields

$$\partial_t^2 F(\boldsymbol{\Sigma}_t)|_{t=0} = \Big\langle (\boldsymbol{T} - \boldsymbol{I})\boldsymbol{\Sigma}_0(\boldsymbol{T} - \boldsymbol{I}), \nabla F(\boldsymbol{\Sigma}_0)\Big\rangle \tag{C.47}$$

$$
\begin{aligned}
&+ \mathbb{E}_{\boldsymbol{x}\boldsymbol{x}^\top \sim Q} 2 \frac{\mathrm{Tr}(\boldsymbol{x}\boldsymbol{x}^\top(\boldsymbol{T}-\boldsymbol{I})\boldsymbol{\Sigma}_0)^2}{(\mathrm{Tr}(\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\Sigma}_0))^{3/2}} \\
&\geq -d_{\mathsf{BW}}^2(\boldsymbol{\Sigma}_0,\boldsymbol{\Sigma}_1)\|\nabla F(\boldsymbol{\Sigma}_0)\|_F \\
&+ \mathbb{E}_{\boldsymbol{x}\boldsymbol{x}^\top \sim Q} 2 \frac{\mathrm{Tr}(\boldsymbol{x}\boldsymbol{x}^\top(\boldsymbol{T}-\boldsymbol{I})\boldsymbol{\Sigma}_0)^2}{(\mathrm{Tr}(\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\Sigma}_0))^{3/2}}.
\end{aligned}
$$

Following the proof of Lemma 11 last term satisfies the lower bound

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}\boldsymbol{x}^\top \sim Q} 2 \frac{\mathrm{Tr}(\boldsymbol{x}\boldsymbol{x}^\top(\boldsymbol{T}-\boldsymbol{I})\boldsymbol{\Sigma}_0)^2}{(\mathrm{Tr}(\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\Sigma}_0))^{3/2}} \\
&\qquad \geq \frac{c_1\beta^2}{3M^{3/2}}\|(\boldsymbol{T}-\boldsymbol{I})\boldsymbol{\Sigma}_0\|_F^2 \\
&\qquad \geq \frac{c_1\beta^2 m}{3M^{3/2}}\|(\boldsymbol{T}-\boldsymbol{I})\boldsymbol{\Sigma}_0^{1/2}\|_F^2 \\
&\qquad = \frac{c_1\beta^2 m}{3M^{3/2}} d_{\mathsf{BW}}^2(\boldsymbol{\Sigma}_0,\boldsymbol{\Sigma}_1)
\end{aligned}
\tag{C.48}
$$

On the other hand, by our Euclidean arguments, $\boldsymbol{S}$ is the unique rank r point such that $\nabla F = \boldsymbol{0}$. We can upper bound $\|\nabla F(\boldsymbol{\Sigma})\|_F$ on the set $\{\boldsymbol{\Sigma} \in \mathbb{S}_+^{d,r} : \|\boldsymbol{\Sigma}-\boldsymbol{S}\|_F \leq \rho\}$ using Lemma 12 by

$$
\|\nabla F(\boldsymbol{\Sigma})\|_F \leq \frac{d\sqrt{r}}{m\sqrt{r-2}}\sqrt{\rho}. \tag{C.49}
$$

Thus, if

$$
\rho \leq \frac{m^2(r-2)}{d^2 r}\frac{c_1^2\beta^4 m^2}{36M^3},
$$

then

$$
\partial_t^2 F(\boldsymbol{\Sigma}_t)|_{t=0} \geq \frac{c_1\beta^2 m}{6M^{3/2}} d_{\mathsf{BW}}^2(\boldsymbol{\Sigma}_0,\boldsymbol{\Sigma}_1). \tag{C.50}
$$

for all $\boldsymbol{\Sigma}_0 \in \{\boldsymbol{\Sigma} \in \mathbb{S}_+^{d,r} : \|\boldsymbol{\Sigma}-\boldsymbol{S}\|_F \leq \rho\}$. We thus have local strong geodesic convexity in the set

$$
\mathcal{S}(m,M) \cap \{\boldsymbol{\Sigma} \in \mathbb{S}_+^{d,r} : \|\boldsymbol{\Sigma}-\boldsymbol{S}\|_F \leq \rho\}. \tag{C.51}
$$

□

Remark 19. *We note that the same proof extends this to the sample setting with high probability in an analogous way to how Lemma 12 extends Lemma 11 to the sample setting.*

C.2.7 Proof of Theorem 4

We first restate the theorem for ease of reference.

Theorem 4 *Suppose that we observe $y_i = \langle \boldsymbol{x}_i\boldsymbol{x}_i^\top, \boldsymbol{S}\rangle$, $i = 1,\dots,n$, where $\boldsymbol{x}_i$ and $\boldsymbol{S}$ satisfy Assumption 1. Suppose further that $r = \mathrm{rank}(\boldsymbol{S}) \geq 5$, and let $\boldsymbol{S}_n = \boldsymbol{C}_n^{1/2}\boldsymbol{S}\boldsymbol{C}_n^{1/2}$. Then, for constants c_1, c_2, if $n \gtrsim dr$, with probability at least $1-\exp(-c_2 n)$,*

1. *$\boldsymbol{S}_n$ is the unique global minimizer of F over $\mathbb{S}_+^d$.*
2. *Let $\boldsymbol{U}_0$ be the initial iterate of BWGD. If $\lambda_1(\boldsymbol{U}_0\boldsymbol{U}_0^\top) \leq M$, and $F(\boldsymbol{U}_0\boldsymbol{U}_0^\top) - F(\boldsymbol{S}_n) \leq \frac{c_1^5 m^8 (r-2)^2\beta^{10}}{6^5 M^{15/2} d^3 r^2}$ for an $\boldsymbol{S}$ dependent constant β, then BWGD with step size 1 satisfies the following bound for $\mathcal{C} = O\big(c_1 m/M^{3/2}\big)$.*

$$
\begin{aligned}
&d_{\mathsf{BW}}^2(\boldsymbol{U}_k\boldsymbol{U}_k^\top, \boldsymbol{C}_n^{1/2}\boldsymbol{S}\boldsymbol{C}_n^{1/2}) \\
&\qquad \leq (1-\mathcal{C})^k (F(\boldsymbol{U}_0\boldsymbol{U}_0^\top) - F(\boldsymbol{S})).
\end{aligned}
$$

Theorem 4. The first statement follows by the Euclidean strong convexity in Lemma 12.

To prove the second statement, suppose that we initialize such that $m \leq \lambda_r(\boldsymbol{U}_0\boldsymbol{U}_0^\top) \leq \lambda_1(\boldsymbol{U}_0\boldsymbol{U}_0^\top) \leq M$, and

$$\begin{aligned} F(\boldsymbol{U}_0\boldsymbol{U}_0^\top) - F(\boldsymbol{S}) &\leq \Big(\frac{m^2(r-2)}{d^2 r}\frac{c_1^2\beta^4 m^2}{36M^3}\Big)^2 \cdot \frac{c_1\beta^2 d}{6M^{3/2}} \\ &= \frac{c_1^5 m^8 (r-2)^2 \beta^{10}}{6^5 M^{15/2} d^3 r^2}. \end{aligned} \tag{C.52}$$

First, by Euclidean strong convexity, we have

$$\|\boldsymbol{\Sigma} - \boldsymbol{S}\|_F \leq \sqrt{(F(\boldsymbol{\Sigma}) - F(\boldsymbol{S}))\frac{6M^{3/2}}{c_1\beta^2 d}}. \tag{C.53}$$

Therefore, the initialization condition on $F(\boldsymbol{U}_0\boldsymbol{U}_0^\top) - F(\boldsymbol{S})$ is enough to guarantee that

$$\|\boldsymbol{U}_0\boldsymbol{U}_0^\top - \boldsymbol{S}\| \leq \frac{m^2(r-2)}{d^2 r}\frac{c_1^2\beta^4 m^2}{36M^3}, \tag{C.54}$$

and therefore Proposition 18 holds. Furthermore, by Lemma 17, $F(\boldsymbol{U}_k\boldsymbol{U}_k^\top) \leq F(\boldsymbol{U}_{k-1}\boldsymbol{U}_{k-1}^\top)$ for all k, and thus

$$\|\boldsymbol{U}_k\boldsymbol{U}_k^\top - \boldsymbol{S}\| \leq \frac{m^2(r-2)}{d^2 r}\frac{2c_1^2\beta^4 m^2}{36M^3} \tag{C.55}$$

for all k, and the iterates remain in the ball of strong geodesic convexity.

Using the strong geodesic convexity of Proposition 18 and Lemma 17, it is then a standard argument to show linear convergence. We can apply, for example, [Zhang and Sra, 2016, Theorem 15] to yield the result.

□

C.3 Proof of Theorem 7

We give a proof of this theorem by following Ghadimi and Lan [2013].

Let $T_{\boldsymbol{\Sigma}_0\to\boldsymbol{\Sigma}_1} = T_{N(\mathbf{0},\boldsymbol{\Sigma}_0)\to N(\mathbf{0},\boldsymbol{\Sigma}_1)}$ be the optimal transport map between the measures $N(\mathbf{0},\boldsymbol{\Sigma}_0)$ and $N(\mathbf{0},\boldsymbol{\Sigma}_1)$. For simplicity, we will write $\gamma_{\boldsymbol{\Sigma}} = N(\mathbf{0},\boldsymbol{\Sigma})$. Consider a observing a sequence of rank one matrices $y_1\boldsymbol{x}_1\boldsymbol{x}_1^\top, \ldots, y_n\boldsymbol{x}_n\boldsymbol{x}_n^\top$. Let

$$\begin{aligned} \boldsymbol{\Sigma}_1 = &\Big((1-\eta)I + \eta\frac{\sqrt{y_1}\boldsymbol{x}_1\boldsymbol{x}_1^\top}{\mathrm{Tr}(\boldsymbol{x}_1\boldsymbol{x}_1^\top\boldsymbol{\Sigma}_0)^{1/2}}\Big)\boldsymbol{\Sigma}_0 \\ &\cdot\Big((1-\eta)I + \eta\frac{\sqrt{y_1}\boldsymbol{x}_1\boldsymbol{x}_1^\top}{\mathrm{Tr}(\boldsymbol{x}_1\boldsymbol{x}_1^\top\boldsymbol{\Sigma}_0)^{1/2}}\Big). \end{aligned} \tag{C.56}$$

or written in terms of measures,

$$\gamma_{\boldsymbol{\Sigma}_1} = [(1-\eta)\,\mathrm{id} + \eta T_{\boldsymbol{\Sigma}_0\to y_1\boldsymbol{x}_1\boldsymbol{x}_1^\top}]_\#\gamma_{\boldsymbol{\Sigma}_0}. \tag{C.57}$$

Due to nonnegative curvature, we have for any rank one matrix $\boldsymbol{w}\boldsymbol{w}^\top$

$$\begin{aligned} &W_2^2(\gamma_{\boldsymbol{\Sigma}_1}, \gamma_{\boldsymbol{w}\boldsymbol{w}^\top}) \leq \|T_{\boldsymbol{\Sigma}_0\to\boldsymbol{\Sigma}_1} - T_{\boldsymbol{\Sigma}_0\to\boldsymbol{w}\boldsymbol{w}^\top}\|_{\boldsymbol{\Sigma}_0}^2 \\ &= \|(1-\eta)\boldsymbol{I} + \eta T_{\boldsymbol{\Sigma}_0\to\boldsymbol{x}_1\boldsymbol{x}_1^\top} - T_{\boldsymbol{\Sigma}_0\to\boldsymbol{w}\boldsymbol{w}^\top}\|_{\boldsymbol{\Sigma}_0}^2 \\ &= \|(\boldsymbol{I} - T_{\boldsymbol{\Sigma}_0\to\boldsymbol{w}\boldsymbol{w}^\top}) + \eta(T_{\boldsymbol{\Sigma}_0\to\boldsymbol{x}_1\boldsymbol{x}_1^\top} - \boldsymbol{I})\|_{\boldsymbol{\Sigma}_0}^2 \\ &= \|\boldsymbol{I} - T_{\boldsymbol{\Sigma}_0\to\boldsymbol{w}\boldsymbol{w}^\top}\|_{\boldsymbol{\Sigma}_0}^2 + \eta^2\|\boldsymbol{I} - T_{\boldsymbol{\Sigma}_0\to\boldsymbol{x}_1\boldsymbol{x}_1^\top}\|_{\boldsymbol{\Sigma}_0}^2 \\ &- 2\eta\langle\boldsymbol{I} - T_{\boldsymbol{\Sigma}_0\to\boldsymbol{w}\boldsymbol{w}^\top}, \boldsymbol{I} - T_{\boldsymbol{\Sigma}_0\to\boldsymbol{x}_1\boldsymbol{x}_1^\top}\rangle. \end{aligned} \tag{C.58}$$

Taking the expectation with respect to $\boldsymbol{w}\boldsymbol{w}^\top = y\boldsymbol{x}\boldsymbol{x}^\top$, where $y = \boldsymbol{x}^\top\boldsymbol{S}\boldsymbol{x}$,

$$F(\boldsymbol{\Sigma}_1) - F(\boldsymbol{\Sigma}_0) \leq -2\eta\langle\nabla F(\boldsymbol{\Sigma}_0), \boldsymbol{I} - T_{\boldsymbol{\Sigma}_0\to\boldsymbol{x}_1\boldsymbol{x}_1^\top}\rangle_{\boldsymbol{\Sigma}_0} \tag{C.59}$$

$$+ \eta^2 d^2_{\mathsf{BW}}(\boldsymbol{\Sigma}_0, y_1 \boldsymbol{x}_1 \boldsymbol{x}_1^\top).$$

Summing over iterations, we find

$$\begin{aligned} &F(\boldsymbol{\Sigma}_K) - F(\boldsymbol{\Sigma}_0) \leq \\ &- 2\eta \sum_{k=1}^K \langle \nabla F(\boldsymbol{\Sigma}_{k-1}), \boldsymbol{I} - T_{\boldsymbol{\Sigma}_{k-1} \to y_k \boldsymbol{x}_k \boldsymbol{x}_k^\top} \rangle_{\boldsymbol{\Sigma}_{k-1}} \\ &+ \eta^2 \sum_{k=1}^K d^2_{\mathsf{BW}}(\boldsymbol{\Sigma}_{k-1}, y_k \boldsymbol{x}_k \boldsymbol{x}_k^\top). \end{aligned} \tag{C.60}$$

Taking the expectation with respect to $y_1 \boldsymbol{x}_1 \boldsymbol{x}_1^\top, \dots, y_K \boldsymbol{x}_K \boldsymbol{x}_K^\top$ conditioned on $\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_{K-1}$, and using $\mathbb{E}[d^2_{\mathsf{BW}}(\boldsymbol{\Sigma}_{k-1}, y_k \boldsymbol{x}_k \boldsymbol{x}_k^\top) | \boldsymbol{\Sigma}_{k-1}] \leq \sigma^2$,

$$\mathbb{E} F(\boldsymbol{\Sigma}_K) - F(\boldsymbol{\Sigma}_0) \leq -2\eta \sum_{k=1}^K \|\nabla F(\boldsymbol{\Sigma}_k)\|^2_{\boldsymbol{\Sigma}_k} + \eta^2 K \sigma^2. \tag{C.61}$$

With this,

$$\sum_{k=1}^K \|\nabla F(\boldsymbol{\Sigma}_k)\|^2_{\boldsymbol{\Sigma}_k} \leq \mathbb{E} \frac{F(\boldsymbol{\Sigma}_K) - F(\boldsymbol{\Sigma}_0)}{2\eta} + \frac{\eta}{2} K \sigma^2. \tag{C.62}$$

Choosing $\eta = 1/\sqrt{K}$,

$$\frac{1}{K} \sum_{k=1}^K \|\nabla F(\boldsymbol{\Sigma}_k)\|^2_{\boldsymbol{\Sigma}_k} \leq \mathbb{E} \frac{F(\boldsymbol{\Sigma}_K) - F(\boldsymbol{\Sigma}_0)}{2\sqrt{K}} + \frac{\eta \sigma^2}{2\sqrt{K}}. \tag{C.63}$$

Within the sequence of iterates $\boldsymbol{\Sigma}_0, \dots, \boldsymbol{\Sigma}_K$, at least one element must have gradient bounded by $O(1/\sqrt{K})$.

C.4 Suboptimal Stationary Points

There potentially exist stationary points that are not optimal in the low-rank case. In particular, with the parametrization $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{U}^\top$, these are points such that

$$\mathbb{E} \frac{\sqrt{\boldsymbol{x}^\top \boldsymbol{V}\boldsymbol{V}^\top \boldsymbol{x}}}{\sqrt{\boldsymbol{x}^\top \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{x}}} \boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{U} = \boldsymbol{U}, \tag{C.64}$$

where $\boldsymbol{V}\boldsymbol{V}^\top = \boldsymbol{S}$.

In the following, we will show that at least in the $r = 1$ case, there are no local minima $\boldsymbol{u}\boldsymbol{u}^\top$ that are not orthogonal to $\boldsymbol{v}\boldsymbol{v}^\top$.

C.4.1 No Local Minima in 1D Case

Theorem 20. *Consider the observation model with the rank one matrix $\boldsymbol{S} = \boldsymbol{v}\boldsymbol{v}^\top$ and $y = \langle \boldsymbol{x}\boldsymbol{x}^\top, \boldsymbol{S} \rangle$. Then, the only fixed points of the iteration*

$$\mathbb{E} \sqrt{\frac{\boldsymbol{x}^\top \boldsymbol{v}\boldsymbol{v}^\top \boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{u}\boldsymbol{u}^\top \boldsymbol{x}}} \boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{u} = \boldsymbol{u} \tag{C.65}$$

are orthogonal to $\boldsymbol{v}$ or $\boldsymbol{u} = \boldsymbol{v}$. Since the points orthogonal to $\boldsymbol{v}$ are local maxima, population gradient descent converges to $\boldsymbol{v}$.

Proof. In the 1-dimensional case, the stationary points are

$$\mathbb{E} \frac{|\boldsymbol{v}^\top \boldsymbol{x}|}{|\boldsymbol{u}^\top \boldsymbol{x}|} \boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{u} = \boldsymbol{u}. \tag{C.66}$$

Obviously, $\boldsymbol{u} = \boldsymbol{v}$ is a stationary point.

On the other hand, suppose that $u \perp v$. Then, we can write

$$\begin{aligned}
&\mathbb{E}\frac{|\boldsymbol{v}^\top \boldsymbol{x}|}{|\boldsymbol{u}^\top \boldsymbol{x}|}\boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{u} = \\
&\mathbb{E}\frac{|\boldsymbol{v}^\top \boldsymbol{x}|}{|\boldsymbol{u}^\top \boldsymbol{x}|}(\boldsymbol{u}\boldsymbol{u}^\top/\|\boldsymbol{u}\|^2 + \boldsymbol{v}\boldsymbol{v}^\top/\|\boldsymbol{v}\|^2 + \boldsymbol{w}\boldsymbol{w}^\top)\boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{u} \\
&= \frac{1}{\|\boldsymbol{u}\|^2}\mathbb{E}\frac{|\boldsymbol{v}^\top \boldsymbol{x}|}{|\boldsymbol{u}^\top \boldsymbol{x}|}\boldsymbol{u}\boldsymbol{u}^\top \boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{u} + \frac{1}{\|\boldsymbol{v}\|^2}E\frac{|\boldsymbol{v}^\top \boldsymbol{x}|}{|\boldsymbol{u}^\top \boldsymbol{x}|}\boldsymbol{v}\boldsymbol{v}^\top \boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{u} \\
&= \Big(\frac{1}{\|\boldsymbol{u}\|^2}\mathbb{E}|\boldsymbol{v}^\top \boldsymbol{x}|\mathbb{E}|\boldsymbol{u}^\top \boldsymbol{x}|\Big)\boldsymbol{u} + \frac{1}{\|\boldsymbol{v}\|^2}\Big(E(\boldsymbol{v}^\top \boldsymbol{x})^2\mathbb{E}\frac{\boldsymbol{x}^\top \boldsymbol{u}}{|\boldsymbol{u}^\top \boldsymbol{x}|}\Big)\boldsymbol{v} \\
&= \frac{2}{\pi}\|\boldsymbol{v}\|\frac{\boldsymbol{u}}{\|\boldsymbol{u}\|}.
\end{aligned}$$

Therefore, for this to be a stationary point, we need

$$\frac{2}{\pi}\frac{\|\boldsymbol{v}\|}{\|\boldsymbol{u}\|} = 1, \text{ or } \|\boldsymbol{u}\| = \frac{2\|\boldsymbol{v}\|}{\pi}. \tag{C.67}$$

Thus, any orthogonal vector with this length is a stationary point.

Finally, we show that there are no other stationary points. For u to be a fixed point, we need to have that

$$\mathbb{E}|\boldsymbol{v}^\top \boldsymbol{x}|\text{sign}(\boldsymbol{u}^\top \boldsymbol{x})\boldsymbol{x} = u. \tag{C.68}$$

Due to the rotational symmetry of the $\boldsymbol{x}$'s, we can assume without loss of generality that only the first two coordinates of $\boldsymbol{v}$, $\boldsymbol{u}$ are nonzero. Furthermore, we can reduce to the two dimensional case, since the coordinates $x_3, \ldots, x_d$ do not contribute to the expectation. Finally, we can assume without loss of generality that u and v are rotated so that $\boldsymbol{u} = [a, 0]^\top$. In this way, if $\boldsymbol{v} = (v_1, v_2)$, (C.68) becomes

$$\mathbb{E}|v_1x_1 + v_2x_2|\text{sign}(x_1)[x_1, x_2]^\top. \tag{C.69}$$

The first coordinate is obviously positive. If we can show that the second coordinate is nonzero, then $\boldsymbol{u}$ cannot be a fixed point. We assume without loss of generality that $v_2 > 0$.

Thus we consider

$$\mathbb{E}|v_1x_1 + v_2x_2|\text{sign}(x_1)x_2 \tag{C.70}$$

for i.i.d. $N(0, 1)$ random variables x_1, x_2. By symmetry, (x_1, x_2) occurs with the same probability as $(-x_1, -x_2)$, and

$$\begin{aligned}
&|v_1(-x_1) + v_2(-x_2)|\text{sign}(-x_1)(-x_2) = \\
&|v_1x_1 + v_2x_2|\text{sign}(x_1)x_2.
\end{aligned} \tag{C.71}$$

Therefore, we can integrate over any half-plane, and so

$$\begin{aligned}
&\mathbb{E}|v_1x_1 + v_2x_2|\text{sign}(x_1)x_2 = \\
&\mathbb{E}_{(x_1,x_2)|x_1>0}|v_1x_1 + v_2x_2|x_2.
\end{aligned} \tag{C.72}$$

For all fixed $x_1 > 0$, it is easy to see that

$$\mathbb{E}_{x_2|x_1>0}|v_1x_1 + v_2x_2|x_2 > 0, \tag{C.73}$$

since $|v_1x_1 + |v_2x_2|| > |v_1x_1 - |v_2x_2||$ when $v_2 > 0$. Therefore the second coordinate cannot be zero, and this means that u is not a fixed point.

Together with the monotonicity of gradient descent, which implies convergence to a fixed point, we conclude that population gradient descent in the 1D case converges to the underlying vector $\boldsymbol{v}$. □

C.5 Highly Local Recovery in Discrete 1D Case

Suppose that we have a discrete set of sensing vectors $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ that satisfy the ℓ_2/ℓ_1-RIP condition, and that $\frac{1}{n}\sum_i \boldsymbol{x}_i \boldsymbol{x}_i^T = \boldsymbol{I}$. Suppose that $\boldsymbol{S} = \boldsymbol{v}\boldsymbol{v}^\top$. Define the set

$$\mathcal{B} = \{\boldsymbol{u} : |\boldsymbol{u}^\top \boldsymbol{x}_i| > \epsilon \; \forall i\}. \tag{C.74}$$

This is an open set. Over this set, we can bound

$$\begin{aligned} \|\nabla F(\boldsymbol{u}) - \nabla F(\boldsymbol{v})\| &\lesssim \|\boldsymbol{v}\boldsymbol{v}^\top - \boldsymbol{u}\boldsymbol{u}^\top\|_F^{1/2} \frac{1}{n}\sum_i \frac{\|\boldsymbol{x}_i\|}{|\boldsymbol{u}^\top \boldsymbol{x}_i|} \\ &\leq C\|\boldsymbol{v}\boldsymbol{v}^\top - \boldsymbol{u}\boldsymbol{u}^\top\|_F^{1/2}, \end{aligned} \tag{C.75}$$

for some C that depends on all parameters. Assume that $\boldsymbol{v} \in \mathcal{B}$. This implies that there exists a ball around $\boldsymbol{v}$ that is contained within $\mathcal{B}$. In this ball, we can get the local Euclidean smoothness bound given in Section C.2.3. In turn, this implies local geodesic strong convexity over a small subset of this ball, which implies local linear convergence.

C.6 Lack of Strong Geodesic Convexity

We illustrate here that the functional F, while being locally strongly convex about $\boldsymbol{S}$ when restricted to rank r matrices, is not locally strongly convex around $\boldsymbol{S}$ for higher rank matrices.

Let $\boldsymbol{S}$ be the matrix

$$\boldsymbol{S} = \begin{bmatrix} a & 0 \\ 0 & 0 \end{bmatrix}, \tag{C.76}$$

$\boldsymbol{\Sigma}_0$ be the matrix

$$\boldsymbol{\Sigma}_0 = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}, \tag{C.77}$$

and let

$$\boldsymbol{T} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \tag{C.78}$$

be the transport map from $\boldsymbol{\Sigma}_0$ to $\boldsymbol{S}$. Let $\boldsymbol{\Sigma}_t$ be the geodesic from $\boldsymbol{S}$ Using the first display in (C.47), we find

$$\begin{aligned} \partial_t^2 F(\boldsymbol{\Sigma}_t)|_{t=0} &= \left\langle (\boldsymbol{T}-\boldsymbol{I})\boldsymbol{\Sigma}_0(\boldsymbol{T}-\boldsymbol{I}), \nabla F(\boldsymbol{\Sigma}_0)\right\rangle \\ &+ \mathbb{E}_{\boldsymbol{x}\boldsymbol{x}^\top \sim Q} 2\frac{\mathrm{Tr}(\boldsymbol{x}\boldsymbol{x}^\top(\boldsymbol{T}-\boldsymbol{I})\boldsymbol{\Sigma}_0)^2}{(\mathrm{Tr}(\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\Sigma}_0))^{3/2}} \\ &= \left\langle \begin{bmatrix} 0 & 0 \\ 0 & b \end{bmatrix}, \nabla F(\boldsymbol{\Sigma}_0)\right\rangle + \mathbb{E}_{\boldsymbol{x}\boldsymbol{x}^\top \sim Q} 2\frac{\mathrm{Tr}(\boldsymbol{x}\boldsymbol{x}^\top(\boldsymbol{T}-\boldsymbol{I})\boldsymbol{\Sigma}_0)^2}{(\mathrm{Tr}(\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\Sigma}_0))^{3/2}} \end{aligned} \tag{C.79}$$

By a trace inequality,

$$2\mathbb{E}_Q \frac{(\mathrm{Tr}[\boldsymbol{x}\boldsymbol{x}^\top(\boldsymbol{T}-\boldsymbol{I})\boldsymbol{\Sigma}_0^{1/2}])^2)}{\mathrm{Tr}[\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\Sigma}_0]^{3/2}} \lesssim \|(\boldsymbol{T}-\boldsymbol{I})\boldsymbol{\Sigma}_0\|_F^2. \tag{C.80}$$

To have geodesic strong convexity, we need to lower bound $\partial_t^2 F(\boldsymbol{\Sigma}_t)|_{t=0}$ by $c\|(\boldsymbol{T}-\boldsymbol{I})\boldsymbol{\Sigma}_0^{1/2}\|_F^2 = c d_{\mathsf{BW}}(\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1)^2$ for some $c > 0$. On the other hand, using the result of Section C.2.3, $\|\nabla F(\boldsymbol{\Sigma}_0)\|_F \lesssim \|\boldsymbol{\Sigma}_0 - \boldsymbol{S}\|_F = |b|$, and so

$$\partial_t^2 F(\boldsymbol{\Sigma}_t)|_{t=0} \lesssim b^2 + \|(\boldsymbol{T}-\boldsymbol{I})\boldsymbol{\Sigma}_0\|_F^2. \tag{C.81}$$

We note that

$$\|(\boldsymbol{T}-\boldsymbol{I})\boldsymbol{\Sigma}_0\|_F^2 = b^2.$$

On the other hand,

$$d_{\mathsf{BW}}^2(\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1) = \|(\boldsymbol{T}-\boldsymbol{I})\boldsymbol{\Sigma}_0^{1/2}\|_F^2 = b,$$

There is no $c > 0$ such that

$$b^2 \geq cb \tag{C.82}$$

for all $b > 0$. Therefore, F is not strongly geodesically convex at full-rank $\boldsymbol{\Sigma}_0$ that are close the boundary. In particular, if the true barycenter is low-rank, then as the full-rank $\boldsymbol{\Sigma}$ approach $\overline{\boldsymbol{\Sigma}}$, we cannot expect strong convexity.

D SUPPLEMENTAL EXPERIMENTS

D.1 Convergence of BWGD $M/m = 1$

We first repeat the experiment in Figure 2, except we generate the data using $\boldsymbol{S} = \boldsymbol{V}\boldsymbol{V}^\top$, where $\boldsymbol{V}$ now has orthogonal columns. The results are displayed in Figure 7.

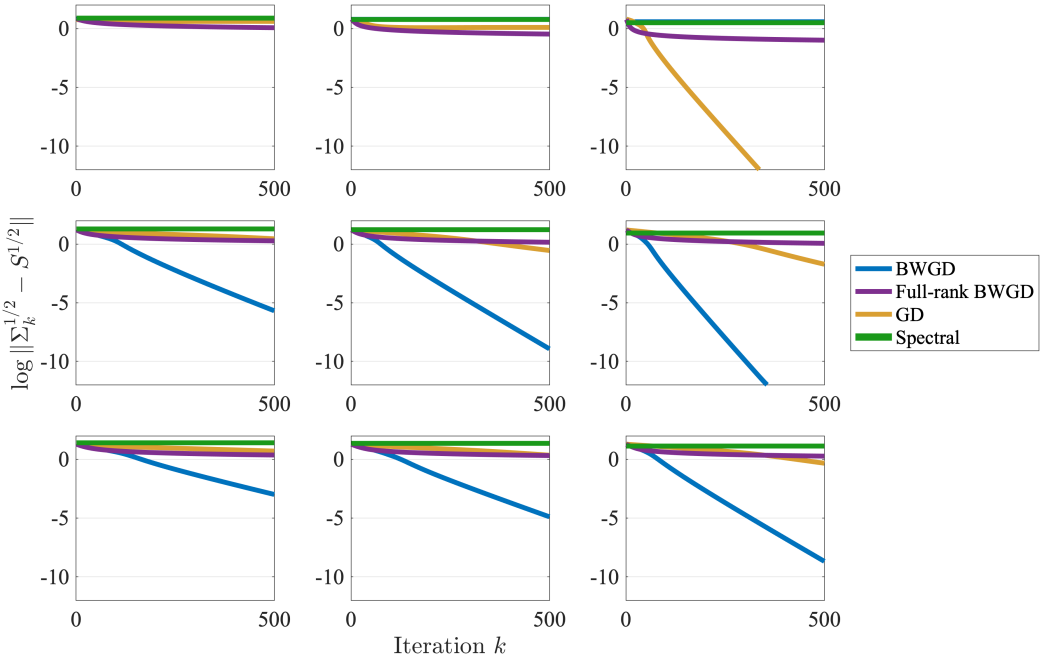

Figure 7: Convergence of BWGD (2.15) with the full gradient and GD [Li et al., 2019]. Here, the model is the same as in Figure 2, except the columns of $\boldsymbol{V}$ are orthogonal and length $\sqrt{d}$. We set $d = 32$, and down the columns we use $r = 1, 4, 16$, respectively. Across the rows we vary the number of points, with $n = 3dr$, $n = 10dr$, and $n = 20dr$, respectively. As we can see, for low to moderate ranks, the BWGD method converges much faster than the standard GD method.

D.2 Phase Retreival on CelebA Data

Here, we repeat the experiment of Section 4.2 on 20 images taken from the CelebA dataset Liu et al. [2015]. We give two plots. First, we plot error versus iteration for both BWGD and Wirtinger Flow. Second, we give some examples of the generated images. As we can see, BWGD recovers sharp images much faster than Wirtinger flow across a range of examples.

D.3 Convergence of BWGD Versus BWSGD

In the plot for SGD, we also give lines to show the different rates. Here, it appears that SGD is converging to the true barycenter. Also, it appears to be converging at a faster than anticipated rate. An explanation of this phenomenon will be explored in future work. In the left plot of Figure 9, we set $d = 20$ and $n = d^2$, and we plot the error versus iteration for BWGD for the various ranks. As we can see, the convergence takes longer as the rank increases. In the right plot of Figure 9, we plot the error versus iteration for BWSGD using the single sample gradient of (2.14), where at each iteration we draw a new sample. As we can see, the BWSGD interpolates between two convergence regimes: a slow regime where the rate is $k^{-1/4}$ and a fast regime where the rate is $k^{-1/2}$. The latter rate is typical of cases where there is local strong convexity or a Polyak-Łojasiewicz inequality.

D.4 Abalone Dataset

We also present an experiment with real data. This example is one where we attempt to measure the heterogeneity in a regression dataset. Here, we pick the classical Abalone dataset available from the UCI machine learning repository [Blake

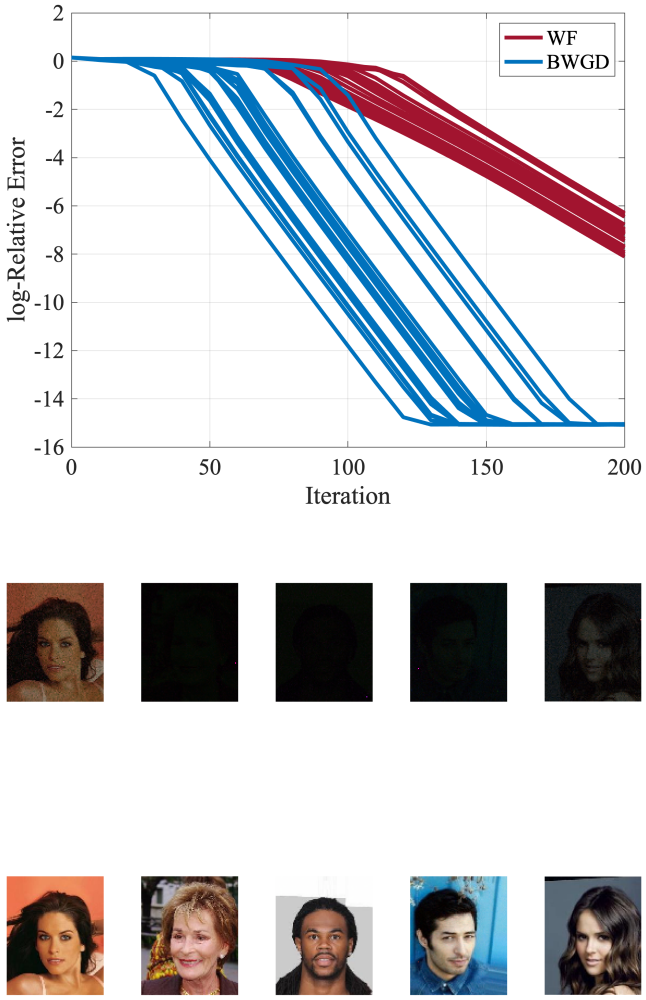

Figure 8: Convergence and example images of BWGD and WF on the CelebA Phase Retrieval experiment.

and Merz, 1998]. In this dataset, one attempts to predict the age of abalone from certain covariates. Linear regression models exhibit poor fit on this data for a variety of reasons. One reason, which we demonstrate here, is the presence of heterogeneity in the measurements.

Assuming the covariance recovery model $y_i = \langle \boldsymbol{x}_i, \boldsymbol{w}_i \rangle + \epsilon_i$, where $\boldsymbol{w}_i \sim N(\boldsymbol{0}, \boldsymbol{S})$, we could try to measure heterogeneity in the data by recovering the covariance of the regression vectors, and then plotting how $\boldsymbol{x}$ relates to y in its top principal space. Here, we recover such a covariance, and then project $\boldsymbol{x}$ onto the top two principal directions. We then plot y against these two directions. The resulting plots show varying degrees of heterogenity depending on the method employed. Here, we compare BWGD, GD, PCA on the $\boldsymbol{x}$'s, and the spectral method, which finds the top principal directions of the matrix

$$S_n = \frac{1}{2n} \sum_{i=1}^{n} y_i (\boldsymbol{x}_i \boldsymbol{x}_i^\top - \boldsymbol{I}) \tag{D.1}$$

As we can see, the spectral methods completely fails here. The BWGD method gives a similar but qualitatively different result from the Euclidean gradient descent method. In particular, there appear to be two primary directions of variation, which would indicate that this may be a mixture of two different regression components. Both BWGD and GD recover a stretched direction, but the secondary direction (in the PX_1 direction) appears to be more pronounced for BWGD.

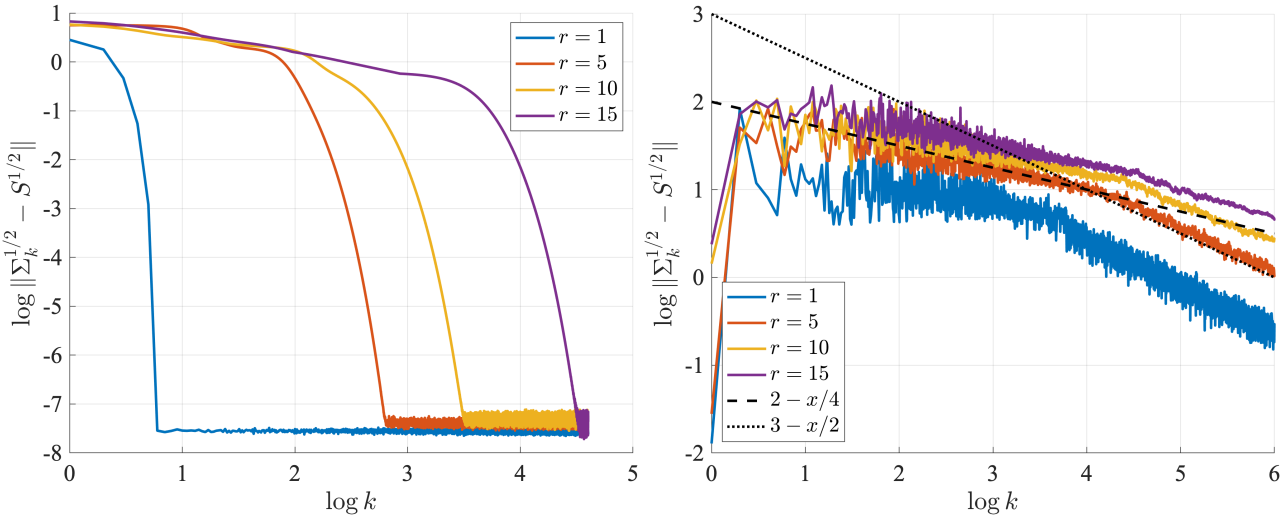

Figure 9: Convergence of BWGD and WF on the CelebA Phase Retrieval experiment.

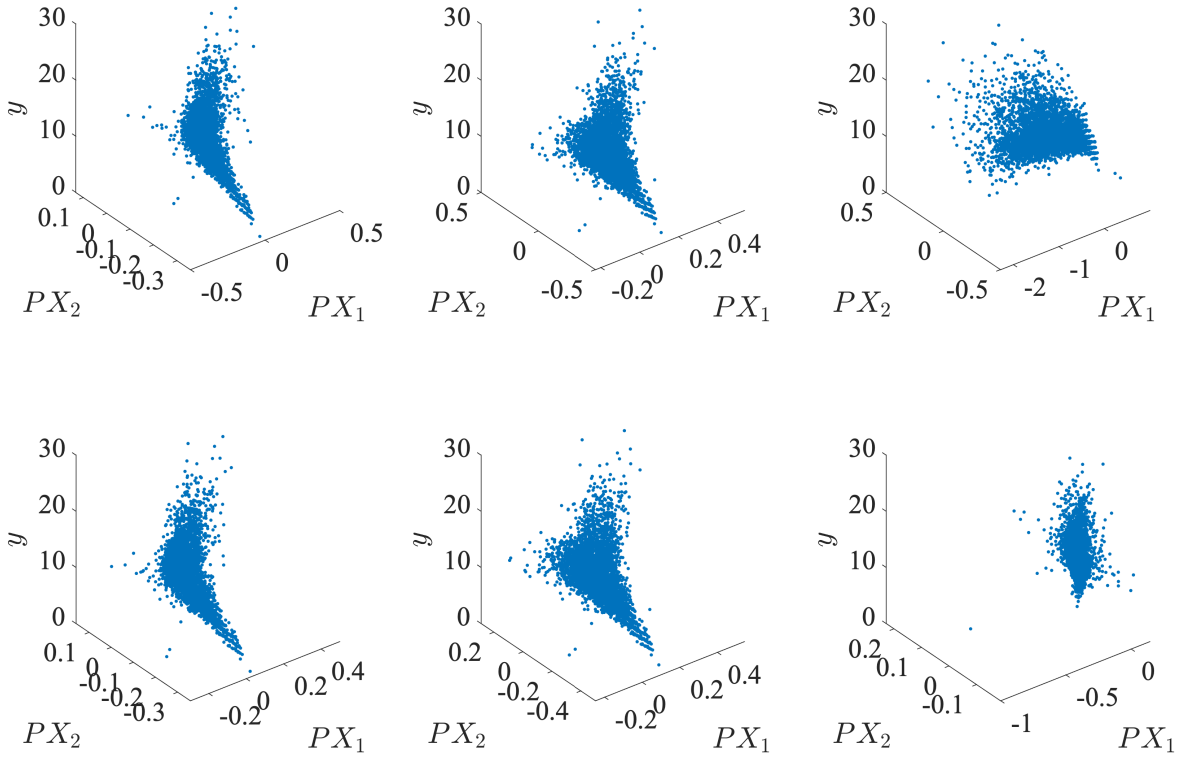

Figure 10: Projections for abalone data. The left column are projections for a fixed rank BW and full-rank BW barycenter, the middle column corresponds to low-rank and full-rank Euclidean gradient descent on (2.2), and the right column corresponds to the top principal subspace of $\boldsymbol{x}$ and the result of the spectral method.