Local Linear Convergence of Gradient Methods for Subspace Optimization via Strict Complementarity

Anonymous Author(s)
Affiliation
Address
email

Abstract

We consider optimization problems in which the goal is find a k-dimensional subspace of $\mathbb{R}^n$, $k << n$, which minimizes a convex and smooth loss. Such problems generalize the fundamental task of principal component analysis (PCA) to include robust and sparse counterparts, and logistic PCA for binary data, among others. This problem could be approached either via nonconvex gradient methods with highly-efficient iterations, but for which arguing about fast convergence to a global minimizer is difficult or, via a convex relaxation for which arguing about convergence to a global minimizer is straightforward, but the corresponding methods are often inefficient. In this work we bridge these two approaches under a strict complementarity assumption, which in particular implies that the optimal solution to the convex relaxation is unique and is also the optimal solution to the original nonconvex problem. Our main result is a proof that a natural nonconvex gradient method which is *SVD-free* and requires only a single QR-factorization of an $n \times k$ matrix per iteration, converges locally with a linear rate. We also establish linear convergence results for the nonconvex projected gradient method, and the Frank-Wolfe method when applied to the convex relaxation.

1 Introduction

We consider the problem of finding a k-dimensional subspace of $\mathbb{R}^n$, $k << n$, which minimizes a given objective function, where we identify a subspace with its corresponding projection matrix. That is, we consider the following optimization problem:

$$\min f(\mathbf{X}) \quad \text{subject to} \quad \mathbf{X} \in \mathcal{P}_{n,k} := \{\mathbf{Q}\mathbf{Q}^\top \mid \mathbf{Q} \in \mathbb{R}^{n\times k},\ \mathbf{Q}^\top\mathbf{Q} = \mathbf{I}\}. \tag{1}$$

Throughout this work and unless stated otherwise, we assume that $f(\cdot)$ is convex, β-smooth (gradient Lipschitz) and, for ease of presentation, we also assume that the gradient $\nabla f(\cdot)$ is a symmetric matrix over the space of $n \times n$ symmetric matrices $\mathbb{S}^n$[1].

Problems of interest that fall into this model include among other robust counterparts of PCA which are based on the smooth and convex Huber loss (see concrete examples in Section 4), logistic PCA [15], and sparse PCA [26]. Note that, in problem (1) we are interested in the low-dimensional subspace itself (as opposed to the projection of the data onto it, as in many other formulations), which is important for instance when the end goal is to perform dimension reduction, which is one of the most important applications of PCA-style methods.

[1] in case the the gradient is not a symmetric matrix at some point $\mathbf{X} \in \mathbb{S}^n$, then denoting it by $\nabla_{\text{nonsym}} f(\mathbf{X})$, we can always take its symmetric counterpart $\nabla f(\mathbf{X}) = \frac{1}{2}(\nabla_{\text{nonsym}} f(\mathbf{X}) + \nabla_{\text{nonsym}} f(\mathbf{X})^\top)$ and, unless stated otherwise, our derivations throughout this work will remain the same

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Motivated by high-dimensional problems, we are interested in highly efficient (in particular in terms of the dimension n) first-order methods for Problem (1). Moreover, we are interested in establishing, at least locally, fast convergence to the global minimizer despite the fact that Problem (1) is nonconvex. Subspace recovery/optimization problems similar to Problem (1) have received significant interest in recent years, see for instance [25, 28, 4, 17, 18, 11, 16, 22, 23] however, different from these works, our approach will not assume that $f(\cdot)$ admits a very specific structure (e.g., a linear or quadratic function), or will be based on a specific underlying statistical model. Instead, we will be interested in deterministic conditions that may hold for quite general $f(\cdot)$ (which is convex and smooth), and may render quite a wide variety of problems "well-posed" for efficient optimization.

We begin by briefly describing two natural dimension-efficient first-order methods for tackling Problem (1). One such method is the nonconvex projected gradient method which follows the dynamics:

$$\mathbf{X}_{t+1} \leftarrow \Pi_{\mathcal{P}_{n,k}}[\mathbf{X}_t - \eta_t \nabla f(\mathbf{X}_t)], \tag{2}$$

where $\Pi_{\mathcal{P}_{n,k}}[\cdot]$ denotes the Euclidean projection onto the set $\mathcal{P}_{n,k}$ (note that since this set is nonconvex, in general, the projection need not be unique). Given the gradient $\nabla f(\mathbf{X}_t)$, the runtime to compute $\mathbf{X}_{t+1}$ is dominated by the computation of the projection. It is not hard to show that the Euclidean projection is given by the projection matrix which corresponds to the span of the top k eigenvectors of the matrix $\mathbf{X}_t - \eta \nabla f(\mathbf{X}_t)$. While accurate computation of this projection requires a (thin) singular value decomposition (SVD) of a $n \times n$ matrix, which amounts to $O(n^3)$ runtime, it can also be approximated up to sufficiently small error using fast iterative methods, such as the well-known orthogonal iteration method [10] (aka subspace iteration method [24]) [2], whose every iteration takes in worst case only $O(kn^2)$ time. When the gradient $\nabla f(\mathbf{X}_t)$ admits a favorable structure such as sparsity or low-rank factorization, this runtime could be significantly improved.

Another natural approach to tackle Problem (1) is to exploit the fact that each $\mathbf{X} \in \mathcal{P}_{n,k}$ could be factored as $\mathbf{X} = \mathbf{Q}\mathbf{Q}^\top$, $\mathbf{Q} \in \mathbb{R}^{n \times k}$ having orthonormal columns, and to apply gradient steps w.r.t. this factorization. This leads to the following dynamics, which we refer to as *Gradient Orthogonal Iteration*:

$$\begin{aligned} &\mathbf{Z}_{t+1} \leftarrow \mathbf{Q}_t - \eta_t \frac{\partial f(\mathbf{Q}\mathbf{Q}^\top)}{\partial \mathbf{Q}}\Big|_{\mathbf{Q}_t} = \mathbf{Q}_t - \eta_t \nabla f(\mathbf{Q}_t\mathbf{Q}_t^\top)\mathbf{Q}_t, \\ &(\mathbf{Q}_{t+1}, \mathbf{R}_{t+1}) \leftarrow \text{QR-FACTORIZE}(\mathbf{Z}_{t+1}), \end{aligned} \tag{3}$$

where QR-FACTORIZE$(\cdot)$ denotes the QR factorization operation, i.e., in the above $\mathbf{Q}_{t+1} \in \mathbb{R}^{n \times k}$ has orthonormal columns. This step is required so $\mathbf{Q}_{t+1}$ has orthonormal columns (and so $\mathbf{Q}_{t+1}\mathbf{Q}_{t+1}^\top$ remains a projection matrix) for all $t \geq 1$.

As opposed to the Dynamics (2), which as discussed, an efficient implementation of will require to run a QR-based iterative method to compute the Euclidean projection onto $\mathcal{P}_{n,k}$ on each iteration, the Dynamics (3) only require a single QR factorization, and thus, given the gradient matrix $\nabla f(\mathbf{Q}_t\mathbf{Q}_t^\top)$, the next iterate $\mathbf{Q}_{t+1}$ can be computed in overall $O(n^2k)$ time. As mentioned above, this runtime could be further significantly improved if the multiplication $\nabla f(\mathbf{Q}_t\mathbf{Q}_t)\mathbf{Q}_t$ could be carried out faster than $O(n^2k)$ (for instance when the gradient is sparse or admits low-rank factorization), since all other operations require only $O(k^2n)$ time (e.g., the QR-factorization of $\mathbf{Z}_{t+1}$).

Obtaining provable guarantees on the fast local convergence of the Dynamics (3) to a global optimal solution of Problem (1) is the main contribution of this work.

While both Dynamics (2), (3) apply efficient iterations, since they are inherently nonconvex, arguing about their convergence to a global optimal solution of (1) is difficult in general. An alternative is to replace Problem (1) with a convex counterpart, for which, arguing about the convergence of first-order methods to a global optimal solution is well understood. Consider the convex set $\mathcal{F}_{n,k} = \text{conv}(\mathcal{P}_{n,k})$, where conv$(\cdot)$ denotes the convex-hull operation. $\mathcal{F}_{n,k}$ is also called the *Fantope* and it is known to admit the following important characterization: $\mathcal{F}_{n,k} = \{\mathbf{X} \in \mathbb{S}^n \mid \mathbf{I} \succeq \mathbf{X} \succeq 0, \text{Tr}(\mathbf{X}) = k\}$, where

[2] This method finds a $n \times k$ matrix $\mathbf{Q}$ with orthonormal columns which approximately span the subspace spanned by the k leading eigenvectors of a symmetric $n \times n$ matrix $\mathbf{A}$, by repeatedly applying the iterations: $(\mathbf{Q}, \mathbf{R}) \leftarrow$ QR-FACTORIZE$(\mathbf{A}\mathbf{Q})$, where QR-FACTORIZE$(\cdot)$ returns the QR factorization.

$\mathbf{A} \succeq 0$ denotes that $\mathbf{A}$ is a positive semidefinite (PSD) matrix, see for instance [20]. This leads to the convex problem:

$$\min f(\mathbf{X}) \quad \text{subject to} \quad \mathbf{X} \in \mathcal{F}_{n,k} = \{\mathbf{X} \in \mathbb{S}^n \mid \mathbf{I} \succeq \mathbf{X} \succeq 0, \mathrm{Tr}(\mathbf{X}) = k\}. \tag{4}$$

A well known first-order method applicable to (4) is the Frank-Wolfe method (aka conditional gradient) [12], which for the convex Problem (4) follows the dynamics:

$$\mathbf{V}_t \leftarrow \arg\min_{\mathbf{V} \in \mathcal{P}_{n,k}} \mathrm{Tr}(\mathbf{V} \nabla f(\mathbf{X}_t)), \quad \mathbf{X}_{t+1} \leftarrow (1-\eta_t)\mathbf{X}_t + \eta_t \mathbf{V}_t, \ \eta_t \in [0,1]. \tag{5}$$

It follows from Ky Fan's maximum principle [7] that computing $\mathbf{V}_t$ amounts to computing the projection matrix onto the span of the k eigenvectors corresponding to the k smallest eigenvalues of $\nabla f(\mathbf{X}_t)$, and hence can be carried out efficiently using the orthogonal iterations method or similar methods, similarly to the computation of the projection in (2)) discussed above. [3] Note however, that the Frank-Wolfe iterates will not be, in general, low rank, and only yield a $O(1/t)$ convergence rate [12].

1.1 The eigengap assumption and strict complementarity

We now turn to discuss our only non-completely standard assumption on Problems (1), (4), which will underly all of our contributions and in particular facilitate our local linear convergence rates.

Assumption 1 (Main assumption)**.** *An optimal solution $\mathbf{X}^*$ to the convex Problem (4) is said to satisfy the eigen-gap assumption with parameter $\delta > 0$, if $\lambda_{n-k}(\nabla f(\mathbf{X}^*)) - \lambda_{n-k+1}(\nabla f(\mathbf{X}^*)) \geq \delta$.*

Assumption 1 in particular implies the following theorem which states that the convex relaxation (4) exactly recovers the unique and optimal solution to the nonconvex Problem (1). This is one aspect in which Assumption 1 captures "well-posed" instances of Problem (1). The proof is in the appendix.

Theorem 1. *If an optimal solution $\mathbf{X}^*$ to Problem (4) satisfies Assumption 1 with some parameter $\delta > 0$, then it has rank k, i.e., $\mathbf{X}^* \in \mathcal{P}_{n,k}$, and it is the unique optimal solution to both Problem (4) and Problem (1).*

Assumption 1 is tightly related to the convex Problem (4) through the concept of *strict-complementarity* which is a classical concept in constrained continuous optimization theory [1]. A similar connection between an eigengap in the gradient at an optimal solution and strict complementarity has been already established in [5] for low-rank matrix optimization problems, where the underlying convex set is either the nuclear norm ball of matrices or the set of PSD matrices with unit trace. Now we establish a similar relationship for the convex relaxation (4) and the Fantope, which is slightly more involved. Let us write the Lagrangian of the convex Problem (4):

$$L(\mathbf{X}, \mathbf{Z}_1, \mathbf{Z}_2, s) = (\mathbf{X}) - \langle \mathbf{Z}_1, \mathbf{X} \rangle - \langle \mathbf{Z}_2, \mathbf{I} - \mathbf{X} \rangle - s(\mathrm{Tr}(\mathbf{X}) - k),$$

where the dual matrix variables $\mathbf{Z}_1, \mathbf{Z}_2$ are constrained to be PSD, i.e., $\mathbf{Z}_1 \succeq 0, \mathbf{Z}_2 \succeq 0$.

The KKT conditions state that $\mathbf{X}^*$, $(\mathbf{Z}_1^*, \mathbf{Z}_2^*, s^*)$ are corresponding optimal primal-dual solutions if and only if the following conditions hold:

$$1.\ \mathbf{I} \succeq \mathbf{X}^* \succeq 0, \mathrm{Tr}(\mathbf{X}^*) = k, \mathbf{Z}_1^* \succeq 0, \mathbf{Z}_2^* \succeq 0, \quad 2.\ \nabla f(\mathbf{X}^*) = \mathbf{Z}_1^* - \mathbf{Z}_2^* + s^* \mathbf{I},$$
$$3.\ \langle \mathbf{Z}_1^*, \mathbf{X}^* \rangle = \langle \mathbf{Z}_2^*, \mathbf{I} - \mathbf{X}^* \rangle = 0.$$

Condition 3 is known as *complementarity*. Since $\mathbf{Z}_1^*, \mathbf{Z}_2^*$ are PSD and $0 \preceq \mathbf{X}^* \preceq \mathbf{I}$, this further implies that $\mathbf{Z}_1^* \mathbf{X}^* = \mathbf{0}, \mathbf{Z}_2^*(\mathbf{I} - \mathbf{X}^*) = \mathbf{0}$, which in turn implies that

$$\mathrm{range}(\mathbf{X}^*) \subseteq \mathrm{nullspace}(\mathbf{Z}_1^*) \wedge \mathrm{range}(\mathbf{I} - \mathbf{X}^*) \subseteq \mathrm{nullspace}(\mathbf{Z}_2^*).$$

Definition 1. *A pair of primal-dual solutions $\mathbf{X}^*$, $(\mathbf{Z}_1^*, \mathbf{Z}_2^*, s^*)$ for Problem (4) is said to satisfy strict complementarity if range$(\mathbf{X}^*)$ = nullspace$(\mathbf{Z}_1^*) \vee$ range$(\mathbf{I} - \mathbf{X}^*)$ = nullspace$(\mathbf{Z}_2^*)$, which is the same as: rank$(\mathbf{Z}_1^*) = n -$ rank$(\mathbf{X}^*) \vee$ rank$(\mathbf{Z}_2^*) =$ rank$(\mathbf{X}^*)$.*

[3] We note that one can also consider projection-based first-order methods for Problem (4), such as the projected gradient method, however in general, the projection onto the Fantope $\mathcal{F}_{n,k}$ will not be a low-rank matrix and hence its computation will require an expensive SVD computation (see details in the sequel).

Theorem 2. *If an optimal solution $\mathbf{X}^*$ for Problem (4) with rank$(\mathbf{X}^*) = k$ satisfies strict complementarity for some corresponding dual solution, then $\lambda_{n-k}(\nabla f(\mathbf{X}^*)) - \lambda_{n-k+1}(\nabla f(\mathbf{X}^*)) > 0$. Conversely, if an optimal solution $\mathbf{X}^*$ for Problem (4) satisfies $\lambda_{n-k}(\nabla f(\mathbf{X}^*)) - \lambda_{n-k+1}(\nabla f(\mathbf{X}^*)) > 0$, then it satisfies strict complementarity for every corresponding dual solution.*

The proof is given in the appendix. Strict complementarity has played a central role in several recent works, both for establishing linear convergence rates for first-order methods, e.g., [29, 6, 8, 5], and improving the complexity of projected gradient methods, due to SVD computations, for low-rank matrix optimization problems, e.g., [9, 13].

1.2 Additional related work

Efficient gradient methods for low-rank nonconvex optimization have received significant interest in recent years, here we mention only a few. [2, 21] gave deterministic guarantees on the local convergence to a global minimizer of factorized gradient descent for certain low-rank optimization problems, under the mild assumption that a low-rank global minimizer exists. However, these results cannot capture constraints such as those in our Problem (1) which encode projection matrices. [4] which considers nonconvex gradient methods for low-rank statistical estimation also considers constraints that cannot capture projection matrices as in Problem (1). An exception is a specific case they consider of linear objective functions. Moreover, even for linear functions such as the specific sparse PCA objective they consider, their analysis requires several non-trivial conditions to hold (e.g. local descent, local smoothness etc), which they only show to hold under Gaussian data.

1.3 Notation

Throughout this work we let $\|\cdot\|$ denote the Euclidean norm for vectors in $\mathbb{R}^n$ and the spectral norm (largest singular value) for matrices in $\mathbb{R}^{m\times n}$ or $\mathbb{S}^n$. We let $\|\cdot\|_F$ denote the Frobenius (Euclidean) norm for matrices. For a matrix $\mathbf{X} \in \mathbb{S}^n$, we let $\lambda_i(\mathbf{X})$ denote the ith largest eigenvalue of $\mathbf{X}$. We let $\langle \cdot, \cdot \rangle$ denote the standard inner-product for both spaces $\mathbb{R}^n$ and $\mathbb{S}^n$.

2 Overview of Results

2.1 Main result

Our main novel contribution is the proof of the following theorem regarding the local linear convergence of the gradient orthogonal iteration (3) to the optimal solution of Problems (4), (1).

Theorem 3. *[Local linear convergence of gradient orthogonal iteration] Suppose Assumption 1 holds true for some optimal solution $\mathbf{X}^*$ for Problem (4) with some parameter $\delta > 0$. Let $G \geq \sup_{\mathbf{X}\in\mathcal{F}_{n,k}} \|\nabla f(\mathbf{X})\|$. Let $\mathbf{Q}_1 \in \mathbb{R}^{n\times k}$, rank$(\mathbf{Q}_1) = k$ and consider the sequence $\{\mathbf{Q}_t\}_{t\geq 1}$ generated by Dynamics (3) with fixed step-size $\eta_t = \eta = \frac{1}{5\max\{\beta, G\}}$ for all $t \geq 1$, and when initialized with $\mathbf{Q}_1 \in \mathcal{P}_{n,k}$ such that $\|\mathbf{Q}_1\mathbf{Q}_1^\top - \mathbf{X}^*\|_F \leq \min\{1, \sqrt{\frac{\delta}{2}}\frac{\eta\delta}{2(1+\eta\beta)}\}$. Then, we have that*

$$\forall t \geq 1: \quad f(\mathbf{Q}_t\mathbf{Q}_t^\top) - f(\mathbf{X}^*) \leq \big(f(\mathbf{Q}_1\mathbf{Q}_1^\top) - f(\mathbf{X}^*)\big)\exp\Big(-\frac{\delta(t-1)}{40\max\{\beta, G\}}\Big).$$

While, as stated above, this is not the first work to consider strict complementarity conditions for bridging convex and nonconvex methods for low-rank optimization, previous works such as [8, 5, 9, 13] consider gradient methods that rely on (nearly) exact (low-rank) SVD computations on each iteration, whereas Theorem 3 considers the more efficient *SVD-free* Dynamics (3), that requires only a single QR-factorization of an $n \times k$ matrix per iteration, which is much faster and simpler to implement. Accordingly, the proof is also considerably more challenging and requires new ideas.

2.2 Additional results

We also prove the following two theorems regarding the local linear convergence of the projected gradient Dynamics (2) and the Frank-Wolfe Dynamics (5). These extend the results in [9, 8] from optimization over the set of positive semidefinite matrices with unit trace to the Fantope.

Theorem 4. *[Local linear convergence of nonconvex PGD] Suppose Assumption 1 holds true for some optimal solution $\mathbf{X}^*$ for Problem (4) with some parameter $\delta > 0$. Consider the sequence $\{\mathbf{X}_t\}_{t\geq 1}$ generated by Dynamics (2) with a fixed step-size $\eta_t = \eta = 1/\beta$ for all $t \geq 1$, and when initialized with $\mathbf{X}_1 \in \mathcal{F}_{n,k}$ such that $\|\mathbf{X}_1 - \mathbf{X}^*\|_F \leq \frac{\delta}{4\beta}$. Then, for all $t \geq 1$ it holds that*

1. *$rank(\mathbf{X}_{t+1}) = k$, and thus, given $\mathbf{X}_t, \nabla f(\mathbf{X}_t)$, $\mathbf{X}_{t+1}$ can be computed via a rank-k SVD,*
2. *$f(\mathbf{X}_t) - f(\mathbf{X}^*) = O(\exp(-\Theta(\delta/\beta)(t-1)))$.*

Theorem 5. *[Local linear convergence of Frank-Wolfe] Suppose Assumption 1 holds true for some optimal solution $\mathbf{X}^*$ for Problem (4) with some parameter $\delta > 0$. Consider the sequences $\{(\mathbf{X}_t, \mathbf{V}_t)\}_{t\geq 1}$ generated by Dynamics (5) when η_t is chosen via line-search. Then, there exists $T_0 = O\left(k(\beta/\delta)^3\right)$ such that,*

$$\forall t \geq T_0 : \; f(\mathbf{X}_{t+1}) - f(\mathbf{X}^*) \leq \big(f(\mathbf{X}_t) - f(\mathbf{X}^*)\big)\Big(1 - \min\{\frac{\delta}{12\beta}, \frac{1}{2}\}\Big).$$

Moreover, for all $t \geq 1$, the rank-k matrix $\mathbf{V}_t$ satisfies $\|\mathbf{V}_t - \mathbf{X}^*\|_F^2 = O\Big(\frac{\beta^2}{\delta^3}\big(f(\mathbf{X}_t) - f(\mathbf{X}^*)\big)\Big)$.

What if Assumption 1 fails? In case Assumption 1 does not hold or holds with negligible parameter δ, not all is lost, since by considering weaker versions of Assumption 1, which consider eigen-gaps in lower eigenvalues, we can still guarantee $\mathbf{X}^*$ has low rank, and that at least the projected gradient method, locally, will require only a low-rank SVD to compute the projection onto the Fantope, while guaranteeing the standard convergence rate of $O(1/t)$ (not linear rate as when Assumption 1 holds). See the appendix for more details.

The complete proofs of Theorems 3, 4, 5, as well as additional results, are given in the appendix. Below we give the main ideas in the proof of Theorem 3.

3 Proof Sketch of Theorem 3

3.1 Preliminaries

Lemma 1 (Euclidean projection onto the Fantope)**.** *Let $\mathbf{X} \in \mathbb{R}^{n\times n}$ and consider its eigen decomposition $\mathbf{X} = \sum_{i=1}^n \gamma_i \mathbf{u}_i\mathbf{u}_i^\top$. The Euclidean projection $\Pi_{\mathcal{F}_{n,k}}[\mathbf{X}]$ is given by: $\Pi_{\mathcal{F}_{n,k}}[\mathbf{X}] = \sum_{i=1}^n \gamma_i^+(\theta)\mathbf{u}_i\mathbf{u}_i^\top$, where $\gamma_i^+(\theta) = \min(\max(\gamma_i - \theta, 0), 1)$, and θ satisfies the equation $\sum_{i=1}^n \gamma_i^+(\theta) = k$. Moreover, $\forall r \in \{k, ..., n-1\}$ it holds that $rank(\Pi_{\mathcal{F}_k}(\mathbf{X})) \leq r$ if and only if $\sum_{i=1}^r \min(\gamma_i - \gamma_{r+1}, 1) \geq k$.*

Remark 1. *Lemma 1 implies that if rank$(\mathbf{X}) \leq r$,only the top r component of the SVD of $\mathbf{X}$ are needed to compute $\Pi_{\mathcal{F}_{n,k}}[\mathbf{X}]$, i.e., a rank-$r$ SVD of $\mathbf{X}$. Moreover, given the rank-$(r+1)$ SVD, we can check the condition $\sum_{i=1}^r \min(\gamma_i - \gamma_{r+1}, 1) \geq k$, to verify whether the projection has rank $\leq r$.*

The following lemma lower-bounds, under Assumption 1, the radius of the ball around the unique optimal solution $\mathbf{X}^*$ inside-which, the PGD mapping w.r.t. the Fantope $\mathcal{F}_{n,k}$ with a fixed step-size, is guaranteed to produce rank-k matrices, i.e., matrices in $\mathcal{P}_{n,k}$, which means that it coincides precisely with the PGD mapping w.r.t. the nonconvex set $\mathcal{P}_{n,k}$ given by the Dynamics (2).

Lemma 2. *Let $\mathbf{X}^* \in \mathcal{F}_{n,k}$ be an optimal solution to Problem (4) which satisfies Assumption 1 with some parameter $\delta > 0$, and let $\eta > 0$. For any $\mathbf{X} \in \mathcal{F}_k$ which satisfies $\|\mathbf{X} - \mathbf{X}^*\|_F \leq \dfrac{\eta\delta}{2(1+\eta\beta)}$, it holds that rank$(\Pi_{\mathcal{F}_{n,k}}[\mathbf{X} - \eta\nabla f(\mathbf{X})]) = k$.*

The following lemma shows that under Assumption 1, Problem (4) has a quadratic growth property. This property is well known to facilitate linear convergence rates of gradient methods [19, 14].

Lemma 3 (Quadratic Growth)**.** *Let $\mathbf{X} \in \mathcal{F}_{n,k}$ be an optimal solution to Problem (4) for which Assumption 1 holds with some $\delta > 0$. Then, $\forall \mathbf{X} \in \mathcal{F}_{n,k} : \|\mathbf{X} - \mathbf{X}^*\|_F^2 \leq \dfrac{2}{\delta}(f(\mathbf{X}) - f(\mathbf{X}^*))$.*

3.2 Gradient Orthogonal Iteration Analysis

We outline the proof of our main algorithmic result — the local linear convergence result of the gradient orthogonal iteration (3) given in Theorem 3. For convenience, we rewrite the Dynamics (3) as Algorithm 1 below which also introduces notation that will be helpful throughout the analysis. Throughout this section we also introduce the auxiliary sequence $\{\mathbf{X}_t\}_{t\geq 1} \subset \mathcal{F}_{n,k}$ given by $\mathbf{X}_1 = \mathbf{Y}_1$ and $\mathbf{X}_{t+1} = \Pi_{\mathcal{F}_{n,k}}[\mathbf{Y}_t - \eta\nabla f(\mathbf{Y}_t)]$.

At a very high-level our analysis of Algorithm 1 relies on the following two components:

1. Using Lemma 2 we can argue that, in the proximity of $\mathbf{X}^*$, $\text{rank}(\mathbf{X}_t) = k$, i.e., $\mathbf{X}_t \in \mathcal{P}_{n,k}$. This implies that $\mathbf{X}_t$ is the projection matrix onto the span of top k eigenvectors of $\mathbf{W}_t$.
2. We view $\mathbf{Q}_t$ as the outcome of applying one iteration of the orthogonal iterations method [10, 24] to $\mathbf{W}_t$[4]. Combined with the previous point, this allows to argue that $\mathbf{Y}_t = \mathbf{Q}_t\mathbf{Q}_t^\top$ is sufficiently close to the projected gradient update $\mathbf{X}_t$, which drives the convergence.

Algorithm 1 Gradient Orthogonal Iteration

initialization: $\mathbf{Y}_1 = \mathbf{Q}_1\mathbf{Q}_1^\top$ for some $\mathbf{Q}_1 \in \mathbb{R}^{n\times k}$ such that $\mathbf{Q}_1^\top\mathbf{Q}_1 = \mathbf{I}$
for $t = 1...$ **do**
  $\mathbf{W}_{t+1} = \mathbf{Y}_t - \eta\nabla f(\mathbf{Y}_t)$
  $\mathbf{Q}_{t+1}\mathbf{R}_{t+1} = \mathbf{W}_{t+1}\mathbf{Q}_t$ (QR factorization)
  $\mathbf{Y}_{t+1} = \mathbf{Q}_{t+1}\mathbf{Q}_{t+1}^\top$
end for

The following key lemma establishes the connection between the sequence $\{\mathbf{Y}_t\}_{t\geq 1}$ produced by Algorithm 1 and the corresponding sequence of exact projected gradient steps $\{\mathbf{X}_t\}_{t\geq 1}$. The proof relies on an original extension of the classical orthogonal iteration method (see [10]).

Lemma 4. *Fix some $t \geq 1$. Suppose that $\eta < 1/G$, $\mathbf{X}_{t+1} \in \mathcal{P}_{n,k}$, and $\|\mathbf{X}_{t+1} - \mathbf{Y}_t\|_F < \sqrt{2}$. It holds that, $\|\mathbf{X}_{t+1} - \mathbf{Y}_{t+1}\|_F^2 \leq \frac{1}{1-\frac{1}{2}\|\mathbf{X}_{t+1}-\mathbf{Y}_t\|_F^2}\left(\frac{\eta G}{1-\eta G}\right)^2\|\mathbf{X}_{t+1} - \mathbf{Y}_t\|_F^2$.*

Proof of Lemma 4. Let us write the eigen decomposition of $\mathbf{W}_{t+1} = \mathbf{Y}_t - \eta\nabla f(\mathbf{Y}_t)$ as:

$$\mathbf{W}_{t+1} = \mathbf{V}\Lambda\mathbf{V}^\top = [\mathbf{V}_1 \quad \mathbf{V}_2]\begin{bmatrix}\Lambda_1 & 0\\ 0 & \Lambda_2\end{bmatrix}\begin{bmatrix}\mathbf{V}_1^\top\\ \mathbf{V}_2^\top\end{bmatrix},$$

where $\mathbf{V}_1 \in \mathbb{R}^{n\times k}, \Lambda_1 \in \mathbb{R}^{k\times k}$ correspond to the largest k eigenvalues.

The main part of the proof will be to prove that $\|\mathbf{V}_2^\top\mathbf{Q}_{t+1}\|_F^2 \leq \frac{1}{\sigma_{\min}^2(\mathbf{V}_1^\top\mathbf{Q}_t)}\left(\frac{\eta G}{1-\eta G}\right)^2\|\mathbf{V}_2^\top\mathbf{Q}_t\|_F^2$.

Note that by definition of $\mathbf{X}_{t+1}$ we have that,

$$\mathbf{X}_{t+1} = \underset{\mathbf{X}\in\mathcal{F}_{n,k}}{\arg\min}\|\mathbf{X} - \mathbf{W}_{t+1}\|_F^2 \underset{(a)}{=} \underset{\mathbf{X}\in\mathcal{P}_{n,k}}{\arg\min}\|\mathbf{X} - \mathbf{W}_{t+1}\|_F^2 \underset{(b)}{=} \underset{\mathbf{X}\in\mathcal{P}_{n,k}}{\arg\max}\langle\mathbf{X}, \mathbf{W}_{t+1}\rangle = \mathbf{V}_1\mathbf{V}_1^\top,$$

where (a) follows from the assumption of the lemma that $\mathbf{X}_{t+1} \in \mathcal{P}_{n,k}$, and (b) follows since all matrices in $\mathcal{P}_{n,k}$ have the same Frobenius norm.

This further implies that

$$\begin{aligned}\sigma_{\min}^2(\mathbf{V}_1^\top\mathbf{Q}_t) &= \lambda_k(\mathbf{V}_1^\top\mathbf{Q}_t\mathbf{Q}_t^\top\mathbf{V}_1) = \sum_{i=1}^{k}\lambda_i(\mathbf{V}_1^\top\mathbf{Q}_t\mathbf{Q}_t^\top\mathbf{V}_1) - \sum_{j=1}^{k-1}\lambda_j(\mathbf{V}_1^\top\mathbf{Q}_t\mathbf{Q}_t^\top\mathbf{V}_1)\\ &\geq \text{Tr}((\mathbf{V}_1^\top\mathbf{Q}_t\mathbf{Q}_t^\top\mathbf{V}_1) - (k-1)\lambda_1(\mathbf{V}_1^\top\mathbf{Q}_t\mathbf{Q}_t^\top\mathbf{V}_1) \geq \text{Tr}(\mathbf{X}_{t+1}\mathbf{Y}_t) - (k+1)\\ &= \left(k - \frac{1}{2}\|\mathbf{X}_{t+1} - \mathbf{Y}_t\|_F^2\right) - (k-1) = 1 - \frac{1}{2}\|\mathbf{X}_{t+1} - \mathbf{Y}_t\|_F^2. \qquad (6)\end{aligned}$$

Thus, under the assumption that $\|\mathbf{X}_{t+1} - \mathbf{Y}_t\|_F < \sqrt{2}$, we have that $(\mathbf{V}_1^\top\mathbf{Q}_t)$ is invertible.

[4] orthogonal iterations compute a matrix $\mathbf{Q} \in \mathbb{R}^{n\times k}$ with orthonormal columns, such that $\mathbf{Q}\mathbf{Q}^\top$ is the projection matrix onto the span of top k eigenvectors of $\mathbf{W}_t$

Since $(\mathbf{Q}_{t+1}, \mathbf{R}_{t+1})$ is the QR factorization of $\mathbf{W}_{t+1}\mathbf{Q}_t$, using the eigen-decomposition of $\mathbf{W}_{t+1}$ we can write $\mathbf{Q}_{t+1}\mathbf{R}_{t+1} = \mathbf{V}\Lambda\mathbf{V}^T\mathbf{Q}_t$. Multiplying both sides by $\mathbf{V}^\top$ we get

$$\begin{bmatrix}\mathbf{V}_1^\top\mathbf{Q}_{t+1}\\ \mathbf{V}_2^\top\mathbf{Q}_{t+1}\end{bmatrix}\mathbf{R}_{t+1} = \begin{bmatrix}\Lambda_1 & 0\\ 0 & \Lambda_2\end{bmatrix}\begin{bmatrix}\mathbf{V}_1^\top\mathbf{Q}_t\\ \mathbf{V}_2^\top\mathbf{Q}_t\end{bmatrix},$$

which leads to the two equations:

$$\Lambda_1\mathbf{V}_1^\top\mathbf{Q}_t = \mathbf{V}_1^\top\mathbf{Q}_{t+1}\mathbf{R}_{t+1}, \tag{7}$$

$$\Lambda_2\mathbf{V}_2^\top\mathbf{Q}_t = \mathbf{V}_2^\top\mathbf{Q}_{t+1}\mathbf{R}_{t+1}. \tag{8}$$

Under the assumption that $\eta < 1/G$, using Weyl's inequality we have that $\lambda_k(\mathbf{W}_{t+1}) \geq \lambda_k(\mathbf{Y}_t) - \eta\lambda_1(\nabla f(\mathbf{Y}_t)) > 0$, and so Λ_1 is invertible. Since following (6) we have that $\sigma_{\min}(\mathbf{V}_1^\top\mathbf{Q}_t) > 0$, it follows that $\text{rank}(\Lambda_1\mathbf{V}_1^\top\mathbf{Q}_t) = k$ and thus, from the equation (7) we have that $\mathbf{V}_1^\top\mathbf{Q}_{t+1}$ and $\mathbf{R}_{t+1}$ are both invertible and we can write $\mathbf{R}_{t+1} = (\mathbf{V}_1^\top\mathbf{Q}_{t+1})^{-1}\Lambda_1\mathbf{V}_1^\top\mathbf{Q}_t$.

Multiplying equation (8) on both sides from the right with $\mathbf{R}_{t+1}^{-1}$, we get

$$\mathbf{V}_2^\top\mathbf{Q}_{t+1} = \Lambda_2\mathbf{V}_2^\top\mathbf{Q}_t(\mathbf{V}_1^\top\mathbf{Q}_t)^{-1}\Lambda_1^{-1}\mathbf{V}_1^\top\mathbf{Q}_{t+1}.$$

Now we can use this to bound $\|\mathbf{V}_2^\top\mathbf{Q}_{t+1}\|_F^2$:

$$\begin{aligned}\|\mathbf{V}_2^\top\mathbf{Q}_{t+1}\|_F^2 &= \|\Lambda_2\mathbf{V}_2^\top\mathbf{Q}_t(\mathbf{V}_1^\top\mathbf{Q}_t)^{-1}\Lambda_1^{-1}\mathbf{V}_1^\top\mathbf{Q}_{t+1}\|_F^2\\ &\underset{(a)}{\leq} \|(\mathbf{V}_1^\top\mathbf{Q}_t)^{-1}\|_2^2\|\Lambda_1^{-1}\|_2^2\|\mathbf{V}_1^\top\mathbf{Q}_{t+1}\|_2^2\|\Lambda_2\|_2^2\|\mathbf{V}_2^\top\mathbf{Q}_t\|_F^2 \underset{(b)}{\leq} \frac{\|\mathbf{V}_2^\top\mathbf{Q}_t\|_F^2}{\sigma_{\min}^2(\mathbf{V}_1^\top\mathbf{Q}_t)}\left(\frac{\lambda_{k+1}(\mathbf{W}_{t+1})}{\lambda_k(\mathbf{W}_{t+1})}\right)^2,\end{aligned} \tag{9}$$

where (a) follows from the inequalities $\|\mathbf{AB}\|_F \leq \min\{\|\mathbf{A}\|_F\|\mathbf{B}\|_2, \|\mathbf{A}\|_2\|\mathbf{B}\|_F\}$, $\|\mathbf{AB}\|_2 \leq \|\mathbf{A}\|_2\|\mathbf{B}\|_2$, and (b) follows from the eigen-decomposition of $\mathbf{W}_{t+1}$ and by noting that since $\mathbf{V}_1, \mathbf{Q}_{t+1}$ both have orthonormal columns, it holds that $\|\mathbf{V}_1^\top\mathbf{Q}_{t+1}\|_2 \leq 1$.

We upper-bound $\lambda_{k+1}(\mathbf{W}_{t+1})/\lambda_k(\mathbf{W}_{t+1})$ by using Weyl's inequality as follows:

$$\frac{\lambda_{k+1}(\mathbf{W}_{t+1})}{\lambda_k(\mathbf{W}_{t+1})} \leq \frac{\lambda_{k+1}(\mathbf{Y}_t) + \eta\lambda_1(-\nabla f(\mathbf{Y}_t))}{\lambda_k(\mathbf{Y}_t) + \eta\lambda_n(-\nabla f(\mathbf{Y}_t))} \leq \frac{\eta G}{1-\eta G}, \tag{10}$$

where we have used the fact that $\mathbf{Y}_t \in \mathcal{P}_{n,k}$ and so $\lambda_k(\mathbf{Y}_t) = 1, \lambda_{k+1}(\mathbf{Y}_t) = 0$.

Plugging (10) into (9) we indeed obtain

$$\|\mathbf{V}_2^\top\mathbf{Q}_{t+1}\|_F^2 \leq \frac{1}{\sigma_{\min}^2(\mathbf{V}_1^\top\mathbf{Q}_t)}\left(\frac{\eta G}{1-\eta G}\right)^2\|\mathbf{V}_2^\top\mathbf{Q}_t\|_F^2. \tag{11}$$

Now, for the final part of the proof, we note that $\|\mathbf{V}_2^\top\mathbf{Q}_{t+1}\|_F^2 = \text{Tr}(\mathbf{V}_2\mathbf{V}_2^\top\mathbf{Y}_{t+1}) = \text{Tr}((\mathbf{I} - \mathbf{X}_{t+1})\mathbf{Y}_{t+1}) = k - \text{Tr}(\mathbf{X}_{t+1}\mathbf{Y}_{t+1}) = \frac{1}{2}\|\mathbf{X}_{t+1} - \mathbf{Y}_{t+1}\|_F^2$, and similarly, $\|\mathbf{V}_2^\top\mathbf{Q}_t\|_F^2 = \frac{1}{2}\|\mathbf{X}_{t+1} - \mathbf{Y}_t\|_F^2$. Plugging these observations and (6) into (11), we obtain the lemma. □

The following lemma is the main step in the proof of the convergence rate of Algorithm 1.

Lemma 5. *Let us denote $h_t = f(\mathbf{Y}_t) - f(\mathbf{X}^*)$ for all $t \geq 1$. Suppose that $\eta \leq \frac{1}{5\max\{\beta,G\}}$ and suppose that $\mathbf{X}_{t+1} \in \mathcal{P}_{n,k}$ and that $\|\mathbf{X}_{t+1} - \mathbf{Y}_t\|_F \leq 1$. Denote the constants $C_0 = 2\left(\frac{\eta G}{1-\eta G}\right)^2$, $C_1 = \frac{2(1+\eta G)C_0}{1-2\eta\beta-2C_0(1+\eta G)}$. It holds that, $h_{t+1} \leq \left(1 - \frac{\eta\delta}{4(1+C_1)}\right)h_t$.*

Proof. Using the β-smoothness of $f(\mathbf{X})$, for any $\mathbf{X} \in \mathcal{F}_{n,k}$ and $\eta \leq \dfrac{1}{\beta}$ it holds that

$$\begin{aligned}f(\mathbf{X}) &\leq f(\mathbf{Y}_t) + \langle\mathbf{X} - \mathbf{Y}_t, \nabla f(\mathbf{Y}_t)\rangle + \frac{1}{2\eta}\|\mathbf{X} - \mathbf{Y}_t\|_F^2\\ &\underset{(a)}{\leq} f(\mathbf{Y}_t) + \langle\mathbf{X} - \mathbf{Y}_t, \nabla f(\mathbf{Y}_t)\rangle + \eta^{-1}\langle\mathbf{Y}_t, \mathbf{Y}_t - \mathbf{X}\rangle\\ &= f(\mathbf{Y}_t) + \eta^{-1}\langle\mathbf{Y}_t - \mathbf{X}, \mathbf{Y}_t - \eta\nabla f(\mathbf{Y}_t)\rangle,\end{aligned} \tag{12}$$

where (a) follows since using the fact that $\mathbf{Y}_t \in \mathcal{P}_{n,k}$, we have that for any $\mathbf{X} \in \mathcal{F}_{n,k}$ it holds that $\|\mathbf{X}\|_F^2 \leq k = \|\mathbf{Y}_t\|_F^2 = \langle \mathbf{Y}_t, \mathbf{Y}_t \rangle$.

Since by the assumption of the lemma $\mathbf{X}_{t+1} = \Pi_{\mathcal{F}_{n,k}}[\mathbf{Y}_t - \eta \nabla f(\mathbf{Y}_t)] \in \mathcal{P}_{n,k}$, using the optimality condition in Lemma 6 we have that for all $\mathbf{Z} \in \mathcal{F}_{n,k}$ it holds that $\langle \mathbf{X}_{t+1} - \mathbf{Z}, \mathbf{Y}_t - \eta_t \nabla f(\mathbf{Y}_t) \rangle \geq 0$. This implies that for all $\mathbf{Z} \in \mathcal{F}_{n,k}$:

$$\begin{aligned}
&\langle \mathbf{Y}_{t+1}, \mathbf{Y}_t - \eta \nabla f(\mathbf{Y}_t) \rangle = \langle \mathbf{X}_{t+1}, \mathbf{Y}_t - \eta \nabla f(\mathbf{Y}_t) \rangle - \langle \mathbf{X}_{t+1} - \mathbf{Y}_{t+1}, \mathbf{Y}_t - \eta \nabla f(\mathbf{Y}_t) \rangle \geq \\
&\langle \mathbf{Z}, \mathbf{Y}_t - \eta \nabla f(\mathbf{Y}_t) \rangle - \langle \mathbf{X}_{t+1} - \mathbf{Y}_{t+1}, \mathbf{Y}_t - \eta \nabla f(\mathbf{Y}_t) \rangle \geq \\
&\langle \mathbf{Z}, \mathbf{Y}_t - \eta \nabla f(\mathbf{Y}_t) \rangle - \|\mathbf{X}_{t+1} - \mathbf{Y}_{t+1}\|^2 \|\mathbf{W}_{t+1}\|_2, \qquad (13)
\end{aligned}$$

where the last inequality is due to Lemma 8 and recalling that $\mathbf{W}_{t+1} = \mathbf{Y}_t - \eta \nabla f(\mathbf{Y}_t)$.

Setting $\mathbf{X} = \mathbf{Y}_{t+1}$ in (12) and plugging-in (13), we have that for any $\mathbf{Z} \in \mathcal{F}_{n,k}$ it holds that

$$\begin{aligned}
f(\mathbf{Y}_{t+1}) &\leq f(\mathbf{Y}_t) + \eta^{-1} \left(\langle \mathbf{Y}_t - \mathbf{Z}, \mathbf{Y}_t - \eta \nabla f(\mathbf{Y}_t) \rangle + \|\mathbf{X}_{t+1} - \mathbf{Y}_{t+1}\|^2 \|\mathbf{W}_{t+1}\|_2 \right) \\
&= f(\mathbf{Y}_t) + \langle \mathbf{Z} - \mathbf{Y}_t, \nabla f(\mathbf{Y}_t) \rangle + \frac{1}{2\eta} \|\mathbf{Z} - \mathbf{Y}_t\|_F^2 + \frac{1}{\eta} \|\mathbf{X}_{t+1} - \mathbf{Y}_{t+1}\|^2 \|\mathbf{W}_{t+1}\|_2 \\
&\leq f(\mathbf{Y}_t) + \langle \mathbf{Z} - \mathbf{Y}_t, \nabla f(\mathbf{Y}_t) \rangle + \frac{1}{2\eta} \|\mathbf{Z} - \mathbf{Y}_t\|_F^2 + \frac{1 + \eta G}{\eta} \|\mathbf{X}_{t+1} - \mathbf{Y}_{t+1}\|_F^2, \qquad (14)
\end{aligned}$$

where the last inequality is due to the following upper-bound on $\|\mathbf{W}_{t+1}\|_2$:

$$\|\mathbf{W}_{t+1}\|_2 = \|\mathbf{Y}_t - \eta \nabla f(\mathbf{Y}_t)\|_2 \leq \|\mathbf{Y}_t\|_2 + \eta \|\nabla f(\mathbf{Y}_t)\|_2 \leq 1 + \eta G.$$

In particular, setting $\mathbf{Z} = (1 - \alpha)\mathbf{Y}_t + \alpha \mathbf{X}^*$ for some $\alpha \in [0, 1]$, we get that

$$f(\mathbf{Y}_{t+1}) \leq f(\mathbf{Y}_t) + \alpha \langle \mathbf{X}^* - \mathbf{Y}_t, \nabla f(\mathbf{Y}_t) \rangle + \frac{\alpha^2}{2\eta} \|\mathbf{X}^* - \mathbf{Y}_t\|_F^2 + \frac{1 + \eta G}{\eta} \|\mathbf{X}_{t+1} - \mathbf{Y}_{t+1}\|_F^2.$$

Subtracting $f(\mathbf{X}^*)$ from both sides, using the convexity of $f(\cdot)$, and Lemma 3 gives

$$h_{t+1} \leq \left(1 - \alpha + \frac{\alpha^2}{\eta \delta}\right) h_t + \frac{1 + \eta G}{\eta} \|\mathbf{X}_{t+1} - \mathbf{Y}_{t+1}\|_F^2.$$

Setting $\alpha = \eta \delta / 2$ (note that since $\eta \leq 1/G$, we have that $\alpha \in [0, 1]$), gives

$$h_{t+1} \leq \left(1 - \frac{\eta \delta}{4}\right) h_t + \frac{1 + \eta G}{\eta} \|\mathbf{X}_{t+1} - \mathbf{Y}_{t+1}\|_F^2. \qquad (15)$$

We now continue to upper-bound the term $\|\mathbf{X}_{t+1} - \mathbf{Y}_{t+1}\|_F^2$. Using Lemma 9 we have that

$$\|\mathbf{X}_{t+1} - \mathbf{Y}_t\|_F^2 \leq \frac{\eta}{1 - \eta \beta} \left(f(\mathbf{Y}_t) - f(\mathbf{X}_{t+1}) \right). \qquad (16)$$

Let us set $\mathbf{Z} = \mathbf{X}_{t+1}$ in (14) to obtain that

$$\begin{aligned}
f(\mathbf{Y}_{t+1}) &\leq f(\mathbf{Y}_t) + \langle \mathbf{X}_{t+1} - \mathbf{Y}_t, \nabla f(\mathbf{Y}_t) \rangle + \frac{1}{2\eta} \|\mathbf{X}_{t+1} - \mathbf{Y}_t\|_F^2 + \frac{1 + \eta G}{\eta} \|\mathbf{X}_{t+1} - \mathbf{Y}_{t+1}\|_F^2 \\
&\leq f(\mathbf{X}_{t+1}) + \frac{1}{2\eta} \|\mathbf{X}_{t+1} - \mathbf{Y}_t\|_F^2 + \frac{1 + \eta G}{\eta} \|\mathbf{X}_{t+1} - \mathbf{Y}_{t+1}\|_F^2,
\end{aligned}$$

where the last inequality is due to convexity of $f(\cdot)$. Rearranging and using Lemma 4 along with the notation $C_0 = 2\left(\frac{\eta G}{1 - \eta G}\right)^2$, we have $f(\mathbf{X}_{t+1}) \geq f(\mathbf{Y}_{t+1}) - \frac{1}{\eta}\left(\frac{1}{2} + C_0(1 + \eta G)\right) \|\mathbf{X}_{t+1} - \mathbf{Y}_t\|_F^2$. Plugging into (16) we obtain

$$\|\mathbf{X}_{t+1} - \mathbf{Y}_t\|_F^2 \leq \frac{\eta}{1 - \eta \beta} \left(f(\mathbf{Y}_t) - f(\mathbf{Y}_{t+1}) + \frac{1}{\eta}\left(\frac{1}{2} + C_0(1 + \eta G)\right) \|\mathbf{X}_{t+1} - \mathbf{Y}_t\|_F^2 \right),$$

and rearranging we obtain

$$\|\mathbf{X}_{t+1} - \mathbf{Y}_t\|_F^2 \leq \frac{1}{1 - \frac{1 + 2C_0(1 + \eta G)}{2(1 - \eta \beta)}} \frac{\eta}{1 - \eta \beta} \left(f(\mathbf{Y}_t) - f(\mathbf{Y}_{t+1}) \right) = \frac{2\eta(h_t - h_{t+1})}{2(1 - \eta \beta) - 1 - 2C_0(1 + \eta G)}.$$

Using Lemma 4 again we have, $\|\mathbf{X}_{t+1} - \mathbf{Y}_{t+1}\|_F^2 \leq \frac{2\eta C_0}{1 - 2\eta\beta - 2C_0(1 + \eta G)} (h_t - h_{t+1})$. Plugging back into (15) we obtain $h_{t+1} \leq \left(1 - \frac{\eta \delta}{4}\right) h_t + \frac{2(1 + \eta G) C_0}{1 - 2\eta\beta - 2C_0(1 + \eta G)} (h_t - h_{t+1})$. Denoting $C_1 = \frac{2(1 + \eta G) C_0}{1 - 2\eta\beta - 2C_0(1 + \eta G)}$, we finally obtain $h_{t+1} \leq \frac{1}{1 + C_1}\left(1 - \frac{\eta \delta}{4} + C_1\right) h_t = \left(1 - \frac{\eta \delta}{4(1 + C_1)}\right) h_t$, as required. The only thing left is to choose a feasible step size. We have to require: $1 - 2\eta\beta - 2C_0(1 + \eta G) > 0$. The latter holds for any $\eta \leq \frac{1}{5 \max\{\beta, G\}}$. $\square$

4 Numerical Simulations

We briefly discuss our numerical simulations. Additional information and results, including support for our warm-start assumption, are given in the appendix. We consider two models for robust subspace recovery from noisy samples: 1. a *spiked covariance* model, and 2. a *sparsely corrupted entries* model. In both models we minimize a robust loss based on the Huber function, which is convex and smooth, over the Fantope. We generate random instances and solve them to high precision (duality gap $< 10^{-10}$) and take the resulting point $\mathbf{X}^*$ as the optimal solution. We measure the eigen-gap (as in Assumption 1) in $\nabla f(\mathbf{X}^*)$ and we compare the recovery error w.r.t. the naive PCA solution $\mathbf{X}_{PCA}$ which simply computes the leading subspace of the empirical covariance. The results are given in Table 1. As can be seen, for both models the recovery error is significantly lower than that of naive PCA, which demonstrates the usefulness of the chosen models . We see that the eigen-gap assumption indeed holds with substantial values of δ.

Noise prob. (p)	0.05	0.1	0.2	0.3	0.4	0.5
$\downarrow$ Model 1: spiked covariance $\downarrow$						
Eigen-gap (δ)	3.21	2.87	2.36	2.04	1.501	1.03
$\|\mathbf{X}^* - \mathbf{P}\|_F$	0.0047	0.0075	0.012	0.016	0.022	0.0298
$\|\mathbf{X}_{PCA} - \mathbf{P}\|_F$	0.045	0.072	0.115	0.157	0.212	0.292
$\downarrow$ Model 2: sparsely corrupted entries $\downarrow$						
Eigen-gap (δ)	5.72	5.49	5.15	4.81	4.38	3.79
$\|\mathbf{X}^* - \mathbf{P}\|_F$	0.049	0.067	0.097	0.111	0.134	0.148
$\|\mathbf{X}_{PCA} - \mathbf{P}\|_F$	0.148	0.199	0.291	0.335	0.401	0.439

Table 1: Recovery and eigen-gap results for spiked covariance sparsely corrupted entries models with varying noise probabilities. $\mathbf{P}$ is the projection matrix onto the ground truth subspace. $n = 100$, $k = 10$, sample size $m = 500$. Results are averages of 20 i.i.d. experiments.

We additionally test the empirical convergence of nonconvex PGD (Dynamics (2)) and the gradient orthogonal iteration method (Dynamics (3)) on the above models. We initialize both methods with the PCA solution $\mathbf{X}_{PCA}$ and use the same fixed step-size for both. We indeed observe the rapid convergence of both methods. In particular, we verify that the iterates of nonconvex PGD over $\mathcal{P}_{n,k}$ coincide precisely with those of PGD w.r.t. the convex Fantope $\mathcal{F}_{n,k}$ (see Remark 1), which suggests that nonconvex PGD indeed converges to the global minimum. Figure 1 (left panel) also suggests that indeed both GD and GOI converge with a linear rate.

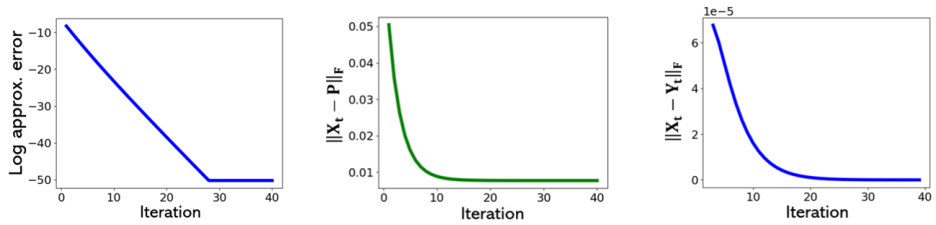

Figure 1: Convergence of PGD and GOI for the spiked covariance model with $p = 0.1$. Left panel shows the approximation error w.r.t. function value of PGD (in log scale). Middle panel shows recovery error of PGD ($\mathbf{P}$ is the ground truth projection matrix). Convergence of GOI from is omitted left and middle panels since it nearly coincides with that of PGD. Right panel shows distance (in Frobenius norm) between iterates of PGD ($\mathbf{X}_t$) and those of GOI ($\mathbf{Y}_t$).

References

[1] Farid Alizadeh, Jean-Pierre A Haeberly, and Michael L Overton. Complementarity and nondegeneracy in semidefinite programming. *Mathematical programming*, 77(1):111–128, 1997.

[2] Srinadh Bhojanapalli, Anastasios Kyrillidis, and Sujay Sanghavi. Dropping convexity for faster semi-definite optimization. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors,

29th Annual Conference on Learning Theory, volume 49 of *Proceedings of Machine Learning Research*, pages 530–582, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR.

[3] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *arXiv preprint arXiv:1405.4980*, 2014.

[4] Yudong Chen and Martin J Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

[5] Lijun Ding, Yingjie Fei, Qiantong Xu, and Chengrun Yang. Spectral frank-wolfe algorithm: Strict complementarity and linear convergence. In *International Conference on Machine Learning*, pages 2535–2544. PMLR, 2020.

[6] Dmitriy Drusvyatskiy and Adrian S Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018.

[7] Ky Fan. On a theorem of weyl concerning eigenvalues of linear transformations i. *Proceedings of the National Academy of Sciences of the United States of America*, 35(11):652, 1949.

[8] Dan Garber. Linear convergence of frank-wolfe for rank-one matrix recovery without strong convexity. *arXiv preprint arXiv:1912.01467*, 2019.

[9] Dan Garber. On the convergence of projected-gradient methods with low-rank projections for smooth convex minimization over trace-norm balls and related problems. *SIAM Journal on Optimization*, 31(1):727–753, 2021.

[10] Gene H Golub and Charles F Van Loan. Matrix computations. johns hopkins studies in the mathematical sciences.

[11] Moritz Hardt and Ankur Moitra. Algorithms and hardness for robust subspace recovery. In *Conference on Learning Theory*, pages 354–375. PMLR, 2013.

[12] Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *International Conference on Machine Learning*, pages 427–435. PMLR, 2013.

[13] Atara Kaplan and Dan Garber. Low-rank extragradient method for nonsmooth and low-rank matrix optimization problems. *Advances in Neural Information Processing Systems*, 34, 2021.

[14] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

[15] Andrew J Landgraf and Yoonkyung Lee. Dimensionality reduction for binary data through the projection of natural parameters. *Journal of Multivariate Analysis*, 180:104668, 2020.

[16] Gilad Lerman and Tyler Maunu. Fast, robust and non-convex subspace recovery. *Information and Inference: A Journal of the IMA*, 7(2):277–336, 2018.

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

[18] Tyler Maunu, Teng Zhang, and Gilad Lerman. A well-tempered landscape for non-convex robust subspace recovery. *Journal of Machine Learning Research*, 20(37), 2019.

[19] Ion Necoara, Yu Nesterov, and Francois Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175(1):69–107, 2019.

[20] Michael L Overton and Robert S Womersley. On the sum of the largest eigenvalues of a symmetric matrix. *SIAM Journal on Matrix Analysis and Applications*, 13(1):41–45, 1992.

[21] Dohyung Park, Anastasios Kyrillidis, Constantine Caramanis, and Sujay Sanghavi. Finding low-rank solutions via nonconvex matrix factorization, efficiently and provably. *SIAM Journal on Imaging Sciences*, 11(4):2165–2204, 2018.

[22] Prasad Raghavendra and Morris Yau. List decodable subspace recovery. In *Conference on Learning Theory*, pages 3206–3226. PMLR, 2020.

[23] Mostafa Rahmani and George K Atia. Randomized robust subspace recovery and outlier detection for high dimensional data matrices. *IEEE Transactions on Signal Processing*, 65(6):1580–1594, 2016.

[24] Yousef Saad. *Numerical methods for large eigenvalue problems: revised edition*. SIAM, 2011.

[25] Ohad Shamir. Fast stochastic algorithms for svd and pca: Convergence properties and convexity. In *International Conference on Machine Learning*, pages 248–256. PMLR, 2016.

[26] Vincent Q Vu, Juhee Cho, Jing Lei, and Karl Rohe. Fantope projection and selection: A near-optimal convex relaxation of sparse pca. *Advances in neural information processing systems*, 26, 2013.

[27] Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

[28] Dejiao Zhang and Laura Balzano. Global convergence of a grassmannian gradient descent algorithm for subspace estimation. In *Artificial Intelligence and Statistics*, pages 1460–1468. PMLR, 2016.

[29] Zirui Zhou and Anthony Man-Cho So. A unified approach to error bounds for structured convex optimization problems. *Mathematical Programming*, 165(2):689–728, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The details in the experimental section allow to reproduce the results exactly
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]