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

Ron Fisher
Technion - Israel Institute of Technology
Haifa, Israel 3200003
ronfisher@campus.technion.ac.il

Dan Garber
Technion - Israel Institute of Technology
Haifa, Israel 3200003
dangar@technion.ac.il

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 high dimensions. 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 others 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 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

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

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), and $\eta_t > 0$ is the step-size. 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 well known that the Euclidean projection is given by the projection matrix which corresponds to the span of the top k eigenvectors of the matrix $\mathbf{X}_t - \eta_t \nabla f(\mathbf{X}_t)$. 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]). The orthogonal iteration 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 given positive semidefinite $n \times n$ matrix $\mathbf{A}$, by repeatedly applying the iterations: $(\mathbf{Q}, \mathbf{R}) \leftarrow$ QR-FACTORIZE$(\mathbf{AQ})$, where QR-FACTORIZE$(\cdot)$ denotes the QR factorization of a matrix, i.e., $\mathbf{Q} \in \mathbb{R}^{n \times k}$ has orthonormal columns. Every iteration of this method 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 a low-rank factorization, the runtime to approximate the projection onto $\mathcal{P}_{n,k}$ using the orthogonal iteration method 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 the QR-factorization step is required to ensure that $\mathbf{Q}_{t+1}\mathbf{Q}_{t+1}^\top$ is also a projection matrix.

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 requires a single QR factorization per iteration, 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 a low-rank factorization), since all other operations require only $O(k^2n)$ time (e.g., factorizing 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 $\mathbf{A} \succeq 0$ denotes that $\mathbf{A}$ is a positive semidefinite matrix (PSD), 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:

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

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. [2] 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 will 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):

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

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$(\mathbf{X}^*)$.*

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.*

[2] 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).

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 runtime 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 $\mathbf{X}^*$ to Problem* (4) *with some parameter $\delta>0$. Let $G\geq \sup_{\mathbf{X}\in\mathcal{F}_{n,k}}\|\nabla f(\mathbf{X})\|$. Consider the sequence $\{\mathbf{Q}_t\}_{t\geq 1}$ generated by Dynamics* (3) *with a 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 $\mathbf{X}^*$ to 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* $\mathbf{X}_t$ *and* $\nabla f(\mathbf{X}_t)$, $\mathbf{X}_{t+1}$ *can be computed via a rank-*k *SVD,*

2. $f(\mathbf{X}_t) - f(\mathbf{X}^*) \leq (f(\mathbf{X}_1) - f(\mathbf{X}^*)) \cdot \exp(-\Theta(\delta/\beta)(t-1)))$.

Theorem 5. *[Local linear convergence of Frank-Wolfe] Suppose Assumption 1 holds true for some optimal solution* $\mathbf{X}^*$ *to 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 between higher eigenvalues, we can still guarantee that $\mathbf{X}^*$ (an optimal solution to Problem (4)) has low rank, and that at least the projected gradient method (when applied to Problem (4)), 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).

Theorem 6. *Let* $\mathbf{X}^* \in \mathcal{F}_{n,k}$ *be some optimal solution to Problem* (4) *and let* $\mu_1 \geq \mu_2 \geq ...\mu_n$ *denote the eigenvalues of* $-\nabla f(\mathbf{X}^*)$. *Let* r *be the smallest integer such that* $r \geq k$ *and* $\mu_r - \mu_{r+1} > 0$. *Then, it holds that rank*$(\mathbf{X}^*) \leq r$. *Moreover, consider the projected gradient dynamics w.r.t. Problem* (4) *given by,* $\mathbf{X}_{t+1} \leftarrow \Pi_{\mathcal{F}_{n,k}}[\mathbf{X}_t - \beta^{-1}\nabla f(\mathbf{X}_t)]$. *For any* $r' \in \{r, \ldots, n-1\}$, *if* $\|\mathbf{X}_1 - \mathbf{X}^*\|_F \leq \frac{\mu_k - \mu_{r'+1}}{4\beta}$, *then it holds that,*

1. $\forall t \geq 1$, *rank*$(\mathbf{X}_{t+1}) \leq r'$, *i.e., given* $\mathbf{X}_t$ *and* $\nabla f(\mathbf{X}_t)$, $\mathbf{X}_{t+1}$ *can be computed via a rank-*r' *SVD.*

2. $\{\mathbf{X}_t\}_{t\geq 1}$ *converges with the standard PGD rate:* $f(\mathbf{X}_t) - f(\mathbf{X}^*) = O(\beta\|\mathbf{X}_1 - \mathbf{X}^*\|_F^2/t)$.

Remark 1. *Note that via the parameter* r', *Theorem 6 offers a flexible tradeoff between the radius of the ball in which PGD needs to be initialized in (increasing* r' *increases the radius), and the rank of the iterates which in turn, implies an upper-bound on the rank of SVD computations required for the projection, which controls the runtime of each iteration.*

Remark 2. *Theorem 6 may be in particular interesting when* $f(\cdot)$ *is subspace-monotone in the sense that for any two subspaces* $\mathcal{S}_1 \subseteq \mathcal{S}_2 \subseteq \mathbb{R}^n$ *and their corresponding projection matrices* $\mathbf{P}_1, \mathbf{P}_2 \in \mathbb{S}^n$, *it holds that* $f(\mathbf{P}_2) \leq f(\mathbf{P}_1)$. *In this case, given an optimal solution* $\mathbf{X}^*$ *to the convex Problem* (4) *with eigen-decomposition* $\mathbf{X}^* = \sum_{i=1}^r \lambda_i \mathbf{u}_i\mathbf{u}_i^\top$, *when* $k < r << n$, *using a projection matrix* $\mathbf{P}^* = \sum_{i=1}^r \mathbf{u}_i\mathbf{u}_i^\top$ *which satisfies* $f(\mathbf{P}^*) \leq \min_{\mathbf{X}\in\mathcal{P}_{n,k}} f(\mathbf{X})$ *may be of interest. For instance, it is not hard to show that* $f(\cdot)$ *of the form* $f(\mathbf{X}) = \sum_{i=1}^m g_i(\|\mathbf{q}_i - \mathbf{X}\mathbf{q}_i\|)$, *where* $g_i(\cdot)$ *is monotone non-decreasing and* $\{\mathbf{q}_i\}_{i=1}^m \subset \mathbb{R}^n$, *is subspace-monotone.*

The complete proofs of Theorems 3, 4, 5, 6, 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* $\mathbf{X} \in \mathbb{S}^n$ *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* $rank(\Pi_{\mathcal{F}_{n,k}}(\mathbf{X})) \leq r$ *if and only if* $\sum_{i=1}^r \min(\gamma_i - \gamma_{r+1}, 1) \geq k$.

Remark 3. *Lemma 1 implies that if rank*$(\mathbf{X}) \leq r$, *then only the top* r *components in 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 $\mathbf{X} \in \mathcal{F}_{n,k}$ 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 establishes that under Assumption 1, Problem (4) has a quadratic growth property. This property is 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 $\mathbf{X}_{t+1} = \Pi_{\mathcal{F}_{n,k}}[\mathbf{Y}_t - \eta\nabla f(\mathbf{Y}_t)]$ for all $t \geq 1$.

At a 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 $\mathbf{X}^*$, 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 $\mathbf{W}_t$ (see also discussion in the Introduction). 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 $t = 1, 2...$ **do**
  $\mathbf{W}_{t+1} \leftarrow \mathbf{Y}_t - \eta\nabla f(\mathbf{Y}_t)$
  $(\mathbf{Q}_{t+1}, \mathbf{R}_{t+1}) \leftarrow$ QR-FACTORIZE$(\mathbf{W}_{t+1}\mathbf{Q}_t)$ (that is $\mathbf{Q}_{t+1}\mathbf{R}_{t+1} = \mathbf{W}_{t+1}\mathbf{Q}_t$)
  $\mathbf{Y}_{t+1} \leftarrow \mathbf{Q}_{t+1}\mathbf{Q}_{t+1}^\top$
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 iteration $t \geq 1$. 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

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

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.

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 $\mathbf{Q}_{t+1}\mathbf{R}_{t+1} = \mathbf{V}\Lambda\mathbf{V}^\top \mathbf{Q}_t$. Multiplying both sides from the left 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 from (6) we have that $\sigma_{\min}(\mathbf{V}_1^\top \mathbf{Q}_t) > 0$, it follows that $\mathrm{rank}(\Lambda_1 \mathbf{V}_1^\top \mathbf{Q}_t) = k$ and thus, from 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 both sides of Equation (8) from the right with $\mathbf{R}_{t+1}^{-1}$, we get

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

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

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

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,

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

Now, for the final part of the proof, we note that $\|\mathbf{V}_2^\top \mathbf{Q}_{t+1}\|_F^2 = \mathrm{Tr}(\mathbf{V}_2\mathbf{V}_2^\top \mathbf{Y}_{t+1}) =$ $\mathrm{Tr}((\mathbf{I} - \mathbf{X}_{t+1})\mathbf{Y}_{t+1}) = k - \mathrm{Tr}(\mathbf{X}_{t+1}\mathbf{Y}_{t+1}) = \frac{1}{2}\|\mathbf{X}_{t+1} - \mathbf{Y}_{t+1}\|_F^2$, 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 $t \geq 1$. Fix some iteration t of Algorithm 1, and suppose that $\eta \leq \frac{1}{5\max\{\beta, G\}}$, $\mathbf{X}_{t+1} \in \mathcal{P}_{n,k}$, and that $\|\mathbf{X}_{t+1} - \mathbf{Y}_t\|_F \leq 1$. Denote the constants $C_0 = 2\Big(\frac{\eta G}{1-\eta G}\Big)^2$, $C_1 = \frac{2(1+\eta G)C_0}{1-2\eta\beta-2C_0(1+\eta G)}$. It holds that, $h_{t+1} \leq \Big(1 - \frac{\eta\delta}{4(1+C_1)}\Big)h_t$, where $\delta > 0$ is the constant from Assumption 1.*

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 $\mathbf{X}_{t+1} = \Pi_{\mathcal{F}_{n,k}}[\mathbf{Y}_t - \eta\nabla f(\mathbf{Y}_t)] = \arg\min_{\mathbf{X}\in\mathcal{F}_{n,k}} \|\mathbf{X} - (\mathbf{Y}_t - \eta\nabla f(\mathbf{Y}_t))\|_F^2$, and by the assumption of the lemma that $\mathbf{X}_{t+1} \in \mathcal{P}_{n,k}$, using the first-order optimality condition, it can be shown that for all $\mathbf{Z} \in \mathcal{F}_{n,k}$: $\langle \mathbf{X}_{t+1} - \mathbf{Z}, \mathbf{Y}_t - \eta_t\nabla f(\mathbf{Y}_t)\rangle \geq 0$, see Lemma 6. This implies that for all $\mathbf{Z} \in \mathcal{F}_{n,k}$:

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

where the last inequality is due to Lemma 8, which uses again the facts that $\mathbf{X}_{t+1} \in \mathcal{P}_{n,k}$ and $\mathbf{X}_{t+1} = \arg\min_{\mathbf{X}\in\mathcal{F}_{n,k}} \|\mathbf{X} - (\mathbf{Y}_t - \eta\nabla f(\mathbf{Y}_t))\|_F^2$, which in turn imply that $\mathbf{X}_{t+1} = \arg\max_{\mathbf{X}\in\mathcal{P}_{n,k}} \langle \mathbf{X}, \mathbf{W}_{t+1}\rangle$, 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,

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

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

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

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

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

We now continue to upper-bound the term $\|\mathbf{X}_{t+1} - \mathbf{Y}_{t+1}\|_F^2$. Using Lemma 9, which apply standard arguments in the analysis of first-order methods, that rely only on the facts that $\mathbf{X}_{t+1} = \Pi_{\mathcal{F}_{n,k}}[\mathbf{Y}_t - \eta\nabla f(\mathbf{Y}_t)]$ and that $f(\cdot)$ is smooth and convex, we have that

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

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

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

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 $f(\mathbf{X}_{t+1}) \geq f(\mathbf{Y}_{t+1}) - \frac{1}{\eta}\Big(\frac{1}{2} + C_0(1+\eta G)\Big)\|\mathbf{X}_{t+1} - \mathbf{Y}_t\|_F^2$. Plugging into (16) we obtain

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

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 $h_{t+1} \leq \Big(1 - \frac{\eta\delta}{4}\Big)h_t + \frac{2(1+\eta G)C_0}{1-2\eta\beta-2C_0(1+\eta G)}(h_t - h_{t+1})$. Denoting $C_1 = \frac{2(1+\eta G)C_0}{1-2\eta\beta-2C_0(1+\eta G)}$, we finally obtain $h_{t+1} \leq \frac{1}{1+C_1}\Big(1 - \frac{\eta\delta}{4} + C_1\Big)h_t = \Big(1 - \frac{\eta\delta}{4(1+C_1)}\Big)h_t$, 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 $\eta \leq \frac{1}{5\max\{\beta, G\}}$. □

4 Numerical Simulations

We turn to discuss our numerical simulations. Due to lack of space, some of the implementation details and results are deferred to the appendix. We consider two models for robust recovery of a low-dimensional subspace 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 in $\nabla f(\mathbf{X}^*)$ (as in Assumption 1), 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
↓ Model 1: spiked covariance ↓						
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
↓ Model 2: sparsely corrupted entries ↓						
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 the spiked covariance and 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 (GOI, Dynamics (3)) on the two models. We initialize both methods

with the PCA solution $\mathbf{X}_{PCA}$ and use the same fixed step-size for both. We examine the convergence of both methods in terms of recovery error and approximation error (w.r.t. the objective function). Additionally, to showcase the benefit of avoiding exact SVD computations (as employed by nonconvex PGD) and using only a single QR factorization per iteration (as in GOI), we compare the runtimes of GOI and nonconvex PGD, but we exclude the time it takes to compute the gradient on each iteration and only account for the time it takes to perform either a rank $-$ k SVD or a QR factorization, where both algorithms were implemented in Python and we have used the built-in functions NUMPY.LINALG.EIGH and NUMPY.LINALG.QR to compute thin-SVDs and QR factorizations, respectively. Finally, we verify during the run of nonconvex PGD, that on each iteration, the projection onto $\mathcal{P}_{n,k}$ is indeed the same as the projection onto the Fantope $\mathcal{F}_{n,k}$ (see Remark 3), which suggests that nonconvex PGD indeed converges to the global minimum.

The results for the spiked covariance model are given in Figure 1 (the results for the sparsely corrupted entries model are very similar and given in the appendix). It can be seen that indeed the distance between the iterates of the two methods decays very quickly and so the graphs of the recovery and approximation errors of both methods coincide. We see that both methods demonstrate a linear convergence rate (w.r.t. the objective). We also see the significant savings in runtime when replacing a thin-SVD computation (in nonconvex PGD) with only a single QR factorization (in GOI).

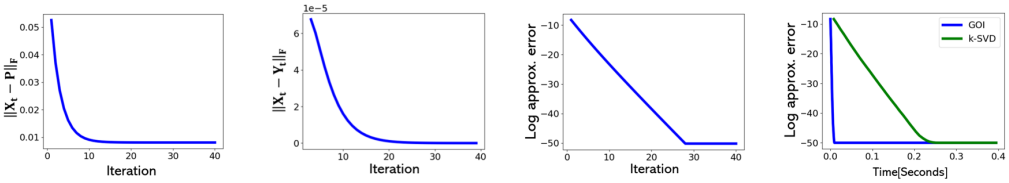

Figure 1: Convergence of PGD and GOI for the spiked covariance model with $p = 0.1$. 1st and 3rd panels from the left show the recovery error ($\mathbf{P}$ is the ground truth projection matrix) and approximation error w.r.t. objective value of PGD, respectively. Convergence of GOI is omitted since it coincides with that of PGD. 2nd panel from the left shows the distance (in Frobenius norm) between the iterates of PGD ($\mathbf{X}_t$) and those of GOI ($\mathbf{Y}_t$). The rightmost panel shows the approximation error (in log scale) vs. time, when only the time to compute matrix factorizations is taken into account.

Importance of warm-start initalization: We examine the performance of nonconvex PGD over $\mathcal{P}_{n,k}$ (Dynamics (2)) for the spiked covariance model considered above, but this time, when initialized with a random (uniformly distributed) projection matrix. We compare it with convex PGD which optimizes over the Fantope $\mathcal{F}_{n,k}$ and uses a full-rank SVD to compute the projection. We use the same step-size as before. We see in Figure 2 (right panel) two trends. First, we clearly see that PGD w.r.t. $\mathcal{P}_{n,k}$ and $\mathcal{F}_{n,k}$ produce very different iterates which in particular implies that, as opposed to the case of warm-start initializaion, the projections of convex PGD onto the Fantope, throughout most of the run are not rank-k. Second, we see that nonconvex PGD is significantly slower than convex PGD. Thus, while both methods eventually converge to the same error, this suggests that far from a global minimizer, the behaviour of nonconvex gradient methods is indeed significantly different than in the local proximity of one, which supports the fact that our theoretical guarantees only hold in a local neighbourhood of a minimizer.

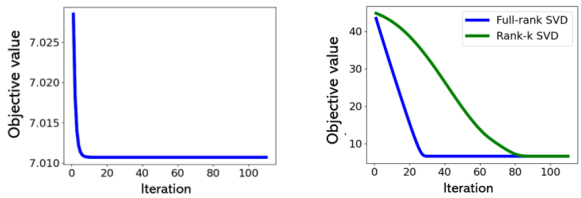

Figure 2: Convergence of PGD for the spiked covariance model with $p = 0.1$ over the Fantope $\mathcal{F}_{n,k}$ with a full-rank SVD, and over $\mathcal{P}_{n,k}$ with rank-k SVD, when initialized with the PCA solution (left panel) and with random initialization (right panel). In the left panel the plots exactly coincide.

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]

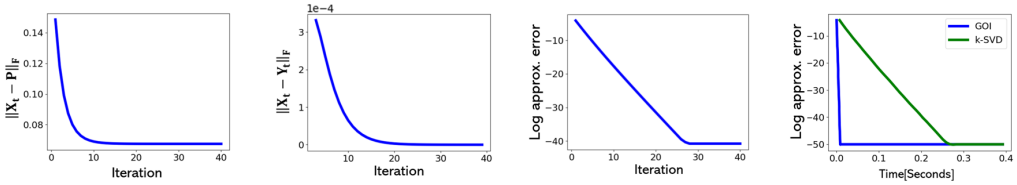

Figure 3: Convergence of PGD and GOI for the sparsely corrupted entries model with $p = 0.1$. 1st and 3rd panels from the left show the recovery error ($\mathbf{P}$ is the ground truth projection matrix) and approximation error w.r.t. objective value of PGD, respectively. Convergence of GOI is omitted since it coincides with that of PGD. 2nd panel from the left shows the distance (in Frobenius norm) between the iterates of PGD ($\mathbf{X}_t$) and those of GOI ($\mathbf{Y}_t$). The rightmost panel shows the approximation error (in log scale) vs. time, when only the time to compute matrix factorizations is taken into account.

A Additional Details on Experiments

The first robust recovery model we consider is a spiked covariance model, in which we draw a uniformly distributed projection matrix onto a k-dimensional subspace $\mathbf{P} \in \mathcal{P}_{n,k}$, and we generate m samples $\mathbf{q}_1, \ldots, \mathbf{q}_m \in \mathbb{R}^n$ such that for each $i \in [m]$, we set $\mathbf{q}_i = \mathbf{P}\mathbf{z}_i / \|\mathbf{P}\mathbf{z}_i\|$ with probability $1-p$, and $\mathbf{q}_i = \mathbf{z}_i$ with probability p, where $p \in (0, 0.5]$, and $\mathbf{z}_1, \ldots, \mathbf{z}_m$ are i.i.d. uniformly distributed unit vectors. The goal is to recover $\mathbf{P}$ by minimizing the following objective function over $\mathcal{F}_{n,k}$:

$$f(\mathbf{X}) = \sum_{i=1}^{m} \text{Huber}_\gamma(\|\mathbf{q}_i - a\mathbf{X}\mathbf{q}_i\|), \quad \text{Huber}_\gamma(x) := \begin{cases} \frac{1}{2}x^2 & \text{if } |x| \le \gamma \\ \gamma(|x| - \frac{1}{2}\gamma) & \text{else.} \end{cases}$$

Here $a \in (0, 1]$ is a regularization parameter and we set it to slightly less than one.

The second model we consider is that of sparsely corrupted entries in which we again draw a uniformly distributed projection matrix $\mathbf{P}$. This time the data points $\mathbf{q}_1, \ldots, \mathbf{q}_m$ are generated by taking $\mathbf{q}_i = \mathbf{P}\mathbf{z}_i / \|\mathbf{P}\mathbf{z}_i\|$ for each $i \in [m]$, where as before $\mathbf{z}_1, \ldots, \mathbf{z}_m$ are i.i.d. uniformly distributed unit vectors, but for each $i \in [m]$, with probability p, we pick a uniformly distributed entry $j \in [n]$ and set it to -1 or $+1$ (with equal probability). The goal is to recover $\mathbf{P}$ by minimizing the following objective function over $\mathcal{F}_{n,k}$:

$$f(\mathbf{X}) = \sum_{i=1}^{m} \sum_{j=1}^{n} \text{Huber}_\gamma([\mathbf{q}_i]_j - [a\mathbf{X}\mathbf{q}_i]_j),$$

where here also $a \in (0, 1]$ is a regularization parameter.

For both models we set the Huber loss parameter to $\gamma = 0.1$. For the first model we set $a = 0.9$ and for the second $a = 0.8$. For a given projection matrix $\mathbf{X} \in \mathcal{P}_{n,k}$, we measure the recovery error according to $\|\mathbf{X} - \mathbf{P}\|_F^2$. For both models we let $\mathbf{X}_{PCA} \in \mathcal{P}_{n,k}$ denote the standard PCA solution, i.e., the projection matrix onto the span of the top k eigenvectors of the empirical covariance $\frac{1}{m}\sum_{i=1}^{m} \mathbf{q}_i \mathbf{q}_i^\top$. For both models we set $n = 100$, $k = 10$, and $m = 500$. For both models we use the projected gradient method to find a projection matrix $\mathbf{X}^* \in \mathcal{P}_{n,k}$ which has negligible dual gap ($< 10^{-10}$). [3] For this $\mathbf{X}^*$ we measure the corresponding eigen-gap $\lambda_{n-k}(\nabla f(\mathbf{X}^*)) - \lambda_{n-k+1}(\nabla f(\mathbf{X}^*))$, and the recovery error $\|\mathbf{X}^* - \mathbf{P}\|_F$. The results are given in Table 1. For each set of parameters the results are the average of 20 i.i.d. experiments.

In a second experiment we fix for both models $p = 0.1$ and vary the dimension n (while keeping k, m fixed as before). The results are given in Table 2. In particular, we see that the eigen-gap δ does not change substantially with the dimension.

We turn to demonstrate the empirical performance of the projected gradient method w.r.t. to the nonconvex set $\mathcal{P}_{n,k}$ (PGD), as given in Dynamics (2), and and gradient orthogonal iteration (GOI), as given in Dynamics (3), for the two models discussed above. We fix $n = 100$ and $p = 0.1$ (keeping all other parameters unchanged). For both methods we use the fixed step-size $\eta = 1/\lambda$, where

[3] For $\mathbf{X} \in \mathcal{F}_{n,k}$ the dual gap is defined as $\text{dg}(\mathbf{X}) = \langle \mathbf{X} - \mathbf{V}, \nabla f(\mathbf{X}) \rangle$, were $\mathbf{V} \in \arg\min_{\mathbf{Z} \in \mathcal{P}_{n,k}} \langle \mathbf{Z}, \nabla f(\mathbf{X}) \rangle$. Since $f(\cdot)$ is convex, we in particular have $f(\mathbf{X}) - \min_{\mathbf{Y} \in \mathcal{F}_{n,k}} f(\mathbf{Y}) \le \text{dg}(\mathbf{X})$.

dim. (n)	100	200	300	400
$\downarrow$ Model 1: spiked covariance $\downarrow$				
Eigen-gap (δ)	2.87	3.02	2.96	3.04
$\|\mathbf{X}^* - \mathbf{P}\|_F$	0.0071	0.005	0.0043	0.0035
$\|\mathbf{X}_{PCA} - \mathbf{P}\|_F$	0.068	0.049	0.043	0.036
$\downarrow$ Model 2: sparsely corrupted entries $\downarrow$				
Eigen-gap (δ)	5.49	5.902	6.06	6.1
$\|\mathbf{X}^* - \mathbf{P}\|_F$	0.067	0.0617	0.058	0.055
$\|\mathbf{X}_{PCA} - \mathbf{P}\|_F$	0.199	0.208	0.206	0.202

Table 2: Recovery and eigen-gap results for the spiked covariance model and corrupted entries model with varying dimension. Each result is the average of 20 i.i.d. experiments.

$\lambda = \lambda_1(\sum_{i=1}^m \mathbf{q}_i\mathbf{q}_i^\top)$, i.e., the largest eigenvalue of the (unnormalized) empirical covariance. We note that smaller values of η seem too conservative in practice from our experimentations. We initialize both methods with the k-PCA projection matrix $\mathbf{X}_{PCA}$. We examine the convergence of both methods in terms of recovery error and approximation error (w.r.t. the objective function). Additionally, to showcase the benefit of avoiding exact SVD computations (as employed by nonconvex PGD) and using only a single QR factorization per iteration (as in GOI), we compare the runtimes of GOI and nonconvex PGD, but we exclude the time it takes to compute the gradient on each iteration and only account for the time it takes to perform either a rank $-$ k SVD or a QR factorization, where both algorithms were implemented in Python and we have used the built-in functions NUMPY.LINALG.EIGH and NUMPY.LINALG.QR to compute thin-SVDs and QR factorizations, respectively.

The results for the spiked covariance model are given in Figure 1, and the results for the sparsely corrupted entries model, which are very similar, are given in Figure 3. It can be seen that indeed the distance between the iterates of the two methods decays very quickly and so the graphs of the recovery and approximation errors of both methods coincide. We in particular see that both methods indeed demonstrate a linear convergence rate (w.r.t. the objective value). We also see the significant savings in runtime when replacing a thin-SVD computation (as used by nonconvex PGD) with only a single QR factorization (as used by GOI).

Moreover, in order to verify the convergence of nonconvex PGD to the global optimal solution (and not just a stationary point of the nonconvex Problem (1)), we verify using the procedure suggested in Remark 3, that on each iteration t, the projection step onto the Fantope $\mathcal{F}_{n,k}$ is also of rank k, i.e., identical to the projection onto $\mathcal{P}_{n,k}$. This means that the iterates of PGD w.r.t. the nonconvex Problem (1) and the iterates of PGD w.r.t. the convex relaxation (4), coincide. Indeed, for all random instances generated and for all iterations executed, we observe that the projection onto the Fantope is of rank k. This suggests that the nonconvex PGD (and consequently also GOI) in particular converges to the global optimal solution of the convex relaxation (4).

B Proof of Theorem 2

Proof. First, observe that for any dual solution $(\mathbf{Z}_1^*, \mathbf{Z}_2^*, s^*)$, it holds that $\mathbf{Z}_1^*$ and $\mathbf{Z}_2^*$ are orthogonal to each other. This is true since, denoting by $\mathbf{X}^*$ the corresponding primal solution, we have that,

$$\begin{aligned}\langle \mathbf{Z}_1^*, \mathbf{Z}_2^* \rangle &= \text{Tr}(\mathbf{Z}_1^*\mathbf{Z}_2^*) = \text{Tr}(\mathbf{Z}_1^*(\mathbf{X}^* + (\mathbf{I} - \mathbf{X}^*))\mathbf{Z}_2^*) \\ &= \text{Tr}(\mathbf{Z}_1^*\mathbf{X}^*\mathbf{Z}_2^*) + \text{Tr}(\mathbf{Z}_1^*(\mathbf{I} - \mathbf{X}^*)\mathbf{Z}_2^*) = 0,\end{aligned}$$

where the last equality follows from the complementarity conditions $\mathbf{Z}_1^*\mathbf{X}^* = \mathbf{0}$ and $(\mathbf{I} - \mathbf{X}^*)\mathbf{Z}_2^* = \mathbf{0}$.

For a given dual solution $(\mathbf{Z}_1^*, \mathbf{Z}_2^*, s^*)$, let us denote $r_1 = \text{rank}(\mathbf{Z}_1^*)$ and $r_2 = \text{rank}(\mathbf{Z}_2^*)$.

Let us write the eigen decompositions $\mathbf{Z}_1^* = \sum_{i=1}^{r_1} \rho_i \mathbf{u}_i\mathbf{u}_i^T$ and $\mathbf{Z}_2^* = \sum_{j=1}^{r_2} \mu_j \mathbf{v}_j\mathbf{v}_j^T$.

From the orthogonality of $\mathbf{Z}_1^*$ and $\mathbf{Z}_2^*$ established above, we get an orthonormal set of vectors $\{\mathbf{u}_1, ..., \mathbf{u}_{r_1}, \mathbf{v}_1, ..., \mathbf{v}_{r_2}\}$ and we can complete it to an orthonormal basis of $\mathbb{R}^n$:

$$B = \{\mathbf{u}_1, ..., \mathbf{u}_{r_1}, \mathbf{v}_1, ..., \mathbf{v}_{r_2}, \mathbf{w}_1, ..., \mathbf{w}_{n-r_1-r_2}\},$$

where $\mathbf{Z}_1^*\mathbf{w}_i = \mathbf{0}$ and $\mathbf{Z}_2^*\mathbf{w}_i = \mathbf{0}$ for any $i \in \{1, \ldots, n - r_1 - r_2\}$.

From the KKT conditions for Problem (4), we have that $\nabla f(\mathbf{X}^*) = \mathbf{Z}_1^* - \mathbf{Z}_2^* + s^*\mathbf{I}$, and so it follows that any $\mathbf{v} \in B$ is an eigenvector of $\nabla f(\mathbf{X}^*)$. Thus, we can write the eigenvalues of $\nabla f(\mathbf{X}^*)$ in non-increasing order from left to right as:

$$\rho_1 + s^*, ..., \rho_{r_1} + s^*, \underbrace{s^*, ..., s^*}_{n-r_1-r_2 \text{ times}}, s^* - \mu_{r_2}, ..., s^* - \mu_1. \tag{17}$$

For the first direction of the theorem, let us assume $\mathbf{X}^*$ satisfies strict complementarity, so for some dual solution $(\mathbf{Z}_1^*, \mathbf{Z}_2^*, s^*)$ we have that $r_1 = n - k$ or $r_2 = k$.

Now, if $r_1 = n - k$, using (17) we get that $\lambda_{n-k}(\nabla f(\mathbf{X}^*)) = \rho_{n-k} + s^*$ and $\lambda_{n-k+1}(\nabla f(\mathbf{X}^*)) \leq s^*$, and so there is a gap of

$$\lambda_{n-k}(\nabla f(\mathbf{X}^*)) - \lambda_{n-k+1}(\nabla f(\mathbf{X}^*)) \geq \rho_{n-k} + s^* - s^* = \rho_{n-k} > 0.$$

Otherwise, if $r_2 = k$, then using (17) we have that $\lambda_{n-k+1}(\nabla f(\mathbf{X}^*)) = s^* - \mu_k$ and $\lambda_{n-k}(\nabla f(\mathbf{X}^*)) > s^*$, and so there is a gap of

$$\lambda_{n-k}(\nabla f(\mathbf{X}^*)) - \lambda_{n-k+1}(\nabla f(\mathbf{X}^*)) \geq s^* - (s^* - \mu_k) = \mu_k > 0.$$

In both cases we get a positive eigen-gap, which proves the first direction of the theorem.

For the reversed direction, let us assume that $\mathbf{X}^*$ satisfies the eigen-gap assumption, and recall that according to Theorem 1 it follows that $\text{rank}(\mathbf{X}^*) = k$. Suppose by way of contradiction that there exists a dual solution $(\mathbf{Z}_1^*, \mathbf{Z}_2^*, s^*)$ for which $r_1 < n - k$ and $r_2 < k$. In this case we have from (17) that,

$$\lambda_{n-k}(\nabla f(\mathbf{X}^*)) = \lambda_{n-k+1}(\nabla f(\mathbf{X}^*)) = s^*,$$

which contradicts the existence of an eigen-gap and so, it must be that $r_1 = n - k$ or $r_2 = k$. □

C Details Missing from Section 3.1 and Proof of Theorem 1

C.1 Proof of Lemma 1

Proof. The first part of the lemma is a known fact, see for instance [26]. For the second part, let us prove that if $\sum_{i=1}^{r} \min(\gamma_i - \gamma_{r+1}, 1) \geq k$, then θ must satisfy $\theta \geq \gamma_{r+1}$. Assume by way contradiction that $\theta < \gamma_{r+1}$. Then,

$$k = \sum_{i=1}^{n} \min(\max(\gamma_i - \theta, 0), 1) > \sum_{i=1}^{n} \min(\max(\gamma_i - \gamma_{r+1}, 0), 1) = \sum_{i=1}^{r} \min(\gamma_i - \gamma_{r+1}, 1),$$

which is a contradiction, and so it must be that $\theta \geq \gamma_{r+1}$, and in that case the projection sets all the bottom $n - r$ components of the eigen-decomposition of $\mathbf{X}$ to zero. Hence, $\text{rank}(\Pi_{\mathcal{F}_{n,k}}[\mathbf{X}]) \leq r$. The reversed direction holds from similar reasoning. □

C.2 Proof of Theorem 1

Before we prove Theorem 1 we need the following lemma which is central to our analysis and connects between an optimal solution and the eigen-decomposition of its corresponding gradient.

Lemma 6. *Let $\mathbf{X}^* \in \mathcal{F}_{n,k}$ be an optimal solution to Problem* (4) *and write the eigen-decomposition of $-\nabla f(\mathbf{X}^*)$ as $-\nabla f(\mathbf{X}^*) = \sum_{i=1}^{n} \mu_i \mathbf{u}_i \mathbf{u}_i^\top$. Let r be the smallest integer such that $r \geq k$ and $\mu_k - \mu_{r+1} > 0$. Then, for all $n \geq i \geq r + 1$, $\mathbf{X}^*$ is orthogonal to $\mathbf{u}_i \mathbf{u}_i^\top$, and rank$(\mathbf{X}^*) \leq r$.*
In particular, if $r = k$, then $\mathbf{X}^* \in \mathcal{P}_{n,k}$ is the unique projection matrix onto the span of the k leading eigenvectors of $-\nabla f(\mathbf{X}^*)$.

Proof. Assume by way of contradiction that $\mathbf{X}^*$ is not orthogonal $\mathbf{u}_{r+1}\mathbf{u}_{r+1}^\top, \ldots, \mathbf{u}_n\mathbf{u}_n^\top$. In this case, $\sum_{i=r+1}^n \mathbf{u}_i^\top \mathbf{X}^* \mathbf{u}_i > 0$, and we can write,

$$\begin{aligned}
\langle \mathbf{X}^*, -\nabla f(\mathbf{X}^*) \rangle &= \sum_{i=1}^r \mu_i \mathbf{u}_i^\top \mathbf{X}^* \mathbf{u}_i + \sum_{i=r+1}^n \mu_i \mathbf{u}_i^\top \mathbf{X}^* \mathbf{u}_i \\
&\underset{(a)}{<} \sum_{i=1}^r \mu_i \mathbf{u}_i^\top \mathbf{X}^* \mathbf{u}_i + \mu_r \sum_{i=r+1}^n \mathbf{u}_i^\top \mathbf{X}^* \mathbf{u}_i \\
&\underset{(b)}{=} \sum_{i=1}^{k-1} \mu_i \mathbf{u}_i^\top \mathbf{X}^* \mathbf{u}_i + \mu_k \sum_{i=k}^n \mathbf{u}_i^\top \mathbf{X}^* \mathbf{u}_i \\
&\underset{(c)}{=} \sum_{i=1}^{k-1} \mu_i \mathbf{u}_i^\top \mathbf{X}^* \mathbf{u}_i + \mu_k \left(k - \sum_{i=1}^{k-1} \mathbf{u}_i^\top \mathbf{X}^* \mathbf{u}_i \right),
\end{aligned}$$

where both (a) and (b) follow from the definition of r, and (c) follows since $\sum_{i=1}^n \mathbf{u}_i^\top \mathbf{X}^* \mathbf{u}_i = \text{Tr}(\mathbf{X}^* \sum_{i=1}^n \mathbf{u}_i \mathbf{u}_i^*) = \text{Tr}(\mathbf{X}^* \mathbf{I}) = k$.

Let us denote the projection matrix onto the span of the top k eigenvectors of $-\nabla f(\mathbf{X}^*)$ by $\mathbf{P}^* = \sum_{i=1}^k \mathbf{u}_i \mathbf{u}_i^\top$, and note that $\langle \mathbf{P}^*, -\nabla f(\mathbf{X}^*) \rangle = \sum_{i=1}^k \mu_i$. It follows that

$$\begin{aligned}
\langle \mathbf{P}^* - \mathbf{X}^*, \nabla f(\mathbf{X}^*) \rangle &= \langle \mathbf{X}^* - \mathbf{P}^*, -\nabla f(\mathbf{X}^*) \rangle \\
&< \sum_{i=1}^{k-1} \mu_i \mathbf{u}_i^\top \mathbf{X}^* \mathbf{u}_i + \mu_k \left(k - \sum_{i=1}^{k-1} \mathbf{u}_i^\top \mathbf{X}^* \mathbf{u}_i \right) - \sum_{i=1}^k \mu_i \\
&= \sum_{i=1}^{k-1} \mu_i \left(\mathbf{u}_i^\top \mathbf{X}^* \mathbf{u}_i - 1 \right) + \mu_k \sum_{i=1}^{k-1} \left(1 - \mathbf{u}_i^\top \mathbf{X}^* \mathbf{u}_i \right) \\
&= \sum_{i=1}^{k-1} \left(1 - \mathbf{u}_i^\top \mathbf{X}^* \mathbf{u}_i \right) (\mu_k - \mu_i) \leq 0,
\end{aligned}$$

where the last inequality follows since for all i, $\mathbf{u}_i^\top \mathbf{X}^* \mathbf{u}_i \in [0, 1]$.

Thus, we have that $\mathbf{X}^*$ violates the first-order optimality condition which contradicts that assumption that it is an optimal solution, and thus we have that $\mathbf{X}^*$ must indeed be orthogonal to $\mathbf{u}_{r+1}\mathbf{u}_{r+1}^\top, \ldots, \mathbf{u}_n\mathbf{u}_n^\top$.

An immediate consequence is that the eigenvectors of $\mathbf{X}^*$ which correspond to non-zero eigenvalues must lie in span$\{\mathbf{u}_1, ..., \mathbf{u}_r\}$ and thus, it must be that rank$(\mathbf{X}^*) \leq r$.

For the final part of the lemma, in case $r = k$, since for all $\mathbf{X} \in \mathcal{F}_{n,k}$, rank$(\mathbf{X}) \geq k$, we have that rank$(\mathbf{X}^*) = k$. In particular, $\mathbf{X}^*$ is a projection matrix, i.e., $\mathbf{X} \in \mathcal{P}_{n,k}$. By the orthogonality result above, it follows that the eigenvectors of $\mathbf{X}^*$ lie in span$\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$, which means that $\mathbf{X}^*$ is indeed the projection matrix onto span$\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$, as stated in the lemma. Note that when $r = k$, this projection matrix is indeed unique (i.e., the subspace spanned by the top k eigenvectors of $-\nabla f(\mathbf{X}^*)$ is unique). □

We now prove Theorem 1.

Proof of Theorem 1. Let $\mathbf{X}^*$ be an optimal solution to the convex relaxation (4) which satisfies Assumption 1 with some $\delta > 0$. It follows directly from Lemma 6 that rank$(\mathbf{X}^*) = k$. From Lemma 6 it further follows that $\mathbf{X}^*$ is the unique projection matrix onto the span of top k eigenvectors of $-\nabla f(\mathbf{X}^*)$, i.e., it is the unique matrix in $\mathbf{X} \in \mathcal{P}_{n,k}$ such that $\langle \mathbf{X}, -\nabla f(\mathbf{X}^*) \rangle = \sum_{i=1}^k \mu_i$, where we write the eigen-decomposition of $-\nabla f(\mathbf{X}^*)$ as $-\nabla f(\mathbf{X}^*) = \sum_{i=1}^n \mu_i \mathbf{u}_i \mathbf{u}_i^\top$. From the von Neumann trace inequality it follows that for any matrix $\mathbf{X} \in \mathcal{P}_{n,k}$ it holds that $\langle \mathbf{X}, -\nabla f(\mathbf{X}^*) \rangle \leq \sum_{i=1}^n \lambda_i(\mathbf{X}) \mu_i = \sum_{i=1}^k \lambda_i(\mathbf{X}) \mu_i = \sum_{i=1}^k \mu_i$. Thus, we have that for all $\mathbf{X} \in \mathcal{P}_{n,k} \setminus \{\mathbf{X}^*\}$: $\langle \mathbf{X} - \mathbf{X}^*, \nabla f(\mathbf{X}^*) \rangle = \langle \mathbf{X}^* - \mathbf{X}, -\nabla f(\mathbf{X}^*) \rangle > 0$. Since $\mathcal{F}_{n,k} = \text{conv}\{\mathcal{P}_{n,k}\}$, this further implies that for all $\mathbf{X} \in \mathcal{F}_{n,k} \setminus \{\mathbf{X}^*\}$: $\langle \mathbf{X} - \mathbf{X}^*, \nabla f(\mathbf{X}^*) \rangle > 0$. Since $f(\cdot)$ is convex, it further holds that

for all $\mathbf{X} \in \mathcal{F}_{n,k} \setminus \{\mathbf{X}^*\}$: $f(\mathbf{X}^*) - f(\mathbf{X}) \leq \langle \mathbf{X}^* - \mathbf{X}, \nabla f(\mathbf{X}^*) \rangle < 0$, and thus, we conclude that $\mathbf{X}^*$ is indeed the unique optimal solution to Problem (4), which also implies that it is the unique optimal solution to Problem (1). □

C.3 Proof of Lemma 3

Proof. Let us write the eigen-decomposition of the gradient $\nabla f(\mathbf{X}^*)$ as $\nabla f(\mathbf{X}^*) = \sum_{i=1}^n \lambda_i \mathbf{u}_i \mathbf{u}_i^\top$. For any $\mathbf{X} \in \mathcal{F}_{n,k}$ it holds that:

$$\begin{aligned}
f(\mathbf{X}) - f(\mathbf{X}^*) &\underset{(a)}{\geq} \langle \mathbf{X} - \mathbf{X}^*, \nabla f(\mathbf{X}^*) \rangle \underset{(b)}{=} \sum_{i=1}^n \lambda_i \mathbf{u}_i^\top \mathbf{X} \mathbf{u}_i - \sum_{i=n-k+1}^n \lambda_i \\
&\underset{(c)}{\geq} (\lambda_{n-k+1} + \delta) \sum_{i=1}^{n-k} \mathbf{u}_i^\top \mathbf{X} \mathbf{u}_i + \sum_{i=n-k+1}^n \lambda_i \mathbf{u}_i^\top \mathbf{X} \mathbf{u}_i - \sum_{i=n-k+1}^n \lambda_i \\
&= (\lambda_{n-k+1} + \delta) \sum_{i=1}^{n-k} \mathbf{u}_i^\top \mathbf{X} \mathbf{u}_i - \sum_{i=n-k+1}^n \lambda_i (1 - \mathbf{u}_i^\top \mathbf{X} \mathbf{u}_i) \\
&\underset{(d)}{\geq} (\lambda_{n-k+1} + \delta) \sum_{i=1}^{n-k} \mathbf{u}_i^\top \mathbf{X} \mathbf{u}_i - \lambda_{n-k+1} \sum_{i=n-k+1}^n (1 - \mathbf{u}_i^\top \mathbf{X} \mathbf{u}_i) \\
&= \lambda_{n-k+1} \sum_{i=1}^n \mathbf{u}_i^\top \mathbf{X} \mathbf{u}_i - k\lambda_{n-k+1} + \delta \sum_{i=1}^{n-k} \mathbf{u}_i^\top \mathbf{X} \mathbf{u}_i, \qquad (18)
\end{aligned}$$

where (a) follows from the convexity of $f(\mathbf{X})$, (b) follows since according Lemma 6 $\mathbf{X}^* = \sum_{i=n-k+1}^n \mathbf{u}_i \mathbf{u}_i^\top$ and so, $\langle \mathbf{X}^*, \nabla f(\mathbf{X}^*) \rangle = \sum_{i=n-k+1}^n \lambda_i$, (c) follows from Assumption 1, and (d) follows since $\mathbf{X} \preceq \mathbf{I}$, which implies that $\mathbf{u}_i^\top \mathbf{X} \mathbf{u}_i \leq \mathbf{u}_i^\top \mathbf{u}_i = 1$.

Using $\sum_{i=1}^n \mathbf{u}_i^\top \mathbf{X} \mathbf{u}_i = \text{Tr}(\mathbf{X} \sum_{i=1}^n \mathbf{u}_i \mathbf{u}_i^\top) = \text{Tr}(\mathbf{X}\mathbf{I}) = k$ and Eq. (18), we have,

$$f(\mathbf{X}) - f(\mathbf{X}^*) \geq \delta \sum_{i=1}^{n-k} \mathbf{u}_i^\top \mathbf{X} \mathbf{u}_i = \delta \left(k - \sum_{i=n-k+1}^n \mathbf{u}_i^\top \mathbf{X} \mathbf{u}_i \right). \qquad (19)$$

Also, using again the fact that $\mathbf{X}^* = \sum_{i=n-k+1}^n \mathbf{u}_i \mathbf{u}_i^\top$, we have that,

$$\|\mathbf{X} - \mathbf{X}^*\|_F^2 = \|\mathbf{X}\|_F^2 + \|\mathbf{X}^*\|_F^2 - 2 \sum_{i=n-k+1}^n \mathbf{u}_i^\top \mathbf{X} \mathbf{u}_i \leq 2 \left(k - \sum_{i=n-k+1}^n \mathbf{u}_i^\top \mathbf{X} \mathbf{u}_i \right), \qquad (20)$$

where the last inequality follows since for any $\mathbf{X} \in \mathcal{F}_{n,k}$ it holds that $\|\mathbf{X}\|_F^2 = \sum_{i=1}^n \lambda_i^2(\mathbf{X}) \leq \sum_{i=1}^n \lambda_i(\mathbf{X}) = k$. Combining Eq. (19) and (20) we finally have that,

$$f(\mathbf{X}) - f(\mathbf{X}^*) \geq \frac{\delta}{2} \|\mathbf{X} - \mathbf{X}^*\|_F^2.$$

□

D Projected Gradient Descent Analysis

In this section we turn to analyze the local convergence of the projected gradient method w.r.t. the sets $\mathcal{P}_{n,k}$ and $\mathcal{F}_{n,k}$, and to prove Theorems 4 and 6.

We first provide the proof of Lemma 2 which is fairly simple, and then prove a more general version of the lemma, which in particular allows to relax Assumption 1.

Proof of Lemma 2. Denote $\mathbf{Y}^* = \mathbf{X}^* - \eta \nabla f(\mathbf{X}^*)$ and denote the eigenvalues of $\mathbf{Y}^*$ in non-increasing order $\sigma_i = \lambda_i(\mathbf{Y}^*), i = 1, \ldots, n$. Denote also $\mathbf{Y} = \mathbf{X} - \eta \nabla f(\mathbf{X})$ with its eigenvalues $\gamma_i = \lambda_i(\mathbf{Y}), i = 1, \ldots, n$. Let us write the eigen-decomposition of $-\nabla f(\mathbf{X}^*)$ as

$-\nabla f(\mathbf{X}^*) = \sum_{i=1}^n \mu_i \mathbf{u}_i \mathbf{u}_i^\top$. From Lemma 6, we have that under Assumption 1, it holds that $\mathbf{X}^* = \sum_{i=1}^k \mathbf{u}_i \mathbf{u}_i^\top$. Thus, we can deduce that

$$\sigma_i = \begin{cases} 1 + \eta\mu_i & \text{if } i \in \{1, ..., k\}; \\ \eta\mu_i & \text{else.} \end{cases} \tag{21}$$

From Lemma 1 we have that $\text{rank}(\Pi_{\mathcal{F}_{n,k}}(\mathbf{Y})) = k$ if and only if $\sum_{i=1}^k \min(\gamma_i - \gamma_{k+1}, 1) \geq k$. Thus, a sufficient condition so that $\text{rank}(\Pi_{\mathcal{F}_{n,k}}(\mathbf{Y})) = k$ is,

$$\gamma_k - \gamma_{k+1} \geq 1. \tag{22}$$

By Weyl's inequality for the eigenvalues and Eq. (21) we have,

$$\begin{aligned} \gamma_k - \gamma_{k+1} &= (\sigma_k - \sigma_{k+1}) + (\gamma_k - \sigma_k) + (\sigma_{k+1} - \gamma_{k+1}) \\ &\geq 1 + \eta(\mu_k - \mu_{k+1}) - 2\|\mathbf{Y} - \mathbf{Y}^*\|_F \\ &= 1 + \eta(\mu_k - \mu_{k+1}) - 2\|\mathbf{X} - \mathbf{X}^* - \eta\nabla f(\mathbf{X}) + \eta\nabla f(\mathbf{X}^*)\|_F \\ &\geq 1 + \eta(\mu_k - \mu_{k+1}) - 2(1 + \eta\beta)\|\mathbf{X} - \mathbf{X}^*\|_F. \end{aligned}$$

Thus, we see that a sufficient condition so that (22) holds, is that $\mathbf{X}$ satisfies

$$\|\mathbf{X} - \mathbf{X}^*\|_F \leq \frac{\eta\delta}{2(1 + \eta\beta)} \leq \frac{\eta(\mu_k - \mu_{k+1})}{2(1 + \eta\beta)},$$

and so the lemma follows. □

The following lemma generalizes Lemma 2 and offers a natural trade-off between the rank of the projected gradient mapping and the size of the ball around an optimal solution $\mathbf{X}^*$ in which it is guaranteed to be upper-bounded.

Lemma 7. *Let $\mathbf{X}^* \in \mathcal{F}_{n,k}$ be an optimal solution to Problem* (4)*, and let $\mu_1 \geq \mu_2 \geq \ldots \mu_n$ denote the eigenvalues of $-\nabla f(\mathbf{X}^*)$. Let r be the smallest integer such that $r \geq k$ and $\mu_k > \mu_{r+1}$. Fix some $\eta > 0$. For any $\mathbf{X} \in \mathcal{F}_{n,k}$ which satisfies*

$$\|\mathbf{X} - \mathbf{X}^*\|_F \leq \frac{\eta(\mu_k - \mu_{r+1})}{2(1 + \eta\beta)}, \tag{23}$$

it holds that rank$(\Pi_{\mathcal{F}_k}(\mathbf{X} - \eta\nabla f(\mathbf{X}))) \leq r$.
More generally, for any $r' \in \{r, r+1, ..., n-1\}$ and for any $\eta > 0$, if $\mathbf{X} \in \mathcal{F}_{n,k}$ satisfies

$$\|\mathbf{X} - \mathbf{X}^*\|_F \leq \frac{\eta(\mu_k - \mu_{r'+1})}{2(1 + \eta\beta)}, \tag{24}$$

then rank$(\Pi_{\mathcal{F}_k}(\mathbf{X} - \eta\nabla f(\mathbf{X}))) \leq r'$.

Proof. From Lemma 6 we have that $r^* := \text{rank}(\mathbf{X}^*) \leq r$. Denote $\mathbf{Y}^* = \mathbf{X}^* - \eta\nabla f(\mathbf{X}^*)$ and denote the eigenvalues of $\mathbf{Y}^*$ as $\sigma_i = \lambda_i(\mathbf{Y}^*), i = 1, \ldots, n$. Denote also $\mathbf{Y} = \mathbf{X} - \eta\nabla f(\mathbf{X})$ with its eigenvalues $\gamma_i = \lambda_i(\mathbf{Y}), i = 1, \ldots, n$.
From the min-max principle for the eigenvalues, letting $\mathcal{V} \subseteq \mathbb{R}^n$ denote some subspace of $\mathbb{R}^n$, we have that for any $i \in \{1, ..., r\}$,

$$\sigma_i = \min_{\mathcal{V}:\dim(\mathcal{V}) = n-i+1} \max_{\mathbf{v} \in \mathcal{V}:\|\mathbf{v}\|=1} \mathbf{v}^\top(\mathbf{X}^* + \eta(-\nabla f(\mathbf{X}^*)))\mathbf{v}. \tag{25}$$

Let us write the eigen-decomposition of $-\nabla f(\mathbf{X}^*)$ as $-\nabla f(\mathbf{X}^*) = \sum_{i=1}^n \mu_i \mathbf{u}_i \mathbf{u}_i^\top$. Note that in Eq. (25) we minimize over all the subspaces $\mathcal{V}$ of dimension $n - i + 1$, $i \leq r$, and so,

$$\mathcal{V} \cap \text{span}\{\mathbf{u}_1, ..., \mathbf{u}_r\} \neq \emptyset, \tag{26}$$

otherwise the direct sum $\mathcal{V} \oplus \text{span}\{\mathbf{u}_1, ..., \mathbf{u}_r\} \subseteq \mathbb{R}^n$ would have dimension $n - i + 1 + r > n$.

Any unit vector $\mathbf{v} \in \mathcal{V}$ can be written as $\mathbf{v} = a\mathbf{u} + b\mathbf{w}$ such that $\mathbf{u} \in \text{span}\{\mathbf{u}_1, ..., \mathbf{u}_r\}$, $\|\mathbf{u}\| = 1$, $\mathbf{w} \in \text{span}\{\mathbf{u}_{r+1}, ..., \mathbf{u}_n\}$, $\|\mathbf{w}\| = 1$, and $a^2 + b^2 = 1$. Thus, for any such unit vector $\mathbf{v}$, using Lemma 6, we have that,

$$\mathbf{v}^\top(\mathbf{X}^* + \eta(-\nabla f(\mathbf{X}^*)))\mathbf{v} = a^2\mathbf{u}^\top\mathbf{X}^*\mathbf{u} + a^2\eta\mathbf{u}^\top(-\nabla f(\mathbf{X}^*))\mathbf{u} + b^2\eta\mathbf{w}^\top(-\nabla f(\mathbf{X}^*))\mathbf{w}. \tag{27}$$

Note that

$$\mathbf{u}^\top(-\nabla f(\mathbf{X}^*))\mathbf{u} \geq \mu_r > \mu_{r+1} \geq \mathbf{w}^\top(-\nabla f(\mathbf{X}^*))\mathbf{w}. \tag{28}$$

This implies that the inner maximum in (25) can only be obtained by vectors in $\mathcal{V} \cap \text{span}\{\mathbf{u}_1, ..., \mathbf{u}_r\}$ (note (26) guarantees such vectors exist). Thus, plugging this observation into (25) we have that for any $i \in \{1, \dots, r\}$,

$$\begin{aligned}
\sigma_i &= \min_{\mathcal{V}:\dim(\mathcal{V})=n-i+1} \max_{\mathbf{v}\in\mathcal{V}\cap\text{span}\{\mathbf{u}_1,\dots,\mathbf{u}_r\},\|\mathbf{v}\|=1} \mathbf{v}^\top(\mathbf{X}^* + \eta(-\nabla f(\mathbf{X}^*)))\mathbf{v} \\
&\geq \min_{\mathcal{V}:\dim(\mathcal{V})=n-i+1} \max_{\mathbf{v}\in\mathcal{V}\cap\text{span}\{\mathbf{u}_1,\dots,\mathbf{u}_r\},\|\mathbf{v}\|=1} \mathbf{v}^\top\mathbf{X}^*\mathbf{v} + \eta\mu_r \\
&\underset{(a)}{=} \min_{\mathcal{V}:\dim(\mathcal{V})=n-i+1} \max_{\mathbf{v}\in\mathcal{V},\|\mathbf{v}\|=1} \mathbf{v}^\top\mathbf{X}^*\mathbf{v} + \eta\mu_r \underset{(b)}{=} \lambda_i(\mathbf{X}^*) + \eta\mu_k,
\end{aligned} \tag{29}$$

where (a) follows from the orthogonality of $\mathbf{X}^*$ to $\mathbf{u}_{r+1}\mathbf{u}_{r+1}^\top, ..., \mathbf{u}_n\mathbf{u}_n^\top$ (see Lemma 6), and (b) follows from the min-max principle for the eigenvalues, and since by definition $\mu_r = \mu_k$.

Using the max-min principle for the eigenvalues, we can write for any $j \in \{r+1, ..., n\}$,

$$\sigma_j = \max_{\mathcal{V}:\dim(\mathcal{V})=j} \min_{\mathbf{v}\in\mathcal{V}:\|\mathbf{v}\|=1} \mathbf{v}^\top(\mathbf{X}^* + \eta(-\nabla f(\mathbf{X}^*)))\mathbf{v}. \tag{30}$$

This time we maximize over all subspaces of dimension j, $j \geq r+1$. Thus, it must hold that for each such subspace $\mathcal{V}$,

$$\mathcal{V} \cap \text{span}\{\mathbf{u}_{r+1}, ..., \mathbf{u}_n\} \neq \emptyset,$$

otherwise the direct sum $\mathcal{V} \oplus \text{span}\{\mathbf{u}_{r+1}, ..., \mathbf{u}_n\} \subseteq \mathbb{R}^n$ would have dimension $j + n - r > n$. Thus, using (27) and (28), we have that the inner minimum in (30) is obtained by vectors in $\mathcal{V} \cap \text{span}\{\mathbf{u}_{r+1}, \dots, \mathbf{u}_n\}$, which is not an empty set. Using this observation we have that for any $j \in \{r+1, \dots, n\}$,

$$\begin{aligned}
\sigma_j &= \max_{\mathcal{V}:\dim(\mathcal{V})=j} \min_{\mathbf{v}\in\mathcal{V}\cap\text{span}\{\mathbf{u}_{r+1},\dots,\mathbf{u}_n\},\|\mathbf{v}\|=1} \mathbf{v}^\top(\mathbf{X}^* + \eta(-\nabla f(\mathbf{X}^*)))\mathbf{v} \\
&\underset{(a)}{=} \max_{\mathcal{V}:\dim(\mathcal{V})=j} \min_{\mathbf{v}\in\mathcal{V}\cap\text{span}\{\mathbf{u}_{r+1},\dots,\mathbf{u}_n\},\|\mathbf{v}\|=1} \mathbf{v}^\top(\eta(-\nabla f(\mathbf{X}^*)))\mathbf{v} \\
&\underset{(b)}{=} \max_{\mathcal{V}:\dim(\mathcal{V})=j} \min_{\mathbf{v}\in\mathcal{V},\|\mathbf{v}\|=1} \mathbf{v}^\top(\eta(-\nabla f(\mathbf{X}^*)))\mathbf{v} = \eta\mu_j,
\end{aligned} \tag{31}$$

where (a) follows since $\mathbf{X}^*$ is orthogonal to $\mathbf{u}_{r+1}\mathbf{u}_{r+1}^\top, \dots, \mathbf{u}_n\mathbf{u}_n^\top$ (see Lemma 6),
and (b) follows since by the eigen-decomposition of $-\nabla f(\mathbf{X}^*)$, restricting $\mathbf{v}$ to the intersection $\mathcal{V} \cap \text{span}\{\mathbf{u}_{r+1}, ..., \mathbf{u}_n\}$ does not increase the inner minimum.

From Lemma 1 we have the sufficient condition so that $\text{rank}(\Pi_{\mathcal{F}_{n,k}}(\mathbf{Y})) \leq r$:

$$\sum_{i=1}^{r} \min(\gamma_i - \gamma_{r+1}, 1) \geq k \implies \text{rank}(\Pi_{\mathcal{F}_{n,k}}(\mathbf{Y})) \leq r. \tag{32}$$

By Weyl's inequality we have that for any $i \in \{1, \dots, r\}$,

$$\begin{aligned}
\gamma_i - \gamma_{r+1} &\geq \sigma_i - \sigma_{r+1} - 2\|\mathbf{Y} - \mathbf{Y}^*\|_F \\
&= \sigma_i - \sigma_{r+1} - 2\|\mathbf{X} - \eta\nabla f(\mathbf{X}) - \mathbf{X}^* + \eta\nabla f(\mathbf{X}^*)\|_F \\
&\geq \sigma_i - \sigma_{r+1} - 2(1+\eta\beta)\|\mathbf{X} - \mathbf{X}^*\|_F.
\end{aligned} \tag{33}$$

Thus, we have that

$$\begin{aligned}
\sum_{i=1}^{r} \min(\gamma_i - \gamma_{r+1}, 1) &\underset{(a)}{\geq} \sum_{i=1}^{r} \min(\sigma_i - \sigma_{r+1} - 2(1+\eta\beta)\|\mathbf{X} - \mathbf{X}^*\|_F, 1) \\
&\underset{(b)}{\geq} \sum_{i=1}^{r} \min(\lambda_i(\mathbf{X}^*) + \eta(\mu_i - \mu_{r+1}) - 2(1+\eta\beta)\|\mathbf{X} - \mathbf{X}^*\|_F, 1) \\
&\geq \sum_{i=1}^{r} \min(\lambda_i(\mathbf{X}^*) + \eta(\mu_r - \mu_{r+1}) - 2(1+\eta\beta)\|\mathbf{X} - \mathbf{X}^*\|_F, 1),
\end{aligned} \tag{34}$$

where (a) follows from (33), and (b) follows from (29) and (31).

Thus, we indeed see that if

$$\|\mathbf{X}-\mathbf{X}^*\|_F \leq \frac{\eta(\mu_r-\mu_{r+1})}{2(1+\eta\beta)},$$

then $\sum_{i=1}^{r}\min(\gamma_i-\gamma_{r+1},1) \geq \sum_{i=1}^{r^*}\lambda_i(\mathbf{X}^*)=k$, which by (32) implies that $\text{rank}(\Pi_{\mathcal{F}_{n,k}}(\mathbf{Y})) \leq r$, as needed.

For the second part of the lemma let us fix some $r' \in \{r,...,n-1\}$. If we have that

$$\|\mathbf{X}-\mathbf{X}^*\|_F \leq \frac{\eta(\mu_r-\mu_{r'+1})}{2(1+\eta\beta)}, \tag{35}$$

then similarly to (34), we will have that,

$$\begin{aligned}
\sum_{i=1}^{r'}\min(\gamma_i-\gamma_{r'+1},1) &\geq \sum_{i=1}^{r}\min(\gamma_i-\gamma_{r'+1},1)\\
&\underset{(a)}{\geq} \sum_{i=1}^{r}\min(\sigma_i-\sigma_{r'+1}-2(1+\eta\beta)\|\mathbf{X}-\mathbf{X}^*\|_F,1)\\
&\underset{(b)}{\geq} \sum_{i=1}^{r}\min(\lambda_i(\mathbf{X}^*)+\eta(\mu_i-\mu_{r'+1})-2(1+\eta\beta)\|\mathbf{X}-\mathbf{X}^*\|_F,1)\\
&\geq \sum_{i=1}^{r}\min(\lambda_i(\mathbf{X}^*)+\eta(\mu_r-\mu_{r'+1})-2(1+\eta\beta)\|\mathbf{X}-\mathbf{X}^*\|_F,1)\\
&\underset{(c)}{\geq} \sum_{i=1}^{r}\min(\lambda_i(\mathbf{X}^*),1)=\sum_{i=1}^{r^*}\lambda_i(\mathbf{X}^*)=k,
\end{aligned}$$

where (a) follows from the same reasoning as (33), (b) follows from (29) and (31), and (c) follows from (35).

Thus, from Lemma 1 we have that (35) indeed implies that $\text{rank}(\Pi_{\mathcal{F}_{n,k}}(\mathbf{Y})) \leq r'$, which proves the second part of the lemma. □

We can now easily prove Theorems 4 and 6 by proving the following unifying theorem.

Theorem 7. *Let $\{\mathbf{X}_t\}_{t\geq 1}$ be a sequence produced by the projected gradient dynamics w.r.t. the convex Problem* (4) *with a fixed step-size $\eta \in (0,1/\beta]$:*

$$\mathbf{X}_{t+1}=\Pi_{\mathcal{F}_{n,k}}(\mathbf{X}_t-\eta\nabla f(\mathbf{X}_t)).$$

Fix some optimal solution $\mathbf{X}^*$ and let $\mu_1 \geq \mu_2 \geq ...\mu_n$ denote the eigenvalues of $-\nabla f(\mathbf{X}^*)$. Let r be the smallest integer such that $r \geq k$ and $\mu_r-\mu_{r+1}>0$. If the initialization $\mathbf{X}_1 \in \mathcal{F}_{n,k}$ satisfies $\|\mathbf{X}_1-\mathbf{X}^*\|_F \leq \frac{\eta(\mu_k-\mu_{r+1})}{2(1+\eta\beta)}$, then for all $t \geq 1$, rank$(\mathbf{X}_{t+1}) \leq r$.

More generally, for every $r' \in \{r,\ldots,n\}$, if $\|\mathbf{X}_1-\mathbf{X}^*\|_F \leq \frac{\eta(\mu_k-\mu_{r'+1})}{2(1+\eta\beta)}$, then for all $t \geq 1$, rank$(\mathbf{X}_{t+1}) \leq r'$.

In particular, if $r=k$, i.e., Assumption 1 holds with some $\delta>0$, and $\|\mathbf{X}_1-\mathbf{X}^*\|_F \leq \frac{\delta}{4\beta}$, setting $\eta=1/\beta$ guarantees that for all $t \geq 1$, rank$(\mathbf{X}_{t+1})=k$, and the sequence $\{\mathbf{X}_t\}_{t\geq 1}$ converges linearly with rate:

$$\forall t \geq 1: \quad f(\mathbf{X}_t)-f(\mathbf{X}^*) \leq (f(\mathbf{X}_1)-f(\mathbf{X}^*))\exp\left(-\Theta\left(\delta/\beta\right)(t-1)\right).$$

Proof. It is a well known fact that the distances of the iterates generated by the projected gradient method with step-size $\eta \in (0,1/\beta]$ to any optimal solution are monotone non-increasing, i.e., the

sequence $\{\|\mathbf{X}_t - \mathbf{X}^*\|_F\}_{t\geq 1}$ is monotone non-increasing, see for instance [3]. Thus, all results of the theorem regarding the rank of the iterates $\mathbf{X}_t, t \geq 1$, follow immediately from this observation, the initialization conditions listed in the theorem, and Lemma 7.

The linear convergence rate under Assumption 1 follows from the quadratic growth result — Lemma 3, and the known linear convergence rate of the projected gradient method for smooth functions that satisfy the quadratic growth property, see for instance [19]. □

E Details Missing from Section 3.2

We prove two auxiliary lemmas and then prove Theorem 3.

Lemma 8. *Let $\mathbf{M} \in \mathbb{S}^n$, and let $\mathbf{X} \in \mathcal{P}_{n,k}$ be the projection matrix onto the span of the top k eigenvectors of $\mathbf{M}$. Then, for any $\mathbf{Z} \in \mathcal{P}_{n,k}$ it holds that,*

$$\langle \mathbf{X} - \mathbf{Z}, \mathbf{M} \rangle \leq \|\mathbf{Z} - \mathbf{X}\|_F^2 \|\mathbf{M}\|_2.$$

Proof. Let us denote by $\mathbf{X}_\perp$ the projection matrix onto the orthogonal subspace, i.e., $\mathbf{X}_\perp = \mathbf{I} - \mathbf{X}$. It holds that,

$$\begin{aligned}\langle \mathbf{X} - \mathbf{Z}, \mathbf{M} \rangle &= \langle \mathbf{X} - \mathbf{Z}, \mathbf{X}\mathbf{M} \rangle + \langle \mathbf{X} - \mathbf{Z}, \mathbf{X}_\perp \mathbf{M} \rangle \\ &= \langle \mathbf{X} - \mathbf{Z}, \mathbf{X}\mathbf{M} \rangle - \langle \mathbf{Z}, \mathbf{X}_\perp \mathbf{M} \rangle. \end{aligned} \tag{36}$$

We consider each of the two terms on the RHS separately.

$$\begin{aligned}\langle \mathbf{X} - \mathbf{Z}, \mathbf{X}\mathbf{M} \rangle &= \mathrm{Tr}((\mathbf{X} - \mathbf{Z})\mathbf{X}\mathbf{M}) = \mathrm{Tr}(\mathbf{X}(\mathbf{X} - \mathbf{Z})\mathbf{X}\mathbf{M}) \\ &\underset{(a)}{\leq} \mathrm{Tr}(\mathbf{X}(\mathbf{X} - \mathbf{Z})\mathbf{X}) \cdot \lambda_1(\mathbf{M}) \\ &= \langle \mathbf{X} - \mathbf{Z}, \mathbf{X} \rangle \cdot \lambda_1(\mathbf{M}) = (k - \langle \mathbf{Z}, \mathbf{X} \rangle) \cdot \lambda_1(\mathbf{M}) \\ &\leq \frac{1}{2}\|\mathbf{Z} - \mathbf{X}\|_F^2 \cdot \|\mathbf{M}\|_2, \end{aligned} \tag{37}$$

where (a) holds since $\mathbf{X}(\mathbf{X} - \mathbf{Z})\mathbf{X}$ is positive semidefinite.

$$\begin{aligned}\langle \mathbf{Z}, \mathbf{X}_\perp \mathbf{M} \rangle &= \mathrm{Tr}(\mathbf{Z}\mathbf{X}_\perp \mathbf{M}) = \mathrm{Tr}(\mathbf{X}_\perp \mathbf{Z}\mathbf{X}_\perp \mathbf{M}) \\ &\underset{(b)}{\geq} \mathrm{Tr}(\mathbf{X}_\perp \mathbf{Z}\mathbf{X}_\perp) \cdot \lambda_n(\mathbf{M}) = \langle \mathbf{Z}, \mathbf{X}_\perp \rangle \cdot \lambda_n(\mathbf{M}) \\ &= \langle \mathbf{Z}, \mathbf{I} - \mathbf{X} \rangle \cdot \lambda_n(\mathbf{M}) = (k - \langle \mathbf{Z}, \mathbf{X} \rangle) \cdot \lambda_n(\mathbf{M}) \\ &\geq -\frac{1}{2}\|\mathbf{Z} - \mathbf{X}^*\|_F^2 \cdot \|\mathbf{M}\|_2, \end{aligned} \tag{38}$$

where (b) holds since $\mathbf{X}_\perp \mathbf{Z}\mathbf{X}_\perp$ is positive semidefinite.

The lemma follows from plugging (37) and (38) into (36). □

Lemma 9. *Fix some $t \geq 1$ and suppose $\eta < 1/\beta$. Then it holds that,*

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

Proof. Define the following function

$$\phi(\mathbf{Z}) := \langle \mathbf{Z}, \nabla f(\mathbf{Y}_t) \rangle + \frac{1}{2\eta}\|\mathbf{Z} - \mathbf{Y}_t\|_F^2,$$

and note that it is $1/\eta$ strongly convex, and that by definition, $\mathbf{X}_{t+1}$ is its minimizer over $\mathcal{F}_{n,k}$. Thus,

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

Rearranging we get,

$$\begin{aligned}\|\mathbf{X}_{t+1}-\mathbf{Y}_t\|_F^2 &\le \eta\langle \mathbf{Y}_t-\mathbf{X}_{t+1}, \nabla f(\mathbf{Y}_t)\rangle \\ &= \eta\langle \mathbf{Y}_t-\mathbf{X}_{t+1}, \nabla f(\mathbf{X}_{t+1})\rangle + \eta\langle \mathbf{Y}_t-\mathbf{X}_{t+1}, \nabla f(\mathbf{Y}_t)-\nabla f(\mathbf{X}_{t+1})\rangle \\ &\underset{(a)}{\le} \eta\langle \mathbf{Y}_t-\mathbf{X}_{t+1}, \nabla f(\mathbf{X}_{t+1})\rangle + \eta\beta\|\mathbf{X}_{t+1}-\mathbf{Y}_t\|_F^2 \\ &\underset{(b)}{\le} \eta\left(f(\mathbf{Y}_t)-f(\mathbf{X}_{t+1})\right) + \eta\beta\|\mathbf{X}_{t+1}-\mathbf{Y}_t\|_F^2,\end{aligned}$$

where (a) follows from the β-smoothness of $f(\cdot)$, and (b) follows from the convexity of $f(\cdot)$. Rearranging, we get the lemma. □

Proof of Theorem 3. The theorem follows from Lemma 5, it only remains to prove that the necessary conditions hold for all $t \ge 1$, i.e., that for all $t \ge 1$, it holds that $\|\mathbf{X}_{t+1}-\mathbf{Y}_t\|_F \le 1$, and $\mathbf{X}_{t+1} \in \mathcal{P}_{n,k}$, i.e., $\text{rank}(\mathbf{X}_{t+1}) = k$. The proof is by induction. For the base case $t = 1$, we first note that using Lemma 9 we have that,

$$\begin{aligned}\|\mathbf{X}_2-\mathbf{Y}_1\|_F^2 &\le \frac{\eta}{1-\eta\beta}\left(f(\mathbf{Y}_1)-f(\mathbf{X}_2)\right) \le \frac{\eta}{1-\eta\beta}\left(f(\mathbf{Y}_1)-f(\mathbf{X}^*)\right) \\ &\underset{(a)}{\le} \frac{\eta}{1-\eta\beta}\left(\langle \mathbf{X}^*-\mathbf{Y}_1, -\nabla f(\mathbf{X}^*)\rangle + \frac{\beta}{2}\|\mathbf{Y}_1-\mathbf{X}^*\|_F^2\right) \\ &\underset{(b)}{\le} \frac{\eta}{1-\eta\beta}\left(G+\frac{\beta}{2}\right)\|\mathbf{Y}_1-\mathbf{X}^*\|_F^2, \qquad (39)\end{aligned}$$

where (a) follows from the β-smoothness of $f(\cdot)$, and (b) follows from Lemma 8 and recalling that under Assumption 1, $\mathbf{X}^*$ is the projection matrix onto the span of the top k eigenvectors of $-\nabla f(\mathbf{X}^*)$ (see Lemma 6).

Note that for $\eta \in (0, \beta)$, $\frac{\eta}{1-\eta\beta}\left(G+\frac{\beta}{2}\right) \le 1 \iff \eta\left(G+\frac{3\beta}{2}\right) \le 1$, which clearly holds for our choice of step-size $\eta = \frac{1}{5\max\{\beta, G\}}$.

Thus, noting that for our choice of step-size and initialization assumption it holds that $\|\mathbf{Y}_1-\mathbf{X}^*\|_F \le 1$, we indeed have that $\|\mathbf{X}_2-\mathbf{X}_1\|_F \le 1$. Also, combining the initialization condition for $\mathbf{Y}_1$ listed in the theorem, together with Lemma 2, immediately implies that $\text{rank}(\mathbf{X}_2) = k$. Thus, the base case $t = 1$ of the induction holds.

Suppose now the induction holds for all $i \in \{1, \dots, t-1\}$, for some $t \ge 2$, and we will prove it for t. Using Lemma 9 we have that,

$$\begin{aligned}\|\mathbf{X}_{t+1}-\mathbf{Y}_t\|_F^2 &\le \frac{\eta}{1-\eta\beta}\left(f(\mathbf{Y}_t)-f(\mathbf{X}_{t+1})\right) \le \frac{\eta}{1-\eta\beta}\left(f(\mathbf{Y}_t)-f(\mathbf{X}^*)\right) \\ &\underset{(a)}{\le} \frac{\eta}{1-\eta\beta}\left(f(\mathbf{Y}_1)-f(\mathbf{X}^*)\right) \underset{(b)}{\le} \frac{\eta}{1-\eta\beta}\left(G+\frac{\beta}{2}\right)\|\mathbf{Y}_1-\mathbf{X}^*\|_F^2, \qquad (40)\end{aligned}$$

where (a) follows by using the induction hypothesis for all $i \le t$ together with Lemma 5, which guarantees that $f(\mathbf{Y}_t) \le f(\mathbf{Y}_1)$, and (b) follows from the same steps as in (39).

Since we have already established that the RHS of (40) is upper-bounded by 1 in the base case of the induction, it follows that $\|\mathbf{X}_{t+1}-\mathbf{Y}_t\|_F \le 1$.

Using the quadratic growth of $f(\cdot)$ (Lemma 3) we have that,

$$\begin{aligned}\|\mathbf{Y}_t-\mathbf{X}^*\|_F^2 &\le \frac{2}{\delta}\left(f(\mathbf{Y}_t)-f(\mathbf{X}^*)\right) \underset{(a)}{\le} \frac{2}{\delta}\left(f(\mathbf{Y}_1)-f(\mathbf{X}^*)\right) \\ &\underset{(b)}{\le} \frac{2}{\delta}\frac{\eta}{1-\eta\beta}\left(G+\frac{\beta}{2}\right)\|\mathbf{Y}_1-\mathbf{X}^*\|_F^2,\end{aligned}$$

where again, (a) follows by using the induction hypothesis for all $i \le t$ together with Lemma 5 which guarantees that $f(\mathbf{Y}_t) \le f(\mathbf{Y}_1)$, and (b) follows from (39).

Thus, using the fact that for our choice of step-size it holds that $\frac{\eta}{1-\eta\beta}\left(G+\frac{\beta}{2}\right) \le 1$, using the initalization assumption on $\mathbf{Y}_1$, and invoking Lemma 2, it follows that indeed $\text{rank}(\mathbf{X}_{t+1}) = k$, i.e., $\mathbf{X}_{t+1} \in \mathcal{P}_{n,k}$, and thus the induction holds for step t as well. □

F Frank-Wolfe Analysis

In this section we prove Theorem 5. Our analysis extends the one in [8] which only considered the case $k = 1$.

We begin with a lemma, whose proof is similar to the arguments used in the proof of Lemma 3, which will be essential to proving the local linear convergence of Frank-Wolfe under Assumption 1.

Lemma 10. *Let $\mathbf{X} \in \mathcal{F}_{n,k}$ and assume that $\lambda_{n-k}(\nabla f(\mathbf{X})) - \lambda_{n-k+1}(\nabla f(\mathbf{X})) \geq \delta_{\mathbf{X}} > 0$. Then, for $\mathbf{V} \in \arg\min_{\mathbf{P} \in \mathcal{P}_{n,k}} \langle \mathbf{P}, \nabla f(\mathbf{X}) \rangle$ it holds that,*

$$\langle \mathbf{X} - \mathbf{V}, \nabla f(\mathbf{X}) \rangle \geq \frac{\delta_{\mathbf{X}}}{2} \|\mathbf{X} - \mathbf{V}\|_F^2.$$

Proof. The proof follows the same lines as the proof of Lemma 3, but replacing $\mathbf{X}^*$ with $\mathbf{X}$, and noting that $\mathbf{V}$ is the (unique) projection matrix onto the span of the top k eigenvectors of $-\nabla f(\mathbf{X})$, similarly to the use of $\mathbf{X}^*$ as the projection matrix onto the span of the top k eigenvectors of $-\nabla f(\mathbf{X}^*)$ in the proof of Lemma 3. □

Algorithm 2 Frank-Wolfe with line-search for the Fantope

1: $\mathbf{X}_1 \leftarrow$ arbitrary point in $\mathcal{F}_{n,k}$
2: **for** $t = 1...$ **do**
3: $\quad \mathbf{V}_t \leftarrow \arg\min_{\mathbf{V} \in \mathcal{P}_{n,k}} \langle \mathbf{V}, \nabla f(\mathbf{X}_t) \rangle$
4: $\quad$ Choose step size $\eta_t \in [0, 1]$ using one of the two options:
5: $\quad$ First Option : $\eta_t \leftarrow \arg\min_{\eta \in [0,1]} f((1-\eta)\mathbf{X}_t + \eta \mathbf{V}_t)$
6: $\quad$ Second Option : $\eta_t \leftarrow \arg\min_{\eta \in [0,1]} f(\mathbf{X}_t) + \eta \langle \mathbf{V}_t - \mathbf{X}_t, \nabla f(\mathbf{X}_t) \rangle + \frac{\eta^2 \beta}{2} \|\mathbf{V}_t - \mathbf{X}_t\|_F^2$
7: $\quad \mathbf{X}_{t+1} \leftarrow (1 - \eta_t)\mathbf{X}_t + \eta_t \mathbf{V}_t$
8: **end for**

Theorem 8 (Formal version of Theorem 5)**.** *Let $\{\mathbf{X}_t\}_{t \geq 1}$ be a sequence produced by Algorithm 2 and denote $\forall t \geq 1$, $h_t := f(\mathbf{X}_t) - f(\mathbf{X}^*)$. Then,*

$$\forall t \geq 1 : \quad h_t = O(k\beta/t). \tag{41}$$

In addition, if Assumption 1 holds with parameter $\delta > 0$, then there exists $T_0 = O\left(k(\beta/\delta)^3\right)$ such that,

$$\forall t \geq T : \quad h_{t+1} \leq h_t \left(1 - \min\{\frac{\delta}{12\beta}, \frac{1}{2}\}\right)). \tag{42}$$

Finally, under Assumption 1, we have that,

$$\forall t \geq 1 : \quad \|\mathbf{V}_t - \mathbf{X}^*\|_F^2 = O\left(\frac{\beta^2}{\delta^3} h_t\right). \tag{43}$$

Proof. Result (41) follows from standard convergence results for the Frank-Wolfe method with line-search [12], and the fact that the Euclidean diameter of the Fantope $\mathcal{F}_{n,k}$ is $\sqrt{2k}$.

For the second part, observe that under Assumption 1, using the the β-smoothness of $f(\cdot)$, the quadratic growth result (Lemma 3) and (41), we have that for all $t \geq 1$,

$$\|\nabla f(\mathbf{X}_t) - \nabla f(\mathbf{X}^*)\|_F \leq \beta \|\mathbf{X}_t - \mathbf{X}^*\| \leq \beta \sqrt{\frac{2h_t}{\delta}} = O\left(\sqrt{\frac{k\beta^3}{t\delta}}\right).$$

Thus, for some $T_0 = O\left(k(\beta/\delta)^3\right)$ we have that,

$$\forall t \geq T_0 : \quad \|\nabla f(\mathbf{X}_t) - \nabla f(\mathbf{X}^*)\|_F \leq \frac{\delta}{3}.$$

Let us write the eigen-decomposition of $\nabla f(\mathbf{X}_t)$ as $\nabla f(\mathbf{X}_t) = \sum_{i=1}^n \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$. Using Weyl's inequality for the eigenvalues we can write for every $t \geq T_0$,

$$\begin{aligned}\lambda_{n-k} - \lambda_{n-k+1} &\geq \lambda_{n-k}(\nabla f(\mathbf{X}^*)) - \lambda_{n-k+1}(\nabla f(\mathbf{X}^*)) - 2\|\nabla f(\mathbf{X}) - \nabla f(\mathbf{X}^*)\|_F \\ &\geq \delta - \frac{2\delta}{3} = \frac{\delta}{3}.\end{aligned}$$

Thus, for all $t \geq T_0$, $\lambda_{n-k+1} < \lambda_{n-k}$ and the matrix $\mathbf{V}_t$ is uniquely defined and given by $\mathbf{V}_t = \sum_{i=n-k+1}^n \mathbf{v}_i \mathbf{v}_i^\top$. Using $\mathbf{X}_{t+1} = (1-\eta_t)\mathbf{X}_t + \eta_t \mathbf{V}_t$, the smoothness of $f(\cdot)$, and the fact that η_t is chosen via line-search, we have that,

$$\forall \eta \in [0,1] : f(\mathbf{X}_{t+1}) \leq f(\mathbf{X}_t) + \eta \langle \mathbf{V}_t - \mathbf{X}_t, \nabla f(\mathbf{X}_t)\rangle + \frac{\eta^2 \beta}{2}\|\mathbf{V}_t - \mathbf{X}_t\|_F^2.$$

Subtracting $f(\mathbf{X}^*)$ from both sides and using Lemma 10 with gap $\delta_{\mathbf{X}} = \delta/3$, we have that for all $t \geq T_0$,

$$\begin{aligned}\forall \eta \in [0,1] : \quad h_{t+1} &\leq h_t + \frac{\eta}{2}\langle \mathbf{V}_t - \mathbf{X}_t, \nabla f(\mathbf{X}_t)\rangle + (\frac{\eta^2\beta}{2} - \frac{\eta\delta}{12})\|\mathbf{V}_t - \mathbf{X}_t\|_F^2 \\ &\leq (1 - \frac{\eta}{2})h_t + (\frac{\eta^2\beta}{2} - \frac{\eta\delta}{12})\|\mathbf{V}_t - \mathbf{X}_t\|_F^2,\end{aligned}$$

where the last inequality follows from the convexity of $f(\cdot)$.

Now, if $\frac{\delta}{6\beta} \leq 1$, by setting $\eta = \frac{\delta}{6\beta}$ we have that $h_{t+1} \leq (1 - \frac{\delta}{12\beta})h_t$. Otherwise, $\delta > 6\beta$ and so, setting $\eta = 1$, we get that $h_{t+1} \leq \frac{1}{2}h_t$, which proves Result (42).

Finally, for the third part of the lemma, recalling that $\mathbf{V}_t$ and $\mathbf{X}^*$ are the projection matrices onto the span of the top k eigenvectors of $-\nabla f(\mathbf{X}_t)$ and $-\nabla f(\mathbf{X}^*)$, respectively, using the well known Davis-Kahan sinθ theorem (see for instance [27]), we have that for all $t \geq 1$,

$$\|\mathbf{V}_t - \mathbf{X}^*\|_F^2 \leq \frac{8\|\nabla f(\mathbf{X}_t) - \nabla f(\mathbf{X}^*)\|_F^2}{(\lambda_{n-k}(\nabla f(\mathbf{X}^*)) - \lambda_{n-k+1}(\nabla f(\mathbf{X}^*)))^2} \leq \frac{8\beta^2\|\mathbf{X}_t - \mathbf{X}^*\|_F^2}{\delta^2} \leq \frac{16\beta^2 h_t}{\delta^3},$$

where the last inequality follows from the quadratic growth result, Lemma 3. Thus, Result (43) follows. □