Quadratic Optimization with Orthogonality Constraints: Explicit Łojasiewicz Exponent and Linear Convergence of Line-Search Methods

Huikang Liu HKLIU@SE.CUHK.EDU.HK
Weijie Wu WWU@SE.CUHK.EDU.HK
Anthony Man-Cho So MANCHOSO@SE.CUHK.EDU.HK
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong, Shatin NT, Hong Kong SAR, China

Abstract

A fundamental class of matrix optimization problems that arise in many areas of science and engineering is that of quadratic optimization with orthogonality constraints. Such problems can be solved using line-search methods on the Stiefel manifold, which are known to converge globally under mild conditions. To determine the convergence rates of these methods, we give an explicit estimate of the exponent in a Łojasiewicz inequality for the (non-convex) set of critical points of the aforementioned class of problems. This not only allows us to establish the linear convergence of a large class of line-search methods but also answers an important and intriguing problem in mathematical analysis and numerical optimization. A key step in our proof is to establish a local error bound for the set of critical points, which may be of independent interest.

1. Introduction

Quadratic optimization problems with orthogonality constraints constitute an important class of matrix optimization problems that have found applications in many areas of science and engineering, such as combinatorial optimization, data mining, dynamical systems, multivariate statistical analysis, and signal processing, just to mention a few (see, e.g., (Bolla et al., 1998; Manton, 2002; Absil et al., 2008; Journée et al., 2010; Kokiopoulou et al., 2011; Saad, 2011; So, 2011; Yger et al., 2012)). A prototypical form of such problems is

$$\min_{X \in \mathrm{St}(m,n)} \left\{ F(X) = \mathrm{tr}\left(X^T A X B\right) \right\}, \qquad \text{(QP-OC)}$$

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

where $\mathrm{St}(m,n) = \left\{ X \in \mathbb{R}^{m \times n} \mid X^T X = I_n \right\}$ (with $m \geq n$ and I_n being the $n \times n$ identity matrix) is the compact Stiefel manifold and $A \in \mathcal{S}^m$, $B \in \mathcal{S}^n$ are given symmetric matrices. The algorithmic aspects of Problem (QP-OC) have been extensively investigated in the literature; see, e.g., the references above. One approach is to exploit the manifold structure of the constraint set $\mathrm{St}(m,n)$ and apply retraction-based line-search methods. Specifically, the update formulae of these methods take the form

$$X_{k+1} = R\left(X_k, \alpha_k \xi_k\right) \quad \text{for } k = 0, 1, \ldots, \qquad (1)$$

where $\alpha_k > 0$ is the step size, ξ_k is a search direction in the tangent space to $\mathrm{St}(m,n)$ at X_k, and $R(X_k, \cdot)$ is a function that maps a vector in the tangent space to $\mathrm{St}(m,n)$ at X_k into a point on $\mathrm{St}(m,n)$. In particular, the iterates produced by (1) are all feasible for Problem (QP-OC). Naturally, the choice of step sizes, search directions, and the retraction will affect the convergence and efficiency of the resulting method. For the general problem of optimizing a smooth function over the Stiefel manifold (which includes Problem (QP-OC) as a special case), various choices have been proposed over the years, and the convergence properties of the resulting methods are relatively well understood; see, e.g., (Abrudan et al., 2008; Absil et al., 2008; Absil & Malick, 2012; Wen & Yin, 2013; Jiang & Dai, 2015) and the references therein. However, very little is known about the *convergence rates* of these methods, even when they are applied to the much more structured problem (QP-OC). In an early work, Smith (Smith, 1994) showed that when used to optimize a smooth function over a Riemannian manifold, the method of steepest descent (which is a special case of (1)) will converge linearly to a critical point if the function is *strongly convex on the manifold*. However, such a notion of convexity is much stronger than that on the Euclidean space. In particular, it is known that every smooth function that is convex on a compact Riemannian manifold (such as the Stiefel manifold) is constant (Bishop & O'Neill, 1969). Therefore, one cannot hope to obtain linear convergence results for Problem (QP-OC) using the

convexity-based approach in (Smith, 1994). Later, Absil et al. considered Problem (QP-OC) with $n = 1$ and $B = I_n = 1$ (which corresponds to minimizing the Rayleigh quotient on the unit sphere in $\mathbb{R}^m$) and showed that a certain line-search method will converge linearly to an eigenvector associated with the smallest eigenvalue λ of A, provided that λ has multiplicity one; see Theorem 4.6.3 of (Absil et al., 2008). However, it is not clear how to extend this result to cover even the case where $n > 1$ and/or the multiplicity of λ is greater than one.

Part of the difficulty in analyzing the convergence rates of line-search methods of the form (1) is due to the fact that optimization problems over the Stiefel manifold (such as Problem (QP-OC)) is non-convex in general. Indeed, much of the existing analysis machinery relies on convexity in a crucial manner. Recently, two different approaches have been developed in an attempt to circumvent such difficulty. The first proceeds by showing that the objective function, when restricted to a suitable neighborhood of a (globally) optimal solution, possesses nice properties, and then using such properties to establish the rate at which a properly initialized iterative method converges to the optimal solution. Such an approach has attracted much attention lately and has been successfully employed to tackle a wide variety of structured non-convex optimization problems; see, e.g., (Sun & Luo, 2015) and the references therein. In the context of Problem (QP-OC), this approach was first pursued by Shamir (Shamir, 2015a) (see also (Shamir, 2015b)), who considered the case where A is negative semidefinite and $B = I_n$ (which corresponds to the Principal Component Analysis (PCA) problem). He proposed a stochastic line-search method for solving the problem and showed that under certain assumptions on the multiplicities of the eigenvalues of A and on the boundedness of A, the method, when properly initialized, will converge linearly to a matrix whose columns are the bottom n eigenvectors of A with high probability. However, Shamir's approach does not apply to Problem (QP-OC) in its full generality (i.e., when A is not negative semidefinite and/or $B \neq I_n$). Moreover, the assumptions on A are quite strong, and it is not clear whether they are necessary for linear convergence or are simply artifacts of the analysis.

The second approach to analyzing the convergence rates of iterative methods in the non-convex setting is to use a so-called *Łojasiewicz inequality*; see, e.g., (Absil et al., 2005; Merlet & Nguyen, 2013; Schneider & Uschmajew, 2015). Roughly speaking, a Łojasiewicz inequality holds at a point if the growth of the objective function around that point can be bounded by an exponent (called the *Łojasiewicz exponent*) of the norm of the objective function gradient. In particular, a Łojasiewicz inequality can be regarded as a regularity condition similar to an error bound—the latter has featured prominently in the convergence rate analysis of iterative methods; see, e.g., (Luo & Tseng, 1993; Luo, 2000; Wang & Lin, 2014; So & Zhou, 2015; Zhou & So, 2015; Zhou et al., 2015). For the problem of optimizing a real-analytic function over a compact real-analytic submanifold (such as Problem (QP-OC)), it is well known that a Łojasiewicz inequality holds at each of the critical points, with possibly different Łojasiewicz exponents at different critical points. Moreover, the iterates generated by a host of retraction-based line-search methods will converge to a critical point, and the convergence rate can be inferred directly from the Łojasiewicz exponent at that particular critical point. (We refer the reader to (Schneider & Uschmajew, 2015) for an account of these results.) Compared with the first approach, this second, Łojasiewicz inequality-based approach can potentially provide more insights into the convergence behavior of iterative methods, as it gives the rate of convergence not just to the optimal solution but to *any* critical point. In particular, it opens up the possibility of determining the convergence rate of an iterative method even if it is initialized arbitrarily. However, powerful as it may seem, the Łojasiewicz inequality-based approach has a severe limitation: Most existing proofs of the Łojasiewicz inequality only guarantee the existence of the Łojasiewicz exponent but do not offer any clue on how to estimate its value. Without such an estimate, one cannot even determine whether a given iterative method converges sublinearly or linearly. To the best of our knowledge, estimates of the Łojasiewicz exponent are available only for non-convex quadratic optimization problems with simple convex constraints (such as a ball or a polyhedron) (Luo & Pang, 1994; Luo & Sturm, 2000; Forti et al., 2006) and general polynomial optimization problems (Li et al., 2015). However, these two classes of results do not shed any light on Problem (QP-OC), as the former is concerned with convex constraints, while the latter gives estimates that depend on the dimensions of the problem and lead to very weak convergence rate guarantees.

With the above two approaches in mind, our goal in this paper is to develop a more refined convergence rate analysis of iterative methods for solving Problem (QP-OC) by enriching the second approach and to strengthen and extend the results obtained using the first approach. Our main contribution is to show that all critical points of Problem (QP-OC) have the same Łojasiewicz exponent and to give a sharp estimate of its value. Such a result is significant, as it expands the currently very limited repertoire of optimization problems for which the Łojasiewicz exponent is known. Moreover, when combined with the convergence analysis framework in (Schneider & Uschmajew, 2015), it immediately implies the linear convergence of various retraction-based line-search methods to a critical point of Problem (QP-OC). A crucial step in our technical development is to establish a local Lipschitzian error bound for

the *non-convex* set of critical points of Problem (QP-OC). Once such an error bound is available, it is rather straightforward to obtain a Łojasiewicz inequality with an explicitly given exponent. We should point out that the aforementioned error bound result is considerably more difficult to establish than those in (Luo & Tseng, 1993; Wang & Lin, 2014; So & Zhou, 2015; Zhou & So, 2015; Zhou et al., 2015), as neither the objective function nor the constraint of Problem (QP-OC) is convex. In addition, our linear convergence result does not require any assumptions on A and B. Thus, it yields a qualitative improvement upon the results in (Absil et al., 2008; Shamir, 2015a;b). As our final contribution, we consider Problem (QP-OC) with $B = I_n$ (which corresponds to finding the bottom n eigenvectors of A) and show, for the first time, that if there is a gap between the n-th and $(n+1)$-st smallest eigenvalues of A, then various retraction-based line-search methods will converge linearly to an optimal solution when properly initialized.

Besides the notations introduced earlier, we shall use $\mathcal{O}^n$ to denote the set of $n \times n$ orthogonal matrices (in particular, we have $\mathcal{O}^n = \mathrm{St}(n,n)$); $\mathrm{Diag}(x_1,\ldots,x_n)$ to denote the diagonal matrix with $x_1,\ldots,x_n$ on the diagonal; $\mathrm{BlkDiag}(A_1,\ldots,A_n)$ to denote the block diagonal matrix whose diagonal blocks are $A_1,\ldots,A_n$. Given a matrix $Y \in \mathbb{R}^{m\times n}$ and a non-empty closed set $\mathcal{X} \subset \mathbb{R}^{m\times n}$, we shall use $\mathrm{dist}(Y,\mathcal{X})$ to denote the distance of Y to $\mathcal{X}$; i.e., $\mathrm{dist}(Y,\mathcal{X}) = \min_{X\in\mathcal{X}} \|X - Y\|_F$. Other notations are standard.

2. Preliminaries

2.1. First-Order Optimality Condition and Descent Directions

To begin, let us introduce some basic definitions and concepts. We view $\mathrm{St}(m,n)$ as an embedded submanifold of $\mathbb{R}^{m\times n}$ with the inherited *Riemannian metric* $\langle\cdot,\cdot\rangle$ given by $\langle X, Y\rangle = \mathrm{tr}\left(X^TY\right)$. For any $X \in \mathrm{St}(m,n)$, the *tangent space* to $\mathrm{St}(m,n)$ at X is given by $T(X) = \left\{Y \in \mathbb{R}^{m\times n} \mid X^TY + Y^TX = \mathbf{0}\right\}$. The *Euclidean gradient* of $F(X) = \mathrm{tr}\left(X^TAXB\right)$ is $\nabla F(X) = 2AXB$. Its orthogonal projection onto $T(X)$, called the *projected gradient* of $F(X)$ and denoted by $\mathrm{grad}\, F(X)$, can be calculated as

$$\begin{aligned}\mathrm{grad}\, F(X) &= \left(I_m - XX^T\right)\nabla F(X)\\ &\quad + \frac{1}{2}X\left(X^T\nabla F(X) - \nabla F(X)^TX\right)\\ &= 2AXB - XX^TAXB - XBX^TAX;\end{aligned}$$

see Example 3.6.2 of (Absil et al., 2008). The set of *critical points* of Problem (QP-OC) is then defined as

$$\mathcal{X} = \{X \in \mathrm{St}(m,n) \mid \mathrm{grad}\, F(X) = \mathbf{0}\}.$$

The following proposition gives a characterization of $\mathcal{X}$:

Proposition 1. *Let $X \in \mathrm{St}(m,n)$ be given. Then, the following are equivalent:*

(i) $\mathrm{grad}\, F(X) = \mathbf{0}$.

(ii) $\nabla F(X) - X\nabla F(X)^TX = \mathbf{0}$.

(iii) For any $\rho > 0$, $D_\rho(X) = \nabla F(X) - X\left(2\rho\nabla F(X)^TX + (1-2\rho)X^T\nabla F(X)\right) = \mathbf{0}$.

Proof. The equivalence between (ii) and (iii) is established in Lemma 2.1 of (Jiang & Dai, 2015). To prove the equivalence between (i) and (ii), observe that

$$\begin{aligned}\mathrm{grad}\, F(X) &= \left(I_m - \frac{1}{2}XX^T\right)\nabla F(X) - \frac{1}{2}X\nabla F(X)^TX\\ &= \left(I_m - \frac{1}{2}XX^T\right)\left(\nabla F(X) - X\nabla F(X)^TX\right).\end{aligned}$$

Now, it remains to note that $I_m - (1/2)XX^T$ is invertible. □

2.2. Retraction-Based Line-Search Methods

A standard and quite natural idea for finding a critical point of Problem (QP-OC) is to start at an arbitrary point on $\mathrm{St}(m,n)$ and then iteratively move in a search direction defined by a tangent vector while staying on $\mathrm{St}(m,n)$ until a critical point is found. Given any $X \in \mathrm{St}(m,n)$, it can be shown that $-D_\rho(X) \in T(X)$ and $-D_\rho(X)$ is a descent direction at $X \in \mathrm{St}(m,n)$ for any $\rho > 0$; see Lemma 3.1 of (Jiang & Dai, 2015). Thus, we can pick some $\rho > 0$ and use $-D_\rho(X)$ as a candidate search direction. After moving the current iterate in the search direction, however, the resulting point need not lie on $\mathrm{St}(m,n)$. Thus, we need to bring the point back on $\mathrm{St}(m,n)$ to form the next iterate. This can be achieved using a retraction.

Definition 1. *A* retraction *on $\mathrm{St}(m,n)$ is a smooth map $R : \bigcup_{X\in\mathrm{St}(m,n)}(\{X\}\times T(X)) \to \mathrm{St}(m,n)$ satisfying (i) $R(X,\mathbf{0}) = X$ for any $X \in \mathrm{St}(m,n)$ and (ii) for any $X \in \mathrm{St}(m,n)$,*

$$\lim_{T(X)\ni\xi\to\mathbf{0}} \frac{\|R(X,\xi) - (X+\xi)\|_F}{\|\xi\|_F} = 0. \qquad (2)$$

Various retractions on the Stiefel manifold have been studied in the literature. Below are two examples:

- Polar Decomposition-Based Retraction:

$$R(X,\xi) = (X+\xi)(I_n + \xi^T\xi)^{-1/2}.$$

- QR-Decomposition-Based Retraction:

$$R(X,\xi) = \mathrm{qf}(X+\xi),$$

where $\mathrm{qf}(A)$ denotes the Q-factor in the thin QR-decomposition of A; see Section 5.2.6 of (Golub & Van Loan, 1996).

Further examples of retractions on the Stiefel manifold can be found in (Absil & Malick, 2012; Kaneko et al., 2013).

With the above preparations, we are now ready to describe a generic retraction-based line-search method for solving Problem (QP-OC); see Algorithm 1.

Algorithm 1 Line-Search Method on $\mathrm{St}(m,n)$

Input: $X^0 \in \mathrm{St}(m,n)$, $\rho > 0$
1: **for** $k = 0, 1, 2, \ldots$ **do**
2: calculate the descent direction $-D_\rho(X^k)$ at X^k
3: choose a step size $\alpha_k > 0$
4: set $X^{k+1} = R\left(X^k, -\alpha_k D_\rho(X^k)\right)$
5: terminate if convergence criterion is met
6: **end for**

It is known that with suitably chosen step sizes $\{\alpha_k\}_{k\geq 0}$ (such as those computed by an Armijo-type rule), the iterates $\{X^k\}_{k\geq 0}$ generated by Algorithm 1 will converge to a critical point $X^* \in \mathcal{X}$ of Problem (QP-OC); see (Schneider & Uschmajew, 2015). However, the rate of convergence is much less understood. A main obstacle is the need to find a suitable criticality measure that is amenable to analysis. In view of Proposition 1, a candidate measure is $\|D_\rho(\cdot)\|_F$ for any $\rho > 0$. Such a choice has several advantages. First, it is easy to compute. Second, we have $\|D_\rho(X)\|_F = 0$ if and only if $X \in \mathcal{X}$. Third, a deep and far-reaching result of Łojasiewicz implies the existence of constants $\delta, \eta > 0$ and $\theta \in (0, 1/2]$ such that the inequality

$$|F(X) - F(X^*)|^{1-\theta} \leq \eta \|D_\rho(X)\|_F \tag{3}$$

holds for all $X \in \mathrm{St}(m,n)$ satisfying $\|X - X^*\|_F \leq \delta$ (note that in general δ, η, θ depend on X^*); see Section 2.2 of (Schneider & Uschmajew, 2015). In particular, the inequality (3), known as the *Łojasiewicz inequality* for Problem (QP-OC), suggests that when $\|D_\rho(X)\|_F$ is small, the objective value of X will be close to that of X^*. It is well known (see, e.g., (Schneider & Uschmajew, 2015)) that the Łojasiewicz inequality (3) implies the sublinear (resp. linear) convergence of Algorithm 1 if $\theta \in (0, 1/2)$ (resp. $\theta = 1/2$). Unfortunately, the value of θ, known as the *Łojasiewicz exponent* for Problem (QP-OC), is still not known. In fact, it remains an important and intriguing open problem in mathematical analysis and numerical optimization to give a good estimate of θ.

3. Main Results

The main contribution of this paper is the following theorem, which answers the above question:

Theorem 1. *(Łojasiewicz Inequality for Problem (QP-OC)) There exist constants $\delta \in (0, \sqrt{2}/2)$ and $\eta > 0$ such that for all $X \in \mathrm{St}(m,n)$ and $X^* \in \mathcal{X}$ with $\|X - X^*\|_F \leq \delta$,*

$$|F(X) - F(X^*)|^{1/2} \leq \eta \|D_\rho(X)\|_F .$$

Theorem 1 is significant because it not only reveals that the constants δ, η, θ in (3) can be made uniform over all critical points $X^* \in \mathcal{X}$ but also establishes the fact that the Łojasiewicz exponent at any critical point is $1/2$. To prove Theorem 1, our strategy is to first establish a related result, which states that $X \in \mathrm{St}(m,n)$ is in fact close to $\mathcal{X}$ when $\|D_\rho(X)\|_F$ is small. Specifically, we have the following theorem:

Theorem 2. *(Local Error Bound for Problem (QP-OC)) There exist constants $\delta \in (0,1)$ and $\eta > 0$ such that for all $X \in \mathrm{St}(m,n)$ with $\mathrm{dist}(X, \mathcal{X}) \leq \delta$,*

$$\mathrm{dist}(X, \mathcal{X}) \leq \eta \|D_\rho(X)\|_F .$$

The error bound in Theorem 2 is reminiscent of those that have appeared in the recent literature; e.g., (Wang & Lin, 2014; So & Zhou, 2015; Zhou & So, 2015; Zhou et al., 2015). However, the former is for the *non-convex* optimization problem (QP-OC), while the latter is for *convex* optimization problems. As such, the techniques used to establish the former are substantially different from those used to establish the latter.

In the next section, we give the proofs of Theorems 1 and 2.

4. Proofs of the Main Results

4.1. Proof of Theorem 2

Let us begin with the proof of Theorem 2, which can be divided into four steps.

4.1.1. Preliminary Observations

Let $A = U_A \Sigma_A U_A^T$ and $B = U_B \Sigma_B U_B^T$ be spectral decompositions of A and B, respectively. It is straightforward to verify that $\mathrm{tr}\left(X^T A X B\right) = \mathrm{tr}\left(\bar{X}^T \Sigma_A \bar{X} \Sigma_B\right)$, where $\bar{X} = U_A^T X U_B \in \mathrm{St}(m,n)$. Thus, we may assume without loss of generality that $A = \mathrm{Diag}(a_1, \ldots, a_m) \in \mathcal{S}^m$ and $B = \mathrm{Diag}(b_1, \ldots, b_n) \in \mathcal{S}^n$, where $a_1 \geq a_2 \geq \cdots \geq a_m$ and $b_1 \geq b_2 \geq \cdots \geq b_n$. By Proposition 1, we can write

$$\mathcal{X} = \left\{X \in \mathrm{St}(m,n) \mid AXB - XBX^T AX = \mathbf{0}\right\}. \tag{4}$$

Now, it can be verified that

$$D_\rho(X) = \left(I_m - (1-2\rho) X X^T\right)\left(\nabla F(X) - X \nabla F(X)^T X\right).$$

Note that $I_m - (1-2\rho)XX^T$ is invertible for any $\rho > 0$ and

$$\begin{aligned} & \left\|\nabla F(X) - X\nabla F(X)^T X\right\|_F \\ \leq\ & \left\|\left(I_m - (1-2\rho)XX^T\right)^{-1}\right\| \cdot \|D_\rho(X)\|_F \\ \leq\ & \max\left\{1, (2\rho)^{-1}\right\} \cdot \|D_\rho(X)\|_F . \end{aligned}$$

In particular, since $\nabla F(X) = 2AXB$, in order to prove Theorem 2, it suffices to prove the following:

Theorem 2'. *There exist constants $\delta \in (0,1)$ and $\eta > 0$ such that for all $X \in \mathrm{St}(m,n)$ with $\mathrm{dist}(X, \mathcal{X}) \leq \delta$,*

$$\mathrm{dist}(X, \mathcal{X}) \leq \eta \left\|AXB - XBX^TAX\right\|_F .$$

4.1.2. Characterizing the Set of Critical Points when B has Full Rank

Consider first the case where B has full rank; i.e., $b_i \neq 0$ for $i = 1, \ldots, n$. Let n_A and n_B be the number of distinct eigenvalues of A and B, respectively. Then, there exist indices $s_0, s_1, \ldots, s_{n_A}$ and $t_0, t_1, \ldots, t_{n_B}$ such that $0 = s_0 < s_1 < \cdots < s_{n_A} = m$ and $0 = t_0 < t_1 < \cdots < t_{n_B} = n$, and

$$\begin{aligned} & a_{s_0+1} = \cdots = a_{s_1} > a_{s_1+1} = \cdots = a_{s_2} \\ & \quad > \cdots > a_{s_{n_A-1}+1} = \cdots = a_{s_{n_A}}, \\ & b_{t_0+1} = \cdots = b_{t_1} > b_{t_1+1} = \cdots = b_{t_2} \\ & \quad > \cdots > b_{t_{n_B-1}+1} = \cdots = b_{t_{n_B}}. \end{aligned}$$

Let $U_1, \ldots, U_{n_A}$ and $V_1, \ldots, V_{n_B}$ be the eigenspaces of A and B, respectively. Note that $\dim(U_i) = s_i - s_{i-1}$ for $i = 1, \ldots, n_A$ and $\dim(V_j) = t_j - t_{j-1}$ for $j = 1, \ldots, n_B$. Furthermore, let

$$\mathcal{H} = \left\{ (h_1, \ldots, h_{n_A}) \;\middle|\; \sum_{i=1}^{n_A} h_i = n, \right. \\ \left. h_i \in \{0, 1, \ldots, s_i - s_{i-1}\} \text{ for } i = 1, \ldots, n_A \right\}$$

and $\{e_i\}_{i=1}^m$ be the standard basis of $\mathbb{R}^m$. Given any $h = (h_1, \ldots, h_{n_A}) \in \mathcal{H}$, define

$$\begin{aligned} & E_i(h) = [e_{s_{i-1}+1} \ \cdots \ e_{s_{i-1}+h_i}] \in \mathbb{R}^{m \times h_i} \quad \text{for } i = 1, \ldots, n_A, \\ & E(h) = [E_1(h) \ \cdots \ E_{n_A}(h)] \in \mathbb{R}^{m \times n}. \end{aligned} \tag{5}$$

We then have the following characterization of the set $\mathcal{X}$ of critical points of Problem (QP-OC), whose proof can be found in the supplementary material:

Proposition 2. *Every $X \in \mathcal{X}$ can be expressed as*

$$X = \mathrm{BlkDiag}(P_1, \ldots, P_{n_A}) \cdot E(h) \cdot \mathrm{BlkDiag}(Q_1, \ldots, Q_{n_B}) \tag{6}$$

for some $P_i \in \mathcal{O}^{s_i - s_{i-1}}$ $(i = 1, \ldots, n_A)$, $Q_j \in \mathcal{O}^{t_j - t_{j-1}}$ $(j = 1, \ldots, n_B)$, and $h \in \mathcal{H}$.

Remarks. (i) Essentially, Proposition 2 states that every $X \in \mathcal{X}$ can be factorized as $X = PQ$, where $P \in \mathrm{St}(m,n)$ and $Q \in \mathcal{O}^n$, and the columns of P (resp. Q) are the eigenvectors of A (resp. B). Indeed, observe that for $i = 1, \ldots, n_A$, the $(s_{i-1}+1)$-st to s_i-th columns of $\mathrm{BlkDiag}(P_1, \ldots, P_{n_A})$ form an orthonormal basis of U_i. Similarly, for $j = 1, \ldots, n_B$, the $(t_{j-1}+1)$-st to t_j-th columns of $\mathrm{BlkDiag}(Q_1, \ldots, Q_{n_B})$ form an orthonormal basis of V_j. To specify which n of the m eigenvectors of A are chosen to form P, we use the matrix $E(h)$, where $h = (h_1, \ldots, h_{n_A}) \in \mathcal{H}$ and h_i is the number of eigenvectors chosen from the eigenspace U_i.

(ii) A result similar to Proposition 2 has appeared in Section 4.8.2 of (Absil et al., 2008). However, the proof therein contains a small gap. Specifically, from the properties that B is diagonal and commutes with X^TAX, it is claimed in Section 4.8.2 of (Absil et al., 2008) that X^TAX is also diagonal. However, this is not true unless the diagonal entries of B are all distinct.

Proposition 2 suggests that we can partition $\mathcal{X}$ into disjoint subsets $\{\mathcal{X}_h\}_{h \in \mathcal{H}}$, where every $X \in \mathcal{X}_h$ can be expressed as

$$X = \mathrm{BlkDiag}(P_1, \ldots, P_{n_A}) \cdot E(h) \cdot \mathrm{BlkDiag}(Q_1, \ldots, Q_{n_B})$$

for some $P_i \in \mathcal{O}^{s_i - s_{i-1}}$ $(i = 1, \ldots, n_A)$ and $Q_j \in \mathcal{O}^{t_j - t_{j-1}}$ $(j = 1, \ldots, n_B)$. Consequently, in order to prove Theorem 2', it suffices to bound $\mathrm{dist}(X, \mathcal{X}_h)$ for any $X \in \mathrm{St}(m,n)$ and $h \in \mathcal{H}$.

4.1.3. Estimating the Distance to the Set of Critical Points

Let $X \in \mathrm{St}(m,n)$ and $h = (h_1, \ldots, h_{n_A}) \in \mathcal{H}$ be arbitrary. By definition,

$$\begin{aligned} \mathrm{dist}(X, \mathcal{X}_h) = \min \{ & \|X - \mathrm{BlkDiag}\left(P_1, \ldots, P_{n_A}\right) \cdot \\ & E(h) \cdot \mathrm{BlkDiag}\left(Q_1, \ldots, Q_{n_B}\right)\|_F \mid \\ & P_i \in \mathcal{O}^{s_i - s_{i-1}} \text{ for } i = 1, \ldots, n_A; \\ & Q_j \in \mathcal{O}^{t_j - t_{j-1}} \text{ for } j = 1, \ldots, n_B \}. \end{aligned} \tag{7}$$

Let $\left(P_1^*, \ldots, P_{n_A}^*, Q_1^*, \ldots, Q_{n_B}^*\right)$ be an optimal solution to (7). Upon letting $P^* = \mathrm{BlkDiag}\left(P_1^*, \ldots, P_{n_A}^*\right) \in \mathcal{O}^m$, $Q^* = \mathrm{BlkDiag}\left(Q_1^*, \ldots, Q_{n_B}^*\right) \in \mathcal{O}^n$, and $\bar{X} = (P^*)^T X (Q^*)^T$, it is clear that $\mathrm{dist}^2(X, \mathcal{X}_h) = \left\|\bar{X} - E(h)\right\|_F^2$. To bound this quantity, consider the decompositions

$$\bar{X} = \left[\bar{X}_1 \ \cdots \ \bar{X}_{n_B}\right], \quad E(h) = \left[\bar{E}_1(h) \ \cdots \ \bar{E}_{n_B}(h)\right], \tag{8}$$

where $\bar{X}_j, \bar{E}_j(h) \in \mathbb{R}^{m \times (t_j - t_{j-1})}$. We then have the following result, whose proof can be found in the supplementary material:

Proposition 3. *For $j = 1, \dots, n_B$ and $k = 1, \dots, m$, denote the k-th row of $\bar{X}_j$ and $\bar{E}_j(h)$ by $\left[\bar{X}_j\right]_k$ and $\left[\bar{E}_j(h)\right]_k$, respectively. Suppose that $\operatorname{dist}(X, \mathcal{X}_h) < 1$. Then,*

$$\operatorname{dist}^2(X, \mathcal{X}_h) = \sum_{j=1}^{n_B} \sum_{k \in \mathcal{I}_j} \Theta\left(\left\|\left[\bar{X}_j\right]_k\right\|_2^2\right),$$

where $\mathcal{I}_j = \left\{k \in \{1, \dots, m\} : \left[\bar{E}_j(h)\right]_k = \mathbf{0}\right\}$.

To establish the desired error bound, we need to link $\left\|AXB - XBX^TAX\right\|_F$ to the bound on $\operatorname{dist}^2(X, \mathcal{X}_h)$ in Proposition 3. This is achieved in two steps. First, we prove the following result:

Proposition 4. *Consider the decomposition of $\bar{X}$ in (8). Then, $\left\|AXB - XBX^TAX\right\|_F^2$*

$$= \Omega\left(\sum_{j=1}^{n_B} \left\|A\bar{X}_j - \bar{X}_j\bar{X}_j^TA\bar{X}_j\right\|_F^2\right).$$

In view of Proposition 4, we then proceed to prove the following bound:

Proposition 5. *There exists a constant $\delta \in (0, 1)$ such that for all $X \in \operatorname{St}(m, n)$ with $\operatorname{dist}(X, \mathcal{X}_h) \le \delta$,*

$$\sum_{j=1}^{n_B} \left\|A\bar{X}_j - \bar{X}_j\bar{X}_j^TA\bar{X}_j\right\|_F^2 = \sum_{j=1}^{n_B} \sum_{k \in \mathcal{I}_j} \Omega\left(\left\|\left[\bar{X}_j\right]_k\right\|_2^2\right).$$

The proofs of Propositions 4 and 5 can be found in the supplementary material. Now, observe that whenever $X \in \operatorname{St}(m, n)$ and $\operatorname{dist}(X, \mathcal{X}) \le \delta$, there exists an $h \in \mathcal{H}$ such that $\operatorname{dist}(X, \mathcal{X}_h) \le \delta$. Hence, by combining Propositions 3, 4, and 5, we obtain Theorem 2'.

4.1.4. Removing the Full Rank Assumption on B

Consider now the case where B does not have full rank. Without loss of generality, we assume that $B = \operatorname{BlkDiag}(\bar{B}, \mathbf{0})$, where $\bar{B} = \operatorname{Diag}(b_1, \dots, b_p) \in \mathcal{S}^p$ has full rank. Then, using (4), it can be shown that

$$\mathcal{X} = \Big\{X = [X_1\ X_2] \in \operatorname{St}(m, n) \mid X_1 \in \mathbb{R}^{m \times p}, \\ X_2 \in \mathbb{R}^{m \times (n-p)}, AX_1\bar{B} - X_1\bar{B}X_1^TAX_1 = \mathbf{0}\Big\}.$$

It follows that for any $X = [X_1\ X_2] \in \operatorname{St}(m, n)$ with $X_1 \in \mathbb{R}^{m \times p}$ and $X_2 \in \mathbb{R}^{m \times (n-p)}$, we have $\operatorname{dist}(X, \mathcal{X}) = \operatorname{dist}(X_1, \bar{\mathcal{X}})$, where

$$\bar{\mathcal{X}} = \left\{X \in \operatorname{St}(m, p) \mid AX\bar{B} - X\bar{B}X^TAX = \mathbf{0}\right\}.$$

By our previous result, there exist constants $\delta \in (0, 1)$ and $\eta > 0$ such that for all $X_1 \in \operatorname{St}(m, p)$ with $\operatorname{dist}(X_1, \bar{\mathcal{X}}) \le \delta$,

$$\operatorname{dist}(X_1, \bar{\mathcal{X}}) \le \eta \left\|AX_1\bar{B} - X_1\bar{B}X_1^TAX_1\right\|_F.$$

To complete the proof, it remains to observe that

$$\begin{aligned} &\left\|AXB - XBX^TAX\right\|_F^2 \\ &= \left\|AX_1\bar{B} - X_1\bar{B}X_1^TAX_1\right\|_F^2 + \left\|X_1\bar{B}X_1^TAX_2\right\|_F^2 \\ &= \left\|AX_1\bar{B} - X_1\bar{B}X_1^TAX_1\right\|_F^2 \\ &\quad + \left\|X_2^T\left(AX_1\bar{B} - X_1\bar{B}X_1^TAX_1\right)X_1^T\right\|_F^2 \\ &= \Theta\left(\left\|AX_1\bar{B} - X_1\bar{B}X_1^TAX_1\right\|_F^2\right). \end{aligned}$$

4.2. Proof of Theorem 1

Recall that the Łojasiewicz inequality in Theorem 1 is concerned with bounding the change in the objective value around a critical point, while the local error bound in Theorem 2 is concerned with bounding the distance to the set of critical points. Thus, to prove Theorem 1, we need a link between the former and the latter. The following technical result furnishes such a link. Its proof can be found in the supplementary material.

Proposition 6. *There exists a constant $\eta > 0$ such that for all $X \in \operatorname{St}(m, n)$ and $X^* \in \mathcal{X}$,*

$$|F(X) - F(X^*)| \le \eta\|X - X^*\|_F^2.$$

Now, let $X \in \operatorname{St}(m, n)$ and $X^* \in \mathcal{X}$ be such that $\|X - X^*\|_F < \delta_0 = \min\{\delta, \sqrt{2}/2\}$, where $\delta \in (0, 1)$ is the constant for which Theorem 2 holds. Furthermore, let $\bar{X}^* \in \mathcal{X}$ be such that $\operatorname{dist}(X, \mathcal{X}) = \|X - \bar{X}^*\|_F < \delta_0$. We claim that $X^*, \bar{X}^* \in \mathcal{X}_h$ for some $h \in \mathcal{H}$. Indeed, if $X^* \in \mathcal{X}_h$ and $\bar{X}^* \in \mathcal{X}_{h'}$ with distinct $h, h' \in \mathcal{H}$, then Proposition 2 and the discussion in Section 4.1.4 imply that $\|X^* - \bar{X}^*\|_F \ge \sqrt{2}$. However, our assumption yields $\|X^* - \bar{X}^*\|_F \le \|X - X^*\|_F + \|X - \bar{X}^*\|_F < 2\delta_0 \le \sqrt{2}$, which is a contradiction. This establishes the claim.

Since the function F is constant on $\mathcal{X}_h$ for any given $h \in \mathcal{H}$, we have $F(X^*) = F(\bar{X}^*)$. Hence, by Proposition 6 and Theorem 2, we obtain

$$\begin{aligned} |F(X) - F(X^*)| &= |F(X) - F(\bar{X}^*)| \\ &\le \eta_1\|X - \bar{X}^*\|_F^2 \\ &\le \eta_1\eta_2\|D_\rho(X)\|_F^2 \end{aligned}$$

for some constants $\eta_1, \eta_2 > 0$. This completes the proof of Theorem 1.

5. Convergence to Critical Points vs. Global Optima

Based on the discussion in Section 2.2, we see that Theorem 1 implies the linear convergence of Algorithm 1 to a critical point of Problem (QP-OC), regardless of how it is initialized. Thus, our result gives a rather complete picture of the convergence behavior of retraction-based line-search methods for solving Problem (QP-OC). Of course,

we are mostly interested in finding an optimal solution to Problem (QP-OC). Thus, it is natural to ask whether one can find a suitable initialization for Algorithm 1 so that it converges to such a solution. In this section, we consider Problem (QP-OC) with $B = I_n$ (which corresponds to finding the bottom n eigenvectors of A) and show that the answer to the above question is positive under the following assumption:

Assumption 1. *The eigenvalues of $A \in \mathcal{S}^m$, given by $a_1 \geq a_2 \geq \cdots \geq a_m$, satisfy $\lambda = a_{m-n} - a_{m-n+1} > 0$.*

We remark that the above assumption is similar to the one used in (Shamir, 2015b). For simplicity of exposition, we shall fix R to be the polar-decomposition-based retraction (see Section 2.2) and use $-D_{1/2}(X)$ as the search direction at $X \in \mathrm{St}(m,n)$ (i.e., $\rho = 1/2$) in the sequel. However, it should be noted that our result can be extended to handle more general retractions and search directions.

To begin, let $A = U_A \Sigma_A U_A^T$ be a spectral decomposition of A and $\mathcal{X}^*$ be the set of optimal solutions to Problem (QP-OC). Using Proposition 2 and Assumption 1, we can express any $X^* \in \mathcal{X}^*$ as

$$X^* = U_A \cdot \mathrm{BlkDiag}(\bar{P}_1, \bar{P}_2) \cdot \begin{bmatrix} \mathbf{0} & I_n \end{bmatrix}^T \cdot \bar{Q}$$

for some $\bar{P}_1 \in \mathcal{O}^{m-n}$, $\bar{P}_2 \in \mathcal{O}^n$, and $\bar{Q} \in \mathcal{O}^n$. In particular, every $X^* \in \mathcal{X}^*$ to Problem (QP-OC) has the form $X^* = U_A \cdot \begin{bmatrix} \mathbf{0} & X_2^* \end{bmatrix}^T$ for some $X_2^* \in \mathbb{R}^{n\times n}$. This motivates us to consider the potential function $g : \mathrm{St}(m,n) \to \mathbb{R}_+$, which is defined as

$$g(X) = \left\| [U_A^T X]_1 \right\|_F^2,$$

where $U_A^T X = \begin{bmatrix} [U_A^T X]_1^T & [U_A^T X]_2^T \end{bmatrix}^T$ with $[U_A^T X]_1 \in \mathbb{R}^{(m-n)\times n}$ and $[U_A^T X]_2 \in \mathbb{R}^{n\times n}$. It is not hard to verify that $g(X^*) = 0$ for any $X^* \in \mathcal{X}^*$ and $g(X) \geq 1$ for any $X \in \mathcal{X} \setminus \mathcal{X}^*$. Our goal now is to show that if a particular iterate $X^k \in \mathrm{St}(m,n)$ satisfies, say, $1/4 \leq g(X^k) \leq 3/4$, then with a suitable choice of the step size, the next iterate $X^{k+1} \in \mathrm{St}(m,n)$ will satisfy $g(X^{k+1}) < g(X^k)$; i.e., the potential value will decrease. Thus, if the initial iterate $X^0 \in \mathrm{St}(m,n)$ satisfies $g(X^0) \leq 3/4$, then subsequent iterates will move away from the critical points in $\mathcal{X} \setminus \mathcal{X}^*$. This, together with the fact that the iterates generated by Algorithm 1 is globally convergent to a critical point, allows us to conclude that the iterates will converge to an optimal solution to Problem (QP-OC).

To achieve the goal, let $X \in \mathrm{St}(m,n)$ be fixed and set $\xi = -D_{1/2}(X)$. Furthermore, define the function $\tilde{g} : \mathbb{R}_+ \to \mathbb{R}_+$ by $\tilde{g}(\alpha) = g(R(X, \alpha\xi))$; i.e.,

$$\tilde{g}(\alpha) = \left\| [U_A^T (X + \alpha\xi)]_1 (I_n + \alpha^2 \xi^T \xi)^{-1/2} \right\|_F^2.$$

Clearly, $\tilde{g}$ is continuously differentiable. The following result reveals the local behavior of $\tilde{g}$. Its proof can be found in the supplementary material.

Proposition 7. *The following hold:*

(i) $\tilde{g}'(0) \leq -4\lambda \left(g(X) - g^2(X) \right)$.

(ii) For $\alpha \in (0, (2\sqrt{n}\|A\|)^{-1}]$, where $\|A\|$ denotes the spectral norm of A, $|\tilde{g}''(\alpha)| \leq 8n\|A\|^2(5\sqrt{n}+24)$.

Proposition 7(i) shows that when $1/4 \leq g(X) \leq 3/4$, we have $\tilde{g}'(0) \leq -3\lambda/4$. It follows from Proposition 7(ii) and the Lipschitz continuity of $\tilde{g}'$ over compact intervals that if $\alpha \in (0, \bar{\alpha}]$, where

$$\bar{\alpha} = \min \left\{ (2\sqrt{n}\|A\|)^{-1}, 3\lambda(32n\|A\|^2(5\sqrt{n}+24))^{-1} \right\},$$

then $\tilde{g}(\alpha) < \tilde{g}(0)$. To summarize, we have the following theorem:

Theorem 3. *(Linear Convergence of Line-Search Methods to Global Optima) Consider Problem (QP-OC) with $B = I_n$. Under Assumption 1, if Algorithm 1 is initialized with a point $X^0 \in \mathrm{St}(m,n)$ that satisfies $g(X^0) \leq 3/4$ and the step sizes $\{\alpha_k\}_{k\geq 0}$ satisfy $\alpha_k \leq \bar{\alpha}$ for all $k \geq 0$, then the iterates generated by Algorithm 1 will converge to an optimal solution to Problem (QP-OC). Moreover, the asymptotic convergence rate is at least linear.*

6. Numerical Results

In this section, we report numerical results on both synthetic and real-world datasets to illustrate the convergence rate of the retraction-based line-search algorithm (Algorithm 1) for solving Problem (QP-OC). As we shall see, the results are consistent with our theoretical analysis. In our experiments, we use $\xi = -D_{1/2}(X)$ as the search direction at $X \in \mathrm{St}(m,n)$. Besides the two retractions mentioned in Section 2.2, we also use the *Cayley transform-based retraction*, which is given by

$$R(X, \alpha\xi) = (I + \alpha H)^{-1}(I - \alpha H)X,$$

where $H = X\nabla F(X)^T - \nabla F(X) X^T$ satisfies $H = -H^T$ and $\xi = HX$. We compute the step sizes using the following Armijo-type rule with parameter $\beta \in (0,1)$ (see (Schneider & Uschmajew, 2015)):

$$\alpha_k = \max \left\{ \beta^\ell \mid F\left(R(X^k, -\beta^\ell D_{1/2}(X^k))\right) - F(X^k) \leq -10^{-3}\beta^\ell \nabla F(X^k)^T D_{1/2}(X^k),\ \ell \geq 0 \right\}. \quad (9)$$

We stop the algorithm when $F(X^k) - F(X^{k+1}) < 10^{-8}$.

6.1. Synthetic Datasets

First, we generate a matrix $A \in \mathcal{S}^m$ whose elements are sampled randomly from the uniform distribution. The initial point $X^0 \in \mathrm{St}(m,n)$ is chosen according to the criterion in Theorem 3. We set $\beta = 0.5$ (resp. $\beta = 0.4$) in (9)

for the polar decomposition-based retraction (resp. QR-decomposition-based and Cayley transform-based retractions). Figure 1 corresponds to the setting where $m = 5000$, $n = 1$, and $B = 1$, while Figure 2 corresponds to $m = 500$, $n = 10$ and $B = I$. In both figures, we observe that the objective value converges linearly to the optimal value. It is worth noting that the rate at which the objective value decreases depends on the retraction used. This is consistent with the results in (Schneider & Uschmajew, 2015), which suggest that the convergence rate of a line-search method is affected by the rate at which the limit (2) tends to zero.

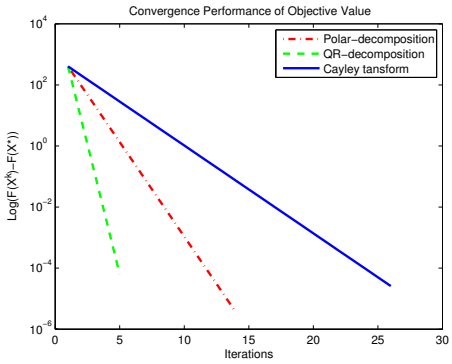

Figure 1. $m = 5000$, $n = 1$, A random, $B = I_1$

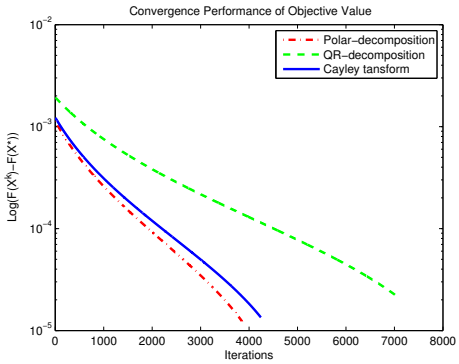

Figure 2. $m = 500$, $n = 10$, A random, $B = I_{10}$

Next, we include a diagonal matrix $B \in \mathcal{S}^n$ whose diagonal elements are sampled randomly from the uniform distribution. Figure 3 corresponds to $m = 500$ and $n = 20$. Again, linear convergence to the optimal value is observed.

6.2. Real-World Dataset

In this section, we conduct a similar numerical study using the training data of the well-known MNIST dataset. We extract 5000 observations of the digit '0' to get a 5000×784 matrix W. We then form $A = W^T W$ and $B = I_2$ (i.e., $m = 784$ and $n = 2$). The first two observations are used to form our initial point $X^0 \in \mathrm{St}(m, n)$, which will converge to the optimal solution based on the result of Theorem 3. We set $\beta = 0.3$ (resp. $\beta = 0.2$) in (9) for the polar

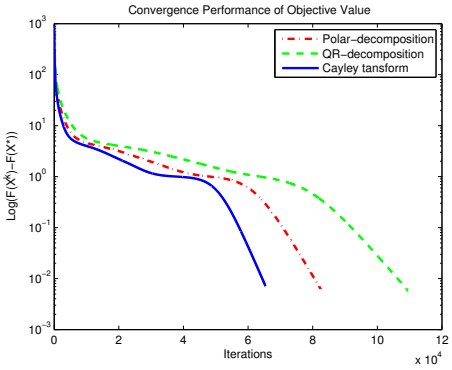

Figure 3. $m = 500$, $n = 20$, A random, B diagonal

decomposition-based retraction (resp. QR-decomposition-based and Cayley transform-based retractions). Figure 4 shows the convergence performance. As seen from the figure, the objective value converges linearly to the optimal value.

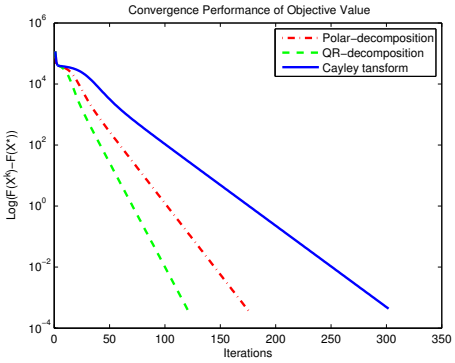

Figure 4. 5000 observations of '0', $m = 784$, $n = 2$

7. Conclusion

In this paper, we gave an explicit estimate of the exponent in a Łojasiewicz inequality for the (non-convex) set of critical points of Problem (QP-OC). Such an estimate was obtained by establishing a local error bound for the aforementioned set of critical points. Together with known arguments, our result implies the linear convergence of a large class of line-search methods on the Stiefel manifold. An interesting future direction would be to extend our techniques to analyze the convergence rates of iterative methods for solving structured non-convex optimization problems.

References

Abrudan, Traian E., Eriksson, Jan, and Koivunen, Visa. Steepest Descent Algorithms for Optimization under Unitary Matrix Constraint. *IEEE Transactions on Signal Processing*, 56(3):1134–1147, 2008.

Absil, P.-A. and Malick, Jérôme. Projection–Like Retrac-

tions on Matrix Manifolds. *SIAM Journal on Optimization*, 22(1):135–158, 2012.

Absil, P.-A., Mahony, R., and Andrews, B. Convergence of the Iterates of Descent Methods for Analytic Cost Functions. *SIAM Journal on Optimization*, 16(2):531–547, 2005.

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

Bishop, R. L. and O'Neill, B. Manifolds of Negative Curvature. *Transactions of the American Mathematical Society*, 145:1–49, 1969.

Bolla, Marianna, Michaletzky, György, Tusnády, Gábor, and Ziermann, Margit. Extrema of Sums of Heterogeneous Quadratic Forms. *Linear Algebra and Its Applications*, 269(1–3):331–365, 1998.

Forti, Mauro, Nistri, Paolo, and Quincampoix, Marc. Convergence of Neural Networks for Programming Problems via a Nonsmooth Łojasiewicz Inequality. *IEEE Transactions on Neural Networks*, 17(6):1471–1486, 2006.

Golub, Gene H. and Van Loan, Charles F. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, Maryland, third edition, 1996.

Jiang, Bo and Dai, Yu-Hong. A Framework of Constraint Preserving Update Schemes for Optimization on Stiefel Manifold. *Mathematical Programming, Series A*, 153 (2):535–575, 2015.

Journée, Michel, Nesterov, Yurii, Richtárik, Peter, and Sepulchre, Rodolphe. Generalized Power Method for Sparse Principal Component Analysis. *Journal of Machine Learning Research*, 11(Feb.):517–553, 2010.

Kaneko, Tetsuya, Fiori, Simone, and Tanaka, Toshihisa. Empirical Arithmetic Averaging over the Compact Stiefel Manifold. *IEEE Transactions on Signal Processing*, 61(4):883–894, 2013.

Kokiopoulou, E., Chen, J., and Saad, Y. Trace Optimization and Eigenproblems in Dimension Reduction Methods. *Numerical Linear Algebra with Applications*, 18(3): 565–602, 2011.

Li, G., Mordukhovich, B. S., and Phạm, T. S. New Fractional Error Bounds for Polynomial Systems with Applications to Hölderian Stability in Optimization and Spectral Theory of Tensors. *Mathematical Programming, Series A*, 153(2):333–362, 2015.

Luo, Zhi-Quan. New Error Bounds and Their Applications to Convergence Analysis of Iterative Algorithms. *Mathematical Programming, Series B*, 88(2):341–355, 2000.

Luo, Zhi-Quan and Pang, Jong-Shi. Error Bounds for Analytic Systems and Their Applications. *Mathematical Programming*, 67(1):1–28, 1994.

Luo, Zhi-Quan and Sturm, Jos F. Error Bounds for Quadratic Systems. In Frenk, Hans, Roos, Kees, Terlaky, Tamás, and Zhang, Shuzhong (eds.), *High Performance Optimization*, volume 33 of *Applied Optimization*, pp. 383–404. Springer Science+Business Media, Dordrecht, 2000.

Luo, Zhi-Quan and Tseng, Paul. Error Bounds and Convergence Analysis of Feasible Descent Methods: A General Approach. *Annals of Operations Research*, 46(1):157–178, 1993.

Manton, Jonathan H. Optimization Algorithms Exploiting Unitary Constraints. *IEEE Transactions on Signal Processing*, 50(3):635–650, 2002.

Merlet, Benoît and Nguyen, Thanh Nhan. Convergence to Equilibrium for Discretizations of Gradient–Like Flows on Riemannian Manifolds. *Differential and Integral Equations*, 26(5–6):571–602, 2013.

Nesterov, Yurii. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, Boston, 2004.

Saad, Yousef. *Numerical Methods for Large Eigenvalue Problems*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, revised edition, 2011.

Schneider, Reinhold and Uschmajew, André. Convergence Results for Projected Line–Search Methods on Varieties of Low–Rank Matrices via Łojasiewicz Inequality. *SIAM Journal on Optimization*, 25(1):622–646, 2015.

Schönemann, Peter H. A Generalized Solution of the Orthogonal Procrustes Problem. *Psychometrika*, 31(1):1–10, 1966.

Shamir, Ohad. A Stochastic PCA and SVD Algorithm with an Exponential Convergence Rate. In *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*, pp. 144–152, 2015a.

Shamir, Ohad. Fast Stochastic Algorithms for SVD and PCA: Convergence Properties and Convexity. Manuscript, available at http://arxiv.org/abs/1507.08788, 2015b.

Smith, Steven T. Optimization Techniques on Riemannian Manifolds. In Bloch, Anthony (ed.), *Hamiltonian and Gradient Flows, Algorithms and Control*, Fields Institue Communications, pp. 113–136. American Mathematical Society, Providence, Rhode Island, 1994.

So, Anthony Man-Cho. Moment Inequalities for Sums of Random Matrices and Their Applications in Optimization. *Mathematical Programming, Series A*, 130(1):125–151, 2011.

So, Anthony Man-Cho and Zhou, Zirui. Non–Asymptotic Convergence Analysis of Inexact Gradient Methods for Machine Learning Without Strong Convexity. Manuscript, available at http://www.se.cuhk.edu.hk/~manchoso/papers/inexact_GM_conv.pdf, 2015.

Sun, Ruoyu and Luo, Zhi-Quan. Guaranteed Matrix Completion via Non–convex Factorization. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2015)*, 2015.

Wang, Po-Wei and Lin, Chih-Jen. Iteration Complexity of Feasible Descent Methods for Convex Optimization. *Journal of Machine Learning Research*, 15:1523–1548, 2014.

Wen, Zaiwen and Yin, Wotao. A Feasible Method for Optimization with Orthogonality Constraints. *Mathematical Programming, Series A*, 142(1–2):397–434, 2013.

Yger, Florian, Berar, Maxime, Gasso, Gilles, and Rakotomamonjy, Alain. Adaptive Canonical Correlation Analysis Based on Matrix Manifolds. In *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*, pp. 1071–1078, 2012.

Zhou, Zirui and So, Anthony Man-Cho. A Unified Approach to Error Bounds for Structured Convex Optimization Problems. Manuscript, available at http://arxiv.org/abs/1512.03518, 2015.

Zhou, Zirui, Zhang, Qi, and So, Anthony Man-Cho. $\ell_{1,p}$–Norm Regularization: Error Bounds and Convergence Rate Analysis of First–Order Methods. In *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*, pp. 1501–1510, 2015.

8. Supplementary Material

8.1. Proof of Proposition 2

Let $X \in \mathcal{X}$ be arbitrary. Using (4) and the fact that $X^TX = I_n$, we have $X^TAXB = BX^TAX$. Since both X^TAX and B are symmetric, this implies that X^TAX and B are simultaneously diagonalizable. In particular, there exist orthogonal matrices $Q_j \in \mathcal{O}^{t_j - t_{j-1}}$ and diagonal matrices $\Sigma_j \in \mathcal{S}^{t_j - t_{j-1}}$, where $j = 1, \ldots, n_B$, such that the columns of $\mathrm{BlkDiag}(Q_1, \ldots, Q_{n_B})$ are the eigenvectors of B, and that

$$X^TAX = \mathrm{BlkDiag}\left(Q_1^T\Sigma_1Q_1, \ldots, Q_{n_B}^T\Sigma_{n_B}Q_{n_B}\right). \tag{10}$$

Now, using (4) again, we have $\left(AX - XX^TAX\right)B = \mathbf{0}$. Since B has full rank and hence invertible, this yields $AX = XX^TAX$. Upon letting $Y = X \cdot \mathrm{BlkDiag}\left(Q_1^T, \ldots, Q_{n_B}^T\right) \in \mathrm{St}(m,n)$ and using (10), we obtain $AY = Y \cdot \mathrm{BlkDiag}(\Sigma_1, \ldots, \Sigma_{n_B})$. As $\Sigma_1, \ldots, \Sigma_{n_B}$ are diagonal, this implies that each of the n columns of Y is an eigenvector of A. To see that X can be expressed in the form given on the right-hand side of (6), it remains to note that A has m eigenvectors in total, and that any set of m eigenvectors of A can be expressed as $\mathrm{BlkDiag}(P_1, \ldots, P_{n_A})$ for some $P_i \in \mathcal{O}^{s_i - s_{i-1}}$, where $i = 1, \ldots, n_A$.

The converse is rather easy to verify. Hence, the proof is completed.

8.2. Proof of Proposition 3

Using (7) and (8), it can be verified that

$$\begin{aligned}
\mathrm{dist}^2(X, \mathcal{X}_h) &= \left\|\bar{X} - E(h)\right\|_F^2 \\
&= \min\left\{ \left\|\bar{X} - E(h) \cdot \mathrm{BlkDiag}(Q_1, \ldots, Q_{n_B})\right\|_F^2 \;\middle|\; Q_j \in \mathcal{O}^{t_j - t_{j-1}} \text{ for } j = 1, \ldots, n_B \right\} \\
&= \sum_{j=1}^{n_B} \min\left\{ \left\|\bar{X}_j - \bar{E}_j(h)Q_j\right\|_F^2 \;\middle|\; Q_j \in \mathcal{O}^{t_j - t_{j-1}} \right\}.
\end{aligned}$$

From the definitions of $E(h)$ in (5) and $\bar{E}_j(h)$ in (8), we see that up to a rearrangement of the rows, $\bar{E}_j(h)$ takes the form $\bar{E}_j(h) = \begin{bmatrix} I_{t_j - t_{j-1}} \\ \mathbf{0} \end{bmatrix}$. Thus, to obtain the desired bound on $\mathrm{dist}^2(X, \mathcal{X}_h)$, it remains to prove the following:

Lemma 1. *Let* $S = \begin{bmatrix} S_1 \\ S_2 \end{bmatrix} \in \mathrm{St}(p,q)$ *be given, with* $S_1 \in \mathbb{R}^{q \times q}$ *and* $S_2 \in \mathbb{R}^{(p-q)\times q}$. *Consider the following problem:*

$$v^* = \min\left\{ \left\| S - \begin{bmatrix} I_q \\ \mathbf{0} \end{bmatrix} X \right\|_F^2 \;\middle|\; X \in \mathcal{O}^q \right\}.$$

Suppose that $v^* < 1$. *Then, we have* $v^* = \Theta\left(\|S_2\|_F^2\right)$.

Proof. Since

$$\left\| S - \begin{bmatrix} I_q \\ \mathbf{0} \end{bmatrix} X \right\|_F^2 = \|S_1 - X\|_F^2 + \|S_2\|_F^2,$$

it suffices to consider the problem

$$\min\left\{ \|S_1 - X\|_F^2 \mid X \in \mathcal{O}^q \right\}. \tag{11}$$

Problem (11) is an instance of the orthogonal Procrustes problem, whose optimal solution is given by $X^* = UV^T$, where $S_1 = U\Sigma V^T$ is the singular value decomposition of S_1 (Schönemann, 1966). It follows that

$$v^* = \|\Sigma - I_q\|_F^2 + \|S_2\|_F^2.$$

Now, since $S \in \mathrm{St}(p,q)$, we have $S^TS = S_1^TS_1 + S_2^TS_2 = I_q$, or equivalently,

$$\Sigma^2 + V^TS_2^TS_2V = I_q.$$

This implies that $\mathbf{0} \preceq \Sigma \preceq I_q$ and

$$I_q - \Sigma = (I_q + \Sigma)^{-1} \left(V^T S_2^T S_2 V\right).$$

It follows that

$$\frac{1}{4}\|S_2\|_F^4 + \|S_2\|_F^2 \leq v^* \leq \|S_2\|_F^4 + \|S_2\|_F^2.$$

This, together with the fact that $\|S_2\|_F^2 \leq v^* < 1$, yields $v^* = \Theta\left(\|S_2\|_F^2\right)$, as desired. □

8.3. Proof of Proposition 4

Recall that $P^* = \mathrm{BlkDiag}\left(P_1^*, \ldots, P_{n_A}^*\right) \in \mathcal{O}^m, Q^* = \mathrm{BlkDiag}\left(Q_1^*, \ldots, Q_{n_B}^*\right) \in \mathcal{O}^n$, $\bar{X} = (P^*)^T X (Q^*)^T$. Upon observing that $AP^* = P^* A$, $BQ^* = Q^* B$, $B = \mathrm{BlkDiag}\left(b_{t_1} I_{t_1 - t_0}, \ldots, b_{t_{n_B}} I_{t_{n_B} - t_{n_B - 1}}\right)$ and using (8), we compute

$$\begin{aligned}
\left\|AXB - XBX^TAX\right\|_F^2 &= \left\|AP^*\bar{X}Q^*B - P^*\bar{X}Q^*B(Q^*)^T\bar{X}^T(P^*)^TAP^*\bar{X}Q^*\right\|_F^2 \\
&= \left\|P^*\left(A\bar{X}B - \bar{X}B\bar{X}^TA\bar{X}\right)Q^*\right\|_F^2 \\
&= \left\|A\bar{X}B - \bar{X}B\bar{X}^TA\bar{X}\right\|_F^2 \\
&= \sum_{j=1}^{n_B}\left\|b_{t_j}A\bar{X}_j - \sum_{k=1}^{n_B} b_{t_k}\bar{X}_k\left(\bar{X}_k^TA\bar{X}_j\right)\right\|_F^2. \qquad (12)
\end{aligned}$$

Now, observe that the columns of $\bar{X}$ are orthonormal and span an n-dimensional subspace $\mathcal{L}$. In particular, for $j = 1, \ldots, n_B$, each column of $A\bar{X}_j$ can be decomposed as $u + v$, where u is a linear combination of the columns of $\bar{X}$ and $v \in \mathcal{L}^\perp$, the orthogonal complement of $\mathcal{L}$. In view of the structure of $\bar{X}$ in (8), this leads to

$$A\bar{X}_j = \sum_{k=1}^{n_B}\bar{X}_k\left(\bar{X}_k^TA\bar{X}_j\right) + T_j,$$

where $T_j \in \mathbb{R}^{m\times(t_j - t_{j-1})}$ is formed by projecting the columns of $A\bar{X}_j$ onto $\mathcal{L}^\perp$. Hence,

$$\begin{aligned}
\left\|b_{t_j}A\bar{X}_j - \sum_{k=1}^{n_B} b_{t_k}\bar{X}_k\left(\bar{X}_k^TA\bar{X}_j\right)\right\|_F^2 &= \sum_{k\neq j}(b_{t_j} - b_{t_k})^2\left\|\bar{X}_k\left(\bar{X}_k^TA\bar{X}_j\right)\right\|_F^2 + b_{t_j}^2\|T_j\|_F^2 \\
&= \Omega\left(\sum_{k\neq j}\left\|\bar{X}_k\left(\bar{X}_k^TA\bar{X}_j\right)\right\|_F^2 + \|T_j\|_F^2\right) \\
&= \Omega\left(\left\|A\bar{X}_j - \bar{X}_j\bar{X}_j^TA\bar{X}_j\right\|_F^2\right), \qquad (13)
\end{aligned}$$

where (13) follows from the fact that $b_{t_j} \neq b_{t_k}$ whenever $j \neq k$ and $b_{t_j} \neq 0$ since B is assumed to have full rank. By combining the above with (12), the proof is completed.

8.4. Proof of Proposition 5

Consider a fixed $j \in \{1, \ldots, n_B\}$. Let $\bar{x}_k$ be the k-th column of $\bar{X}_j$ and $(\bar{x}_k)_\alpha$ be the α-th entry of $\bar{x}_k$, where $k = 1, \ldots, t_j - t_{j-1}$ and $\alpha = 1, \ldots, m$. Suppose that $\mathrm{dist}(X, \mathcal{X}_h) = \left\|\bar{X} - E(h)\right\|_F = \tau$ for some $\tau \in (0, 1)$. Using the definition of $E(h)$ in (5), we have

$$(\bar{x}_k)_\alpha = \begin{cases} 1 + O(\tau) & \text{if } \alpha = \pi(k), \\ O(\tau) & \text{otherwise,} \end{cases}$$

where $\pi(k)$ is the coordinate of the k-th column of $\bar{E}_j(h)$ that equals 1. Since $\pi(k) \neq \pi(\ell)$ whenever $k \neq \ell$, it follows that

$$\bar{x}_k^TA\bar{x}_\ell = \begin{cases} a_{\pi(k)} + O(\tau) & \text{if } k = \ell, \\ O(\tau) & \text{otherwise.} \end{cases}$$

Now, let Δ_k be the k-th column of $A\bar{X}_j - \bar{X}_j\bar{X}_j^T A\bar{X}_j$, where $k = 1, \ldots, t_j - t_{j-1}$. Then,

$$\Delta_k = A\bar{x}_k - \sum_{\ell=1}^{t_j - t_{j-1}} \bar{x}_\ell \left(\bar{x}_\ell^T A \bar{x}_k\right) = \left(A - a_{\pi(k)} I_m\right) \bar{x}_k - O(\tau) \cdot \left(\sum_{\ell=1}^{t_j - t_{j-1}} \bar{x}_\ell \right).$$

Let $\Pi_{\mathcal{I}_j}$ be the projector onto the coordinates in $\mathcal{I}_j$. By Proposition 3 and the assumption that $\operatorname{dist}(X, \mathcal{X}_h) = \tau$, we have

$$\sum_{\ell=1}^{t_j - t_{j-1}} \left\|\Pi_{\mathcal{I}_j}(\bar{x}_\ell)\right\|_2^2 = \sum_{k \in \mathcal{I}_j} \left\|\left[\bar{X}_j\right]_k\right\|_2^2 = O(\tau).$$

Hence,

$$\begin{aligned} \left\|\Pi_{\mathcal{I}_j}(\Delta_k)\right\|_2 &\geq \left\|\Pi_{\mathcal{I}_j}\left(\left(A - a_{\pi(k)} I_m\right)\bar{x}_k\right)\right\|_2 - O(\tau) \cdot \left(\sum_{\ell=1}^{t_j - t_{j-1}} \left\|\Pi_{\mathcal{I}_j}(\bar{x}_\ell)\right\|_2 \right) \\ &\geq \left\|\Pi_{\mathcal{I}_j}\left(\left(A - a_{\pi(k)} I_m\right)\bar{x}_k\right)\right\|_2 - O(\tau^2). \end{aligned} \tag{14}$$

Let $i' \in \{0, 1, \ldots, n_A - 1\}$ be such that $s_{i'} + 1 \leq \pi(k) \leq s_{i'+1}$. Then, we have

$$\begin{aligned} \left\|\Pi_{\mathcal{I}_j}\left(\left(A - a_{\pi(k)} I_m\right)\bar{x}_k\right)\right\|_2^2 &= \sum_{i \neq i'} \sum_{\alpha \in \mathcal{I}_j \cap \{s_i+1, \ldots, s_{i+1}\}} \left(\left(a_{s_i+1} - a_{\pi(k)}\right)(\bar{x}_k)_\alpha\right)^2 \\ &= \sum_{i \neq i'} \sum_{\alpha \in \mathcal{I}_j \cap \{s_i+1, \ldots, s_{i+1}\}} \Omega\left((\bar{x}_k)_\alpha^2\right) \\ &= \Omega\left(\left\|\Pi_{\mathcal{I}_j}(\bar{x}_k)\right\|_2^2\right) - O\left(\left\|\Pi_{\mathcal{I}_j \cap \{s_{i'}+1, \ldots, s_{i'+1}\}}(\bar{x}_k)\right\|_2^2\right). \end{aligned} \tag{15}$$

To bound the term $\left\|\Pi_{\mathcal{I}_j \cap \{s_{i'}+1, \ldots, s_{i'+1}\}}(\bar{x}_k)\right\|_2^2$, we proceed as follows. Let $Y = X(Q^*)^T \in \operatorname{St}(m, n)$ and decompose it as

$$Y = \begin{bmatrix} Y_{11} & \cdots & Y_{1n_A} \\ \vdots & \ddots & \vdots \\ Y_{n_A 1} & \cdots & Y_{n_A n_A} \end{bmatrix},$$

where $Y_{ii} \in \mathbb{R}^{(s_i - s_{i-1}) \times h_i}$, for $i = 1, \ldots, n_A$. Observe that

$$\begin{aligned} \operatorname{dist}^2(X, \mathcal{X}_h) &= \min\left\{ \|Y - \operatorname{BlkDiag}(P_1, \ldots, P_{n_A}) \cdot E(h)\|_F^2 \;\middle|\; P_i \in \mathcal{O}^{s_i - s_{i-1}} \text{ for } i = 1, \ldots, n_A \right\} \\ &= \sum_{1 \leq i \neq j \leq n_A} \|Y_{ij}\|_F^2 + \sum_{i=1}^{n_A} \min\left\{ \left\| Y_{ii} - P_i \begin{bmatrix} I_{h_i} \\ \mathbf{0} \end{bmatrix} \right\|_F^2 \;\middle|\; P_i \in \mathcal{O}^{s_i - s_{i-1}} \right\}. \end{aligned} \tag{16}$$

The following lemma establishes a bound on the second term in (16):

Lemma 2. *For $i = 1, \ldots, n_A$, let*

$$v_i^* = \min\left\{ \left\| Y_{ii} - P_i \begin{bmatrix} I_{h_i} \\ \mathbf{0} \end{bmatrix} \right\|_F^2 \;\middle|\; P_i \in \mathcal{O}^{s_i - s_{i-1}} \right\}. \tag{17}$$

Suppose that $v_i^* < 1$. Then, we have

$$v_i^* = \Theta\left(\left\| \sum_{j \neq i} Y_{ji}^T Y_{ji} \right\|_F^2 \right).$$

Let us defer the proof of Lemma 2 to the end of this section. Together with (16), Lemma 2 implies that

$$\operatorname{dist}^2(X, \mathcal{X}_h) = \sum_{1 \le i \ne j \le n_A} \|Y_{ij}\|_F^2 + \sum_{i=1}^{n_A} \Theta\left(\left\| \sum_{j \ne i} Y_{ji}^T Y_{ji} \right\|_F^2 \right).$$

Since $\operatorname{dist}(X, \mathcal{X}_h) = \tau$ for some $\tau \in (0,1)$, we have $\sum_{1 \le i \ne j \le n_A} \|Y_{ij}\|_F^2 = O(\tau^2)$. This implies that for $i = 1, \dots, n_A$,

$$v_i^* = O\left(\left(\sum_{j \ne i} \|Y_{ji}\|_F^2 \right)^2 \right) = O(\tau^4)$$

Now, decompose $\bar{X} = (P^*)^T Y$ as

$$\begin{bmatrix} \bar{X}_{11} & \cdots & \bar{X}_{1 n_A} \\ \vdots & \ddots & \vdots \\ \bar{X}_{n_A 1} & \cdots & \bar{X}_{n_A n_A} \end{bmatrix},$$

where $\bar{X}_{ii} = (P_i^*)^T Y_{ii} \in \mathbb{R}^{(s_i - s_{i-1}) \times h_i}$ for $i = 1, \dots, n_A$. Note that for $i = 1, \dots, n_A$, we have

$$v_i^* = \left\| \bar{X}_{ii} - \begin{bmatrix} I_{h_i} \\ \mathbf{0} \end{bmatrix} \right\|_F^2.$$

Moreover, observe that $\Pi_{\mathcal{I}_j \cap \{s_{i'}+1, \dots, s_{i'+1}\}}(\bar{x}_k)$ is part of $\bar{X}_{i'+1, i'+1}$ and does not intersect the diagonal of the top $h_{i'+1} \times h_{i'+1}$ block of $\bar{X}_{i'+1, i'+1}$. Thus, by Lemma 2,

$$\left\| \Pi_{\mathcal{I}_j \cap \{s_{i'}+1, \dots, s_{i'+1}\}}(\bar{x}_k) \right\|_2^2 \le v_{i'+1}^* = O(\tau^4).$$

Together with (14) and (15), this yields

$$\left\| \Pi_{\mathcal{I}_j}(\Delta_k) \right\|_2^2 \ge \Omega\left(\left\| \Pi_{\mathcal{I}_j}(\bar{x}_k) \right\|_2^2 \right) - O(\tau^3).$$

It follows that

$$\begin{aligned} \left\| A\bar{X}_j - \bar{X}_j \bar{X}_j^T A \bar{X}_j \right\|_F^2 &= \sum_{k=1}^{t_j - t_{j-1}} \|\Delta_k\|_2^2 \\ &\ge \sum_{k=1}^{t_j - t_{j-1}} \left\| \Pi_{\mathcal{I}_j}(\Delta_k) \right\|_2^2 \\ &\ge \sum_{k=1}^{t_j - t_{j-1}} \Omega\left(\left\| \Pi_{\mathcal{I}_j}(\bar{x}_k) \right\|_2^2 \right) - O(\tau^3) \\ &= \sum_{k \in \mathcal{I}_j} \Omega\left(\left\| \left[\bar{X}_j \right]_k \right\|_2^2 \right) - O(\tau^3). \end{aligned}$$

Since $\operatorname{dist}(X, \mathcal{X}_h) = \tau$, upon summing both sides of the above inequality over $j = 1, \dots, n_B$ and using Proposition 3, we conclude that for sufficiently small $\tau \in (0,1)$,

$$\sum_{j=1}^{n_B} \left\| A\bar{X}_j - \bar{X}_j \bar{X}_j^T A \bar{X}_j \right\|_F^2 = \Omega\left(\operatorname{dist}(X, \mathcal{X}_h) \right) - O(\tau^3) = \Omega\left(\operatorname{dist}(X, \mathcal{X}_h) \right) = \sum_{j=1}^{n_B} \sum_{k \in \mathcal{I}_j} \Omega\left(\left\| \left[\bar{X}_j \right]_k \right\|_2^2 \right),$$

as desired.

To complete the proof, it remains to prove Lemma 2.

Proof of Lemma 2. Consider a fixed $i \in \{1, \ldots, n_A\}$. Note that Problem (17) is again an instance of the orthogonal Procrustes problem. Hence, by the result in (Schönemann, 1966), an optimal solution to Problem (17) is given by

$$P_i^* = H_i \begin{bmatrix} W_i^T & \mathbf{0} \\ \mathbf{0} & I_{s_i - s_{i-1} - h_i} \end{bmatrix},$$

where $Y_{ii} = H_i \begin{bmatrix} \Sigma_i \\ \mathbf{0} \end{bmatrix} W_i^T$ is a singular value decomposition of Y_{ii} with $H_i \in \mathcal{O}^{s_i - s_{i-1}}$, $W_i \in \mathcal{O}^{h_i}$, and $\Sigma_i \in \mathcal{S}^{h_i}$ being diagonal. It follows from (17) that

$$v_i^* = \left\| Y_{ii} - P_i^* \begin{bmatrix} I_{h_i} \\ \mathbf{0} \end{bmatrix} \right\|_F^2 = \|\Sigma_i - I_{h_i}\|_F^2.$$

Now, since $Y \in \mathrm{St}(m, n)$, we have

$$Y_{ii}^T Y_{ii} + \sum_{j \neq i} Y_{ji}^T Y_{ji} = W_i \Sigma_i^2 W_i^T + \sum_{j \neq i} Y_{ji}^T Y_{ji} = I_{h_i},$$

or equivalently,

$$\Sigma_i^2 + W_i^T \left(\sum_{j \neq i} Y_{ji}^T Y_{ji} \right) W_i = I_{h_i}.$$

By following the arguments in the proof of Lemma 1, we conclude that

$$\|\Sigma_i - I_{h_i}\|_F^2 = \Theta\left(\left\| \sum_{j \neq i} Y_{ji}^T Y_{ji} \right\|_F^2 \right),$$

as desired. □

8.5. Proof of Proposition 6

Observe that F, when viewed as a function on $\mathbb{R}^{m \times n}$, is continuously differentiable with Lipschitz continuous gradient. Thus, we have

$$|F(X) - F(X^*) - \langle \nabla F(X^*), X - X^* \rangle| \leq \frac{L}{2} \|X - X^*\|_F^2, \tag{18}$$

where $L > 0$ is the Lipschitz constant of ∇F; see, e.g., (Nesterov, 2004). Now, by Proposition 1, we have $\nabla F(X^*) = X^* \nabla F(X^*)^T X^*$. This implies that

$$\langle \nabla F(X^*), X - X^* \rangle = \left\langle X^* \nabla F(X^*)^T X^*, X - X^* \right\rangle = \left\langle \nabla F(X^*)^T X^*, (X^*)^T X - I_n \right\rangle. \tag{19}$$

On the other hand,

$$\begin{aligned} \left\langle \nabla F(X^*)^T X^*, I_n - X^T X^* \right\rangle &= \left\langle (X^*)^T \nabla F(X^*), (X^*)^T X^* - X^T X^* \right\rangle \\ &= \left\langle X^* \nabla F(X^*)^T X^*, X^* - X \right\rangle \\ &= -\langle \nabla F(X^*), X - X^* \rangle. \end{aligned} \tag{20}$$

Upon adding (19) and (20) and using the fact that $(X - X^*)^T (X - X^*) = 2I_n - (X^*)^T X - X^T X^*$, we obtain

$$2\langle \nabla F(X^*), X - X^* \rangle = -\left\langle \nabla F(X^*)^T X^*, (X - X^*)^T (X - X^*) \right\rangle.$$

Together with the fact that $\|AB\|_F \leq \|A\| \cdot \|B\|_F$ for any matrices A, B, where $\|\cdot\|$ denotes the spectral norm, this gives

$$\begin{aligned} |\langle \nabla F(X^*), X - X^* \rangle| &\leq \frac{1}{2} \left\| \nabla F(X^*)^T X^* \right\|_F \|X - X^*\|_F^2 \\ &= \left\| B (X^*)^T A X^* \right\|_F \|X - X^*\|_F^2 \\ &\leq \|A\|_F \|B\|_F \|X - X^*\|_F^2. \end{aligned}$$

By combining this with (18), we obtain the desired inequality with $\eta = (L/2) + \|A\|_F \|B\|_F$.

8.6. Proof of Proposition 7

To prove (i), observe that since $\xi = -2(I - XX^T)AX$, we have

$$[U_A^T\xi]_1 = -2\Sigma_{A,1}[U_A^TX]_1 + 2[U_A^TX]_1\left([U_A^TX]_1^T\Sigma_{A,1}[U_A^TX]_1 + [U_A^TX]_2^T\Sigma_{A,2}[U_A^TX]_2\right),$$

where $\Sigma_A = \mathrm{BlkDiag}(\Sigma_{A,1}, \Sigma_{A,2})$ with $\Sigma_{A,1} \in \mathcal{S}^{m-n}$ and $\Sigma_{A,2} \in \mathcal{S}^n$ being diagonal. Moreover,

$$[U_A^TX]_1^T[U_A^TX]_1 + [U_A^TX]_2^T[U_A^TX]_2 = I_n.$$

Hence,

$$\begin{aligned}
\tilde{g}'(0) &= 2\cdot\mathrm{tr}\left([U_A^TX]_1^T[U_A^T\xi]_1\right)\\
&= -4\cdot\mathrm{tr}\left(\left(I - [U_A^TX]_1^T[U_A^TX]_1\right)[U_A^TX]_1^T\Sigma_{A,1}[U_A^TX]_1\right) + 4\cdot\mathrm{tr}\left([U_A^TX]_1^T[U_A^TX]_1[U_A^TX]_2^T\Sigma_{A,2}[U_A^TX]_2\right)\\
&= -4\cdot\mathrm{tr}\left([U_A^TX]_2^T[U_A^TX]_2[U_A^TX]_1^T\Sigma_{A,1}[U_A^TX]_1\right) + 4\cdot\mathrm{tr}\left([U_A^TX]_1^T[U_A^TX]_1[U_A^TX]_2^T\Sigma_{A,2}[U_A^TX]_2\right)\\
&\le -4(a_{m-n} - a_{m-n+1})\,\mathrm{tr}\left([U_A^TX]_1^T[U_A^TX]_1[U_A^TX]_2^T[U_A^TX]_2\right)\\
&\le -4\lambda\left(g(X) - g^2(X)\right),
\end{aligned}$$

as desired.

To prove (ii), we compute

$$\begin{aligned}
\tilde{g}''(\alpha) = {}& 2\cdot\mathrm{tr}\left([U_A^T\xi]_1^T[U_A^T\xi]_1(I_n + \alpha^2\xi^T\xi)^{-1}\right)\\
&- 2\cdot\mathrm{tr}\left(\left([U_A^TX]_1^T[U_A^TX]_1 + 3\alpha([U_A^TX]_1^T[U_A^T\xi]_1 + [U_A^T\xi]_1^T[U_A^TX]_1) + 5\alpha^2[U_A^T\xi]_1^T[U_A^T\xi]_1\right)\xi^T\xi(I_n + \alpha^2\xi^T\xi)^{-2}\right)\\
&+ 8\alpha^2\,\mathrm{tr}\left(([U_A^TX]_1 + \alpha[U_A^T\xi]_1)^T([U_A^TX]_1 + \alpha[U_A^T\xi]_1)(\xi^T\xi)^2(I_n + \alpha^2\xi^T\xi)^{-3}\right).
\end{aligned}$$

Since $I_n + \alpha^2\xi^T\xi \succeq I_n$, we have $\mathbf{0} \preceq (I_n + \alpha^2\xi^T\xi)^{-1} \preceq I_n$. This, together with the fact that $[U_A^T\xi]_1^T[U_A^T\xi]_1 \succeq \mathbf{0}$ and $[U_A^T\xi]_1^T[U_A^T\xi]_1 \preceq \xi^T\xi$, implies that

$$\mathrm{tr}\left([U_A^T\xi]_1^T[U_A^T\xi]_1(I_n + \alpha^2\xi^T\xi)^{-1}\right) \le \mathrm{tr}\left([U_A^T\xi]_1^T[U_A^T\xi]_1\right) \le \|\xi\|_F^2. \tag{21}$$

Next, using the fact that $\|AB\|_F \le \|A\|\cdot\|B\|_F$ for any matrices A, B, we bound

$$\left\|[U_A^TX]_1^T[U_A^TX]_1\right\|_F \le \left\|[U_A^TX]_1\right\|\cdot\left\|[U_A^TX]_1\right\|_F \le \sqrt{n}, \tag{22}$$

$$\left\|[U_A^TX]_1^T[U_A^T\xi]_1\right\|_F \le \left\|[U_A^TX]_1\right\|\cdot\left\|[U_A^T\xi]_1\right\|_F \le \|\xi\|_F. \tag{23}$$

Moreover, we have

$$\left\|[U_A^T\xi]_1^T[U_A^T\xi]_1\right\|_F \le \left\|[U_A^T\xi]_1\right\|_F^2 \le \|\xi\|_F^2. \tag{24}$$

Since $\mathbf{0} \preceq (I_n + \alpha^2\xi^T\xi)^{-2} \preceq I_n$, it follows from (22)–(24) that

$$\begin{aligned}
&\left|\mathrm{tr}\left(\left([U_A^TX]_1^T[U_A^TX]_1 + 3\alpha([U_A^TX]_1^T[U_A^T\xi]_1 + [U_A^T\xi]_1^T[U_A^TX]_1) + 5\alpha^2[U_A^T\xi]_1^T[U_A^T\xi]_1\right)\xi^T\xi(I_n + \alpha^2\xi^T\xi)^{-2}\right)\right|\\
\le\ &\left\|[U_A^TX]_1^T[U_A^TX]_1 + 3\alpha([U_A^TX]_1^T[U_A^T\xi]_1 + [U_A^T\xi]_1^T[U_A^TX]_1) + 5\alpha^2[U_A^T\xi]_1^T[U_A^T\xi]_1\right\|_F\left\|\xi^T\xi(I_n + \alpha^2\xi^T\xi)^{-2}\right\|_F\\
\le\ &\|\xi\|_F^2\left(\sqrt{n} + 6\alpha\|\xi\|_F + 5\alpha^2\|\xi\|_F^2\right).
\end{aligned} \tag{25}$$

Lastly, using similar techniques as above, we bound

$$\mathrm{tr}\left(([U_A^TX]_1 + \alpha[U_A^T\xi]_1)^T([U_A^TX]_1 + \alpha[U_A^T\xi]_1)(\xi^T\xi)^2(I_n + \alpha^2\xi^T\xi)^{-3}\right) \le \|\xi\|_F^4\left(\sqrt{n} + 2\alpha\|\xi\|_F + \alpha^2\|\xi\|_F^2\right). \tag{26}$$

It follows from (21), (25), and (26) that

$$|\tilde{g}''(\alpha)| \le 2\|\xi\|_F^2 + 2\|\xi\|_F^2\left(\sqrt{n} + 6\alpha\|\xi\|_F + 5\alpha^2\|\xi\|_F^2\right) + 8\alpha^2\|\xi\|_F^4\left(\sqrt{n} + 2\alpha\|\xi\|_F + \alpha^2\|\xi\|_F^2\right).$$

Recall that $\xi = -2(I - XX^T)AX$. Hence, we have $\|\xi\|_F^2 \le 4\|AX\|_F^2 \le 4n\|A\|^2$. In particular, for $\alpha \in (0, (2\sqrt{n}\|A\|)^{-1}]$, we have $|\tilde{g}''(\alpha)| \le 8n\|A\|^2(5\sqrt{n} + 24)$. This completes the proof.