Quadratic Optimization with Orthogonality Constraint: Explicit Łojasiewicz Exponent and Linear Convergence of Retraction-Based Line-Search and Stochastic Variance-Reduced Gradient Methods[*]

Huikang Liu[†] Anthony Man-Cho So[‡] Weijie Wu[§]

April 30, 2018

Abstract

The problem of optimizing a quadratic form over an orthogonality constraint (QP-OC for short) is one of the most fundamental matrix optimization problems and arises in many applications. In this paper, we characterize the growth behavior of the objective function around the critical points of the QP-OC problem and demonstrate how such characterization can be used to obtain strong convergence rate results for iterative methods that exploit the manifold structure of the orthogonality constraint (i.e., the Stiefel manifold) to find a critical point of the problem. Specifically, our primary contribution is to show that the Łojasiewicz exponent at any critical point of the QP-OC problem is 1/2. Such a result is significant, as it expands the currently very limited repertoire of optimization problems for which the Łojasiewicz exponent is explicitly known. Moreover, it allows us to show, in a unified manner and for the first time, that a large family of retraction-based line-search methods will converge linearly to a critical point of the QP-OC problem. Then, as our secondary contribution, we propose a stochastic variance-reduced gradient (SVRG) method called Stiefel-SVRG for solving the QP-OC problem and present a novel Łojasiewicz inequality-based linear convergence analysis of the method. An important feature of Stiefel-SVRG is that it allows for general retractions and does not require the computation of any vector transport on the Stiefel manifold. As such, it is computationally more advantageous than other recently-proposed SVRG-type algorithms for manifold optimization.

[*] A preliminary version of this work has appeared in the Proceedings of the 33rd International Conference on Machine Learning (ICML 2016), 2016 [25]. This research is supported in part by the Hong Kong Research Grants Council (RGC) General Research Fund (GRF) Projects CUHK 14205314, CUHK 14206814, and CUHK 14208117.

[†] Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N. T., Hong Kong. E-mail: `hkliu@se.cuhk.edu.hk`

[‡] Department of Systems Engineering and Engineering Management, and, by courtesy, CUHK-BGI Innovation Institute of Trans-omics, The Chinese University of Hong Kong, Shatin, N. T., Hong Kong. E-mail: `manchoso@se.cuhk.edu.hk`

[§] Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N. T., Hong Kong. E-mail: `janewunju@gmail.com`

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 [44], data mining [22], dynamical systems [35], multivariate statistical analysis [6, 57], numerical linear algebra [36], and signal processing [31, 1], just to mention a few. Perhaps the simplest form of such problems is

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

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 $A \in \mathbb{R}^{m\times m}$, $B \in \mathbb{R}^{n\times n}$ are given symmetric matrices. One approach to tackling Problem (QP-OC) is to exploit the manifold structure of the constraint set $\mathrm{St}(m,n)$ and use retraction-based line-search methods. Roughly speaking, these methods generate iterates according to the formula

$$X^{k+1} = R\left(X^k, \alpha_k \xi^k\right) \quad \text{for } k = 0, 1, \ldots, \tag{1}$$

where $\alpha_k \geq 0$ is the step size, ξ^k is a descent 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 $\mathrm{St}(m,n)$. By definition, all the iterates produced by (1) are feasible for Problem (QP-OC). However, the choice of step sizes $\{\alpha_k\}_{k\geq 0}$, search directions $\{\xi^k\}_{k\geq 0}$, and the retraction R 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; see, e.g., [1, 3, 4, 55, 19] and the references therein. Although most of the resulting methods are known to be convergent, very little is known about their *convergence rates*, even when they are applied to the much more structured problem (QP-OC). In some early works (see, e.g., [43, 53, 56]), several line-search methods for optimizing a smooth function over a Riemannian manifold were considered. It was shown that if the methods in question converge to a *non-degenerate* critical point, then the convergence rate is linear. However, a non-degenerate critical point is necessarily isolated. For general instances of Problem (QP-OC), such a critical point may not exist. Even if it does, it is generally impossible to ensure *a priori* that the aforementioned methods will converge to one. Therefore, the linear convergence results obtained for Problem (QP-OC) via this line of argument are not entirely satisfactory. Later, Absil et al. [3, Theorem 4.6.3] 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. However, it is not clear how to extend this result to cover 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, and 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 growth properties and then using such properties to establish the convergence rate of a properly initialized iterative method. This approach has been employed to study a wide variety of structured non-convex optimization problems, including dictionary learning [5, 49, 50], matrix completion and sensing [16, 60, 51], phase retrieval [10, 34, 48], phase synchronization [26, 61], and tensor decomposition and factorization [18, 52]. In the context of Problem (QP-OC), such an approach was first pursued by Shamir [41, 42], 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 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 top 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, it is not clear whether the assumptions on A 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., [2, 32, 38]. Roughly speaking, a Łojasiewicz inequality holds at a point if the growth of the objective function around that point can be bounded by a certain 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 that is related to various forms of error bounds (see, e.g., [7, 24]) – the latter have featured prominently in the convergence rate analysis of iterative methods (see, e.g., [30, 27, 17, 63, 8, 24, 26, 46, 62]). Moreover, it plays a fundamental role in understanding the asymptotic behavior of discrete and continuous dynamical systems; see, e.g., [7, 32, 13] and the references therein. 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 [38] for details of these results.) Compared with the first approach, this second, Łojasiewicz inequality-based approach provides insights into the behavior of the objective function around not just the globally optimal solutions but also the critical points, thus opening 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 certain structured convex optimization problems [24], non-convex quadratic optimization problems with simple convex constraints (such as a ball or a polyhedron) [28, 29, 14, 24], and general polynomial optimization problems [23]. However, these three classes of results do not shed any light on Problem (QP-OC), as the first two apply only to problems with some convexity properties, while the third gives estimates that depend on the dimensions of the problem and lead to very weak convergence rate guarantees.

In view of the above discussion, our goal in this paper is to characterize the growth behavior of the objective function of Problem (QP-OC) around *all* critical points and demonstrate how such characterization can be used to gain a better understanding of the convergence rates of a wide range of iterative methods for solving Problem (QP-OC). Our primary contribution is to show that all critical points of Problem (QP-OC) have the same Łojasiewicz exponent and to determine its exact value. Such a result is significant, as it expands the currently very limited repertoire of optimization problems for which the Łojasiewicz exponent is known and contributes to the growing literature on the geometry of structured non-convex optimization problems. A crucial step in our analysis is to establish a local Lipschitzian error bound for the *non-convex* set of critical points of Problem (QP-OC). Once such error bound is available, it is rather straightforward to derive the value of the Łojasiewicz exponent. We should point out that the aforementioned error bound is considerably more difficult to establish than those in [30, 63, 46, 62], as neither the objective function nor the constraint of Problem (QP-OC) is convex. Thus, it could be of independent interest.

After characterizing the Łojasiewicz exponent at the critical points of Problem (QP-OC), we proceed to study the algorithmic consequences of such characterization. To begin, we specialize the convergence analysis framework in [38] to our setting and show, in a unified manner and for the first time, that various retraction-based line-search methods for solving Problem (QP-OC) will converge linearly to a critical point. Our linear convergence result does not require any assumptions on A and B. In particular, it holds for *all* instances of Problem (QP-OC), even for those whose critical points are not isolated. Thus, it yields a qualitative improvement upon the results in [3, 41, 42]. Furthermore, we provide a quantitative description of how the convergence rate depends on the retraction used, which, to the best of our knowledge, is new. Next, we explore the possibility of analyzing the convergence rates of iterative methods that do not fall under the framework in [38]. This is motivated by the fact that the framework in [38] only covers *deterministic descent* methods, which potentially precludes many efficient iterative schemes; cf. [46] in the context of unconstrained smooth convex minimization. As our secondary contribution, we propose a stochastic variance-reduced gradient (SVRG) method called Stiefel-SVRG for solving Problem (QP-OC) and present a novel Łojasiewicz inequality-based linear convergence analysis of the method. Compared with other recently-proposed SVRG-type algorithms for manifold optimization (such as those in [58, 37]), Stiefel-SVRG is computationally more advantageous, as it allows for more general retractions and does not require the computation of any vector transport on the Stiefel manifold. Moreover, our convergence analysis of Stiefel-SVRG is much stronger than those for other manifold SVRG-type algorithms in prior works, thanks to our characterization of the Łojasiewicz exponent for Problem (QP-OC).

The rest of this paper is organized as follows. In Section 2, we review some basic concepts in manifold optimization that are essential to our study and introduce the Łojasiewicz inequality for Problem (QP-OC). Then, in Section 3, we characterize the Łojasiewicz exponent at all critical points of Problem (QP-OC). In Section 4, we consider different retraction-based methods for solving Problem (QP-OC) and show how our main result implies their linear convergence. Lastly, we end with some closing remarks in Section 5.

Besides the notations introduced earlier, we shall use $\mathcal{S}^n$ to denote the set of $n \times n$ symmetric matrices; $\mathcal{O}^n$ to denote the set of $n \times n$ orthogonal matrices (in particular, we have $\mathcal{O}^n = \mathrm{St}(n, n)$); $\mathcal{P}^n$ to denote the set of $n \times n$ permutation matrices; $\mathrm{Diag}(x_1, \ldots, x_n)$ to denote the diagonal matrix with $x_1, \ldots, x_n$ on the diagonal; $\mathrm{BlkDiag}(Y_1, \ldots, Y_n)$ to denote the block diagonal matrix whose diagonal blocks are $Y_1, \ldots, Y_n$. 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 $\|Y\|$, $\|Y\|_F$, $\|Y\|_*$ to denote the spectral norm, Frobenius norm, Schatten 1-norm (also known as nuclear norm) of Y, respectively, and $\operatorname{dist}(Y,\mathcal{X}) = \min_{X\in\mathcal{X}} \|X-Y\|_F$ to denote the Euclidean distance of Y to $\mathcal{X}$. Other notations are standard.

2 Preliminaries

Let us begin with 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 $\langle X,Y\rangle = \operatorname{tr}\left(X^TY\right)$. 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 $F(X) = \operatorname{tr}\left(X^TAXB\right)$ is $\nabla F(X) = 2AXB$. Its orthogonal projection onto $T(X)$, called the *projected gradient* of $F(X)$ and denoted by $\operatorname{grad} F(X)$, can be calculated as

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

see, e.g., [3, Example 3.6.2]. The set of *critical points* of Problem (QP-OC) is then defined as

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

For our subsequent development, the following alternative descriptions of $\mathcal{X}$ will be useful:

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

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

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

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

(iv) There exists a $\rho\neq 0$ such that $D_\rho(X) = \mathbf{0}$.

Proof Observe that

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

Thus, the equivalence between (i) and (ii) follows from the invertibility of $I_m - (1/2)XX^T$.

Next, suppose that (ii) holds. Since $X\in\mathrm{St}(m,n)$, we have $X^T\nabla F(X) = \nabla F(X)^TX$. It follows that for any $\rho\in\mathbb{R}$,

$$D_\rho(X) = \nabla F(X) - X\left(2\rho\nabla F(X)^TX + (1-2\rho)\nabla F(X)^TX\right) = \mathbf{0};$$

i.e., (iii) holds. Conversely, suppose that (iii) holds. By taking $\rho = 1/2$, we have $D_{1/2}(X) = \nabla F(X) - X\nabla F(X)^TX = \mathbf{0}$; i.e., (ii) holds. We remark that in [19, Lemma 2.1], it is mentioned that (ii) is equivalent to $D_\rho(X) = \mathbf{0}$ for all $\rho > 0$. However, as the proof above shows, the equivalence actually holds for any $\rho\in\mathbb{R}$.

Lastly, observe that

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

and

$$\left\langle (I_m - XX^T)\nabla F(X), X\left(\nabla F(X)^T X - X^T\nabla F(X)\right)\right\rangle = 0.$$

Hence, if $\rho \neq 0$, then $D_\rho(X) = \mathbf{0}$ if and only if $(I_m - XX^T)\nabla F(X) = X\left(\nabla F(X)^T X - X^T \nabla F(X)\right) = \mathbf{0}$. This establishes the equivalence between (iii) and (iv). □

While Proposition 1 provides different characterizations of $\mathcal{X}$, a deep and far-reaching result of Łojasiewicz provides a way to characterize the growth behavior of F around each critical point in $\mathcal{X}$. Specifically, for each $X^* \in \mathcal{X}$, there exist $\delta, \eta > 0$ and $\theta \in (0, 1/2]$ such that

$$|F(X) - F(X^*)|^{1-\theta} \leq \eta \cdot \|\operatorname{grad} F(X)\|_F \tag{3}$$

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 [38, Section 2.2]. In particular, the inequality (3), known as the *Łojasiewicz inequality* for Problem (QP-OC), shows that the growth of F around X^* can be controlled by $\|\operatorname{grad} F\|_F$. Furthermore, Proposition 1 suggests that the said growth can also be controlled by $\|D_\rho\|_F$. Indeed, for any given $\rho \neq 0$, the inequality (3) can be equivalently formulated as

$$|F(X) - F(X^*)|^{1-\theta} \leq \bar{\eta} \cdot \|D_\rho(X)\|_F \tag{4}$$

for some $\bar{\eta} > 0$. This is a simple consequence of the following result:

Proposition 2 *Let $X \in \mathrm{St}(m,n)$ and $\rho \neq 0$ be given. Then,*

$$\frac{1}{2}\min\left\{1, \frac{1}{2|\rho|}\right\} \cdot \|D_\rho(X)\|_F \leq \|\operatorname{grad} F(X)\|_F \leq \max\left\{1, \frac{1}{2|\rho|}\right\} \cdot \|D_\rho(X)\|_F.$$

Proof Let $C_\rho(X) = I_m - (1-2\rho)XX^T$. It is easy to verify that

$$D_\rho(X) = C_\rho(X)\left(\nabla F(X) - X\nabla F(X)^T X\right). \tag{5}$$

Moreover, since the eigenvalues of XX^T are either 0 or 1, we see that $C_\rho(X)$ is invertible for any $\rho \neq 0$. It follows from (2) that for any $\rho \neq 0$,

$$\operatorname{grad} F(X) = \left(I_m - \frac{1}{2}XX^T\right) C_\rho(X)^{-1} D_\rho(X).$$

In particular, we have

$$\|\operatorname{grad} F(X)\|_F = \left\|\left(I_m - \frac{1}{2}XX^T\right) C_\rho(X)^{-1} D_\rho(X)\right\|_F \leq \max\left\{1, \frac{1}{2|\rho|}\right\} \cdot \|D_\rho(X)\|_F$$

and

$$\|D_\rho(X)\|_F = \left\|C_\rho(X)\left(I_m - \frac{1}{2}XX^T\right)^{-1} \operatorname{grad} F(X)\right\|_F \leq 2\max\{1, 2|\rho|\} \cdot \|\operatorname{grad} F(X)\|_F.$$

This completes the proof. □

It is well known (see, e.g., [38]) that the Łojasiewicz inequality (3) implies the sublinear (resp. linear) convergence of a host of retraction-based line-search methods 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 not known. Thus, it is natural to ask whether one can give a good estimate of θ. As is evident from our prior discussion, such an estimate will be of interest to both mathematical analysis and numerical optimization communities.

3 Characterizing the Łojasiewicz Exponent for Problem (QP-OC)

In view of Proposition 2 and the remarks preceding it, to address the above question, it suffices to focus on formulation (4) of the Łojasiewicz inequality for Problem (QP-OC) with $\rho > 0$. Thus, let $\rho > 0$ be fixed throughout this section. The main contribution of this paper is the following theorem:

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

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

Theorem 1 is significant because it not only reveals that the constants δ, η, θ in (4) 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. We remark that the value 1/2 is the best possible, as the Łojasiewicz exponent θ satisfies $\theta \in (0, 1/2]$; see [38, Footnote 2] and (4). 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 $\delta \in (0,1)$ and $\eta > 0$ such that for all $X \in \mathrm{St}(m,n)$ with $\mathrm{dist}(X, \mathcal{X}) \leq \delta$,*

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

The error bound in Theorem 2 is reminiscent of those that have appeared in the recent literature (see, e.g., [63, 26, 46, 62] and the references therein). However, the particular structure of Problem (QP-OC) gives rise to a number of analysis challenges that are not present in prior studies. As such, some new ideas are needed to establish Theorem 2.

Although Theorem 2 is a local result in the sense that it only holds for points lying in a neighborhood of the set of critical points $\mathcal{X}$, it can be globalized with minimal extra effort. Specifically, we have the following corollary, which will come in handy when we study the convergence behavior of Stiefel-SVRG in Section 4.2:

Corollary 1 *(Global Error Bound for Problem* (QP-OC)*) There exists an $\bar{\eta} > 0$ such that for all $X \in \mathrm{St}(m,n)$,*

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

Proof Let $\delta \in (0,1)$ and $\eta > 0$ be the constants given in Theorem 2. Consider the set

$$\mathcal{Y} = \{Y \in \mathrm{St}(m,n) \mid \mathrm{dist}(Y, \mathcal{X}) \geq \delta\}.$$

By the definition of $\mathcal{X}$ and Proposition 1, we have $\|D_\rho(X)\|_F > 0$ for all $X \in \mathcal{Y}$. Hence, by the continuity of $X \mapsto \|D_\rho(X)\|_F$ and the compactness of $\mathcal{Y}$, there exists an $\ell > 0$ such that $\|D_\rho(X)\|_F \geq \ell$ for all $X \in \mathcal{Y}$. On the other hand, we have $\|X - Y\|_F \leq 2\sqrt{n}$ for all $X, Y \in \mathrm{St}(m,n)$. It follows that

$$\mathrm{dist}(X, \mathcal{X}) \leq 2\sqrt{n} \leq \frac{2\sqrt{n}}{\ell} \cdot \|D_\rho(X)\|_F$$

for all $X \in \mathcal{Y}$. This, together with Theorem 2, implies the desired result with $\bar{\eta} = \max\{\eta, 2\sqrt{n}/\ell\}$. □

The proofs of Theorems 1 and 2 will be given in Sections 3.2 and 3.1, respectively. As the proofs are quite technical and tedious, readers who are more interested in the algorithmic consequences of Theorem 1 can skip ahead to Section 4.

3.1 Proof of Theorem 2

3.1.1 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 $\mathrm{tr}\left(X^T AXB\right) = \mathrm{tr}\left(\bar{X}^T \Sigma_A \bar{X} \Sigma_B\right)$, where $\bar{X} = U_A^T X U_B \in \mathrm{St}(m,n)$. Thus, we may assume without loss of generality that

$$A = \mathrm{Diag}(a_1, \ldots, a_m) \in \mathcal{S}^m \quad \text{and} \quad B = \mathrm{Diag}(b_1, \ldots, b_n) \in \mathcal{S}^n, \tag{6}$$

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

$$\mathcal{X} = \left\{X \in \mathrm{St}(m,n) \mid AXB - XBX^T AX = \mathbf{0}\right\}. \tag{7}$$

Since

$$\left\|\nabla F(X) - X\nabla F(X)^T X\right\|_F = \|C_\rho(X)^{-1} D_\rho(X)\|_F \leq \max\left\{1, \frac{1}{2\rho}\right\} \cdot \|D_\rho(X)\|_F$$

by (5) and $\nabla F(X) = 2AXB$, in order to prove Theorem 2, it suffices to prove the following:

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

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

3.1.2 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. We assume that $n_A \geq 2$ (for otherwise the objective function is identically constant and Theorems 1 and 2 will be trivial because then every $X \in \mathrm{St}(m,n)$ is stationary) and $n_B \geq 1$. 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

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

In particular, we see from (6) that $A \in \mathcal{S}^m$ and $B \in \mathcal{S}^n$ can be expressed as

$$A = \mathrm{BlkDiag}\left(a_{s_1} I_{s_1-s_0}, \ldots, a_{s_{n_A}} I_{s_{n_A}-s_{n_A-1}}\right), \tag{8}$$

$$B = \mathrm{BlkDiag}\left(b_{t_1} I_{t_1-t_0}, \ldots, b_{t_{n_B}} I_{t_{n_B}-t_{n_B-1}}\right), \tag{9}$$

respectively. Now, 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 $j = 1, \ldots, n_B$. Define

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

In addition, given any $h = (h_1, \ldots, h_{n_A}) \in \mathcal{H}$, define

$$\begin{aligned} E_i(h) &= [e_{s_{i-1}+1} \cdots e_{s_{i-1}+h_i}] \in \mathbb{R}^{m \times h_i} \quad \text{for } i = 1, \ldots, n_A, \\ E(h) &= [E_1(h) \cdots E_{n_A}(h)] \in \mathbb{R}^{m \times n}, \end{aligned} \tag{10}$$

where $\{e_i\}_{i=1}^m$ is the standard basis of $\mathbb{R}^m$. Informally, we are going to choose h_i eigenvectors from the eigenspace U_i of A and the matrix $E_i(h)$ is used to extract those eigenvectors from U_i, where $i = 1, \ldots, n_A$. When written explicitly, the matrix $E(h)$ takes the form

$$E(h) = \left[\begin{array}{c|c|c|c} \begin{matrix} I_{h_1} \\ \mathbf{0}_{(s_1-s_0-h_1)\times h_1} \end{matrix} & & & \\ \hline & \begin{matrix} I_{h_2} \\ \mathbf{0}_{(s_2-s_1-h_2)\times h_2} \end{matrix} & & \\ \hline & & \ddots & \\ \hline & & & \begin{matrix} I_{h_{n_A}} \\ \mathbf{0}_{(s_{n_A}-s_{n_A-1}-h_{n_A})\times h_{n_A}} \end{matrix} \end{array}\right]. \tag{11}$$

Using the above definitions, we can characterize the set of critical points of Problem (QP-OC) as follows:

Proposition 3 *Every $X \in \mathcal{X}$ can be expressed as*

$$X = \mathrm{BlkDiag}(P_1, \ldots, P_{n_A}) \cdot E(h) \cdot \Pi \cdot \mathrm{BlkDiag}\left(Q_1^T, \ldots, Q_{n_B}^T\right) \tag{12}$$

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

Before we prove Proposition 3, some remarks are in order.

Remark 1

(a) *Essentially, Proposition 3 states that every $X \in \mathcal{X}$ can be factorized as $X = PQ^T$, where $P \in \mathrm{St}(m,n)$, $Q \in \mathcal{O}^n$, and the columns of P (resp. Q) are eigenvectors of A (resp. B). Indeed, observe that for $i = 1,\dots,n_A$, the $(s_{i-1}+1)$-st to s_i-th columns of $\mathrm{BlkDiag}(P_1,\dots,P_{n_A})$ form an orthonormal basis of U_i. Similarly, for $j = 1,\dots,n_B$, the $(t_{j-1}+1)$-st to t_j-th columns of $\mathrm{BlkDiag}(Q_1,\dots,Q_{n_B})$ form an orthonormal basis of V_j. The matrix $E(h)$ extracts n columns from $\mathrm{BlkDiag}(P_1,\dots,P_{n_A})$ to form P.*

(b) *A result similar to Proposition 3 has appeared in [3, Section 4.8.2]. However, it assumes that the diagonal entries of B are all distinct.[1] By contrast, Proposition 3 does not make any assumption on A and B.*

Proof Let $X \in \mathcal{X}$ be arbitrary. Using (7) 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 $j = 1,\dots,n_B$, such that the columns of $\mathrm{BlkDiag}(Q_1,\dots,Q_{n_B})$ are eigenvectors of B, and that

$$X^TAX = \mathrm{BlkDiag}\left(Q_1\Sigma_1Q_1^T,\dots,Q_{n_B}\Sigma_{n_B}Q_{n_B}^T\right). \tag{13}$$

Now, by (7) and the commutativity of X^TAX and B, 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

$$Y = X \cdot \mathrm{BlkDiag}\left(Q_1,\dots,Q_{n_B}\right) \in \mathrm{St}(m,n) \tag{14}$$

and using (13), we obtain $AY = Y \cdot \mathrm{BlkDiag}(\Sigma_1,\dots,\Sigma_{n_B})$. As $\Sigma_1,\dots,\Sigma_{n_B}$ are diagonal, this implies that each of the n columns of Y is an eigenvector of A. In view of the structure of A in (8) and Remark 1(a), we see that up to a permutation of the columns, Y takes the form $\mathrm{BlkDiag}\left(P_1,\dots,P_{n_A}\right) \cdot E(h)$ for some $P_i \in \mathcal{O}^{s_i - s_{i-1}}$, where $i = 1,\dots,n_A$. This, together with (14), yields the expression on the right-hand side of (12).

Conversely, suppose that X takes the form (12) for some $P_i \in \mathcal{O}^{s_i - s_{i-1}}$ $(i = 1,\dots,n_A)$, $Q_j \in \mathcal{O}^{t_j - t_{j-1}}$ $(j = 1,\dots,n_B)$, $h \in \mathcal{H}$, and $\Pi \in \mathcal{P}^n$. Define $Y = \mathrm{BlkDiag}(P_1,\dots,P_{n_A}) \cdot E(h) \cdot \Pi \in \mathrm{St}(m,n)$. Observe that the columns of Y are eigenvectors of A. Hence, we have $AY = Y \cdot \mathrm{Diag}(\lambda_1,\dots,\lambda_n)$ for some $\lambda_1,\dots,\lambda_n \in \mathbb{R}$. Using this and the fact that $X = Y \cdot \mathrm{BlkDiag}\left(Q_1^T,\dots,Q_{n_B}^T\right)$ and $Y^TY = I_n$, we obtain

$$\begin{aligned} X^TAX &= \mathrm{BlkDiag}\left(Q_1,\dots,Q_{n_B}\right) \cdot Y^TAY \cdot \mathrm{BlkDiag}\left(Q_1^T,\dots,Q_{n_B}^T\right) \\ &= \mathrm{BlkDiag}\left(Q_1,\dots,Q_{n_B}\right) \cdot \mathrm{Diag}(\lambda_1,\dots,\lambda_n) \cdot \mathrm{BlkDiag}\left(Q_1^T,\dots,Q_{n_B}^T\right) \end{aligned} \tag{15}$$

and

$$\begin{aligned} AX &= AY \cdot \mathrm{BlkDiag}\left(Q_1^T,\dots,Q_{n_B}^T\right) \\ &= X \cdot \mathrm{BlkDiag}\left(Q_1,\dots,Q_{n_B}\right) \cdot \mathrm{Diag}(\lambda_1,\dots,\lambda_n) \cdot \mathrm{BlkDiag}\left(Q_1^T,\dots,Q_{n_B}^T\right) \\ &= XX^TAX. \end{aligned} \tag{16}$$

[1] Such an assumption is omitted in the original text of [3] but is needed for the result in [3, Section 4.8.2] to hold. The omission is corrected in the online errata at `https://sites.uclouvain.be/absil/amsbook/errata.html`.

In particular, since the columns of $\mathrm{BlkDiag}(Q_1,\dots,Q_{n_B})$ are eigenvectors of B, we see from (15) that X^TAX and B are simultaneously diagonalizable. Consequently, we have $X^TAXB = BX^TAX$, which together with (16) implies that $\mathbf{0} = AXB - XX^TAXB = AXB - XBX^TAX$, or equivalently, $X \in \mathcal{X}$, as desired. $\square$

Proposition 3 suggests that we can express $\mathcal{X}$ as

$$\mathcal{X} = \bigcup_{h\in\mathcal{H},\,\Pi\in\mathcal{P}^n} \mathcal{X}_{h,\Pi},$$

where every $X \in \mathcal{X}_{h,\Pi}$ takes the form

$$X = \mathrm{BlkDiag}(P_1,\dots,P_{n_A}) \cdot E(h) \cdot \Pi \cdot \mathrm{BlkDiag}\left(Q_1^T,\dots,Q_{n_B}^T\right)$$

for some $P_i \in \mathcal{O}^{s_i - s_{i-1}}$ $(i = 1,\dots,n_A)$ and $Q_j \in \mathcal{O}^{t_j - t_{j-1}}$ $(j = 1,\dots,n_B)$. The following result elucidates the structure of the collection $\{\mathcal{X}_{h,\Pi}\}_{h\in\mathcal{H},\,\Pi\in\mathcal{P}^n}$:

Proposition 4 *Let $h,h' \in \mathcal{H}$ and $\Pi,\Pi' \in \mathcal{P}^n$ be arbitrary. Then, either $\mathcal{X}_{h,\Pi} = \mathcal{X}_{h',\Pi'}$ or $\mathcal{X}_{h,\Pi} \cap \mathcal{X}_{h',\Pi'} = \emptyset$. If the latter holds, then for any $X \in \mathcal{X}_{h,\Pi}$ and $X' \in \mathcal{X}_{h',\Pi'}$, we have $\|X - X'\|_F \ge \sqrt{2}$.*

The proof of Proposition 4 can be found in Appendix A. Armed with Proposition 4, in order to prove Theorem 2', it suffices to bound $\mathrm{dist}(X,\mathcal{X}_{h,\Pi})$ for any $X \in \mathrm{St}(m,n)$, $h \in \mathcal{H}$, and $\Pi \in \mathcal{P}^n$.

3.1.3 Estimating the Distance to the Set of Critical Points

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

$$\begin{aligned}\mathrm{dist}(X,\mathcal{X}_{h,\Pi}) = \min\Big\{ &\left\| X - \mathrm{BlkDiag}\left(P_1,\dots,P_{n_A}\right)\cdot E(h)\cdot\Pi\cdot\mathrm{BlkDiag}\left(Q_1^T,\dots,Q_{n_B}^T\right)\right\|_F \;\Big|\\ &P_i \in \mathcal{O}^{s_i-s_{i-1}} \text{ for } i=1,\dots,n_A;\; Q_j \in \mathcal{O}^{t_j - t_{j-1}} \text{ for } j = 1,\dots,n_B\Big\}. \end{aligned} \tag{17}$$

Let $\left(P_1^*,\dots,P_{n_A}^*,Q_1^*,\dots,Q_{n_B}^*\right)$ be an optimal solution to (17). Upon letting

$$P^* = \mathrm{BlkDiag}\left(P_1^*,\dots,P_{n_A}^*\right) \in \mathcal{O}^m, \quad Q^* = \mathrm{BlkDiag}\left(Q_1^*,\dots,Q_{n_B}^*\right) \in \mathcal{O}^n,$$

and $\bar{X} = (P^*)^TXQ^*$, it is clear that $\mathrm{dist}^2(X,\mathcal{X}_{h,\Pi}) = \left\|\bar{X} - E(h)\Pi\right\|_F^2$. To bound this quantity, consider the decompositions

$$\bar{X} = \left[\bar{X}_1 \cdots \bar{X}_{n_B}\right], \quad E(h)\Pi = \left[\bar{E}_1(h) \cdots \bar{E}_{n_B}(h)\right], \tag{18}$$

where $\bar{X}_j, \bar{E}_j(h) \in \mathbb{R}^{m\times(t_j - t_{j-1})}$. It should be noted that $\bar{E}_j(h)$ needs not be the same as $E_j(h)$ in (10), as the former is of dimensions $m \times (t_j - t_{j-1})$, while the latter is of dimensions $m \times h_j$. On the other hand, observe from (11) that every non-zero row and every column of $E(h)$ has exactly one 1. It follows that up to a permutation of the rows, $\bar{E}_j(h)$ takes the form

$$\bar{E}_j(h) = \begin{bmatrix} I_{t_j - t_{j-1}} \\ \mathbf{0}_{(m-t_j+t_{j-1})\times(t_j-t_{j-1})}\end{bmatrix}. \tag{19}$$

Such an observation allows us to establish the following result:

Proposition 5 *For $j = 1, \ldots, n_B$ and $k = 1, \ldots, m$, let $\left[\bar{X}_j\right]_k$ and $\left[\bar{E}_j(h)\right]_k$ be the k-th row of $\bar{X}_j$ and $\bar{E}_j(h)$, respectively. Suppose that* $\operatorname{dist}(X, \mathcal{X}_{h,\Pi}) < 1$. *Then,*

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

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

The proof of Proposition 5 can be found in Appendix B.

To establish the error bound in Theorem 2', we need to link $\left\| AXB - XBX^TAX \right\|_F$ to the bound on $\operatorname{dist}^2(X, \mathcal{X}_{h,\Pi})$ in Proposition 5. This is achieved in two steps. First, we prove the following result:

Proposition 6 *Consider the decomposition of $\bar{X}$ in* (18). *Let $\lambda_{B,g} = \min_{j \in \{1,\ldots,n_B-1\}} (b_{t_j} - b_{t_{j+1}}) > 0$ be the smallest eigengap of B (by convention, we set $\lambda_{B,g} = |b_{t_1}|$ if $n_B = 1$) and $\lambda_{B,s} = \min_{j \in \{1,\ldots,n_B\}} |b_{t_j}| > 0$ be the smallest (in magnitude) eigenvalue of B. Set $\lambda_B = \min\{\lambda_{B,g}, \lambda_{B,s}\}$. Then, we have*

$$\left\| AXB - XBX^TAX \right\|_F^2 \geq \lambda_B^2 \sum_{j=1}^{n_B} \left\| A\bar{X}_j - \bar{X}_j\bar{X}_j^TA\bar{X}_j \right\|_F^2.$$

In view of Propositions 5 and 6, we then prove the following bound:

Proposition 7 *Let $\lambda_{A,g} = \min_{i \in \{1,\ldots,n_A-1\}} (a_{s_i} - a_{s_{i+1}}) > 0$ be the smallest eigengap of A. There exists a $\delta \in (0,1)$ such that for all $X \in \operatorname{St}(m,n)$ with $\operatorname{dist}(X, \mathcal{X}_{h,\Pi}) \leq \delta$,*

$$\sum_{j=1}^{n_B} \left\| A\bar{X}_j - \bar{X}_j\bar{X}_j^TA\bar{X}_j \right\|_F^2 \geq \frac{\lambda_{A,g}^2}{8} \sum_{j=1}^{n_B} \sum_{k \in \mathcal{I}_j} \left\| \left[\bar{X}_j\right]_k \right\|_2^2.$$

The proofs of Propositions 6 and 7 can be found in Appendices C and D, respectively. Now, observe that whenever $X \in \operatorname{St}(m,n)$ and $\operatorname{dist}(X, \mathcal{X}) \leq \delta$, there exist $h \in \mathcal{H}$ and $\Pi \in \mathcal{P}^n$ such that $\operatorname{dist}(X, \mathcal{X}_{h,\Pi}) \leq \delta$. Hence, by combining Propositions 5, 6, and 7, we obtain Theorem 2'.

3.1.4 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 $\bar{B} = \operatorname{Diag}(b_1, \ldots, b_p) \in \mathcal{S}^p$ has full rank. Let $X = [X_1 \; X_2] \in \operatorname{St}(m,n)$ with $X_1 \in \operatorname{St}(m,p)$ and $X_2 \in \operatorname{St}(m, n-p)$. Using (7), we have $X \in \mathcal{X}$ if and only if

$$\begin{aligned} \mathbf{0} &= A \begin{bmatrix} X_1 & X_2 \end{bmatrix} \begin{bmatrix} \bar{B} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} - \begin{bmatrix} X_1 & X_2 \end{bmatrix} \begin{bmatrix} \bar{B} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} X_1^T \\ X_2^T \end{bmatrix} A \begin{bmatrix} X_1 & X_2 \end{bmatrix} \\ &= \begin{bmatrix} AX_1\bar{B} - X_1\bar{B}X_1^TAX_1 & -X_1\bar{B}X_1^TAX_2 \end{bmatrix}. \end{aligned}$$

Observe that if $AX_1\bar{B} - X_1\bar{B}X_1^TAX_1 = \mathbf{0}$, then

$$X_1\bar{B}X_1^TAX_2 = X_1 \left(AX_1\bar{B} \right)^T X_2 = X_1 \left(X_1\bar{B}X_1^TAX_1 \right)^T X_2 = \mathbf{0},$$

where the last equality follows from the fact that $X_1^T X_2 = \mathbf{0}$. Thus, we can express $\mathcal{X}$ as

$$\mathcal{X} = \left\{ X = [X_1 \ X_2] \in \mathrm{St}(m,n) \mid X_1 \in \mathrm{St}(m,p),\ X_2 \in \mathrm{St}(m,n-p),\ AX_1\bar{B} - X_1\bar{B}X_1^T AX_1 = \mathbf{0} \right\}.$$

Now, define

$$\bar{\mathcal{X}} = \left\{ Y \in \mathrm{St}(m,p) \mid AY\bar{B} - Y\bar{B}Y^T AY = \mathbf{0} \right\}.$$

We then have the following result:

Proposition 8 *Given any $X = [X_1 \ X_2] \in \mathrm{St}(m,n)$ with $X_1 \in \mathrm{St}(m,p)$ and $X_2 \in \mathrm{St}(m,n-p)$, we have*

$$\mathrm{dist}(X, \mathcal{X}) \le \sqrt{3} \cdot \mathrm{dist}(X_1, \bar{\mathcal{X}}).$$

Proof Since $\mathrm{St}(m,n)$ and $\bar{\mathcal{X}}$ are compact, we can find

$$\bar{Y}_1 \in \arg\min_{Y \in \bar{\mathcal{X}}} \|X_1 - Y\|_F, \quad \bar{Y}_2 \in \arg \min_{\substack{Y \in \mathrm{St}(m,n-p) \\ [\bar{Y}_1 \ \ Y] \in \mathrm{St}(m,n)}} \|X_2 - Y\|_F.$$

By construction, we have $\bar{Y} = \begin{bmatrix} \bar{Y}_1 & \bar{Y}_2 \end{bmatrix} \in \mathcal{X}$. This implies that

$$\mathrm{dist}^2(X, \mathcal{X}) \le \|X - \bar{Y}\|_F^2 = \|X_1 - \bar{Y}_1\|_F^2 + \|X_2 - \bar{Y}_2\|_F^2.$$

Note that $\|X_1 - \bar{Y}_1\|_F^2 = \mathrm{dist}^2(X_1, \bar{\mathcal{X}})$. Hence, it remains to show that $\|X_2 - \bar{Y}_2\|_F^2 \le 2 \cdot \mathrm{dist}^2(X_1, \bar{\mathcal{X}})$. Towards that end, let $M = \left(I - \bar{Y}_1 \bar{Y}_1^T\right) X_2$. As $\bar{Y}_1^T M = \mathbf{0}$, we have

$$\|X_2 - \bar{Y}_2\|_F^2 = 2(n-p) - 2 \max_{\substack{Y \in \mathrm{St}(m,n-p) \\ [\bar{Y}_1 \ \ Y] \in \mathrm{St}(m,n)}} \mathrm{tr}\left(X_2^T Y\right) = 2(n-p) - 2 \max_{Y \in \mathrm{St}(m,n-p)} \mathrm{tr}\left(M^T Y\right).$$

Using the fact that $\max_{Y \in \mathrm{St}(m,n-p)} \mathrm{tr}(M^T Y) = \|M\|_*$ and $\|M\| \le 1$, we compute

$$\|M\|_* \ge \|M\|_F^2 = \|X_2\|_F^2 - \|X_2^T \bar{Y}_1\|_F^2 = n - p - \|X_2^T \bar{Y}_1\|_F^2.$$

Together with the fact that $\|CD\|_F \le \|C\| \cdot \|D\|_F$ for any matrices C, D, we obtain

$$\|X_2 - \bar{Y}_2\|_F^2 \le 2 \cdot \|X_2^T \bar{Y}_1\|_F^2 = 2 \cdot \|X_2^T (\bar{Y}_1 - X_1)\|_F^2 \le 2 \cdot \|X_1 - \bar{Y}_1\|_F^2 = 2 \cdot \mathrm{dist}^2(X_1, \bar{\mathcal{X}}),$$

as desired. □

By our result in Section 3.1.3, there exist $\delta \in (0,1)$ and $\eta > 0$ such that for all $X_1 \in \mathrm{St}(m,p)$ with $\mathrm{dist}(X_1, \bar{\mathcal{X}}) \le \delta$,

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

This, together with Proposition 8 and the fact that

$$\begin{aligned} \left\| AXB - XBX^T AX \right\|_F^2 &= \left\| AX_1\bar{B} - X_1\bar{B}X_1^T AX_1 \right\|_F^2 + \left\| X_1\bar{B}X_1^T AX_2 \right\|_F^2 \\ &\ge \left\| AX_1\bar{B} - X_1\bar{B}X_1^T AX_1 \right\|_F^2, \end{aligned}$$

yields Theorem 2'.

3.2 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:

Proposition 9 *There exists an $\eta > 0$ such that for all $X \in \mathrm{St}(m,n)$ and $X^* \in \mathcal{X}$,*

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

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

$$|F(X) - F(X^*) - \langle \nabla F(X^*), X - X^* \rangle| \leq \frac{L_F}{2} \cdot \|X - X^*\|_F^2, \tag{20}$$

where $L_F \leq 2 \cdot \|A\| \cdot \|B\|$ is the Lipschitz constant of ∇F; see, e.g., [33]. Now, by Proposition 1, we have $\nabla F(X^*) = X^* \nabla F(X^*)^T X^*$. This implies that

$$\langle \nabla F(X^*), X - X^* \rangle = \left\langle X^* \nabla F(X^*)^T X^*, X - X^* \right\rangle = \left\langle \nabla F(X^*)^T X^*, (X^*)^T X - I_n \right\rangle. \tag{21}$$

On the other hand,

$$\begin{aligned} \left\langle \nabla F(X^*)^T X^*, I_n - X^T X^* \right\rangle &= \left\langle (X^*)^T \nabla F(X^*), (X^*)^T X^* - X^T X^* \right\rangle \\ &= \left\langle X^* \nabla F(X^*)^T X^*, X^* - X \right\rangle \\ &= -\langle \nabla F(X^*), X - X^* \rangle. \end{aligned} \tag{22}$$

Upon adding (21) and (22) and using the fact that $(X - X^*)^T (X - X^*) = 2I_n - (X^*)^T X - X^T X^*$, we obtain

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

which leads to

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

By combining this with (20), we obtain the desired inequality with $\eta = (L_F/2) + \|A\|_F \cdot \|B\|$. □

Now, let $X \in \mathrm{St}(m,n)$ 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 given in Theorem 2. Furthermore, let $\bar{X}^* \in \mathcal{X}$ be such that $\mathrm{dist}(X, \mathcal{X}) = \|X - \bar{X}^*\|_F < \delta_0$. We claim that $X^*, \bar{X}^* \in \mathcal{X}_{h,\Pi}$ for some $h \in \mathcal{H}$ and $\Pi \in \mathcal{P}^n$. Indeed, if $X^* \in \mathcal{X}_{h,\Pi}$ and $\bar{X}^* \in \mathcal{X}_{h',\Pi'}$ with $\mathcal{X}_{h,\Pi} \cap \mathcal{X}_{h',\Pi'} = \emptyset$, then $\|X^* - \bar{X}^*\|_F \geq \sqrt{2}$ by Proposition 4. However, our assumption yields $\|X^* - \bar{X}^*\|_F \leq \|X - X^*\|_F + \|X - \bar{X}^*\|_F < 2\delta_0 \leq \sqrt{2}$, which is a contradiction. This establishes the claim.

Since the function F is constant on $\mathcal{X}_{h,\Pi}$ for any given $h \in \mathcal{H}$ and $\Pi \in \mathcal{P}^n$, we have $F(X^*) = F(\bar{X}^*)$. Hence, by Proposition 9 and Theorem 2, we obtain

$$|F(X) - F(X^*)| = |F(X) - F(\bar{X}^*)| \leq \eta_1 \cdot \|X - \bar{X}^*\|_F^2 \leq \eta_1 \eta_2 \cdot \|D_\rho(X)\|_F^2$$

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

4 Algorithmic Consequences of the Łojasiewicz Inequality for Problem (QP-OC)

One of the main motivations for finding the value of the Łojasiewicz exponent at the critical points of Problem (QP-OC) is that it allows a better understanding of the convergence behavior of a host of iterative methods when applied to solve the problem. As a first illustration, we combine the Łojasiewicz inequality in Theorem 1 with the convergence analysis framework in [38] to show, for the first time, that many retraction-based line-search methods converge linearly when applied to Problem (QP-OC). Then, we propose a new class of retraction-based SVRG methods—called Stiefel-SVRG—for solving Problem (QP-OC) and present a novel Łojasiewicz inequality-based analysis to establish their linear convergence in expectation.

4.1 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. One way of implementing this idea is as follows. Suppose that $X \in \mathrm{St}(m,n)$ is the current iterate. It can be shown that for the matrix $D_\rho(X)$ defined in Proposition 1, we have $-D_\rho(X) \in T(X)$ and $-D_\rho(X)$ is a descent direction at $X \in \mathrm{St}(m,n)$ for any $\rho > 0$; see [19, Lemma 3.1]. 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 needs not lie on $\mathrm{St}(m,n)$. To bring the point back on $\mathrm{St}(m,n)$ to form the next iterate, we use a *retraction* on $\mathrm{St}(m,n)$, which 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 $X \in \mathrm{St}(m,n)$,

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

Various retractions on the Stiefel manifold have been studied in the literature. Some examples include the polar decomposition-based retraction

$$R_{\mathsf{polar}}(X,\xi) = (X+\xi)(I_n + \xi^T\xi)^{-1/2}; \tag{23}$$

the QR-decomposition-based retraction

$$R_{\mathsf{QR}}(X,\xi) = \mathrm{qf}(X+\xi), \tag{24}$$

where $\mathrm{qf}(A)$ denotes the Q-factor in the thin QR-decomposition of A (see [15, Section 5.2.6]); the Cayley transform-based retraction

$$R_{\mathsf{cayley}}(X,\xi) = \left(I_m - \frac{1}{2}W(\xi)\right)^{-1}\left(I_m + \frac{1}{2}W(\xi)\right)X, \tag{25}$$

where

$$W(\xi) = \left(I_m - \frac{1}{2}XX^T\right)\xi X^T - X\xi^T\left(I_m - \frac{1}{2}XX^T\right).$$

Other examples can be found in [4, 21, 19]. In the sequel, we assume that the retraction R satisfies the following additional property:

(P). (Second-Order Boundedness) There exist $\phi \in (0, 1]$ and $M > 0$ such that for all $X \in \mathrm{St}(m, n)$ and $\xi \in T(X)$ satisfying $\|\xi\|_F \leq \phi$,

$$\|R(X, \xi) - (X + \xi)\|_F \leq M \cdot \|\xi\|_F^2.$$

The above property turns out to be rather mild. In particular, we show in Appendix E that all three retractions (23)–(25) satisfy such property.

To guarantee descent, we also need to determine a suitable step size α to move along the search direction $-D_\rho(X)$. This can be achieved, for instance, by performing a line search using the following Armijo-type rule with parameters $\beta, \gamma \in (0, 1)$:

$$\alpha = \max_{\ell \geq 0} \left\{ \beta^\ell \,\middle|\, F\left(R\left(X, -\beta^\ell D_\rho(X)\right)\right) - F(X) \leq -\gamma\beta^\ell \cdot \langle \nabla F(X), D_\rho(X) \rangle \right\}. \tag{26}$$

As the following result shows, the step size α calculated according to (26) satisfies $\alpha > 0$ and hence is well defined:

Fact 1 *(cf. [38, Proposition 2.8]) Given $X \in \mathrm{St}(m, n)$, define*

$$\bar{\alpha}(X) = \inf \left\{ \alpha > 0 \mid F\left(R(X, -\alpha D_\rho(X))\right) - F(X) = -\gamma\alpha \cdot \langle \nabla F(X), D_\rho(X) \rangle \right\}.$$

Then, we have $\bar{\alpha}(X) > 0$. Moreover, for any $\alpha \in [0, \bar{\alpha}(X)]$, we have

$$F\left(R(X, -\alpha D_\rho(X))\right) - F(X) \leq -\gamma\alpha \cdot \langle \nabla F(X), D_\rho(X) \rangle.$$

The above discussion leads us to the generic retraction-based line-search method described in Algorithm 1.

Algorithm 1 Retraction-Based Line-Search Method for Solving Problem (QP-OC)

```
Require: $X^0 \in \mathrm{St}(m, n)$, $\rho > 0$, $\beta, \gamma \in (0, 1)$
for $k = 0, 1, 2, \ldots$ do
    calculate the descent direction $-D_\rho(X^k)$ at $X^k$
    calculate the step size $\alpha_k$ according to the Armijo-type rule (26)
    set $X^{k+1} = R\left(X^k, -\alpha_k D_\rho(X^k)\right)$
    terminate if convergence criterion is met
end for
```

Naturally, we are interested in the convergence behavior of Algorithm 1. Since Theorem 1 implies that the Łojasiewicz exponent at any critical point of Problem (QP-OC) is 1/2, by adapting the convergence analysis framework in [38, Section 2.3], it can be shown that *any* sequence $\{X^k\}_{k \geq 0}$ in $\mathrm{St}(m, n)$ satisfying the following properties will converge at least R-linearly locally to some critical point $X^* \in \mathcal{X}$:[2]

(A1). (Sufficient Descent) There exist a constant $\kappa > 0$ and an index $k_1 \geq 0$ such that for $k \geq k_1$,

$$F(X^{k+1}) - F(X^k) \leq -\kappa \cdot \|D_\rho(X^k)\|_F \cdot \|X^{k+1} - X^k\|_F.$$

[2] That is, there exist constants $r_0 > 0$, $r_1 \in (0, 1)$ and index $K \geq 0$ such that $\|X^k - X^*\|_F \leq r_0 r_1^k$ for all $k \geq K$.

(A2). (Stationarity) There exists an index $k_2 \geq 0$ such that for $k \geq k_2$,

$$\|D_\rho(X^k)\|_F = 0 \quad \Longrightarrow \quad X^{k+1} = X^k.$$

(A3). (Safeguard) There exist a constant $\mu > 0$ and an index $k_3 \geq 0$ such that for $k \geq k_3$,

$$\|D_\rho(X^k)\|_F \leq \mu \cdot \|X^{k+1} - X^k\|_F.$$

Thus, our goal now is to verify that the sequence $\{X^k\}_{k\geq 0}$ generated by Algorithm 1 indeed satisfies (A1)–(A3). First, observe that (A2) follows automatically from line 4 of Algorithm 1 and the definition of a retraction. Next, consider (A1). Without loss of generality, we may assume that $D_\rho(X^k) \neq \mathbf{0}$ for all $k \geq 0$. We claim that

$$\lim_{k\to\infty} \alpha_k \cdot \|D_\rho(X^k)\|_F = 0. \tag{27}$$

To prove (27), we begin with the following inequality from [19, Appendix 2]:

$$-\left\langle \nabla F(X^k), D_\rho(X^k) \right\rangle \leq -\min\{1, \rho\} \cdot \left\| \nabla F(X^k) - X^k \nabla F(X^k)^T X^k \right\|_F^2.$$

Using (5) and the invertibility of $C_\rho(X^k) = I_m - (1-2\rho)X^k(X^k)^T$ for any $\rho > 0$, we compute

$$\left\| \nabla F(X^k) - X^k \nabla F(X^k)^T X^k \right\|_F = \left\| C_\rho(X^k)^{-1} D_\rho(X^k) \right\|_F \geq \min\left\{1, \frac{1}{2\rho}\right\} \cdot \|D_\rho(X^k)\|_F.$$

It follows that

$$-\left\langle \nabla F(X^k), D_\rho(X^k) \right\rangle \leq -\bar{\rho} \cdot \|D_\rho(X^k)\|_F^2, \tag{28}$$

where

$$\bar{\rho} = \min\left\{\rho, \frac{1}{4\rho}, \frac{1}{4\rho^2}\right\}.$$

This, together with lines 3 and 4 of Algorithm 1, implies that

$$F(X^{k+1}) - F(X^k) \leq -\bar{\gamma}\alpha_k \cdot \|D_\rho(X^k)\|_F^2 \tag{29}$$

for all $k \geq 0$, where $\bar{\gamma} = \gamma\bar{\rho}$. In particular, since F is bounded below on $\mathrm{St}(m,n)$ and $\alpha_k \in (0,1]$ for all $k \geq 0$, we have

$$\sum_{k=0}^{\infty} \alpha_k^2 \cdot \|D_\rho(X^k)\|_F^2 \leq \sum_{k=0}^{\infty} \alpha_k \cdot \|D_\rho(X^k)\|_F^2 < \infty.$$

This yields (27), as desired.

Now, by line 4 of Algorithm 1, (27), and Property (P), we have

$$\|X^{k+1} - X^k\|_F = \left\| R\left(X^k, -\alpha_k D_\rho(X^k)\right) - X^k \right\|_F \leq (M+1)\alpha_k \cdot \|D_\rho(X^k)\|_F$$

for all sufficiently large $k \geq 0$. This, together with (29), yields

$$F(X^{k+1}) - F(X^k) \leq -\frac{\bar{\gamma}}{M+1} \cdot \|D_\rho(X^k)\|_F \cdot \|X^{k+1} - X^k\|_F$$

for all sufficiently large $k \geq 0$. It follows that (A1) holds with

$$\kappa = \frac{\bar{\gamma}}{M+1}. \tag{30}$$

Lastly, consider (A3). Since

$$\begin{aligned} X^{k+1} - X^k &= R\left(X^k, -\alpha_k D_\rho(X^k)\right) - X^k \\ &= R\left(X^k, -\alpha_k D_\rho(X^k)\right) - \left(X^k - \alpha_k D_\rho(X^k)\right) - \alpha_k D_\rho(X^k), \end{aligned}$$

we have

$$\|X^{k+1} - X^k\|_F \geq \alpha_k \cdot \|D_\rho(X^k)\|_F - \left\|R\left(X^k, -\alpha_k D_\rho(X^k)\right) - \left(X^k - \alpha_k D_\rho(X^k)\right)\right\|_F$$

and

$$\|X^{k+1} - X^k\|_F \leq \alpha_k \cdot \|D_\rho(X^k)\|_F + \left\|R\left(X^k, -\alpha_k D_\rho(X^k)\right) - \left(X^k - \alpha_k D_\rho(X^k)\right)\right\|_F.$$

Upon dividing both sides of the above inequalities by $\alpha_k \cdot \|D_\rho(X^k)\|_F$, taking $k \to \infty$, and using (27) and Property (P), we get

$$\lim_{k\to\infty} \frac{\|X^{k+1} - X^k\|_F}{\alpha_k \cdot \|D_\rho(X^k)\|_F} = 1.$$

This implies that for all sufficiently large $k \geq 0$,

$$\|X^{k+1} - X^k\|_F \geq \frac{\alpha_k}{2} \cdot \|D_\rho(X^k)\|_F. \tag{31}$$

Hence, to establish (A3), it suffices to show that $\liminf_{k\to\infty} \alpha_k > 0$. Towards that end, let $\bar{\alpha}_k = \bar{\alpha}(X^k) > 0$, where $\bar{\alpha}(X^k)$ is defined in Fact 1. By (26) and Fact 1, we have

$$\alpha_k = 1 \text{ if } \bar{\alpha}_k \geq 1; \quad \alpha_k \geq \beta\bar{\alpha}_k \text{ if } \bar{\alpha}_k < 1. \tag{32}$$

It is clear from (31) that (A3) holds automatically for those indices $k \geq 0$ satisfying $\alpha_k = 1$. Thus, we may assume that $\bar{\alpha}_k < 1$ for all $k \geq 0$. In this case, it remains to show that $\liminf_{k\to\infty} \bar{\alpha}_k > 0$, as we would then have $\liminf_{k\to\infty} \alpha_k > 0$ from (32). To begin, we note that $\bar{\alpha}_k \cdot \|D_\rho(X^k)\|_F \leq (\alpha_k/\beta) \cdot \|D_\rho(X^k)\|_F$ by (32), which, together with (27), implies that

$$\lim_{k\to\infty} \bar{\alpha}_k \cdot \|D_\rho(X^k)\|_F = 0. \tag{33}$$

Now, the rest of the argument is similar to that used in the proof of [38, Theorem 2.10]. Specifically, by the Mean Value Theorem and the definition of $\bar{\alpha}_k$, there exists a $\zeta_k \in (0,1)$ such that $Z^k = \zeta_k\left(R\left(X^k - \bar{\alpha}_k D_\rho(X^k)\right) - X^k\right)$ satisfies

$$\begin{aligned} \left(R\left(X^k - \bar{\alpha}_k D_\rho(X^k)\right) - X^k\right)^T \nabla F(X^k + Z^k) &= F\left(R\left(X^k - \bar{\alpha}_k D_\rho(X^k)\right)\right) - F(X^k) \\ &= -\gamma\bar{\alpha}_k \cdot \left\langle \nabla F(X^k), D_\rho(X^k)\right\rangle. \end{aligned} \tag{34}$$

Using the fact that ∇F is Lipschitz continuous with parameter $L_F \le 2 \cdot \|A\| \cdot \|B\|$, we compute

$$\begin{aligned}
& \|Z^k\|_F \cdot \left\| R\left(X^k - \bar{\alpha}_k D_\rho(X^k)\right) - X^k \right\|_F \\
\ge\ & \frac{1}{L_F} \cdot \left\| \nabla F(X^k) - \nabla F(X^k + Z^k) \right\|_F \cdot \left\| R\left(X^k - \bar{\alpha}_k D_\rho(X^k)\right) - X^k \right\|_F && (35) \\
\ge\ & \frac{1}{L_F} \cdot \left| \left\langle \nabla F(X^k), R\left(X^k - \bar{\alpha}_k D_\rho(X^k)\right) - X^k \right\rangle + \gamma \bar{\alpha}_k \cdot \left\langle \nabla F(X^k), D_\rho(X^k) \right\rangle \right| && (36) \\
\ge\ & \frac{(1-\gamma)\bar{\alpha}_k}{L_F} \cdot \left| \left\langle \nabla F(X^k), D_\rho(X^k) \right\rangle \right| \\
& \quad - \frac{1}{L_F} \cdot \left| \left\langle \nabla F(X^k), R\left(X^k - \bar{\alpha}_k D_\rho(X^k)\right) - \left(X^k - \bar{\alpha}_k D_\rho(X^k)\right) \right\rangle \right|, && (37)
\end{aligned}$$

where (35) follows from the Lipschitz continuity of ∇F; (36) follows from the Cauchy-Schwarz inequality and (34). Using (33) and Property (P), we have

$$\|Z^k\|_F \le \left\| R\left(X^k - \bar{\alpha}_k D_\rho(X^k)\right) - X^k \right\|_F \le (M+1)\bar{\alpha}_k \cdot \|D_\rho(X^k)\|_F$$

for all sufficiently large $k \ge 0$. Moreover, by (28), we have

$$\left| \left\langle \nabla F(X^k), D_\rho(X^k) \right\rangle \right| \ge \bar{\rho} \cdot \|D_\rho(X^k)\|_F^2$$

for all $k \ge 0$. Hence, we obtain from (37) that

$$\bar{\alpha}_k \ge c_0 - c_1 \cdot \|\nabla F(X^k)\|_F \cdot \frac{\left\| R\left(X^k - \bar{\alpha}_k D_\rho(X^k)\right) - \left(X^k - \bar{\alpha}_k D_\rho(X^k)\right) \right\|_F}{\bar{\alpha}_k \cdot \|D_\rho(X^k)\|_F^2} \tag{38}$$

for all sufficiently large $k \ge 0$, where

$$c_0 = \frac{(1-\gamma)\bar{\rho}}{(M+1)^2 L_F} > 0 \quad \text{and} \quad c_1 = \frac{1}{(M+1)^2 L_F} > 0.$$

Since $\|\nabla F(X)\|_F \le 2 \cdot \|A\|_F \cdot \|B\|$ for all $X \in \mathrm{St}(m,n)$ and

$$\left\| R\left(X^k - \bar{\alpha}_k D_\rho(X^k)\right) - \left(X^k - \bar{\alpha}_k D_\rho(X^k)\right) \right\|_F \le M \bar{\alpha}_k^2 \cdot \|D_\rho(X^k)\|_F^2$$

for all sufficiently large $k \ge 0$ by (33) and Property (P), we conclude from (38) that

$$\liminf_{k \to \infty} \bar{\alpha}_k \ge \frac{c_0}{1 + 2c_1 M \cdot \|A\|_F \cdot \|B\|} > 0,$$

as desired. In particular, by combining this estimate with (31) and (32), we may take

$$\mu = 2 \max \left\{ 1, \frac{(M+1)^2 L_F + 2M \cdot \|A\|_F \cdot \|B\|}{(1-\gamma)\beta\bar{\rho}} \right\} \tag{39}$$

in (A3).

In summary, we have proven the following result:

Theorem 3 *Suppose that the retraction* $R : \bigcup_{X \in \mathrm{St}(m,n)} (\{X\} \times T(X)) \to \mathrm{St}(m,n)$ *in Algorithm 1 satisfies Property (P). Then, the sequence* $\{X^k\}_{k \geq 0}$ *generated by Algorithm 1 will converge at least R-linearly to some critical point* $X^* \in \mathcal{X}$ *of Problem* (QP-OC).

The significance of the results developed in this sub-section is threefold. First, Theorem 3 holds without requiring any assumptions on A and B. In particular, it holds for all instances of Problem (QP-OC), even for those whose critical points are not isolated. This is in sharp contrast to existing results concerning the linear convergence of certain line-search methods for solving Problem (QP-OC) (such as those in [3, 41, 42]), which require some assumptions on A and/or B; see the introduction for a discussion. Second, many existing first-order methods for solving Problem (QP-OC) employ one of the retractions in (23)–(25) and one of the search directions in the family $\{-D_\rho(\cdot)\}_{\rho>0}$; see, e.g., [3, 55, 19]. Our results lead to a unified and the first linear convergence analysis of line-search methods that employ these retractions and search directions. Third, by revisiting the proof of [38, Theorem 2.3], it can be shown that the rate of linear convergence of Algorithm 1 is bounded by $1 - \kappa/(2\eta^2\mu)$, where $\eta, \kappa, \mu > 0$ are the constants given in Theorem 1, (A1), and (A3), respectively. This, together with (30) and (39), yields a quantitative description of how the convergence rate of Algorithm 1 depends on the retraction used, which, to the best of our knowledge, is new. In particular, we see that a larger M in Property (P) will result in a slower linear rate of convergence.

4.2 Retraction-Based Stochastic Variance-Reduced Gradient Methods

In applications one is often interested in solving instances of Problem (QP-OC) whose objective functions F have a finite-sum structure. A case in point is the PCA problem, in which F takes the form

$$F(X) = -\mathrm{tr}\left[X^T \left(\frac{1}{N} \sum_{i=1}^{N} a_i a_i^T \right) X \right],$$

where $a_1, \ldots, a_N \in \mathbb{R}^m$ are the data points. Of course, the finite-sum structure of the objective function does not preclude one from tackling the problem using the retraction-based line-search method described in Algorithm 1. However, each iteration requires the evaluation of the full gradient ∇F, which could be expensive when the number of summands in the objective function is large. To improve computational efficiency, a natural idea is to extend stochastic methods for optimization in Euclidean space to the manifold setting. This is first pursued by Bonnabel [9], who developed a stochastic gradient method for optimizing a smooth function over a Riemannian manifold. However, similar to its Euclidean space counterpart, the method suffers from slow convergence; see, e.g., [59]. Here, we consider a different approach. Specifically, we extend the SVRG method of Johnson and Zhang [20]—which is developed for optimization in Euclidean space and has been shown to enjoy fast convergence—and propose a new stochastic method called Stiefel-SVRG for solving the manifold optimization problem (QP-OC). To motivate the development of Stiefel-SVRG, let us briefly review the basic elements of the (Euclidean) SVRG method in [20]. Consider the optimization problem

$$\min_{X \in \mathcal{E}} \left\{ Q(X) = \frac{1}{N} \sum_{i=1}^{N} Q_i(X) \right\}, \tag{40}$$

where $\mathcal{E}$ is a finite-dimensional Euclidean space and $Q_1,\ldots,Q_N:\mathcal{E}\to\mathbb{R}$ are smooth functions. The SVRG method in [20] solves (40) by proceeding iteratively, and the iterations are divided into epochs. In each epoch, the method takes an iterate $\tilde{X}\in\mathcal{E}$ from the previous epoch and performs a sequence of stochastic gradient descent updates of the form

$$X^{k+1} = X^k - \alpha_k\left(\nabla Q_{i_k}(X^k) - \nabla Q_{i_k}(\tilde{X}) + \nabla Q(\tilde{X})\right), \tag{41}$$

where the index i_k is chosen uniformly at random from $\{1,\ldots,N\}$. Informally, the term $\nabla Q_{i_k}(\tilde{X}) - \nabla Q(\tilde{X})$ is introduced to reduce the variance of the random estimate of the full gradient, which can lead to faster convergence of the method. We refer the reader to [20] for a rigorous treatment of this argument.

Now, let us return to our problem of interest; i.e., Problem (QP-OC). Suppose that the matrix $A\in\mathcal{S}^m$ in the objective function admits a finite-sum decomposition $A=(1/N)\sum_{i=1}^N A_i$ for some given matrices $A_1,\ldots,A_N\in\mathcal{S}^m$. Then, we can express Problem (QP-OC) as

$$\min_{X\in\mathrm{St}(m,n)}\left\{F(X)=\frac{1}{N}\sum_{i=1}^N F_i(X)\right\},$$

where $F_i(X)=\mathrm{tr}(X^TA_iXB)$ for $i=1,\ldots,N$. To extend the SVRG method in [20] to solve the above problem, one tempting idea is to mimic the update rule (41) and set

$$X^{k+1} = R\left(X^k, -\alpha_k\left(\mathrm{grad}\, F_{i_k}(X^k) - \mathrm{grad}\, F_{i_k}(\tilde{X}) + \mathrm{grad}\, F(\tilde{X})\right)\right).$$

Unfortunately, since $\mathrm{grad}\, F_{i_k}(X^k)$ and $\mathrm{grad}\, F_{i_k}(\tilde{X})$ belong to different tangent spaces to $\mathrm{St}(m,n)$, the above update rule is ill-defined. Although this can be fixed by using a so-called *vector transport* (see, e.g., [3, Chapter 8] for a definition) to move the vectors $\mathrm{grad}\, F_{i_k}(\tilde{X})$ and $\mathrm{grad}\, F(\tilde{X})$ in the tangent space at $\tilde{X}\in\mathrm{St}(m,n)$ to the tangent space at $X^k\in\mathrm{St}(m,n)$, such an approach incurs the additional cost of computing the vector transport and hence is not desirable. A conceptually simpler and computationally more efficient approach is to first project the vector $\nabla F_{i_k}(X^k) - \nabla F_{i_k}(\tilde{X}) + \nabla F(\tilde{X})$ onto the tangent space at $X^k\in\mathrm{St}(m,n)$ and then apply a retraction to get the next iterate. This leads to our proposed Stiefel-SVRG, which is described in Algorithm 2.[3]

One of the main challenges in analyzing the convergence rate of Stiefel-SVRG is that the iterates it generates are random and do not necessarily satisfy the sufficient descent condition (A1). As a result, we cannot simply apply the convergence analysis framework in Section 4.1. To circumvent this difficulty, we present a novel analysis of Stiefel-SVRG and establish its linear convergence in expectation. Specifically, we prove the following result:

Theorem 4 *Suppose that in Algorithm 2, the initial point is $\tilde{X}^0\in\mathrm{St}(m,n)$, the retraction R satisfies Property (P), and the step size α satisfies*

$$0<\alpha<\min\left\{\frac{\phi}{6\cdot\|A\|_F\cdot\|B\|},\frac{1}{8c_0},\frac{1}{2c_1(\Gamma-1)},\frac{1}{c_1}\left[\left(1+\frac{\delta}{c_2}\right)^{1/\Gamma}-1\right]\right\}, \tag{42}$$

[3] Stiefel-SVRG was first presented by the second author at the 13th Chinese Workshop on Machine Learning and Applications held in Nanjing, China in 2015 [45]. As such, it predates the SVRG methods for manifold optimization developed in [58, 37]. More importantly, Stiefel-SVRG does not require the computation of any vector transport, which makes it computationally more advantageous than the SVRG methods proposed in [58, 37].

Algorithm 2 Stiefel-SVRG: Retraction-Based SVRG Method for Solving Problem (QP-OC)

```
Require: $\tilde{X}^0 \in \mathrm{St}(m,n)$, $\alpha > 0$, $\Gamma \geq 2$
for $s = 0, 1, \ldots$ do
    set $X^0 = \tilde{X}^s$
    for $k = 0, 1, \ldots, \Gamma - 1$ do
        sample $i_k \in \{1, \ldots, N\}$ uniformly at random
        set $G^k = \nabla F_{i_k}(X^k) - \nabla F_{i_k}(X^0) + \nabla F(X^0)$
        set $\xi^k = \left(I_m - X^k(X^k)^T\right) G^k + (1/2)\left((X^k)^T G^k - (G^k)^T X^k\right)$
        set $X^{k+1} = R(X^k, -\alpha\xi^k)$
    end for
    set $\tilde{X}^{s+1} = X^J$, where $J = \arg\min_{k \in \{0,1,\ldots,\Gamma\}} F(X^k)$
    terminate if convergence criterion is met
end for
```

where

$$
\begin{aligned}
c_0 &= (M^2 + 4M + 1) \cdot \|A\| \cdot \|B\|, \\
c_1 &= 2(M+1)\left(\max_{i \in \{1,\ldots,N\}} \|A_i\| + \|A\|\right)\|B\|, \\
c_2 &= \frac{6(M+1) \cdot \|A\|_F \cdot \|B\|}{c_1} = \frac{3 \cdot \|A\|_F}{\max_{i\in\{1,\ldots,N\}} \|A_i\| + \|A\|},
\end{aligned}
$$

$\phi \in (0,1]$ *and* $M > 0$ *are the constants given in Property (P), and* $\delta \in (0, \sqrt{2}/2)$ *is the constant given in Theorem 1. Then, the sequence* $\{\tilde{X}^s\}_{s \geq 0}$ *generated from the epochs of Algorithm 2 satisfies* $F(\tilde{X}^s) \searrow F^*$ *for some* $F^* \in \mathbb{R}$*, and every limit point* X^* *of the sequence* $\{\tilde{X}^s\}_{s\geq 0}$ *is a critical point of Problem* (QP-OC) *with* $F(X^*) = F^*$*. Moreover, we have*

$$
\mathbb{E}\left[F(\tilde{X}^{s+1}) - F^*\right] \leq \frac{2\eta^2}{\alpha\Gamma + 2\eta^2}\mathbb{E}\left[F(\tilde{X}^s) - F^*\right], \tag{43}
$$

where $\eta > 0$ *is the constant given in Theorem 1 and the expectation is taken over all the random choices in Algorithm 2.*

As we shall see, our characterization of the Łojasiewicz exponent at the critical points of Problem (QP-OC) plays a key role in the proof of Theorem 4. This again demonstrates the power and utility of our main result (Theorem 1).

To prove Theorem 4, we begin with the following proposition, which bounds the change in the objective values of successive iterates within an epoch of Algorithm 2 and shows that every limit point of the sequence $\{\tilde{X}^s\}_{s\geq 0}$ has the same objective value and is a critical point of Problem (QP-OC). Its proof can be found in Appendix F.

Proposition 10 *Under the setting of Theorem 4, consider an arbitrary epoch* $s \geq 0$ *and let* $X^0 = \tilde{X}^s, X^1, \ldots, X^\Gamma$ *be the sequence generated by Algorithm 2 in this epoch. Then, we have*

$$
F(X^{k+1}) - F(X^k) + \alpha \cdot \mathrm{tr}\left[\left((X^k)^T A \xi^k + (\xi^k)^T A X^k\right) B\right] \leq c_0 \alpha^2 \cdot \|\xi^k\|_F^2. \tag{44}
$$

Consequently, for all $s \geq 0$,

$$F(\tilde{X}^{s+1}) - F(\tilde{X}^s) \leq -\frac{7\alpha}{8} \cdot \|\mathrm{grad}\, F(\tilde{X}^s)\|_F^2. \tag{45}$$

In particular, we have $F(\tilde{X}^s) \searrow F^*$ for some $F^* \in \mathbb{R}$, and every limit point X^* of the sequence $\{\tilde{X}^s\}_{s\geq 0}$ satisfies $F(X^*) = F^*$ and $\mathrm{grad}\, F(X^*) = \mathbf{0}$.

The following is a simple corollary of Proposition 10, whose proof can be found in Appendix G:

Corollary 2 *Under the setting of Theorem 4, consider an arbitrary epoch $s \geq 0$ and let $X^0 = \tilde{X}^s, X^1, \ldots, X^\Gamma$ be the sequence generated by Algorithm 2 in this epoch. Then, we have*

$$\begin{aligned} \mathbb{E}\left[F(X^{k+1}) - F(X^k)\right] \quad &\leq \quad \left(-\alpha + 2c_0\alpha^2\right) \mathbb{E}\left[\|\mathrm{grad}\, F(X^k)\|_F^2\right] \\ &\quad + 2c_0c_1^2\alpha^4 k \sum_{j=0}^{k-1} (c_1\alpha + 1)^{2(k-1-j)} \mathbb{E}\left[\|\mathrm{grad}\, F(X^j)\|_F^2\right], \end{aligned} \tag{46}$$

where the expectation is taken over the random choices of $i_0, \ldots, i_{\Gamma-1}$ within the epoch.

Next, we have the following proposition, which states that when $s \geq 0$ is sufficiently large, the iterates generated by Algorithm 2 from epoch s onwards will all be close to a certain component of $\mathcal{X}$. It makes use of the definition of $\mathcal{X}_{h,\Pi}$ (see the paragraph following the proof of Proposition 3) and the proof can be found in Appendix H.

Proposition 11 *Under the setting of Theorem 4, there exist $h \in \mathcal{H}$ and $\Pi \in \mathcal{P}^n$ such that for all sufficiently large $s \geq 0$, the sequence $X^0 = \tilde{X}^s, X^1, \ldots, X^\Gamma$ generated in epoch s of Algorithm 2 satisfies $\mathrm{dist}(X^k, \mathcal{X}_{h,\Pi}) \leq \delta/3$ for $k = 0, 1, \ldots, \Gamma$, where $\delta \in (0, \sqrt{2}/2)$ is the constant given in Theorem 1. Consequently, every limit point of the sequence $\{\tilde{X}^s\}_{s\geq 0}$ belongs to $\mathcal{X}_{h,\Pi}$ and $F(X) = F^*$ for all $X \in \mathcal{X}_{h,\Pi}$, where $F^* \in \mathbb{R}$ is the constant given in Proposition 10.*

We are now ready to finish the proof of Theorem 4:

Proof of Theorem 4 Let $X^0 = \tilde{X}^s, X^1, \ldots, X^\Gamma$ be the sequence generated by Algorithm 2 in epoch $s \geq 0$. Upon summing the inequality (46) over $k = 0, 1, \ldots, \Gamma - 1$, we get

$$\begin{aligned} \mathbb{E}\left[F(X^\Gamma) - F(X^0)\right] \quad &\leq \quad (-\alpha + 2c_0\alpha^2) \sum_{k=0}^{\Gamma-1} \mathbb{E}\left[\|\mathrm{grad}\, F(X^k)\|_F^2\right] \\ &\quad + 2c_0c_1^2\alpha^4 \sum_{k=1}^{\Gamma-1}\sum_{j=0}^{k-1} k(c_1\alpha + 1)^{2(k-1-j)} \mathbb{E}\left[\|\mathrm{grad}\, F(X^j)\|_F^2\right] \\ &\leq \quad (-\alpha + 2c_0\alpha^2) \sum_{k=0}^{\Gamma-1} \mathbb{E}\left[\|\mathrm{grad}\, F(X^k)\|_F^2\right] \\ &\quad + 2c_0c_1^2\alpha^4(\Gamma - 1) \sum_{j=0}^{\Gamma-2}\sum_{k=j+1}^{\Gamma-1} (c_1\alpha + 1)^{2(k-1-j)} \mathbb{E}\left[\|\mathrm{grad}\, F(X^j)\|_F^2\right] \\ &\leq \quad \left(-\alpha + 2c_0\alpha^2 + c_0c_1\alpha^3(\Gamma - 1)(c_1\alpha + 1)^{2(\Gamma-1)}\right) \sum_{k=0}^{\Gamma-1} \mathbb{E}\left[\|\mathrm{grad}\, F(X^k)\|_F^2\right], \end{aligned}$$

where the last inequality follows from the fact that for $j = 0, 1, \ldots, \Gamma - 2$,

$$\sum_{k=j+1}^{\Gamma-1} (c_1\alpha + 1)^{2(k-1-j)} \leq \sum_{k=0}^{\Gamma-2} (c_1\alpha + 1)^{2k} \leq \frac{(c_1\alpha + 1)^{2(\Gamma-1)} - 1}{(c_1\alpha + 1)^2 - 1} \leq \frac{(c_1\alpha + 1)^{2(\Gamma-1)}}{2c_1\alpha}.$$

Now, our choice of the step size α implies that

$$-\alpha + c_0\alpha^2 \left(2 + c_1\alpha(\Gamma - 1)(c_1\alpha + 1)^{2(\Gamma-1)}\right) \leq -\alpha + 4c_0\alpha^2 \leq -\frac{\alpha}{2}.$$

Moreover, by Proposition 11, for all sufficiently large $s \geq 0$, we have $\operatorname{dist}(X^k, \mathcal{X}) = \operatorname{dist}(X^k, \mathcal{X}_{h,\Pi}) \leq \delta/3$ for $k = 0, 1, \ldots, \Gamma$. Upon noting $\operatorname{grad} F(X) = D_{1/4}(X)$ and applying the Łojasiewicz inequality for Problem (QP-OC) (Theorem 1), we obtain

$$\mathbb{E}\left[F(X^\Gamma) - F^*\right] + \frac{\alpha}{2\eta^2} \sum_{k=0}^{\Gamma-1} \mathbb{E}\left[F(X^k) - F^*\right] \leq \mathbb{E}\left[F(X^0) - F^*\right].$$

Since $X^0 = \tilde{X}^s$ and $F(\tilde{X}^{s+1}) \leq F(X^k)$ for $k = 0, 1, \ldots, \Gamma$ by lines 2 and 9 of Algorithm 2, respectively, we conclude that

$$\begin{aligned} \left(1 + \frac{\alpha\Gamma}{2\eta^2}\right) \mathbb{E}\left[F(\tilde{X}^{s+1}) - F^*\right] &\leq \mathbb{E}\left[F(X^\Gamma) - F^*\right] + \frac{\alpha}{2\eta^2} \sum_{k=0}^{\Gamma-1} \mathbb{E}\left[F(X^k) - F^*\right] \\ &\leq \mathbb{E}\left[F(\tilde{X}^s) - F^*\right], \end{aligned}$$

as desired. □

Remark 2 *It is worth noting that there is an alternative, much simpler proof of the linear convergence of Algorithm 2. Indeed, by Proposition 11 and Theorem 1, we have*

$$F(\tilde{X}^s) - F^* \leq \eta \cdot \|\operatorname{grad} F(\tilde{X}^s)\|_F^2$$

for all sufficiently large $s \geq 0$. This, together with (45) in Proposition 10, yields the following linear convergence result:

$$F(\tilde{X}^{s+1}) - F^* \leq F(\tilde{X}^s) - F^* - \frac{7\alpha}{8\eta}\left(F(\tilde{X}^s) - F^*\right) = \left(1 - \frac{7\alpha}{8\eta}\right)\left(F(\tilde{X}^s) - F^*\right). \tag{47}$$

Note that the above inequality is deterministic*; i.e., it holds for* any *realization of the sequence $\{\tilde{X}^s\}_{s\geq 0}$. This is in contrast with the inequality (43), which holds only* in expectation. *Nevertheless, as Propositions 10 and 11 are proven under the assumption that the step size α is small (see (42)), the rate of convergence $1 - (7\alpha)/(8\eta)$ in (47) could be very close to 1 and inferior to the rate $2\eta^2/(\alpha\Gamma + 2\eta^2)$ in (43).*

5 Conclusion

In this paper, we showed that the Łojasiewicz exponent at any critical point of Problem (QP-OC) is 1/2. This is achieved by establishing a local error bound for the (non-convex) set of critical points of Problem (QP-OC), which could be of independent interest. Our result expands the currently very limited repertoire of optimization problems for which the Łojasiewicz exponent is known. Moreover, it allows us to analyze the convergence rates of various iterative methods that exploit the manifold structure of $\mathrm{St}(m,n)$ to solve Problem (QP-OC). To illustrate the latter, we first combined our result on the Łojasiewicz exponent with the convergence analysis framework in [38] to show that a large class of retraction-based line-search methods will converge linearly to a critical point of Problem (QP-OC). Then, we proposed a new retraction-based stochastic variance-reduced gradient method called Stiefel-SVRG for solving Problem (QP-OC) and presented a novel Łojasiewicz inequality-based analysis to establish its linear convergence. A natural future direction that stems from our work is to determine the Łojasiewicz exponents for other structured non-convex optimization problems, such as the best rank-one approximation of tensors [54].

Acknowledgements

We thank the associate editor for coordinating the review of our manuscript and the anonymous reviewer for his/her detailed comments.

Appendix

A Proof of Proposition 4

Observe that given any $X \in \mathcal{X}_{h,\Pi}$, we can write

$$\begin{aligned}\mathcal{X}_{h,\Pi} = \big\{ &\mathrm{BlkDiag}(P_1,\ldots,P_{n_A}) \cdot X \cdot \mathrm{BlkDiag}\left(Q_1^T,\ldots,Q_{n_B}^T\right) \mid \\ &P_i \in \mathcal{O}^{s_i - s_{i-1}} \text{ for } i=1,\ldots,n_A;\ Q_j \in \mathcal{O}^{t_j - t_{j-1}} \text{ for } j=1,\ldots,n_B \big\}.\end{aligned}$$

Thus, if $X \in \mathcal{X}_{h,\Pi} \cap \mathcal{X}_{h',\Pi'}$, then

$$\mathrm{BlkDiag}(P_1,\ldots,P_{n_A}) \cdot X \cdot \mathrm{BlkDiag}\left(Q_1^T,\ldots,Q_{n_B}^T\right) \in \mathcal{X}_{h,\Pi} \cap \mathcal{X}_{h',\Pi'}$$

for any $P_i \in \mathcal{O}^{s_i - s_{i-1}}$ $(i=1,\ldots,n_A)$ and $Q_j \in \mathcal{O}^{t_j - t_{j-1}}$ $(j=1,\ldots,n_B)$. This implies that $\mathcal{X}_{h,\Pi} = \mathcal{X}_{h',\Pi'}$.

Now, suppose that $\mathcal{X}_{h,\Pi} \cap \mathcal{X}_{h',\Pi'} = \emptyset$. Let $X \in \mathcal{X}_{h,\Pi}$ and $X' \in \mathcal{X}_{h',\Pi'}$ be arbitrary. Then, there exist $P_i \in \mathcal{O}^{s_i - s_{i-1}}$ $(i=1,\ldots,n_A)$ and $Q_j \in \mathcal{O}^{t_j - t_{j-1}}$ $(j=1,\ldots,n_B)$ such that

$$\|X - X'\|_F^2 = \left\| E(h')\Pi' - \mathrm{BlkDiag}\left(P_1,\ldots,P_{n_A}\right) \cdot E(h) \cdot \Pi \cdot \mathrm{BlkDiag}\left(Q_1^T,\ldots,Q_{n_B}^T\right) \right\|_F^2. \tag{48}$$

Consider the following block decomposition of $E(h)\Pi$ (and similarly for $E(h')\Pi'$):

$$E(h)\Pi = \begin{bmatrix} E_{1,1}(h,\Pi) & \cdots & E_{1,n_B}(h,\Pi) \\ \vdots & \ddots & \vdots \\ E_{n_A,1}(h,\Pi) & \cdots & E_{n_A,n_B}(h,\Pi) \end{bmatrix}$$

where $E_{i,j}(h,\Pi) \in \mathbb{R}^{(s_i-s_{i-1})\times(t_j-t_{j-1})}$ for $i = 1,\dots,n_A$ and $j = 1,\dots,n_B$. Let $|E_{i,j}(h,\Pi)|$ be the number of ones in $E_{i,j}(h,\Pi)$. We then have two cases:

Case 1. There exist $i \in \{1,\dots,n_A\}$ and $j \in \{1,\dots,n_B\}$ such that $|E_{i,j}(h,\Pi)| \neq |E_{i,j}(h',\Pi')|$.

It can be seen from (11) that for any $u \in \mathcal{H}$, every column of $E(u)$ has exactly one 1. Hence, for any $u \in \mathcal{H}$ and $\Phi \in \mathcal{P}^n$, every column of $E(u)\Phi$ also has exactly one 1. In particular, we have

$$\sum_{k=1}^{n_A} |E_{k,j}(h,\Pi)| = \sum_{k=1}^{n_A} |E_{k,j}(h',\Pi')| = t_j - t_{j-1},$$

which implies that $|E_{i',j}(h,\Pi)| \neq |E_{i',j}(h',\Pi')|$ for some $i' \in \{1,\dots,n_A\} \setminus \{i\}$. Now, we compute

$$\begin{aligned} \|X - X'\|_F^2 &\geq \left\|E_{i,j}(h',\Pi') - P_i E_{i,j}(h,\Pi) Q_j^T\right\|_F^2 + \left\|E_{i',j}(h',\Pi') - P_{i'} E_{i',j}(h,\Pi) Q_j^T\right\|_F^2 \\ &\geq \min_{\substack{P\in\mathcal{O}^{s_i-s_{i-1}} \\ Q\in\mathcal{O}^{t_j-t_{j-1}}}} \left\|E_{i,j}(h',\Pi') - P E_{i,j}(h,\Pi) Q^T\right\|_F^2 \\ &\quad + \min_{\substack{P\in\mathcal{O}^{s_{i'}-s_{i'-1}} \\ Q\in\mathcal{O}^{t_j-t_{j-1}}}} \left\|E_{i',j}(h',\Pi') - P E_{i',j}(h,\Pi) Q^T\right\|_F^2. \end{aligned} \tag{49}$$

Both terms in (49) are instances of the two-sided orthogonal Procrustes problem and admit the following characterization [40]:

$$\begin{aligned} \min_{\substack{P\in\mathcal{O}^{s_i-s_{i-1}} \\ Q\in\mathcal{O}^{t_j-t_{j-1}}}} \left\|E_{i,j}(h',\Pi') - P E_{i,j}(h,\Pi) Q^T\right\|_F^2 &= \sum_{k=1}^{K} \left(\sigma_k(E_{i,j}(h',\Pi')) - \sigma_k(E_{i,j}(h,\Pi))\right)^2, \\ \min_{\substack{P\in\mathcal{O}^{s_{i'}-s_{i'-1}} \\ Q\in\mathcal{O}^{t_j-t_{j-1}}}} \left\|E_{i',j}(h',\Pi') - P E_{i',j}(h,\Pi) Q^T\right\|_F^2 &= \sum_{k=1}^{K'} \left(\sigma_k(E_{i',j}(h',\Pi')) - \sigma_k(E_{i',j}(h,\Pi))\right)^2. \end{aligned}$$

Here, $K = \min\{s_i - s_{i-1}, t_j - t_{j-1}\}$, $K' = \min\{s_{i'} - s_{i'-1}, t_j - t_{j-1}\}$, and $\sigma_k(Y)$ is the k-th largest singular value of Y. Observe that for any $\alpha \in \{1,\dots,n_A\}$, $\beta \in \{1,\dots,n_B\}$, $u \in \mathcal{H}$, and $\Phi \in \mathcal{P}^n$, every non-zero row and every non-zero column of $E_{\alpha,\beta}(u,\Phi)$ has exactly one 1. It follows that the singular values of $E_{\alpha,\beta}(u,\Phi)$ are either 0 or 1, and there are $|E_{\alpha,\beta}(u,\Phi)|$ of the latter. Since $|E_{i,j}(h,\Pi)| \neq |E_{i,j}(h',\Pi')|$ and $|E_{i',j}(h,\Pi)| \neq |E_{i',j}(h',\Pi')|$, we conclude from (49) that $\|X - X'\|_F^2 \geq 2$.

Case 2. $|E_{i,j}(h,\Pi)| = |E_{i,j}(h',\Pi')|$ for $i = 1,\dots,n_A$ and $j = 1,\dots,n_B$.

We show that $X = X'$ in this case, which would then contradict the assumption that $\mathcal{X}_{h,\Pi} \cap \mathcal{X}_{h',\Pi'} = \emptyset$. To begin, let $i \in \{1,\dots,n_A\}$ be arbitrary and consider the i-th block row of $E(h)\Pi$ and $E(h')\Pi'$; i.e.,

$$\begin{aligned} \mathrm{BlkRow}_i(E(h)\Pi) &= \left[E_{i,1}(h,\Pi) \cdots E_{i,n_B}(h,\Pi)\right], \\ \mathrm{BlkRow}_i\left(E(h')\Pi'\right) &= \left[E_{i,1}(h',\Pi') \cdots E_{i,n_B}(h',\Pi')\right]. \end{aligned}$$

By (11), every non-zero row of $\mathrm{BlkRow}_i(E(h)\Pi)$ and $\mathrm{BlkRow}_i\left(E(h')\Pi'\right)$ has exactly one 1. Moreover, we have $|E_{i,j}(h,\Pi)| = |E_{i,j}(h',\Pi')|$ for $j = 1,\dots,n_B$ by assumption. Hence, we can find permutation matrices $\Phi_{i,1}, \Phi_{i,2}, \dots, \Phi_{i,n_B} \in \mathcal{P}^{s_i-s_{i-1}}$ such that for $j = 1,\dots,n_B$,

(i) the indices of the rows of $\Phi_{i,j}\,(\Phi_{i,j-1}\Phi_{i,j-2}\cdots\Phi_{i,1}E_{i,j}(h,\Pi))$ that contain a 1 are the same as those of $E_{i,j}(h',\Pi')$ that contain a 1 (i.e., the k-th row of $\Phi_{i,j}\,(\Phi_{i,j-1}\Phi_{i,j-2}\cdots\Phi_{i,1}E_{i,j}(h,\Pi))$ contains a 1 if and only if the k-th row of $E_{i,j}(h',\Pi')$ contains a 1, where $k\in\{1,\ldots,s_i-s_{i-1}\}$);

(ii) the indices of the rows of $(\Phi_{i,j-1}\Phi_{i,j-2}\cdots\Phi_{i,1})\,[E_{i,1}(h,\Pi)\,\cdots\,E_{i,j-1}(h,\Pi)]$ that contain a 1 are fixed by $\Phi_{i,j}$ (i.e., if the k-th row of $(\Phi_{i,j-1}\Phi_{i,j-2}\cdots\Phi_{i,1})\,[E_{i,1}(h,\Pi)\,\cdots\,E_{i,j-1}(h,\Pi)]$ contains a 1, then $\Phi_{i,j}e_k=e_k$, where e_k is the k-th standard basis vector of $\mathbb{R}^{s_i-s_{i-1}}$ and $k\in\{1,\ldots,s_i-s_{i-1}\}$).

Upon letting $\Phi_i=\Phi_{i,n_B}\Phi_{i,n_B-1}\cdots\Phi_{i,1}\in\mathcal{P}^{s_i-s_{i-1}}$ and using properties (i) and (ii) above, we see that the indices of the rows of $\Phi_iE_{i,j}(h,\Pi)$ that contain a 1 are the same as those of $E_{i,j}(h',\Pi')$ that contain a 1 for $j=1,\ldots,n_B$.

Next, let $j\in\{1,\ldots,n_B\}$ be arbitrary and consider the j-th block column of $\mathrm{BlkDiag}(\Phi_1,\ldots,\Phi_{n_A})\cdot E(h)\cdot\Pi$ and $E(h')\Pi'$; i.e.,

$$\begin{aligned}
\mathrm{BlkCol}_j\left(\mathrm{BlkDiag}(\Phi_1,\ldots,\Phi_{n_A})\cdot E(h)\cdot\Pi\right) &= \begin{bmatrix}\Phi_1E_{1,j}(h,\Pi)\\ \vdots\\ \Phi_{n_A}E_{n_A,j}(h,\Pi)\end{bmatrix},\\
\mathrm{BlkCol}_j\left(E(h')\Pi'\right) &= \begin{bmatrix}E_{1,j}(h',\Pi')\\ \vdots\\ E_{n_A,j}(h',\Pi')\end{bmatrix}.
\end{aligned}$$

By (11), each column of $\mathrm{BlkCol}_j\left(\mathrm{BlkDiag}(\Phi_1,\ldots,\Phi_{n_A})\cdot E(h)\cdot\Pi\right)$ and $\mathrm{BlkCol}_j\left(E(h')\Pi'\right)$ has exactly one 1. Since $|E_{i,j}(h,\Pi)|=|E_{i,j}(h',\Pi')|$ for $i=1,\ldots,n_A$ by assumption, we have $|\Phi_iE_{i,j}(h,\Pi)|=|E_{i,j}(h',\Pi')|$. Moreover, by the definition of $\Phi_1,\ldots,\Phi_{n_A}$, the indices of the rows of $\Phi_iE_{i,j}(h,\Pi)$ that contain a 1 are the same as those of $E_{i,j}(h',\Pi')$ that contain a 1. Thus, there exists a permutation matrix $\Psi_j\in\mathcal{P}^{t_j-t_{j-1}}$ such that

$$\mathrm{BlkCol}_j\left(E(h')\Pi'\right)=\mathrm{BlkCol}_j\left(\mathrm{BlkDiag}(\Phi_1,\ldots,\Phi_{n_A})\cdot E(h)\cdot\Pi\right)\cdot\Psi_j.$$

In particular, we obtain $E(h')\Pi'=\mathrm{BlkDiag}(\Phi_1,\ldots,\Phi_{n_A})\cdot E(h)\cdot\Pi\cdot\mathrm{BlkDiag}(\Psi_1,\ldots,\Psi_{n_B})$. Since a permutation matrix is also an orthogonal matrix, we conclude from (48) that $\|X-X'\|_F^2=0$, or equivalently, $X=X'$, as desired.

B Proof of Proposition 5

Using (17) and (18), it can be verified that

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

Since up to a permutation of the rows $\bar{E}_j$ takes the form (19), in order to obtain the desired bound on $\mathrm{dist}^2(X, \mathcal{X}_{h,\Pi})$, 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* $\|S_2\|_F^2 \le v^* \le 2\|S_2\|_F^2$*.*

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

$$\min \left\{ \|S_1 - X\|_F^2 \mid X \in \mathcal{O}^q \right\}. \tag{50}$$

Problem (50) 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 [39]. 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

$$I_q - \Sigma = (I_q + \Sigma)^{-1}\left(V^TS_2^TS_2V\right).$$

It follows that

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

This, together with the fact that $\|S_2\|_F^2 \le v^* < 1$, yields the desired result. □

C Proof of Proposition 6

Recall that

$$P^* = \mathrm{BlkDiag}\left(P_1^*, \ldots, P_{n_A}^*\right) \in \mathcal{O}^m,\; Q^* = \mathrm{BlkDiag}\left(Q_1^*, \ldots, Q_{n_B}^*\right) \in \mathcal{O}^n,\; \bar{X} = (P^*)^TXQ^*.$$

Upon observing that $AP^* = P^*A$, $BQ^* = Q^*B$ and using (9), (18), we compute

$$\begin{aligned}
\left\|AXB - XBX^TAX\right\|_F^2 &= \left\|AP^*\bar{X}(Q^*)^TB - P^*\bar{X}(Q^*)^TBQ^*\bar{X}^T(P^*)^TAP^*\bar{X}(Q^*)^T\right\|_F^2 \\
&= \left\|P^*\left(A\bar{X}B - \bar{X}B\bar{X}^TA\bar{X}\right)(Q^*)^T\right\|_F^2 \\
&= \left\|A\bar{X}B - \bar{X}B\bar{X}^TA\bar{X}\right\|_F^2 \\
&= \sum_{j=1}^{n_B} \left\| b_{t_j}A\bar{X}_j - \sum_{k=1}^{n_B} b_{t_k}\bar{X}_k\left(\bar{X}_k^TA\bar{X}_j\right)\right\|_F^2.
\end{aligned} \tag{51}$$

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 (18), this leads to

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

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,

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

where $\lambda_B = \min\{\lambda_{B,g}, \lambda_{B,s}\}$, $\lambda_{B,g} = \min_{j\in\{1,\ldots,n_B-1\}}(b_{t_j} - b_{t_{j+1}}) > 0$, and $\lambda_{B,s} = \min_{j\in\{1,\ldots,n_B\}} |b_{t_j}| > 0$. By combining the above with (51), the proof is completed.

D Proof of Proposition 7

Consider a fixed $j \in \{1, \ldots, n_B\}$. 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 $k = 1, \ldots, t_j - t_{j-1}$. Since

$$\left\|A\bar{X}_j - \bar{X}_j\bar{X}_j^T A\bar{X}_j\right\|_F^2 = \sum_{k=1}^{t_j - t_{j-1}} \|\Delta_k\|_2^2,$$

our goal is to establish a lower bound on $\|\Delta_k\|_2^2$ for $k = 1, \ldots, t_j - t_{j-1}$. Towards that end, 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 $\alpha = 1, \ldots, m$. Then, we can write

$$\Delta_k = A\bar{x}_k - \sum_{\ell=1}^{t_j - t_{j-1}} \bar{x}_\ell\left(\bar{x}_\ell^T A\bar{x}_k\right). \tag{52}$$

Suppose that $\mathrm{dist}(X, \mathcal{X}_{h,\Pi}) = \left\|\bar{X} - E(h)\Pi\right\|_F = \tau$ for some $\tau \in (0,1)$. Using the representations of $\bar{X}$ and $E(h)\Pi$ in (18), we have

$$(\bar{x}_k)_\alpha \in \begin{cases} [1-\tau, 1+\tau] & \text{if } \alpha = \iota(k), \\ [-\tau, \tau] & \text{otherwise,}\end{cases} \tag{53}$$

where $\iota(k)$ is the coordinate of the k-th column of $\bar{E}_j(h)$ that equals 1. Now, by (52),

$$\begin{aligned}
\Delta_k &= A\bar{x}_k - \bar{x}_k\left(\bar{x}_k^T A\bar{x}_k\right) - \sum_{\ell\neq k}\bar{x}_\ell\left(\bar{x}_\ell^T A\bar{x}_k\right) \\
&= \left(A - a_{\iota(k)}I_m\right)\bar{x}_k + \left(a_{\iota(k)} - \bar{x}_k^T A\bar{x}_k\right)\bar{x}_k - \sum_{\ell\neq k}\bar{x}_\ell\left(\bar{x}_\ell^T A\bar{x}_k\right).
\end{aligned}$$

Let $\mathrm{proj}_{\mathcal{I}_j}$ be the projector onto the coordinates in $\mathcal{I}_j = \left\{k \in \{1,\ldots,m\} \;\middle|\; \left[\bar{E}_j(h)\right]_k = \mathbf{0}\right\}$ (recall that $\left[\bar{E}_j(h)\right]_k$ is the k-th row of $\bar{E}_j(h)$). Clearly, we have

$$\|\Delta_k\|_2 \geq \|\mathrm{proj}_{\mathcal{I}_j}(\Delta_k)\|_2 \geq \left\|\mathrm{proj}_{\mathcal{I}_j}\left(\left(A - a_{\iota(k)}I_m\right)\bar{x}_k\right)\right\|_2 - \sum_{\ell=1}^{t_j - t_{j-1}} |\nu_\ell| \cdot \|\mathrm{proj}_{\mathcal{I}_j}(\bar{x}_\ell)\|_2, \tag{54}$$

where

$$\nu_\ell = \begin{cases} a_{\iota(k)} - \bar{x}_k^T A\bar{x}_k & \text{if } \ell = k, \\ \bar{x}_\ell^T A\bar{x}_k & \text{otherwise.} \end{cases}$$

Let $\lambda_{A,m} = \max_{i\in\{1,\ldots,n_A\}} |a_{s_i}|$ be the largest (in magnitude) eigenvalue of A. Using (53) and the fact that $\iota(k) \neq \iota(\ell)$ whenever $k \neq \ell$, we bound

$$\left|a_{\iota(k)} - \bar{x}_k^T A\bar{x}_k\right| \leq \left|a_{\iota(k)}\left(1 - (\bar{x}_k)^2_{\iota(k)}\right)\right| + \left|\sum_{\alpha\neq\iota(k)} a_\alpha(\bar{x}_k)^2_\alpha\right| \leq \lambda_{A,m}(m\tau^2 + 2\tau)$$

and

$$\left|\bar{x}_\ell^T A\bar{x}_k\right| \leq \lambda_{A,m}\sum_{\alpha=1}^{m} |(\bar{x}_\ell)_\alpha| \cdot |(\bar{x}_k)_\alpha| \leq \lambda_{A,m}(m\tau^2 + 2\tau) \quad \text{for } \ell \neq k.$$

This implies that $|\nu_\ell| \leq \lambda_{A,m}(m\tau^2 + 2\tau)$ for $\ell = 1,\ldots,t_j - t_{j-1}$. Moreover, since $\bar{x}_1,\ldots,\bar{x}_{t_j - t_{j-1}}$ are the columns of $\bar{X}_j$, by Proposition 5, the definition of $\mathcal{I}_j$, and the assumption that $\mathrm{dist}(X,\mathcal{X}_{h,\Pi}) = \tau$, we have

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

It follows from (54) that

$$\|\Delta_k\|_2 \geq \left\|\mathrm{proj}_{\mathcal{I}_j}\left(\left(A - a_{\iota(k)}I_m\right)\bar{x}_k\right)\right\|_2 - \lambda_{A,m}\sqrt{t_j - t_{j-1}}(m\tau^2 + 2\tau)\tau. \tag{55}$$

Next, we bound the first term on the right-hand side of the above inequality. Considering the structure of A in (8), let $i' \in \{0,1,\ldots,n_A - 1\}$ be such that $s_{i'} + 1 \leq \iota(k) \leq s_{i'+1}$ and recall that $\lambda_{A,g} = \min_{i\in\{1,\ldots,n_A-1\}}(a_{s_i} - a_{s_{i+1}}) > 0$. Then, we have

$$\begin{aligned} \left\|\mathrm{proj}_{\mathcal{I}_j}\left(\left(A - a_{\iota(k)}I_m\right)\bar{x}_k\right)\right\|_2^2 &= \sum_{i\neq i'} \sum_{\alpha\in\mathcal{I}_j\cap\{s_i+1,\ldots,s_{i+1}\}} \left(\left(a_{s_i+1} - a_{\iota(k)}\right)(\bar{x}_k)_\alpha\right)^2 \\ &\geq \lambda_{A,g}^2 \sum_{i\neq i'} \sum_{\alpha\in\mathcal{I}_j\cap\{s_i+1,\ldots,s_{i+1}\}} (\bar{x}_k)^2_\alpha \\ &= \lambda_{A,g}^2\left(\left\|\mathrm{proj}_{\mathcal{I}_j}(\bar{x}_k)\right\|_2^2 - \left\|\mathrm{proj}_{\mathcal{I}_j\cap\{s_{i'}+1,\ldots,s_{i'+1}\}}(\bar{x}_k)\right\|_2^2\right). \end{aligned} \tag{56}$$

To bound the term $\left\|\mathrm{proj}_{\mathcal{I}_j\cap\{s_{i'}+1,\ldots,s_{i'+1}\}}(\bar{x}_k)\right\|_2^2$, we proceed as follows. Let $\bar{Y} = XQ^*\Pi^T \in \mathrm{St}(m,n)$. Then, we have $\bar{X} = (P^*)^T XQ^* = (P^*)^T\bar{Y}\Pi$ and

$$\mathrm{dist}(X,\mathcal{X}_{h,\Pi}) = \left\|\bar{X} - E(h)\Pi\right\|_F = \left\|(P^*)^T\bar{Y} - E(h)\right\|_F.$$

We are now interested in locating the entries of $\mathrm{proj}_{\mathcal{I}_j \cap \{s_{i'}+1,\ldots,s_{i'+1}\}}(\bar{x}_k)$ in the matrix $(P^*)^T\bar{Y}$. Towards that end, recall that $P^* = \mathrm{BlkDiag}\left(P_1^*,\ldots,P_{n_A}^*\right)$ and consider the decomposition

$$(P^*)^T\bar{Y} = \begin{bmatrix} (P_1^*)^T\bar{Y}_{1,1} & \cdots & (P_1^*)^T\bar{Y}_{1,n_A} \\ \vdots & \ddots & \vdots \\ (P_{n_A}^*)^T\bar{Y}_{n_A,1} & \cdots & (P_{n_A}^*)^T\bar{Y}_{n_A,n_A} \end{bmatrix}, \tag{57}$$

where $P_i^* \in \mathcal{O}^{s_i - s_{i-1}}$ and $\bar{Y}_{i,i} \in \mathbb{R}^{(s_i - s_{i-1})\times h_i}$, for $i = 1,\ldots,n_A$. Since $\iota(k)$ is the coordinate of the k-th column of $\bar{E}_j(h)$ that equals 1 and $s_{i'}+1 \le \iota(k) \le s_{i'+1}$, we see from (10) and (11) that the k-th column of $\bar{E}_j(h)$ belongs to

$$E_{i'+1}(h) = \left[\begin{array}{c} \mathbf{0}_{s_{i'}\times h_{i'+1}} \\ \hline I_{h_{i'+1}} \\ \mathbf{0}_{(s_{i'+1}-s_{i'}-h_{i'+1})\times h_{i'+1}} \\ \hline \mathbf{0}_{(m-s_{i'+1})\times h_{i'+1}} \end{array}\right]. \tag{58}$$

As $\bar{x}_k$ is the k-th column of $\bar{X}_j$ and $\left\|\bar{X} - E(h)\Pi\right\|_F^2 = \sum_{j=1}^{n_B}\|\bar{X}_j - \bar{E}_j(h)\|_F^2$, it follows that all the entries of $\mathrm{proj}_{\mathcal{I}_j \cap \{s_{i'}+1,\ldots,s_{i'+1}\}}(\bar{x}_k)$ lie in $(P_{i'+1}^*)^T\bar{Y}_{i'+1,i'+1}$. Furthermore, by (58) and the definition of $\mathcal{I}_j$, the entries of $\mathrm{proj}_{\mathcal{I}_j \cap \{s_{i'}+1,\ldots,s_{i'+1}\}}(\bar{x}_k)$ do not intersect the diagonal of the top $h_{i'+1}\times h_{i'+1}$ block of $(P_{i'+1}^*)^T\bar{Y}_{i'+1,i'+1}$. Consequently, we have

$$\left\|\mathrm{proj}_{\mathcal{I}_j \cap \{s_{i'}+1,\ldots,s_{i'+1}\}}(\bar{x}_k)\right\|_2^2 \le \left\|(P_{i'+1}^*)^T\bar{Y}_{i'+1,i'+1} - \begin{bmatrix} I_{h_{i'+1}} \\ \mathbf{0} \end{bmatrix}\right\|_F^2. \tag{59}$$

To obtain an upper bound on the right-hand side of (59), we need the following lemma:

Lemma 2 *Consider the decomposition of $(P^*)^T\bar{Y}$ in (57). For $i = 1,\ldots,n_A$, let*

$$v_i^* = \min\left\{ \left\| P_i^T\bar{Y}_{i,i} - \begin{bmatrix} I_{h_i} \\ \mathbf{0} \end{bmatrix}\right\|_F^2 \,\middle|\, P_i \in \mathcal{O}^{s_i - s_{i-1}} \right\}. \tag{60}$$

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

$$\frac{1}{4}\left\|\sum_{j\ne i}\bar{Y}_{j,i}^T\bar{Y}_{j,i}\right\|_F^2 \le v_i^* \le \left\|\sum_{j\ne i}\bar{Y}_{j,i}^T\bar{Y}_{j,i}\right\|_F^2.$$

Let us defer the proof of Lemma 2 to the end of this section. Now, observe that by (11) and (17),

$$\begin{aligned} \mathrm{dist}^2(X,\mathcal{X}_{h,\Pi}) &= \min\left\{ \left\|\mathrm{BlkDiag}\left(P_1^T,\ldots,P_{n_A}^T\right)\cdot\bar{Y} - E(h)\right\|_F^2 \,\middle|\, P_i \in \mathcal{O}^{s_i - s_{i-1}} \text{ for } i = 1,\ldots,n_A \right\} \\ &= \sum_{i=1}^{n_A}\min\left\{ \left\| P_i^T\bar{Y}_{i,i} - \begin{bmatrix} I_{h_i} \\ \mathbf{0} \end{bmatrix}\right\|_F^2 \,\middle|\, P_i \in \mathcal{O}^{s_i - s_{i-1}} \right\} + \sum_{1\le i\ne j\le n_A}\|\bar{Y}_{i,j}\|_F^2. \end{aligned} \tag{61}$$

Since $\operatorname{dist}(X, \mathcal{X}_{h,\Pi}) = \tau$ for some $\tau \in (0,1)$, we have $\sum_{1 \le i \ne j \le n_A} \|\bar{Y}_{i,j}\|_F^2 \le \tau^2$ from (61). Hence, by Lemma 2 and (59), we have

$$v_i^* \le \left(\sum_{j \ne i} \|\bar{Y}_{j,i}\|_F^2 \right)^2 \le \tau^4 \qquad \text{for } i = 1, \ldots, n_A$$

and

$$\left\| \operatorname{proj}_{\mathcal{I}_j \cap \{s_{i'}+1, \ldots, s_{i'+1}\}} (\bar{x}_k) \right\|_2^2 \le v_{i'+1}^* \le \tau^4.$$

This, together with (55), (56) and the fact that the implications

$$c \ge a - b \quad \Longrightarrow \quad a^2 \le 2(b^2 + c^2) \quad \Longrightarrow \quad c^2 \ge \frac{a^2}{2} - b^2$$

hold for any $a, b, c \in \mathbb{R}$, yields

$$\|\Delta_k\|_2^2 \ge \frac{\lambda_{A,g}^2}{2} \left(\left\| \operatorname{proj}_{\mathcal{I}_j} (\bar{x}_k) \right\|_2^2 - \tau^4 \right) - \lambda_{A,m}^2 (t_j - t_{j-1})(m\tau^2 + 2\tau)^2 \tau^2.$$

It follows that

$$\begin{aligned}
\left\| A\bar{X}_j - \bar{X}_j \bar{X}_j^T A \bar{X}_j \right\|_F^2 &= \sum_{k=1}^{t_j - t_{j-1}} \|\Delta_k\|_2^2 \\
&\ge \frac{\lambda_{A,g}^2}{2} \sum_{k=1}^{t_j - t_{j-1}} \left\| \operatorname{proj}_{\mathcal{I}_j} (\bar{x}_k) \right\|_2^2 - (t_j - t_{j-1}) \left(\frac{\lambda_{A,g}^2 \tau^4}{2} + \lambda_{A,m}^2 (t_j - t_{j-1})(m\tau^2 + 2\tau)^2 \tau^2 \right) \\
&= \frac{\lambda_{A,g}^2}{2} \sum_{k \in \mathcal{I}_j} \left\| \left[\bar{X}_j\right]_k \right\|_2^2 - (t_j - t_{j-1}) \left(\frac{\lambda_{A,g}^2 \tau^4}{2} + \lambda_{A,m}^2 (t_j - t_{j-1})(m\tau^2 + 2\tau)^2 \tau^2 \right)
\end{aligned}$$

(recall that $\left[\bar{X}_j\right]_k$ is the k-th row of $\bar{X}_j$). Upon summing both sides of the above inequality over $j = 1, \ldots, n_B$ and using Proposition 5 and the assumption that $\operatorname{dist}(X, \mathcal{X}_{h,\Pi}) = \tau$, we obtain

$$\begin{aligned}
\sum_{j=1}^{n_B} \left\| A\bar{X}_j - \bar{X}_j \bar{X}_j^T A \bar{X}_j \right\|_F^2 &\ge \frac{\lambda_{A,g}^2}{4} \cdot \operatorname{dist}^2(X, \mathcal{X}_{h,\Pi}) - \frac{n\lambda_{A,g}^2 \tau^4}{2} - n^2 \lambda_{A,m}^2 (m\tau^2 + 2\tau)^2 \tau^2 \\
&\ge \frac{\lambda_{A,g}^2}{8} \sum_{j=1}^{n_B} \sum_{k \in \mathcal{I}_j} \left\| \left[\bar{X}_j\right]_k \right\|_2^2
\end{aligned}$$

whenever $\tau \in (0,1)$ satisfies

$$\left(\frac{n\lambda_{A,g}^2}{2} + n^2 \lambda_{A,m}^2 (m+2)^2 \right) \tau^2 \le \frac{\lambda_{A,g}^2}{8}.$$

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

Proof of Lemma 2. Consider a fixed $i \in \{1, \dots, n_A\}$. Note that Problem (60) is again an instance of the orthogonal Procrustes problem. Hence, by the result in [39], an optimal solution to Problem (60) 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 $\bar{Y}_{i,i} = H_i \begin{bmatrix} \Sigma_i \\ \mathbf{0} \end{bmatrix} W_i^T$ is a singular value decomposition of $\bar{Y}_{i,i}$. It follows from (60) that

$$v_i^* = \left\| (P_i^*)^T \bar{Y}_{i,i} - \begin{bmatrix} I_{h_i} \\ \mathbf{0} \end{bmatrix} \right\|_F^2 = \|\Sigma_i - I_{h_i}\|_F^2.$$

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

$$\bar{Y}_{i,i}^T \bar{Y}_{i,i} + \sum_{j \neq i} \bar{Y}_{j,i}^T \bar{Y}_{j,i} = W_i \Sigma_i^2 W_i^T + \sum_{j \neq i} \bar{Y}_{j,i}^T \bar{Y}_{j,i} = I_{h_i},$$

or equivalently,

$$\Sigma_i^2 + W_i^T \left(\sum_{j \neq i} \bar{Y}_{j,i}^T \bar{Y}_{j,i} \right) W_i = I_{h_i}.$$

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

$$\frac{1}{4} \left\| \sum_{j \neq i} \bar{Y}_{j,i}^T \bar{Y}_{j,i} \right\|_F^2 \leq v_i^* \leq \left\| \sum_{j \neq i} \bar{Y}_{j,i}^T \bar{Y}_{j,i} \right\|_F^2,$$

as desired. □

E Second-Order Boundedness of Some Retractions on $\mathrm{St}(m, n)$

E.1 Second-Order Boundedness of R_{polar}

Let $X \in \mathrm{St}(m, n)$ and $\xi \in T(X)$ be arbitrary. By definition, we have

$$\begin{aligned} \|R_{\mathsf{polar}}(X, \xi) - (X + \xi)\|_F &= \|(X + \xi)(I_n + \xi^T \xi)^{-1/2} - (X + \xi)\|_F \\ &\leq \|X + \xi\| \cdot \|(I_n + \xi^T \xi)^{-1/2} - I_n\|_F. \end{aligned}$$

Let $\xi^T \xi = U \Sigma U^T$ be a spectral decomposition of $\xi^T \xi$ with $\Sigma = \mathrm{Diag}(\lambda_1, \dots, \lambda_n)$ and $\lambda_1, \dots, \lambda_n \geq 0$. Then, a simple calculation yields

$$\|(I_n + \xi^T \xi)^{-1/2} - I_n\|_F^2 = \sum_{i=1}^n ((1 + \lambda_i)^{-1/2} - 1)^2 \leq \frac{1}{4} \sum_{i=1}^n \lambda_i^2 = \frac{1}{4} \cdot \|\xi^T \xi\|_F^2.$$

Since $\|X + \xi\| \leq \|X\| + \|\xi\| \leq 1 + \|\xi\|_F$, we conclude that whenever $\|\xi\|_F \leq 1$,

$$\|R_{\mathsf{polar}}(X, \xi) - (X + \xi)\|_F \leq \|\xi\|_F^2;$$

i.e., R_{polar} satisfies Property (P) with $\phi = M = 1$.

E.2 Second-Order Boundedness of R_{QR}

Let $X \in \mathrm{St}(m,n)$ and $\xi \in T(X)$ be arbitrary. Suppose that $\|\xi\|_F \leq 1/2$. Then, for any $t \in [-1,1]$, the matrix $X(t) = X + t\xi$ has full column rank and hence admits a unique thin QR-decomposition $X(t) = Q(t)R(t)$, where $Q(t) \in \mathrm{St}(m,n)$ and $R(t) \in \mathbb{R}^{n\times n}$ are both differentiable and $R(t)$ is upper triangular with positive diagonal entries; see, e.g., [12]. Since the unique thin QR-decomposition of X is given by $X = XI_n$, we have $R(0) = I_n$. This, together with the fact that $\|Q(t)\| \leq 1$, implies

$$\|R_{\mathsf{QR}}(X,\xi) - (X+\xi)\|_F = \|Q(1)(I_n - R(1))\|_F \leq \|R(1) - R(0)\|_F \leq \int_0^1 \|R'(t)\|_F \, dt. \tag{62}$$

To bound $\|R'(t)\|_F$, we adopt the so-called matrix equation approach in [47, 11]. Using the identity $R(t)^T R(t) = X(t)^T X(t)$ and the fact that $\xi \in T(X)$ implies $X^T\xi + \xi^T X = \mathbf{0}$, we have

$$R(t)^T R(t) = I_n + t^2 \xi^T \xi. \tag{63}$$

Differentiating both sides of (63) with respect to t yields

$$R'(t)^T R(t) + R(t)^T R'(t) = 2t\xi^T\xi.$$

In particular, since $R(t)$ is invertible, we have

$$\left(R'(t)R(t)^{-1}\right)^T + R'(t)R(t)^{-1} = 2t\left(R(t)^{-1}\right)^T (\xi^T\xi)R(t)^{-1}.$$

Now, observe that $R'(t)R(t)^{-1}$ is upper triangular. Thus, the above identity implies that

$$R'(t) = 2t \cdot \mathrm{up}\left[\left(R(t)^{-1}\right)^T (\xi^T\xi)R(t)^{-1}\right] \cdot R(t),$$

where for any $C \in \mathbb{R}^{n\times n}$,

$$[\mathrm{up}(C)]_{ij} = \begin{cases} C_{ij} & \text{if } i < j, \\ C_{ii}/2 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\lambda_1, \ldots, \lambda_n \geq 0$ be the eigenvalues of $\xi^T\xi$. Using (63) and the fact that $2 \cdot \|\mathrm{up}(C)\|_F^2 \leq \|C\|_F^2$ for any $C \in \mathcal{S}^n$, we bound

$$\begin{aligned} 2\left\|\mathrm{up}\left[\left(R(t)^{-1}\right)^T (\xi^T\xi)R(t)^{-1}\right]\right\|_F^2 &\leq \left\|\left(R(t)^{-1}\right)^T (\xi^T\xi)R(t)^{-1}\right\|_F^2 \\ &= \sum_{i=1}^n \left(\frac{\lambda_i}{1+t^2\lambda_i}\right)^2 \\ &\leq \|\xi^T\xi\|_F^2. \end{aligned}$$

On the other hand, we have $\|R(t)\| \leq \sqrt{1 + t^2 \cdot \|\xi\|^2} \leq \sqrt{5}/2$ by (63) and the assumption that $\|\xi\|_F \leq 1/2$ and $t \in [-1,1]$. It follows that

$$\|R'(t)\|_F \leq 2t \cdot \left\|\mathrm{up}\left[\left(R(t)^{-1}\right)^T (\xi^T\xi)R(t)^{-1}\right]\right\|_F \cdot \|R(t)\| \leq \frac{\sqrt{10}t}{2} \cdot \|\xi\|_F^2.$$

Upon substituting this into (62) and integrating, we obtain

$$\|R_{\mathsf{QR}}(X,\xi) - (X+\xi)\|_F \leq \frac{\sqrt{10}}{4} \cdot \|\xi\|_F^2;$$

i.e., R_{QR} satisfies Property (P) with $\phi = 1/2$ and $M = \sqrt{10}/4$.

E.3 Second-Order Boundedness of R_{cayley}

Let $X \in \mathrm{St}(m,n)$ and $\xi \in T(X)$ be arbitrary. Suppose that $\|\xi\|_F \leq 1/2$. Then, we have $\|W(\xi)\|_F \leq 2 \cdot \|\xi\|_F \leq 1$. Hence, we may write

$$\left(I_m - \frac{1}{2}W(\xi)\right)^{-1} = \sum_{i=0}^{\infty} \left(\frac{1}{2}W(\xi)\right)^i.$$

In particular, we have

$$\begin{aligned}
& \|R_{\mathsf{cayley}}(X,\xi) - (X+\xi)\|_F \\
= & \left\| \left(I_m + \frac{1}{2}W(\xi) + \sum_{i=2}^{\infty}\left(\frac{1}{2}W(\xi)\right)^i \right) \left(I_m + \frac{1}{2}W(\xi)\right) X - (X+\xi) \right\|_F \\
= & \left\| (W(\xi)X - \xi) + \frac{1}{4}W(\xi)^2 X + \left(\sum_{i=2}^{\infty}\left(\frac{1}{2}W(\xi)\right)^i \right) \left(I_m + \frac{1}{2}W(\xi)\right) X \right\|_F.
\end{aligned}$$

Now, observe that

$$W(\xi)X - \xi = \left(I_m - \frac{1}{2}XX^T \right)\xi - \frac{1}{2}X\xi^T X - \xi = -\frac{1}{2}X(X^T\xi + \xi^T X) = \mathbf{0},$$

where the last equality follows from the fact that $\xi \in T(X)$. Hence, we obtain

$$\|R_{\mathsf{cayley}}(X,\xi) - (X+\xi)\|_F \leq \frac{1}{4} \cdot \|W(\xi)\|_F^2 + \left[\sum_{i=2}^{\infty} \left(\frac{1}{2^i} + \frac{1}{2^{i+1}} \right) \right] \cdot \|W(\xi)\|_F^2 \leq 4 \cdot \|\xi\|_F^2;$$

i.e., R_{cayley} satisfies Property (P) with $\phi = 1/2$ and $M = 4$.

F Proof of Proposition 10

We first establish the inequality (44). Define $\epsilon^{k+1} = R(X^k, -\alpha\xi^k) - (X^k - \alpha\xi^k) = X^{k+1} - (X^k - \alpha\xi^k)$ for $k = 0, 1, \ldots, \Gamma - 1$. Then,

$$\begin{aligned}
F(X^{k+1}) &= \mathrm{tr}\left[(X^k - \alpha\xi^k + \epsilon^{k+1})^T A (X^k - \alpha\xi^k + \epsilon^{k+1}) B \right] \\
&= F(X^k) - \alpha \cdot \mathrm{tr}\left[\left((X^k)^T A \xi^k + (\xi^k)^T A X^k \right) B \right] \\
&\quad + \mathrm{tr}\left[\left((X^k)^T A \epsilon^{k+1} + (\epsilon^{k+1})^T A X^k \right) B \right] \\
&\quad - \alpha \cdot \mathrm{tr}\left[\left((\xi^k)^T A \epsilon^{k+1} + (\epsilon^{k+1})^T A \xi^k \right) B \right] \\
&\quad + \alpha^2 \cdot \mathrm{tr}\left[(\xi^k)^T A \xi^k B \right] + \mathrm{tr}\left[(\epsilon^{k+1})^T A \epsilon^{k+1} B \right]. \qquad (64)
\end{aligned}$$

Now, let us bound the terms in (64) in turn. Using the fact that ξ^k is the orthogonal projection of G^k onto $T(X^k)$ and $\nabla F_i(X) = 2A_i XB$, $\nabla F(X) = 2AXB$, we have

$$\|\xi^k\|_F \leq \|G^k\|_F \leq 2\left(\|A_{i_k}X^k B\|_F + \|A_{i_k}X^0 B\|_F + \|AX^0B\|_F \right) \leq 6 \cdot \|A\|_F \cdot \|B\|.$$

By our choice of the step size α, we have $\|\alpha\xi^k\|_F \leq \phi \leq 1$. It follows from Property (P) and some simple calculation that

$$\begin{aligned} \mathrm{tr}\left[\left((X^k)^T A\epsilon^{k+1} + (\epsilon^{k+1})^T AX^k\right) B\right] &\leq 2 \cdot \|A\| \cdot \|B\| \cdot \|\epsilon^{k+1}\|_F \\ &\leq 2\alpha^2 M \cdot \|A\| \cdot \|B\| \cdot \|\xi^k\|_F^2, \qquad (65) \\ -\mathrm{tr}\left[\left((\xi^k)^T A\epsilon^{k+1} + (\epsilon^{k+1})^T A\xi^k\right) B\right] &\leq 2 \cdot \|A\| \cdot \|B\| \cdot \|\xi^k\|_F \cdot \|\epsilon^{k+1}\|_F \\ &\leq 2\alpha M \cdot \|A\| \cdot \|B\| \cdot \|\xi^k\|_F^2, \qquad (66) \\ \mathrm{tr}\left[(\epsilon^{k+1})^T A\epsilon^{k+1} B\right] &\leq \|A\| \cdot \|B\| \cdot \|\epsilon^{k+1}\|_F^2 \\ &\leq \alpha^2 M^2 \cdot \|A\| \cdot \|B\| \cdot \|\xi^k\|_F^2. \qquad (67) \end{aligned}$$

Moreover, it is clear that

$$\mathrm{tr}\left[(\xi^k)^T A\xi^k B\right] \leq \|A\| \cdot \|B\| \cdot \|\xi^k\|_F^2. \qquad (68)$$

Upon substituting (65)–(68) into (64) and simplifying, we obtain

$$F(X^{k+1}) - F(X^k) + \alpha \cdot \mathrm{tr}\left[\left((X^k)^T A\xi^k + (\xi^k)^T AX^k\right) B\right] \leq c_0\alpha^2 \cdot \|\xi^k\|_F^2$$

with $c_0 = (M^2 + 4M + 1) \cdot \|A\| \cdot \|B\|$, as desired.

Next, we establish the inequality (45). Since $\xi^0 = \mathrm{grad}\, F(X^0) = \mathrm{proj}_{T(X^0)}(\nabla F(X^0))$, where $\mathrm{proj}_{T(X)}$ is the projector onto $T(X)$, by the idempotence of $\mathrm{proj}_{T(X)}$ and the fact that $\nabla F(X) = 2AXB$, we have

$$\mathrm{tr}\left[\left((X^0)^T A\xi^0 + (\xi^0)^T AX^0\right) B\right] = \|\mathrm{grad}\, F(X^0)\|_F^2.$$

Upon substituting this into (44) and noting that $c_0 \leq 1/(8\alpha)$, we obtain

$$F(X^1) - F(X^0) \leq -\frac{7\alpha}{8} \cdot \|\mathrm{grad}\, F(X^0)\|_F^2.$$

Since $X^0 = \tilde{X}^s$ and $F(\tilde{X}^{s+1}) \leq F(X^1)$ by lines 2 and 9 of Algorithm 2, respectively, the above inequality is equivalent to (45).

The inequality (45) shows that the sequence $\{F(\tilde{X}^s)\}_{s\geq 0}$ is monotonically decreasing, which, together with the fact that F is bounded below on $\mathrm{St}(m,n)$, implies that $F(\tilde{X}^s) \searrow F^*$ for some $F^* \in \mathbb{R}$. By the continuity of F, we conclude that every limit point X^* of the sequence $\{\tilde{X}^s\}_{s\geq 0}$ satisfies $F(X^*) = F^*$ and $\mathrm{grad}\, F(X^*) = \mathbf{0}$. This completes the proof of Proposition 10.

G Proof of Corollary 2

Let $\mathscr{F}_k$ be the σ-algebra generated by $X^0, \ldots, X^k$ for $k = 0, 1, \ldots, \Gamma - 1$. Since $\mathbb{E}[G^k \mid \mathscr{F}_k] = \nabla F(X^k)$, we have $\mathbb{E}[\xi^k \mid \mathscr{F}_k] = \mathrm{grad}\, F(X^k) = \mathrm{proj}_{T(X^k)}(\nabla F(X^k))$. Again, using the idempotence of $\mathrm{proj}_{T(X)}$ and the fact that $\nabla F(X) = 2AXB$, we obtain

$$\mathbb{E}\left[\mathrm{tr}\left[\left((X^k)^T A\xi^k + (\xi^k)^T AX^k\right) B\right] \,\middle|\, \mathscr{F}_k\right] = \|\mathrm{grad}\, F(X^k)\|_F^2. \qquad (69)$$

On the other hand, the non-expansiveness of $\mathrm{proj}_{T(X)}$ yields

$$\begin{aligned}
\|\xi^k\|_F &\leq \|\xi^k - \mathrm{grad}\, F(X^k)\|_F + \|\mathrm{grad}\, F(X^k)\|_F \\
&= \left\|\mathrm{proj}_{T(X^k)}(G^k) - \mathrm{proj}_{T(X^k)}(\nabla F(X^k))\right\|_F + \|\mathrm{grad}\, F(X^k)\|_F \\
&\leq \|G^k - \nabla F(X^k)\|_F + \|\mathrm{grad}\, F(X^k)\|_F \qquad (70)
\end{aligned}$$

and hence

$$\|\xi^k\|_F^2 \leq 2\left(\|G^k - \nabla F(X^k)\|_F^2 + \|\mathrm{grad}\, F(X^k)\|_F^2\right). \qquad (71)$$

By the definition of G^k and the fact that ∇F_i (resp. ∇F) is Lipschitz continuous with parameter $L_{F_i} \leq 2 \cdot \|A_i\| \cdot \|B\|$ for $i = 1, \dots, N$ (resp. $L_F \leq 2 \cdot \|A\| \cdot \|B\|$), we have

$$\begin{aligned}
\|G^k - \nabla F(X^k)\|_F &\leq \left\|\nabla F_{i_k}(X^k) - \nabla F_{i_k}(X^0)\right\|_F + \left\|\nabla F(X^0) - \nabla F(X^k)\right\|_F \\
&\leq c' \cdot \|X^k - X^0\|_F \qquad (72)
\end{aligned}$$

with $c' = 2\left(\max_{i\in\{1,\dots,N\}} \|A_i\| + \|A\|\right)\|B\|$. To bound $\|X^k - X^0\|_F$, observe that

$$\begin{aligned}
\|X^{k+1} - X^k\|_F &= \|\alpha\xi^k + \epsilon^{k+1}\|_F \\
&\leq \alpha \cdot \|\xi^k\|_F + \|\epsilon^{k+1}\|_F \\
&\leq \alpha \cdot \|\xi^k\|_F + \alpha^2 M \cdot \|\xi^k\|_F^2 \\
&\leq \alpha(M+1) \cdot \|\xi^k\|_F \qquad (73) \\
&\leq \alpha(M+1)\left(c' \cdot \|X^k - X^0\|_F + \|\mathrm{grad}\, F(X^k)\|_F\right), \qquad (74)
\end{aligned}$$

where (73) is due to the fact that $\|\alpha\xi^k\|_F \leq \phi \leq 1$ and (74) follows from (70). This yields

$$\begin{aligned}
\|X^{k+1} - X^0\|_F &\leq \|X^{k+1} - X^k\|_F + \|X^k - X^0\|_F \\
&\leq (c_1\alpha + 1) \cdot \|X^k - X^0\|_F + \alpha(M+1) \cdot \|\mathrm{grad}\, F(X^k)\|_F,
\end{aligned}$$

where $c_1 = c'(M+1)$. In particular, we have

$$\|X^{k+1} - X^0\|_F \leq \alpha(M+1)\sum_{j=0}^{k}(c_1\alpha+1)^{k-j} \cdot \|\mathrm{grad}\, F(X^j)\|_F, \qquad (75)$$

which implies that

$$\|X^{k+1} - X^0\|_F^2 \leq \alpha^2(M+1)^2(k+1)\sum_{j=0}^{k}(c_1\alpha+1)^{2(k-j)} \cdot \|\mathrm{grad}\, F(X^j)\|_F^2. \qquad (76)$$

It follows from (71), (72), and (76) that

$$\mathbb{E}\left[\|\xi^k\|_F^2\right] \leq 2c_1^2\alpha^2 k\sum_{j=0}^{k-1}(c_1\alpha+1)^{2(k-1-j)}\mathbb{E}\left[\|\mathrm{grad}\, F(X^j)\|_F^2\right] + 2\mathbb{E}\left[\|\mathrm{grad}\, F(X^k)\|_F^2\right].$$

This, together with (44) and (69), yields the desired result.

H Proof of Proposition 11

By Proposition 10, the global error bound for Problem (QP-OC) (Corollary 1), and the fact that $\mathrm{grad}\, F(X) = D_{1/4}(X)$, we have

$$F(\tilde{X}^{s+1}) - F(\tilde{X}^s) \leq -\frac{7\alpha}{8\bar{\eta}^2} \cdot \mathrm{dist}^2(\tilde{X}^s, \mathcal{X})$$

for all $s \geq 0$. Since $F(\tilde{X}^s) \searrow F^*$, the above inequality implies the existence of $s_0 \geq 0$ such that $\mathrm{dist}(\tilde{X}^s, \mathcal{X}) \leq \delta/3$ for all $s \geq s_0$, where $\delta \in (0, \sqrt{2}/2)$ is the constant given in Theorem 1. Now, consider a fixed $s \geq s_0$ and let $\hat{X}^s, \hat{X}^{s+1} \in \mathcal{X}$ be such that $\mathrm{dist}(\tilde{X}^s, \mathcal{X}) = \|\tilde{X}^s - \hat{X}^s\|_F$ and $\mathrm{dist}(\tilde{X}^{s+1}, \mathcal{X}) = \|\tilde{X}^{s+1} - \hat{X}^{s+1}\|_F$. Suppose that $\hat{X}^s \in \mathcal{X}_{h,\Pi}$ and $\hat{X}^{s+1} \in \mathcal{X}_{h',\Pi'}$ with $\mathcal{X}_{h,\Pi} \cap \mathcal{X}_{h',\Pi'} = \emptyset$. Then, we have $\|\hat{X}^s - \hat{X}^{s+1}\|_F \geq \sqrt{2} \geq 2\delta$ by Proposition 4. On the other hand, using (75), the fact that $\|\mathrm{grad}\, F(X)\|_F \leq \|\nabla F(X)\|_F \leq 2 \cdot \|A\|_F \cdot \|B\|$ for all $X \in \mathrm{St}(m,n)$, and our choice of the step size α, the sequence $X^0 = \tilde{X}^s, X^1, \ldots, X^\Gamma$ generated by Algorithm 2 in epoch s satisfies

$$\begin{aligned} \|X^{k+1} - X^0\|_F &\leq 2\alpha(M+1) \cdot \|A\|_F \cdot \|B\| \sum_{j=0}^{k} (c_1\alpha + 1)^{k-j} \\ &= \frac{2(M+1)\left((c_1\alpha+1)^{k+1} - 1\right) \cdot \|A\|_F \cdot \|B\|}{c_1} \\ &\leq \frac{\delta}{3} \end{aligned}$$

for $k = 0, 1, \ldots, \Gamma - 1$. This implies that

$$\|\hat{X}^s - \hat{X}^{s+1}\|_F \leq \|\hat{X}^s - \tilde{X}^s\|_F + \|\tilde{X}^{s+1} - \tilde{X}^s\|_F + \|\hat{X}^{s+1} - \tilde{X}^{s+1}\|_F \leq \delta,$$

which is a contradiction. Hence, we have $\mathcal{X}_{h,\Pi} \cap \mathcal{X}_{h',\Pi'} \neq \emptyset$, which by Proposition 4 yields $\mathcal{X}_{h,\Pi} = \mathcal{X}_{h',\Pi'}$. Consequently, we have $\mathrm{dist}(\tilde{X}^s, \mathcal{X}) = \mathrm{dist}(\tilde{X}^s, \mathcal{X}_{h,\Pi}) \leq \delta/3$ for all sufficiently large $s \geq 0$. This, together with Proposition 10 and the fact that the function F is constant on $\mathcal{X}_{h,\Pi}$, implies that every limit point of the sequence $\{\tilde{X}^s\}_{s\geq 0}$ belongs to $\mathcal{X}_{h,\Pi}$ and $F(X) = F^*$ for all $X \in \mathcal{X}_{h,\Pi}$. This completes the proof.

References

[1] T. E. Abrudan, J. Eriksson, and V. Koivunen. Steepest Descent Algorithms for Optimization under Unitary Matrix Constraint. *IEEE Transactions on Signal Processing*, 56(3):1134–1147, 2008.

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

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

[4] P.-A. Absil and J. Malick. Projection–Like Retractions on Matrix Manifolds. *SIAM Journal on Optimization*, 22(1):135–158, 2012.

[5] A. Agarwal, A. Anandkumar, P. Jain, and P. Netrapalli. Learning Sparsely Used Overcomplete Dictionaries via Alternating Minimization. *SIAM Journal on Optimization*, 26(4):2775–2799, 2016.

[6] M. Bolla, G. Michaletzky, G. Tusnády, and M. Ziermann. Extrema of Sums of Heterogeneous Quadratic Forms. *Linear Algebra and Its Applications*, 269(1–3):331–365, 1998.

[7] J. Bolte, A. Danilidis, O. Ley, and L. Mazet. Characterizations of Łojasiewicz Inequalities: Subgradient Flows, Talweg, Convexity. *Transactions of the American Mathematical Society*, 362(6):3319–3363, 2010.

[8] J. Bolte, T. P. Ngyuen, J. Peypouquet, and B. W. Suter. From Error Bounds to the Complexity of First–Order Descent Methods for Convex Functions. *Mathematical Programming, Series A*, 165(2):471–507, 2017.

[9] S. Bonnabel. Stochastic Gradient Descent on Riemannian Manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013.

[10] E. J. Candès, X. Li, and M. Soltanolkotabi. Phase Retrieval via Wirtinger Flow: Theory and Algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[11] X.-W. Chang, C. C. Paige, and G. W. Stewart. Perturbation Analyses for the QR Factorization. *SIAM Journal on Matrix Analysis and Applications*, 18(3):775–791, 1997.

[12] L. Dieci and T. Eirola. On Smooth Decompositions of Matrices. *SIAM Journal on Matrix Analysis and Applications*, 20(3):800–819, 1999.

[13] P. M. N. Feehan. Global Existence and Convergence of Solutions to Gradient Systems and Applications to Yang–Mills Gradient Flow. Monograph, available at `https://arxiv.org/abs/1409.1525`, 2014.

[14] M. Forti, P. Nistri, and M. Quincampoix. Convergence of Neural Networks for Programming Problems via a Nonsmooth Łojasiewicz Inequality. *IEEE Transactions on Neural Networks*, 17(6):1471–1486, 2006.

[15] G. H. Golub and C. F. Van Loan. *Matrix Computations.* The Johns Hopkins University Press, Baltimore, Maryland, third edition, 1996.

[16] M. Hardt. Understanding Alternating Minimization for Matrix Completion. In *Proceedings of the 55th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2014)*, pages 651–660, 2014.

[17] K. Hou, Z. Zhou, A. M.-C. So, and Z.-Q. Luo. On the Linear Convergence of the Proximal Gradient Method for Trace Norm Regularization. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: Proceedings of the 2013 Conference*, pages 710–718, 2013.

[18] P. Jain and S. Oh. Provable Tensor Factorization with Missing Data. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Proceedings of the 2014 Conference*, pages 1431–1439, 2014.

[19] B. Jiang and Y.-H. Dai. A Framework of Constraint Preserving Update Schemes for Optimization on Stiefel Manifold. *Mathematical Programming, Series A*, 153(2):535–575, 2015.

[20] R. Johnson and T. Zhang. Accelerating Stochatic Gradient Descent Using Predictive Variance Reduction. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: Proceedings of the 2013 Conference*, pages 315–323, 2013.

[21] T. Kaneko, S. Fiori, and T. Tanaka. Empirical Arithmetic Averaging over the Compact Stiefel Manifold. *IEEE Transactions on Signal Processing*, 61(4):883–894, 2013.

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

[23] G. Li, B. S. Mordukhovich, and T. S. Phạm. 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.

[24] G. Li and T. K. Pong. Calculus of the Exponent of Kurdyka–Łojasiewicz Inequality and Its Applications to Linear Convergence of First–Order Methods. Accepted for publication in *Foundations of Computational Mathematics*, 2017.

[25] H. Liu, W. Wu, and A. M.-C. So. Quadratic Optimization with Orthogonality Constraints: Explicit Łojasiewicz Exponent and Linear Convergence of Line–Search Methods. In *Proceedings of the 33rd International Conference on Machine Learning (ICML 2016)*, pages 1158–1167, 2016.

[26] H. Liu, M.-C. Yue, and A. M.-C. So. On the Estimation Performance and Convergence Rate of the Generalized Power Method for Phase Synchronization. *SIAM Journal on Optimization*, 27(4):2426–2446, 2017.

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

[28] Z.-Q. Luo and J.-S. Pang. Error Bounds for Analytic Systems and Their Applications. *Mathematical Programming*, 67(1):1–28, 1994.

[29] Z.-Q. Luo and J. F. Sturm. Error Bounds for Quadratic Systems. In H. Frenk, K. Roos, T. Terlaky, and S. Zhang, editors, *High Performance Optimization*, volume 33 of *Applied Optimization*, pages 383–404. Springer Science+Business Media, Dordrecht, 2000.

[30] Z.-Q. Luo and P. Tseng. Error Bounds and Convergence Analysis of Feasible Descent Methods: A General Approach. *Annals of Operations Research*, 46(1):157–178, 1993.

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

[32] B. Merlet and T. N. Nguyen. Convergence to Equilibrium for Discretizations of Gradient–Like Flows on Riemannian Manifolds. *Differential and Integral Equations*, 26(5–6):571–602, 2013.

[33] Yu. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course.* Kluwer Academic Publishers, Boston, 2004.

[34] P. Netrapalli, P. Jain, and S. Sanghavi. Phase Retrieval Using Alternating Minimization. *IEEE Transactions on Signal Processing*, 63(18):4814–4826, 2015.

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

[36] H. Sato and T. Iwai. A Riemannian Optimization Approach to the Matrix Singular Value Decomposition. *SIAM Journal on Optimization*, 23(1):188–212, 2013.

[37] H. Sato, H. Kasai, and B. Mishra. Riemannian Stochastic Variance Reduced Gradient. Manuscript, available at `https://arxiv.org/abs/1702.05594`, 2017.

[38] R. Schneider and A. Uschmajew. 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.

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

[40] P. H. Schönemann. On Two–Sided Orthogonal Procrustes Problems. *Psychometrika*, 33(1):19–33, 1968.

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

[42] O. Shamir. Fast Stochastic Algorithms for SVD and PCA: Convergence Properties and Convexity. In *Proceedings of the 33rd International Conference on Machine Learning (ICML 2016)*, pages 248–256, 2016.

[43] S. T. Smith. Optimization Techniques on Riemannian Manifolds. In A. Bloch, editor, *Hamiltonian and Gradient Flows, Algorithms and Control*, Fields Institue Communications, pages 113–136. American Mathematical Society, Providence, Rhode Island, 1994.

[44] A. M.-C. So. Moment Inequalities for Sums of Random Matrices and Their Applications in Optimization. *Mathematical Programming, Series A*, 130(1):125–151, 2011.

[45] A. M.-C. So. Pinning Down the Łojasiewicz Exponent: Towards Understanding the Convergence Behavior of First–Order Methods for Structured Non–Convex Optimization Problems. Slides, available at `http://lamda.nju.edu.cn/conf/mla15/files/suwz.pdf`, 2015.

[46] A. M.-C. So and Z. Zhou. Non–Asymptotic Convergence Analysis of Inexact Gradient Methods for Machine Learning Without Strong Convexity. *Optimization Methods and Software*, 32(4):963–992, 2017.

[47] J. Sun. On Perturbation Bounds for the QR Factorization. *Linear Algebra and Its Applications*, 215:95–111, 1995.

[48] J. Sun, Q. Qu, and J. Wright. A Geometric Analysis of Phase Retrieval. Accepted for publication in *Foundations of Computational Mathematics*, 2017.

[49] J. Sun, Q. Qu, and J. Wright. Complete Dictionary Recovery Over the Sphere I: Overview and the Geometric Picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2017.

[50] J. Sun, Q. Qu, and J. Wright. Complete Dictionary Recovery Over the Sphere II: Recovery by Riemannian Trust–Region Method. *IEEE Transactions on Information Theory*, 63(2):885–914, 2017.

[51] R. Sun and Z.-Q. Luo. Guaranteed Matrix Completion via Non–convex Factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

[52] W. W. Sun, J. Lu, H. Liu, and G. Cheng. Provable Sparse Tensor Decomposition. *Journal of the Royal Statistical Society, Series B*, 79(3):899–916, 2017.

[53] C. Udrişte. *Convex Functions and Optimization Methods on Riemannian Manifolds*, volume 297 of *Mathematics and Its Applications*. Springer Science+Business Media, B.V., Dordrecht, The Netherlands, 1994.

[54] A. Uschmajew. A New Convergence Proof for the Higher–Order Power Method and Generalizations. *Pacific Journal of Optimization*, 11(2):309–321, 2015.

[55] Z. Wen and W. Yin. A Feasible Method for Optimization with Orthogonality Constraints. *Mathematical Programming, Series A*, 142(1–2):397–434, 2013.

[56] Y. Yang. Globally Convergent Optimization Algorithms on Riemannian Manifolds: Uniform Framework for Unconstrained and Constrained Optimization. *Journal of Optimization Theory and Applications*, 132(2):245–265, 2007.

[57] F. Yger, M. Berar, G. Gasso, and A. Rakotomamonjy. Adaptive Canonical Correlation Analysis Based on Matrix Manifolds. In *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*, pages 1071–1078, 2012.

[58] H. Zhang, S. J. Reddi, and S. Sra. Riemannian SVRG: Fast Stochastic Optimization on Riemannian Manifolds. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29: Proceedings of the 2016 Conference*, pages 4592–4600, 2016.

[59] H. Zhang and S. Sra. First–Order Methods for Geodesically Convex Optimization. In V. Feldman, A. Rakhlin, and O. Shamir, editors, *Proceedings of the 29th Annual Conference on Learning Theory (COLT 2016)*, volume 49 of *Proceedings of Machine Learning Research*, pages 1617–1638, 2016.

[60] Q. Zheng and J. Lafferty. A Convergent Gradient Descent Algorithm for Rank Minimization and Semidefinite Programming from Random Linear Measurements. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28: Proceedings of the 2015 Conference*, pages 109–117, 2015.

[61] Y. Zhong and N. Boumal. Near–Optimal Bounds for Phase Synchronization. *SIAM Journal on Optimization*, 28(2):989–1016, 2018.

[62] Z. Zhou and A. M.-C. So. A Unified Approach to Error Bounds for Structured Convex Optimization Problems. *Mathematical Programming, Series A*, 165(2):689–728, 2017.

[63] Z. Zhou, Q. Zhang, and A. M.-C. So. $\ell_{1,p}$–Norm Regularization: Error Bounds and Convergence Rate Analysis of First–Order Methods. In *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*, pages 1501–1510, 2015.