Journal of Computational Mathematics
Vol.xx, No.x, 2023, 1–31.
http://www.global-sci.org/jcm
doi:10.4208/jcm.2307-m2021-0331

SOLVING OPTIMIZATION PROBLEMS OVER THE STIEFEL MANIFOLD BY SMOOTH EXACT PENALTY FUNCTIONS*

Nachuan Xiao
The Institute of Operations Research and Analytics, National University of Singapore, Singapore
Email: xnc@lsec.cc.ac.cn
Xin Liu[1)]
State Key Laboratory of Scientific and Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, and University of Chinese Academy of Sciences, China
Email: liuxin@lsec.cc.ac.cn

Abstract

In this paper, we present a novel penalty model called ExPen for optimization over the Stiefel manifold. Different from existing penalty functions for orthogonality constraints, ExPen adopts a smooth penalty function without using any first-order derivative of the objective function. We show that all the first-order stationary points of ExPen with a sufficiently large penalty parameter are either feasible, namely, are the first-order stationary points of the original optimization problem, or far from the Stiefel manifold. Besides, the original problem and ExPen share the same second-order stationary points. Remarkably, the exact gradient and Hessian of ExPen are easy to compute. As a consequence, abundant algorithm resources in unconstrained optimization can be applied straightforwardly to solve ExPen.

Mathematics subject classification: 90C30, 65K05.
Key words: Orthogonality constraint, Stiefel manifold, Penalty function.

1. Introduction

In this paper, we consider the following optimization problem:

$$\begin{aligned} \min_{X\in\mathbb{R}^{n\times p}} \quad & f(X) \\ \text{s.t.} \quad & X^\top X = I_p, \end{aligned} \tag{OCP}$$

where I_p denotes the $p\times p$ identity matrix, and $f:\mathbb{R}^{n\times p}\mapsto\mathbb{R}$ satisfies the following assumption.

Assumption 1.1 (Blank Assumption on f). *The functions f and ∇f are locally Lipschitz continuous in $\mathbb{R}^{n\times p}$.*

Recall that a mapping $T:\mathbb{R}^{n\times p}\to\mathbb{R}^m$ is locally Lipschitz continuous over $\mathbb{R}^{n\times p}$ if for any $X_0\in\mathbb{R}^{n\times p}$, there exists a constant M and $\delta>0$ such that for any $X\in\mathbb{R}^{n\times p}$ satisfying $\|X-X_0\|_{\mathrm{F}}\le\delta$, it holds that $\|T(X)-T(X_0)\|\le M\|X-X_0\|_{\mathrm{F}}$.

The feasible region of the orthogonality constraints $X^\top X=I_p$ is the Stiefel manifold embedded in the $n\times p$ real matrix space, denoted by

$$\mathcal{S}_{n,p}:=\left\{X\in\mathbb{R}^{n\times p}\mid X^\top X=I_p\right\}.$$

* Received November 20, 2021 / Revised version received November 2, 2022 / Accepted July 26, 2023 / Published online October 27, 2023 /
1) Corresponding author

We also call it as the Stiefel manifold for brevity. Optimization problems with orthogonality constraints have wide applications in statistics [17, 49], scientific computation [37, 43], image processing [6] and many other related areas [25, 41, 68]. Interested readers can refer to some recent works [21, 35, 59, 63], a recent survey [33], and several books [4, 10] for details.

1.1. Motivation

Optimization over the Stiefel manifold, which is a smooth and compact Riemannian manifold, has been discovered to enjoy a close relationship with unconstrained optimization. However, developing optimization approaches over the Stiefel manifold is inherently complicated by the nonconvexity of the manifold. Various existing unconstrained optimization approaches, i.e. the approaches for solving nonconvex unconstrained optimization problems, can be extended to their Riemannian versions by the local diffeomorphisms between the Stiefel manifold and Euclidean space. The approaches, called Riemannian optimization approaches for brevity hereinafter, include gradient descent with line-search [2, 4, 36, 58, 60], conjugate gradient methods [3], Riemannian accelerated gradient method [13, 54, 66, 67], Riemannian adaptive gradient methods [7], etc. With the frameworks and geometrical materials described in [4], theoretical results of these Riemannian optimization approaches have been established by following almost the same proof techniques as their unconstrained prototypes. These results include the global convergence, local convergence rate, worst-case complexity, and saddle-point-escaping properties, [5, 12, 13, 24, 32, 55, 69].

The Riemannian optimization approaches usually consist of two fundamental parts. The first one is the so-called retraction which maps a point from the tangent space to the manifold. Retractions can be further categorized into two classes: the geodesic-like retractions and the projection-like ones. The former ones require to calculate the geodesics along the manifold and hence are expensive. The latter ones enjoy relatively lower computational cost, but as demonstrated in various existing works [21, 61, 63], computing the projection-like retractions is still more expensive than matrix-matrix multiplication. The second part is called parallel transport which moves a tangent vector along a given curve on a Stiefel manifold parallelly. The purpose of parallel transport is to design the manifold version of some advanced unconstrained optimization approaches, such as conjugate gradient methods or gradient methods with momentum. However, as illustrated in [4], computing the parallel transport on Stiefel manifold is equivalent to finding a solution to a differential equation, which is definitely impractical in computation. To this end, the authors of [4] have proposed the concept of vector transport, which can be regarded as an approximation to parallel transport, hence is computationally affordable. Unfortunately, due to the approximation error introduced by vector transports, analyzing the convergence properties of Riemannian optimization algorithms is challenging and usually cannot directly follow the existing results for their unconstrained counterparts, see [35] for instances. As illustrated in various existing works [5, 12, 13, 66, 67, 69], both parallel transports and geodesics play an essential role in establishing convergence properties. It is still difficult to verify whether their theoretical convergence properties is valid when these approaches are built by retractions and vector transports.

To avoid computing the retractions, parallel transports, or vector transports to the Stiefel manifold, some approaches aim to find smooth mappings from the Euclidean space to the Stiefel manifold, which directly reformulates OCP to unconstrained optimization. Among them, [38,39] construct equivalent unconstrained problems for OCP by exponential function for square matrices. To efficiently compute the matrix exponential, they apply the iterative approach proposed by [30]. Therefore, their approaches require $\mathcal{O}(n^3)$ flops in each iteration for computing the matrix exponential and thus are computationally expensive in practice. Inspired by [38,39], several recent works use Cayley transformation [16, 27, 40, 47] to avoid computing the matrices exponential. These approaches require computing the inverse of $n \times n$ matrices in each iterate, which still requires $\mathcal{O}(n^3)$ flops in general. Furthermore, calculating the derivatives through the exponential mapping or the Cayley transformation can be more costly. In particular, as illustrated in the numerical experiments in [1, 16], when applying nonlinear conjugate gradient methods, the computational time of these approaches is usually much higher than existing Riemannian conjugate gradient approaches.

Recently, some infeasible approaches have been verified to be efficient in solving optimization problems over the Stiefel manifold. They utilize a completely different approach from existing Riemannian optimization methods. Based on the framework of the augmented Lagrangian method (ALM) [8,28,48,51], the authors of [21] have proposed the proximal linearized augmented Lagrangian method (PLAM) and its column-wise normalization version (PCAL) for (OCP). Both PLAM and PCAL update the multipliers corresponding to the orthogonality constraints by a closed-form expression. Additionally, [1] have proposed the landing algorithm, which follows a two-step alternative updating framework. Inspired by the closed-form updating scheme in PLAM and PCAL, the authors of [61] have proposed an exact penalty function named PenC,

$$\min_{\|X\|_{\mathrm{F}} \leq K} \quad h_{PenC}(X) = f(X) - \frac{1}{2}\left\langle \Phi\left(X^\top \nabla f(X)\right), X^\top X - I_p \right\rangle + \frac{\beta}{4}\left\| X^\top X - I_p \right\|_{\mathrm{F}}^2,$$

where $K \geq \sqrt{p}$ is a prefixed constant and Φ is the symmetrization operator defined as

$$\Phi(M) := \frac{1}{2}(M + M^\top).$$

In [61], the authors have illustrated the equivalence between OCP and PenC, which further proposed the corresponding infeasible first-order and second-order methods PenCF and PenCS, respectively. Moreover, successive works [34, 62] have illustrated that PenC could be extended to objective function with special structures. The above-mentioned penalty-function-based approaches are verified to enjoy high efficiency and scalability due to avoiding retractions or parallel transport to the Stiefel manifold. However, their penalty functions involve the first-order derivatives of the original objective, which leads to two limitations. Firstly, the smoothness of the penalty function requires higher-order smoothness of the original objective function. Secondly, calculating an exact gradient of these penalty functions is usually expensive in practice. As a result, many existing unconstrained optimization approaches cannot be directly applied to minimize these penalty functions.

1.2. Contributions

The contributions of this paper can be summarized as the following two folds.

A novel penalty function. We propose a novel penalty function

$$h(X) := f\left(X\left(\frac{3}{2}I_p - \frac{1}{2}X^\top X\right)\right) + \frac{\beta}{4}\left\| X^\top X - I_p \right\|_{\mathrm{F}}^2, \tag{1.1}$$

and construct the following unconstrained optimization problem which is abbreviated as ExPen:

$$\min_{X \in \mathbb{R}^{n\times p}} h(X). \tag{ExPen}$$

With a sufficiently large penalty parameter β, we illustrate that any first-order stationary point (FOSP) of ExPen is either feasible and hence a FOSP of OCP, or far away from the Stiefel manifold. Besides, we prove that any eigenvalue of the Riemannian Hessian at any first-order stationary point $X \in \mathcal{S}_{n,p}$ is an eigenvalue of $\nabla^2 h(X)$. Then we show that any second-order stationary point (SOSP) of ExPen is an SOSP of OCP. We call the above two relationships the first-order relationship and second-order relationship, respectively, for brevity. These two relationships imply that ExPen can be regarded as an exact penalty function.

An universal tool. The exact penalty model ExPen builds up a bridge between various existing unconstrained optimization approaches and OCP. Moreover, those rich theoretical results of unconstrained optimization approaches can be directly applied in solving OCP. In particular, some newly developed techniques for unconstrained optimization can be extended to solve optimization over the Stiefel manifold through ExPen. We use the nonlinear conjugate gradient method as an example. It is difficult to find a compromise between computational efficiency and theoretical guarantee if we adopt Riemannian optimization approaches to achieve this extension. Preliminary numerical experiments illustrates that ExPen yields direct and efficient of nonlinear conjugate gradient solver from SciPy package.

1.3. Notations

In this paper, the Euclidean inner product of two matrices $X, Y \in \mathbb{R}^{n\times p}$ is defined as $\langle X, Y\rangle = \mathrm{tr}(X^\top Y)$, where $\mathrm{tr}(A)$ is the trace of the square matrix A. Besides, $\|\cdot\|_2$ and $\|\cdot\|_{\mathrm{F}}$ represent the 2-norm and the Frobenius norm, respectively. The notations $\mathrm{diag}(A)$ and $\mathrm{Diag}(x)$ stand for the vector formed by the diagonal entries of matrix A, and the diagonal matrix with the entries of $x \in \mathbb{R}^n$ to be its diagonal, respectively. We denote the smallest eigenvalue of A by $\lambda_{\min}(A)$. We set the Riemannian metric on Stiefel manifold as the metric inherited from the standard inner product in $\mathbb{R}^{n,p}$. We set $\mathcal{T}_X$ as the tangent space of Stiefel manifold at X, which can be expressed as

$$\mathcal{T}_X := \left\{D \in \mathbb{R}^{n\times p} \mid \Phi(D^\top X) = 0\right\},$$

while $\mathcal{N}_X$ is denoted as the normal space of Stiefel manifold at X

$$\mathcal{N}_X := \left\{D \in \mathbb{R}^{n\times p} \mid D = X\Lambda, \Lambda = \Lambda^\top\right\}.$$

And $\mathrm{grad}\, f(X)$ denotes the Riemannian gradient of f at $X \in \mathcal{S}_{n,p}$ in Riemannian metric that is inherited from the Euclidean metric, namely,

$$\mathrm{grad}\, f(X) := \nabla f(X) - X\Phi\big(X^\top \nabla f(X)\big).$$

Besides, we uses $\nabla^2 f(X)[D]$ to represent the Hessian-matrix product. The Riemannian Hessian of f at $X \in \mathcal{S}_{n,p}$ in Euclidean measure is denoted as $\mathrm{hess}\, f(X) : \mathcal{T}_X \to \mathcal{T}_X$, whose bilinear form can be written as

$$\langle D_1, \mathrm{hess}\, f(X)[D_2]\rangle := \left\langle D_1, \nabla^2 f(X)[D_2] - D_2\Phi\big(X^\top \nabla f(X)\big)\right\rangle, \quad \forall D_1, D_2 \in \mathcal{T}_X.$$

Finally, $\mathcal{P}_{\mathcal{S}_{n,p}}(X)=UV^\top$ denotes the orthogonal projection to Stiefel manifold, where $X=U\Sigma V^\top$ is the economic SVD of X with $U \in \mathcal{S}_{n,p}, V \in \mathcal{S}_{p,p}$ and Σ is $p\times p$ diagonal matrix with the singular values of X on its diagonal.

The rest of this paper is organized as follows. In Section 2, we present several preliminaries and useful lemmas. Then we explore the first-order and second-order relationships between OCP and ExPen, respectively, in Section 3. We show how to solve OCP through unconstrained optimization approaches by an illustrative example in Section 4 and draw a brief conclusion in the last section.

2. Preliminaries

In this section, we provide several preliminary properties of ExPen. We first introduce the definitions, assumptions and define several constants. Then we give some preliminary properties of ExPen. Finally, we present the computational complexity of calculating the derivatives of ExPen.

2.1. Definitions

The first-order optimality condition of problem OCP can be written as

Definition 2.1 ([4]). *Given a point $X \in \mathcal{S}_{n,p}$, we call X a first-order stationary point of OCP if* $\mathrm{grad}\, f(X) = 0$.

According to [20], any $X \in \mathbb{R}^{n\times p}$ is a first-order stationary point of OCP if and only if it satisfies

$$\begin{cases} \nabla f(X) - X\Phi\big(X^\top \nabla f(X)\big) = 0, \\ X^\top X = I_p. \end{cases}$$

Next, we present the definition of the second-order optimality condition of OCP.

Definition 2.2. *Given a point $X \in \mathcal{S}_{n,p}$, if f is twice-differentiable, X is a first-order stationary point of OCP and*

$$\langle D, \mathrm{hess}\, f(X)[D]\rangle \geq 0$$

holds for any $D \in \mathcal{T}_X$, then we call X a second-order stationary point of OCP.

Besides, we present the definitions of first-order and second-order optimality conditions of ExPen. Given a point $X \in \mathbb{R}^{n\times p}$, we say X is a first-order stationary point of a differentiable function $h : \mathbb{R}^{n\times p} \to \mathbb{R}$ if and only if $\nabla h(X) = 0$. And when h is twice-order differentiable, X is a second-order stationary point of h if and only if X is a first-order stationary point of h and

$$\big\langle \nabla^2 h(X)[D], D\big\rangle \geq 0, \quad \forall D \in \mathbb{R}^{n\times p}. \tag{2.1}$$

Next we present the definitions of Łojasiewicz inequality [45, 46], which coincide with the definitions in [9].

Definition 2.3. *Let f be a differentiable function. Then f is said to have the Euclidean Łojasiewicz gradient inequality at $X \in \mathbb{R}^{n\times p}$ if and only if there exists a neighborhood U of X, and constants $\theta \in (0, 1]$, $C > 0$ such that for any $Y \in U$,*

$$\|\nabla f(Y)\|_{\mathrm{F}} \geq C|f(Y) - f(X)|^{1-\theta}.$$

Besides, we present the definitions of Riemannian Łojasiewicz inequality [31].

Definition 2.4. *Let f be a differentiable function. Then f is said to have the Riemannian Łojasiewicz gradient inequality at $X \in \mathcal{S}_{n,p}$ if and only if there exists a neighborhood $U \subset \mathcal{S}_{n,p}$ of X, and constants $\theta \in (0,1]$, $C > 0$ such that for any $Y \in U$,*

$$\|\mathrm{grad}\, f(Y)\|_{\mathrm{F}} \geq C|f(Y) - f(X)|^{1-\theta}.$$

The constant θ is usually named as Łojasiewicz exponent in the gradient inequality.

2.2. Assumptions

In this subsection, we present some additional assumptions on the objective function f in OCP, which are usually optional throughout this paper. Before presenting these additional assumptions used in some parts of this paper, we first define some set and operators:

- $\Omega := \{X \in \mathbb{R}^{n\times p} \mid \|X\|_2 \leq 1 + 1/12\}$,
- $\overline{\Omega}_r := \{X \in \mathbb{R}^{n\times p} \mid \|X^\top X - I_p\|_{\mathrm{F}} \leq r\}$,
- $G(X) := \nabla f(Y)|_{Y = X(3I_p/2 - X^\top X/2)}$,
- $\mathcal{H}(X) := \nabla^2 f(Y)|_{Y = X(3I_p/2 - X^\top X/2)}$,
- $J_X(D) := DX(3I_p/2 - X^\top X/2) - X\Phi(D^\top X)$,
- $g(X) := f\big(X(3I_p/2 - X^\top X/2)\big)$.

Clearly, we have $\overline{\Omega}_{1/12} \subset \overline{\Omega}_{1/6} \subset \Omega$. In addition, we present several constants for the theoretical analysis of ExPen:

- $M_0 := \sup_{X\in\Omega} f(X) - \inf_{X\in\Omega} f(X)$,
- $M_1 := \sup_{X\in\Omega} \|G(X)\|_{\mathrm{F}}$,
- $M_2 := \sup_{X,Y\in\Omega, X\neq Y} \big(\|\nabla g(X) - \nabla g(Y)\|_{\mathrm{F}} / \|X - Y\|_{\mathrm{F}}\big)$,
- $\bar{\beta} := \max\{12M_1, 6M_2\}$.

It is worth mentioning that Assumption 1.1 guarantees that both f and ∇f are locally Lipschitz continuous over $\mathbb{R}^{n\times p}$. Consequently, both G and ∇g are locally Lipschitz continuous over $\mathbb{R}^{n\times p}$. Therefore, together with the fact that Ω is a compact subset of $\mathbb{R}^{n\times p}$, the parameters M_0, M_1 and M_2 are finite and independent from the penalty parameter β.

Assumption 2.1 (The Global Lipschitz Continuity of f). *f is globally Lipschitz continuous in $\mathbb{R}^{n\times p}$.*

Although Assumption 2.1 is restrictive, it is optional in our theoretical analysis. It will be specifically mentioned where applicable.

Moreover, under Assumption 2.1, we define several additional constants for ExPen:

- $\hat{M}_1 := \sup_{X\in\mathbb{R}^{n\times p}} \|G(X)\|_{\mathrm{F}}$,
- $\hat{M}_2 := \sup_{X,Y\in\mathbb{R}^{n\times p}, X\neq Y} \big(\|\nabla g(X) - \nabla g(Y)\|_{\mathrm{F}} / \|X - Y\|_{\mathrm{F}}\big)$,
- $\hat{\beta} := \max\{12\hat{M}_1, 6\hat{M}_2\}$.

We emphasize that parameters $\hat{M}_1$ and $\hat{M}_2$ are independent with the penalty parameter β. Besides, it follows from Assumption 2.1 that $\hat{M}_1 \geq M_1$ and $\hat{M}_2 \geq M_2$.

Furthermore, when we analyze the Hessian at $h(X)$, the objective function f in OCP should be twice-differentiable. As a results, in some cases we assume the objective function f be twice differentiable.

Assumption 2.2 (The Second-Order Differentiability of f). *$\nabla^2 f(X)$ exists at every $X \in \mathbb{R}^{n\times p}$.*

In the rest of this subsection, we present several useful lemmas for further use. We first show that J_X is the Jacobian of the mapping $X \mapsto X(3I_p/2 - X^\top X/2)$ in the following lemma.

Proposition 2.1. *For any $X, Y \in \mathbb{R}^{n\times p}$, let $D = Y - X$, we have*

$$\left\| Y\left(\frac{3}{2}I_p - \frac{1}{2}Y^\top Y\right) - X\left(\frac{3}{2}I_p - \frac{1}{2}X^\top X\right) - J_X(D)\right\|_{\mathrm{F}} = \mathcal{O}\big(\|D_{\mathrm{F}}^2\|\big).$$

Besides,

$$\left\| Y\left(\frac{3}{2}I_p - \frac{1}{2}Y^\top Y\right) - X\left(\frac{3}{2}I_p - \frac{1}{2}X^\top X\right) - J_X(D) - \left[D\Phi(D^\top X) + \frac{1}{2}XD^\top D\right]\right\|_{\mathrm{F}} = \mathcal{O}\big(\|D_{\mathrm{F}}^3\|\big).$$

Proof. Let $D = Y - X$, from the expression of $Y(3I_p/2 - Y^\top Y/2)$ we can conclude that

$$\begin{aligned} Y\left(\frac{3}{2}I_p - \frac{1}{2}Y^\top Y\right) &= X\left(\frac{3}{2}I_p - \frac{1}{2}X^\top X\right) + D\left(\frac{3}{2}I_p - \frac{1}{2}X^\top X\right) \\ &\quad - X\Phi(X^\top D) - D\Phi(D^\top X) - \frac{1}{2}XD^\top D - \frac{1}{2}DD^\top D \\ &= X\left(\frac{3}{2}I_p - \frac{1}{2}X^\top X\right) + J_X(D) - D\Phi(D^\top X) - \frac{1}{2}XD^\top D - \frac{1}{2}DD^\top D, \end{aligned}$$

and thus complete the proof. □

In the following Lemma, we present the expression of $\nabla h(X)$.

Proposition 2.2. *For any $X \in \mathbb{R}^{n\times p}$,*

$$\nabla h(X) = G(X)\left(\frac{3}{2}I_p - \frac{1}{2}X^\top X\right) - X\Phi\big(X^\top G(X)\big) + \beta X\big(X^\top X - I_p\big).$$

Proof. First, we aim to prove that the linear mapping J_X is self-adjoint for any $X \in \mathbb{R}^{n\times p}$. From the expression of J_X, for any $Z, W \in \mathbb{R}^{n\times p}$, we then obtain

$$\begin{aligned} \langle J_X(W), Z\rangle &= \left\langle W\left(\frac{3}{2}I_p - \frac{1}{2}X^\top X\right) - X\Phi(X^\top W), Z\right\rangle \\ &= \mathrm{tr}\left(Z^\top W\left(\frac{3}{2}I_p - \frac{1}{2}X^\top X\right)\right) - \mathrm{tr}\left(Z^\top X\Phi(X^\top W)\right) \\ &\overset{(i)}{=} \mathrm{tr}\left(W^\top Z\left(\frac{3}{2}I_p - \frac{1}{2}X^\top X\right)\right) - \mathrm{tr}\left(W^\top X\Phi(X^\top Z)\right) \\ &= \left\langle Z\left(\frac{3}{2}I_p - \frac{1}{2}X^\top X\right) - X\Phi(X^\top Z), W\right\rangle = \langle J_X(Z), W\rangle. \end{aligned} \tag{2.2}$$

Here (i) follows the fact that $\mathrm{tr}(AB) = \mathrm{tr}(A^\top B)$ holds for any square matrix A and any symmetric matrix B. By Proposition 2.1, for any $X, Y \in \mathbb{R}^{n\times p}$, let $D = Y - X$, we have

$$\begin{aligned}
& f\left(Y\left(\frac{3}{2}I_p - \frac{1}{2}Y^\top Y\right)\right) - f\left(X\left(\frac{3}{2}I_p - \frac{1}{2}X^\top X\right)\right) \\
&= \langle G(X), J_X(D)\rangle + \mathcal{O}\big(\|D_{\mathrm{F}}^2\|\big) = \big\langle D, J_X\big(G(X)\big)\big\rangle + \mathcal{O}\big(\|D_{\mathrm{F}}^2\|\big) \\
&= \left\langle D, G(X)\left(\frac{3}{2}I_p - \frac{1}{2}X^\top X\right) - X\Phi\big(X^\top G(X)\big)\right\rangle + \mathcal{O}\big(\|D_{\mathrm{F}}^2\|\big),
\end{aligned}$$

which illustrates that

$$\nabla g(X) = G(X)\left(\frac{3}{2}I_p - \frac{1}{2}X^\top X\right) - X\Phi\big(X^\top G(X)\big).$$

Then from the fact that

$$h(X) = g(X) + \beta\big\|X^\top X - I_{p}{}_{\mathrm{F}}^{2}\big\|,$$

we conclude that

$$\nabla h(X) = G(X)\left(\frac{3}{2}I_p - \frac{1}{2}X^\top X\right) - X\Phi\big(X^\top G(X)\big) + \beta X\big(X^\top X - I_p\big),$$

and complete the proof. □

We can conclude from the definition of M_1 and Proposition 2.2 that $\|\nabla g(X)\|_{\mathrm{F}} \leq 2M_1$ for any $X \in \Omega$. Besides, from the expression of $h(X)$ illustrated in Lemma 2.2, we can conclude that $\nabla h(X) = \mathrm{grad}\, f(X)$ holds for any $X \in \mathcal{S}_{n,p}$. Furthermore, the following proposition illustrates the expression of $\nabla^2 h(X)$.

Proposition 2.3. *Suppose $f(X)$ satisfies the conditions in Assumption 2.2, then*

$$\nabla^2 g(X)[D] = J_X\big(\mathcal{H}(X)[J_X(D)]\big) - D\Phi\big(X^\top G(X)\big) - X\Phi\big(D^\top G(X)\big) - G(X)\Phi(D^\top X).$$

Moreover,

$$\nabla^2 h(X)[D] = \nabla^2 g(X)[D] + \beta\big(2X\Phi(X^\top D) + D(X^\top X - I_p)\big).$$

Proof. As illustrated in (2.2) from Lemma 2.2, the mapping J_X is self-adjoint for any $X \in \mathbb{R}^{n\times p}$. Then by Proposition 2.1, for any $Y \in \mathbb{R}^{n\times p}$ and let $D = Y - X$, we have

$$\begin{aligned}
& f\left(Y\left(\frac{3}{2}I_p - \frac{1}{2}Y^\top Y\right)\right) - f\left(X\left(\frac{3}{2}I_p - \frac{1}{2}X^\top X\right)\right) \\
&= f\left(X\left(\frac{3}{2}I_p - \frac{1}{2}X^\top X\right) + J_X(D) - \left[D\Phi(D^\top X) + \frac{1}{2}XD^\top D\right]\right) \\
&\quad - f\left(X\left(\frac{3}{2}I_p - \frac{1}{2}X^\top X\right)\right) + \mathcal{O}\big(\|D_{\mathrm{F}}^3\|\big) \\
&= \left\langle G(X), J_X(D) - \left[D\Phi(D^\top X) + \frac{1}{2}XD^\top D\right]\right\rangle \\
&\quad + \frac{1}{2}\big\langle \mathcal{H}(X)[J_X(D)], J_X(D)\big\rangle + \mathcal{O}\big(\|D_{\mathrm{F}}^3\|\big) \\
&= \langle G(X), J_X(D)\rangle - \left\langle G(X), D\Phi(D^\top X) + \frac{1}{2}XD^\top D\right\rangle
\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{2}\left\langle \mathcal{H}(X)[J_X(D)], J_X(D)\right\rangle + \mathcal{O}\left(\|D_{\mathrm{F}}^3\|\right)\\
=&\left\langle D, \nabla g(X)\right\rangle - \left\langle \Phi\left(D^\top G(X)\right), \Phi(D^\top X)\right\rangle - \frac{1}{2}\left\langle D^\top D, X^\top G(X)\right\rangle\\
&+\frac{1}{2}\left\langle \mathcal{H}(X)[J_X(D)], J_X(D)\right\rangle + \mathcal{O}\left(\|D_{\mathrm{F}}^3\|\right).
\end{aligned}
$$

Therefore, the Hessian of $g(X)$ can be expressed as

$$
\nabla^2 g(X)[D] = J_X\left(\mathcal{H}(X)[J_X(D)]\right) - D\Phi\left(X^\top G(X)\right) - X\Phi\left(D^\top G(X)\right) - G(X)\Phi(D^\top X).
$$

Moreover, since

$$
\begin{aligned}
\left\|Y^\top Y - I_p\right\|_{\mathrm{F}}^2 - \left\|X^\top X - I_p\right\|_{\mathrm{F}}^2 = &\left\langle 4D, X\left(X^\top X - I_p\right)\right\rangle + 4\left\langle \Phi(D^\top X), \Phi(D^\top X)\right\rangle\\
&+ 2\left\langle D^\top D, X^\top X - I_p\right\rangle + \mathcal{O}\left(\|D_{\mathrm{F}}^3\|\right),
\end{aligned}
$$

the Hessian of $h(X)$ can be expressed as

$$
\nabla^2 h(X)[D] = \nabla^2 g(X)[D] + \beta\left[2X\Phi(X^\top D) + D\left(X^\top X - I_p\right)\right].
$$

The proof is complete. □

Next, we give an important equality.

Lemma 2.1. *For any $X \in \mathbb{R}^{n\times p}$, we have*

$$
\left\langle X\left(X^\top X - I_p\right), \nabla g(X)\right\rangle = -\frac{3}{2}\left\langle \left(X^\top X - I_p\right)^2, \Phi\left(X^\top G(X)\right)\right\rangle.
$$

Proof. Consider the inner product of $\nabla g(X)$ and $X(X^\top X - I_p)$, the following equality holds for any $X \in \mathbb{R}^{n\times p}$:

$$
\begin{aligned}
\left\langle X\left(X^\top X - I_p\right), \nabla g(X)\right\rangle =& \left\langle X\left(X^\top X - I_p\right), G(X)\left(\frac{3}{2}I_p - \frac{1}{2}X^\top X\right)\right\rangle\\
&- \left\langle X\left(X^\top X - I_p\right), X\Phi\left(X^\top G(X)\right)\right\rangle\\
=& \left\langle \left(X^\top X - I_p\right)\left(\frac{3}{2}I_p - \frac{1}{2}X^\top X\right), \Phi\left(X^\top G(X)\right)\right\rangle\\
&- \left\langle \left(X^\top X - I_p\right)X^\top X, \Phi\left(X^\top G(X)\right)\right\rangle\\
=& -\frac{3}{2}\left\langle \left(X^\top X - I_p\right)^2, \Phi\left(X^\top G(X)\right)\right\rangle.
\end{aligned}
$$

The proof is complete. □

Finally, we arrive at the main proposition in this preliminary section.

Proposition 2.4. *Suppose Assumption* 2.1 *holds, and $\tilde{X}$ is a first-order stationary point of ExPen, then $\|\tilde{X}\|_2 \leq 1 + \hat{M}_1/\beta$. Furthermore, when $\beta \geq \hat{\beta}$, we can conclude that all the first-order stationary points of ExPen are contained in Ω.*

Proof. Let $\tilde{X} = U\Sigma V^\top$ be the singular value decomposition of $\tilde{X}$, namely, $U \in \mathbb{R}^{n\times p}$ and $V \in \mathbb{R}^{n\times p}$ are the orthogonal matrices and Σ is a diagonal matrix with singular values of $\tilde{X}$ on its diagonal and $\sigma_1 \leq \cdots \leq \sigma_p$. Suppose the statement to be proved is not true, we achieve $\sigma_p > 1 + \hat{M}_1/\beta$. Let $\tilde{D} := U\mathrm{Diag}(0, \ldots, 0, 1)V^\top$, then from the first-order optimality condition, we have

$$
\left\langle \nabla h(\tilde{X}), \tilde{D}\right\rangle = 0.
$$

Besides, Assumption 2.1 illustrates that $\|G(X)\|_{\mathrm{F}}$ is bounded and thus

$$
\begin{aligned}
\left|\left\langle \tilde{D}, \nabla g(\tilde{X})\right\rangle\right| &= \left\|\left\langle \tilde{D}, G(X)\left(\frac{3}{2}I_p - \frac{1}{2}X^\top X\right) - \left\langle D, X\Phi\left(X^\top G(X)\right)\right\rangle\right\rangle\right\| \\
&\leq \left|\mathrm{tr}\left(\left(\frac{3}{2}I_p - \frac{1}{2}X^\top X\right) D^\top G(X)\right)\right| \\
&\quad + \frac{1}{2}\left|\mathrm{tr}\left(D^\top X X^\top G(X)\right)\right| + \frac{1}{2}\left|\mathrm{tr}\left(G(X)^\top X D^\top X\right)\right| \\
&\leq \left|\frac{3}{2} - \frac{1}{2}\sigma_p^2\right| \hat{M}_1 + \frac{1}{2}\sigma_p^2 \hat{M}_1 + \frac{1}{2}\sigma_p^2 \hat{M}_1 \leq \frac{3}{2}\left(\sigma_p^2 + 1\right)\hat{M}_1.
\end{aligned}
$$

On the other hand, from the definition of $\tilde{D}$, we can conclude that

$$
\left\langle \tilde{X}\left(\tilde{X}^\top \tilde{X} - I_p\right), \tilde{D}\right\rangle = \sigma_p(\sigma_p^2 - 1).
$$

Notice that when $\beta \geq 3\hat{M}_1$, for any $t \geq 1 + \hat{M}_1/\beta$, it holds that

$$
\frac{(t^2-1)t}{t^2+1} > \frac{(t^2-1)}{t^2+1} \geq 1 - \frac{2}{t^2+1} \geq 1 - \frac{1}{1 + 3\hat{M}_1/\beta} \geq \frac{3\hat{M}_1}{2\beta}. \tag{2.3}
$$

Therefore, when $\beta \geq 3\hat{M}_1$, we achieve

$$
\begin{aligned}
\left\langle \nabla h(\tilde{X}), \tilde{D}\right\rangle &\geq \beta\left\langle \tilde{X}\left(\tilde{X}^\top \tilde{X} - I_p\right), \tilde{D}\right\rangle - \left|\left\langle \tilde{D}, \nabla g(\tilde{X})\right\rangle\right| \\
&\geq (\sigma_p^2 - 1)(\beta\sigma_p) - \frac{3}{2}\left(\sigma_p^2 + 1\right)\hat{M}_1 > 0,
\end{aligned}
$$

which contradicts to the first-order optimality. Therefore, we can conclude that

$$
\left\|\tilde{X}\right\|_2 \leq 1 + \frac{\hat{M}_1}{\beta}.
$$

The proof is complete. □

Additionally, in the following proposition, we illustrate that Assumption 2.1 implies that ExPen is bounded below for any $\beta > 0$.

Proposition 2.5. *Suppose Assumption* 2.1 *holds, then for any $\beta > 0$, ExPen is bounded below over $\mathbb{R}^{n\times p}$.*

Proof. For any $X \in \mathbb{R}^{n\times p}$, it holds from Assumption 2.1 that

$$
\begin{aligned}
|g(X) - g(0)| &= \left|f\left(X\left(\frac{3}{2}I_p - \frac{1}{2}X^\top X\right)\right) - f(0)\right| \\
&\leq \hat{M}_1 \left\|X\left(\frac{3}{2}I_p - \frac{1}{2}X^\top X\right)\right\|_{\mathrm{F}} \\
&\leq p\hat{M}_1\left(\frac{1}{2}\left\|X\right\|_2^3 + \frac{3}{2}\left\|X\right\|_2\right).
\end{aligned}
$$

Moreover, it holds that

$$
\left\|X^\top X - I_p\right\|_{\mathrm{F}} \geq \left(\left\|X\right\|_2^2 - 1\right)^2.
$$

Therefore, for any $X \in \mathbb{R}^{n\times p}$ that satisfies

$$
\left\|X\right\|_2 \geq \frac{128p\hat{M}_1}{\beta} + 2,
$$

it holds that

$$\left(\|X\|_2^2 - 1\right)^2 \geq \frac{1}{16} \|X\|_2^4 .$$

Then we achieve

$$\begin{aligned}
h(X) - h(0) &\geq \frac{\beta}{4} \left\| X^\top X - I_p \right\|_{\mathrm{F}} - |g(X) - g(0)| \\
&\geq \frac{\beta}{4} \left(\|X\|_2^2 - 1\right)^2 - p\hat{M}_1 \left(\frac{1}{2} \|X\|_2^3 + \frac{3}{2} \|X\|_2 \right) \\
&\geq \frac{\beta}{64} \|X\|_2^4 - p\hat{M}_1 \left(\frac{1}{2} \|X\|_2^3 + \frac{3}{2} \|X\|_2 \right) \\
&\geq \frac{\beta}{64} \|X\|_2^4 - 2p\hat{M}_1 \|X\|_2^3 > 0.
\end{aligned}$$

As a result, it holds that

$$\begin{aligned}
\inf_{X \in \mathbb{R}^{n \times p}} h(X) &= \min \left\{ \inf_{\|X\|_2 \leq \frac{128 p \hat{M}_1}{\beta} + 2} h(X), \inf_{\|X\|_2 \geq \frac{128 p \hat{M}_1}{\beta} + 2} h(X) \right\} \\
&\geq \min \left\{ \inf_{\|X\|_2 \leq \frac{128 p \hat{M}_1}{\beta} + 2} h(X), h(0) \right\} > -\infty.
\end{aligned}$$

Hence we complete the proof. □

2.3. Computational complexity of the first-order oracle

In this subsection, we analyze the cost of calculating the first-order derivative of $h(X)$, which takes the main computational cost in each iterate of a first-order algorithm such as gradient descent methods, nonlinear conjugate gradient methods, etc. Then we compare it with the fundamental operations in Riemannian optimization approaches. From the expression for $\nabla h(X)$ illustrated in Lemma 2.2, we find that computing $\nabla h(X)$ only involves computing ∇f and matrix-matrix multiplication. The computational cost of the basic linear algebra operations and the overall costs of computing the gradient of h are listed in Table 2.1, while a comparison between several fundamental operations in Riemannian optimization and their corresponding operations for $h(X)$ are listed in Table 2.2. Here, FO denotes the computational costs of computing the gradient of f, and those terms in bold stand for the operations that cannot be parallelized.

Table 2.1: Computational complexity the first-order oracle in ExPen.

Compute $\nabla f(X(3I_p/2 - X^\top X/2))(3I_p/2 - X^\top X/2)$	$X^\top X$	np^2
	$X(X^\top X - I_p)$	$2np^2$
	$G(X) = \nabla f(Y)\vert_{Y = X(3I_p/2 - X^\top X/2)}$	1FO
	$G(X)(3I_p/2 - X^\top X/2)$	$2np^2$
Compute $X\Phi(X^\top G(X))$	$\Phi(X^\top G(X))$	$2np^2$
	$X\Phi(X^\top G(X))$	$2np^2$
In total	$1\mathrm{FO} + 9\mathrm{np}^2$	

Table 2.2: Comparison on the computational complexity of the first-order oracles among Riemannian optimization approaches, ExPen based approaches and the specialized optimization algorithm PLAM. Here $\mathcal{L}_\beta(X,\Lambda) := f(X) - \langle X^\top X - I_p, \Lambda\rangle/2 + \beta\|X^\top X - I_p\|_{\mathrm{F}}^2/4$.

Riemannian optimization approaches	Riemannian gradient	$\nabla f(X) - X\nabla f(X)^\top X$ $1\mathrm{FO} + 4\mathrm{np}^2$ [21]
	Retraction	Cholesky factorization: $3np^2 + \mathcal{O}(\mathbf{p^3})$ Gram-Schmidt: $\mathbf{2np^2}$
	Vector transport	$\xi_X \in \mathcal{T}_X \to \xi_X - Y\Phi(Y^\top \xi_X) \in \mathcal{T}_Y$ $4np^2$
ExPen based approaches	Euclidean gradient	$\nabla h(X)$ $1\mathrm{FO} + 9\mathrm{np}^2$
	No retraction	— —
	No vector transport	— —
PLAM [21]	Descending direction	$\nabla_x \mathcal{L}_\beta(X, \Phi(\nabla f(X)^\top X))$ $1\mathrm{FO} + 7\mathrm{np}^2$ [21]
	No retraction	— —
	No vector transport	— —

3. Properties of ExPen

In this section, we analyze the theoretical properties of ExPen.

3.1. First-order relationship

In this subsection, we study the first-order relationship between OCP and ExPen. The main theoretical results of this subsection can be summarized in Fig. 3.1. Here "A.", "D.", "P.", "T." are the abbreviations of "Assumption", "Definition", "Proposition", and "Theorem", respectively.

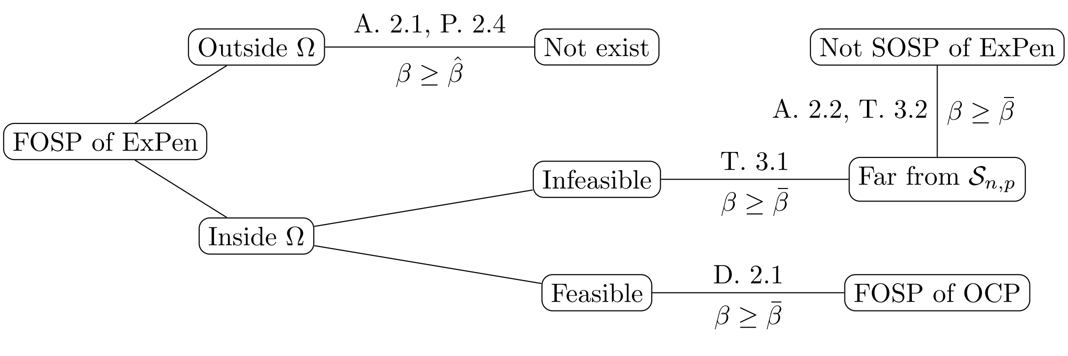


Fig. 3.1. Roadmap of the first-order relationship between OCP and ExPen.

The following theorem categorizes the first-order stationary points of ExPen in Ω.

Theorem 3.1. *Suppose $X^* \in \Omega$ is a first-order stationary point of ExPen, and $\beta \geq \bar{\beta}$, then either X^* is a first-order stationary point of OCP, or $\sigma_{\min}(X^*) \leq \sqrt{2M_1/\beta}$.*

Proof. Suppose $\sigma_{\min}(X^*) > \sqrt{2M_1/\beta}$, then $\beta {X^*}^\top X^* - 2M_1 I_p$ is positive definite. Besides, from Lemma 2.1 we achieve

$$
\begin{aligned}
0 &= \left\langle \nabla h(X^*), X^*\left({X^*}^\top X^* - I_p\right)\right\rangle \\
&\geq \left\langle \beta X^*\left({X^*}^\top X^* - I_p\right), X^*\left({X^*}^\top X^* - I_p\right)\right\rangle \\
&\quad - \left|\left\langle \nabla g(X^*), X^*\left({X^*}^\top X^* - I_p\right)\right\rangle\right| \\
&\geq \left\langle \beta X^*\left({X^*}^\top X^* - I_p\right), X^*\left({X^*}^\top X^* - I_p\right)\right\rangle \\
&\quad - \frac{3}{2}\left\|{X^*}^\top G(X^*)\right\|_2 \mathrm{tr}\left(\left({X^*}^\top X^* - I_p\right)^2\right) \\
&\geq \left\langle \beta X^*\left({X^*}^\top X^* - I_p\right), X^*\left({X^*}^\top X^* - I_p\right)\right\rangle \\
&\quad - \left\langle \frac{3}{2}\left\|X^*\right\|_2 \left\|G(X^*)\right\|_{\mathrm{F}} \cdot I_p, \left({X^*}^\top X^* - I_p\right)^2\right\rangle \\
&\geq \left\langle \beta {X^*}^\top X^* - 2M_1 I_p, \left({X^*}^\top X^* - I_p\right)^2\right\rangle \geq 0,
\end{aligned}
$$

which illustrates that ${X^*}^\top X^* = I_p$. Then we can conclude that $0 = \nabla h(X^*) = \mathrm{grad}\, f(X^*)$ and thus complete the proof. $\square$

As illustrated in Theorem 3.1, any first-order stationary point of ExPen in Ω is either a first-order stationary point of (OCP), or is far from the Stiefel manifold. The following theorem illustrates that any infeasible first-order stationary point of ExPen cannot be a second-order stationary point of $h(X)$.

Theorem 3.2. *Suppose Assumption 2.2 holds, $\beta \geq \bar{\beta}$, then any infeasible first-order stationary point $\tilde{X}$ of ExPen in Ω is not a second-order stationary point of ExPen. More specifically, $\lambda_{\min}(\nabla^2 h(\tilde{X})) \leq -\beta/24$.*

Proof. Suppose the statement is not true, namely, $\tilde{X}$ is a second-order stationary point of OCP. Since $\beta \geq 12M_1$, it holds that $\sigma_{\min}(\tilde{X}^\top \tilde{X}) \leq 1/6$ by Proposition 3.1. Let $\tilde{X} = U\Sigma V^\top$ be the singular value decomposition of $\tilde{X}$, namely, $U \in \mathbb{R}^{n\times p}$ and $V \in \mathbb{R}^{n\times p}$ are the orthogonal matrices and Σ is a diagonal matrix with singular values of $\tilde{X}$ on its diagonal. Without loss of generality, we assume $\sigma_1 \leq 1/\sqrt{6}$ which is the first entry of the diagonal matrix Σ.

Then we denote $D = -u_1 v_1^\top$, where u_1 and v_1 are the first columns of U and V, respectively. It holds that

$$
\begin{aligned}
(\tilde{X} + tD)^\top(\tilde{X} + tD) &= \tilde{X}^\top \tilde{X} + 2tD^\top \tilde{X} + t^2 D^\top D \\
&= V^\top \Sigma^2 V - 2t\sigma_1 v_1 v_1^\top + t^2 v_1 v_1^\top.
\end{aligned}
$$

Due to the first-order stationarity of $\tilde{X}$, it holds $\nabla h(\tilde{X}) = 0$ which implies $D^\top \nabla h(\tilde{X}) = 0$. First, we have

$$
\begin{aligned}
&\left\|(\tilde{X} + tD)^\top(\tilde{X} + tD) - I_p\right\|_{\mathrm{F}}^2 \\
= &\left\|\tilde{X}^\top \tilde{X} - I_p + 2t\Phi(\tilde{X}^\top D) + t^2 D^\top D\right\|_{\mathrm{F}}^2
\end{aligned}
$$

$$
\begin{aligned}
&\geq \left\|\tilde{X}^\top\tilde{X}-I_p\right\|_{\mathrm{F}}^2 + 4t^2\left\langle \Phi(\tilde{X}^\top D), \Phi(\tilde{X}^\top D)\right\rangle \\
&\quad + 2t^2\left\langle D^\top D, \tilde{X}^\top\tilde{X}-I_p\right\rangle + \mathcal{O}(t^3) \\
&\geq \left\|\tilde{X}^\top\tilde{X}-I_p\right\|_{\mathrm{F}}^2 + 4t^2\sigma_1^2 - 2t^2(1-\sigma_1^2) + \mathcal{O}(t^3) \\
&\geq \left\|\tilde{X}^\top\tilde{X}-I_p\right\|_{\mathrm{F}}^2 - t^2 + \mathcal{O}(t^3).
\end{aligned}
$$

As a result,

$$
\begin{aligned}
h(\tilde{X}+tD) &\leq h(\tilde{X}) + t\cdot\left\langle D, \nabla h(\tilde{X})\right\rangle + \frac{t^2}{2}\left\|\nabla^2 g(\tilde{X})\right\|_{\mathrm{F}}\|D\|_{\mathrm{F}}^2 - \frac{\beta}{8}t^2 + \mathcal{O}(t^3) \\
&\leq h(\tilde{X}) + \frac{t^2}{2}M_2 - \frac{\beta t^2}{8} + \mathcal{O}(t^3) \leq h(\tilde{X}) - \frac{t^2\beta}{24} + \mathcal{O}(t^3),
\end{aligned}
$$

which contradicts to the second-order optimality of ExPen. Therefore, we can conclude that any infeasible first-order stationary point of ExPen in Ω is not a second-order stationary point of ExPen, and $\lambda_{\min}(\nabla^2 h(\tilde{X})) \leq -\beta/24$. □

Combining Proposition 2.4 with Theorem 3.1, we arrive at the following corollary.

Corollary 3.1. *Suppose Assumptions* 2.1 *and* 2.2 *hold, and* $\beta \geq \hat{\beta}$. *Let* X^* *be a first-order stationary point of ExPen, then either* X^* *is a first-order stationary point of OCP, or is far from the Stiefel manifold and can not be a second-order stationary point of ExPen.*

The proof of this corollary is straightforward and hence omitted.

3.2. Second-order relationship

In this subsection, we study the first-order relationship between OCP and ExPen. The main theoretical results of this subsection can be summarized in Fig. 3.2.

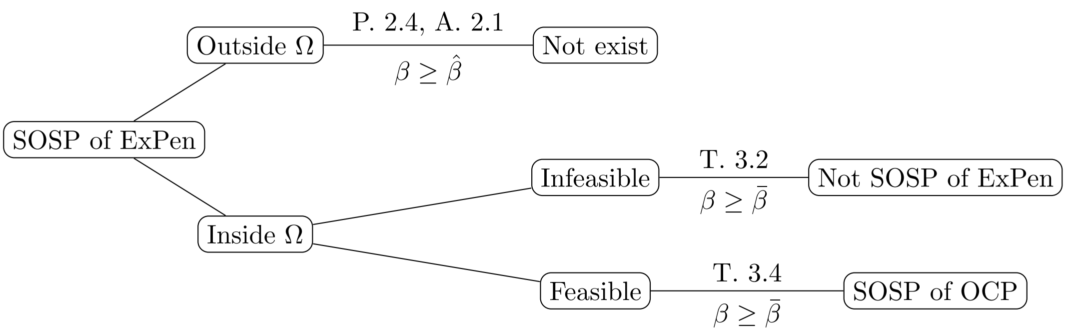


Fig. 3.2. Roadmap of the second-order relationship between OCP and ExPen under Assumption 2.2.

We first analyze the relationship between Riemannian Hessian of the original objective function f and the Euclidean Hessian of the penalty function h.

Lemma 3.1. *Suppose* $f(X)$ *satisfies Assumption* 2.2. *Then for any given* $X \in \mathcal{S}_{n,p}$ *and any* $D_1 \in \mathcal{T}_X$, *the following equality holds:*

$$
\left\langle D_1, \nabla^2 h(X)[D_1]\right\rangle = \left\langle D_1, \nabla^2 f(X)[D_1] - D_1\Phi\big(X^\top\nabla f(X)\big)\right\rangle. \tag{3.1}
$$

Proof. Since $D_1 \in \mathcal{T}_X$, by the definition of $\mathcal{T}_X$ we have $\Phi(D_1^\top X) = 0$. Moreover, the definition of J_X indicates that $J_X(D_1) = D_1$. As a result, from Proposition 2.3 we can conclude that

$$\begin{aligned}\nabla^2 h(X)[D_1] = \nabla^2 g(X)[D_1] &= J_X\big(\mathcal{H}(X)[J_X(D_1)]\big) - D_1\Phi\big(X^\top \nabla f(X)\big)\\ &\quad - \nabla f(X)\Phi\big(X^\top D_1\big) - X\Phi\big(D_1^\top \nabla f(X)\big)\\ &= J_X\left(\nabla^2 f(X)[D_1]\right) - D_1\Phi\big(X^\top \nabla f(X)\big) - X\Phi\big(D_1^\top \nabla f(X)\big).\end{aligned} \tag{3.2}$$

Therefore, for any $D_1 \in \mathcal{T}_X$, we have

$$\begin{aligned}\left\langle D_1, \nabla^2 h(X)[D_1]\right\rangle &= \left\langle D_1, \nabla^2 f(X)[D_1] - D_1\Phi\big(X^\top \nabla f(X)\big)\right\rangle\\ &\quad - \left\langle \Phi(D_1^\top X), \Phi\big(D_1^\top \nabla f(X)\big)\right\rangle\\ &= \left\langle D_1, \nabla^2 f(X)[D_1] - D_1\Phi\big(X^\top \nabla f(X)\big)\right\rangle.\end{aligned}$$

The proof is complete. □

Lemma 3.2. *Suppose $f(X)$ satisfies Assumption 2.2 and $X \in \mathcal{S}_{n,p}$ is a first-order stationary point of $h(X)$. Then for any $D_1 \in \mathcal{T}_X$ and any $D_2 \in \mathcal{N}_X$, we have*

$$\left\langle D_2, \nabla^2 h(X)[D_2]\right\rangle \geq (2\beta - M_2)\left\|D_2\right\|_{\mathrm{F}}^2, \tag{3.3}$$

$$\left\langle D_1, \nabla^2 h(X)[D_2]\right\rangle = 0. \tag{3.4}$$

Proof. Since $D_1 \in \mathcal{T}_X$, by the definition of $\mathcal{T}_X$ we have $\Phi(D_1^\top X) = 0$. Besides, the definition of J_X indicates that $J_X(D_1) = D_1$. Since $D_1 \in \mathcal{T}_X$ and $D_2 \in \mathcal{N}_X$, we have $J_X(D_2) = 0$. Besides, there exists a symmetric matrix $\Lambda_2 \in \mathbb{R}^{p\times p}$ such that $D_2 = X\Lambda_2$. Then we have

$$\begin{aligned}\left\langle D_2, \nabla^2 h(X)[D_1]\right\rangle &= \left\langle D_2, J_X\big(\nabla^2 f(X)[J_X(D_1)]\big) - D_1\Phi\big(X^\top \nabla f(X)\big) - X\Phi\big(D_1^\top \nabla f(X)\big)\right\rangle\\ &= -\left\langle D_2, D_1\Phi\big(X^\top \nabla f(X)\big)\right\rangle - \left\langle D_2, X\Phi\big(D_1^\top \nabla f(X)\big)\right\rangle\\ &= -\mathrm{tr}\left(\Lambda_2 X^\top D_1 X^\top \nabla f(X)\right) - \mathrm{tr}\left(\Lambda_2 \Phi\big(D_1^\top X X^\top \nabla f(X)\big)\right)\\ &\overset{(i)}{=} \mathrm{tr}\big(\Lambda_2 D_1^\top X X^\top \nabla f(X)\big) - \mathrm{tr}\left(\Lambda_2 \Phi\big(D_1^\top X X^\top \nabla f(X)\big)\right)\\ &\overset{(ii)}{=} \mathrm{tr}\left(\Lambda_2\left[D_1^\top X X^\top \nabla f(X) - \Phi\big(D_1^\top X X^\top \nabla f(X)\big)\right]\right) = 0.\end{aligned}$$

Here (i) follows the fact that $D_1^\top X$ is skew-symmetric, and (ii) directly uses the fact that for any $T \in \mathbb{R}^{p\times p}$, $T - \Phi(T)$ is skew-symmetric.

Moreover, notice that $D_2 \in \mathcal{N}_X$ implies that $\|\Phi(D_2^\top X)\|_{\mathrm{F}} = \|D\|_{\mathrm{F}}$. Then by the expression of $\nabla^2 h(X)$ in Proposition 2.3, we can conclude that

$$\left\langle D_2, \nabla^2 h(X)[D_2]\right\rangle = \left\langle D_2, \nabla^2 g(X)[D_2]\right\rangle + 2\beta\left\|\Phi(D_2^\top X)\right\|_{\mathrm{F}}^2 \geq (2\beta - M_2)\left\|D_2\right\|_{\mathrm{F}}^2,$$

which completes the proof. □

Theorem 3.3. *Suppose $f(X)$ satisfies Assumption 2.2, for any first-order stationary point $X \in \mathcal{S}_{n,p}$ of OCP, any eigenvalue of* $\mathrm{hess}\, f(X)$ *is an eigenvalue of $\nabla^2 h(X)$. In turn, any eigenvalue of $\nabla^2 h(X)$ is either an eigenvalue of* $\mathrm{hess}\, f(X)$*, or greater than $2\beta - M_2$.*

Proof. Let σ_1 be an eigenvalue of $\mathrm{hess}\, f(X)$, it follows from Definition 2.2 that there exists an $\hat{D}_1 \in \mathcal{T}_X$ such that

$$J_X\big(\nabla^2 f(X)[\hat{D}_1] - \hat{D}_1 \Phi\big(X^\top \nabla f(X)\big)\big) = \sigma D_1.$$

In addition, Lemma 3.1 indicates that

$$\nabla^2 h(X)[\hat{D}_1] = J_X\big(\nabla^2 f(X)[\hat{D}_1] - \hat{D}_1 \Phi\big(X^\top \nabla f(X)\big)\big) = \sigma D_1.$$

Therefore, any eigenvalue of $\mathrm{hess}\, f(X)$ is an eigenvalue of $\nabla^2 h(X)$.

On the other hand, Lemmas 3.1 and 3.2 verify that the linear operator $\nabla^2 h(X)$ maps a vector in $\mathcal{T}_X$ or $\mathcal{N}_X$ to $\mathcal{T}_X$ or $\mathcal{N}_X$, respectively. Then any eigenvector of $\nabla^2 h(X)$ is either in $\mathcal{T}_X$ or $\mathcal{N}_X$. For any $D_2 \in \mathcal{N}_X$, (3.3) implies that

$$\big\langle D_2, \nabla^2 h(X)[D_2]\big\rangle \geq (2\beta - M_2)\, \|D_2\|_{\mathrm{F}}^2,$$

from which we can conclude that any eigenvalue of $\nabla^2 h(X)$ is either an eigenvalue of $\mathrm{hess}\, f(X)$, or greater than $2\beta - M_2$. □

Based on Theorem 3.3, we can establish the second-order relationship between OCP and ExPen.

Theorem 3.4. *Suppose $f(X)$ satisfies Assumption* 2.2 *and $\beta \geq \bar{\beta}$, then any second-order stationary point of $h(X)$ in Ω is a second-order stationary point of OCP. Moreover, OCP and ExPen have exactly the same second-order stationary points in Ω.*

Proof. Let $X \in \Omega$ be a second-order stationary point of $h(X)$, then all eigenvalues of $\nabla^2 h(X)$ are nonnegative. It follows from Theorems 3.1 and 3.2 that X is feasible. We can further conclude X is a first-order stationary point of OCP by Proposition 2.2. In addition, Theorem 3.3 shows that all eigenvalues of $\mathrm{hess}\, f(X)$ consist of a subset of the spectra of $\nabla^2 h(X)$, and thus are nonnegative. Namely, X is a second-order stationary point of OCP.

In turn, let $X \in \mathcal{S}_{n,p}$ be a second-order stationary point of OCP, naturally, all the eigenvalues of $\mathrm{hess}\, f(X)$ are nonnegative. Then we can immediately obtain that all the eigenvalues of $\nabla^2 h(X)$ are nonnegative resulting from Theorem 3.3 and the fact that $2\beta \geq M$. Hence, X is a second-order stationary point of $h(X)$. □

Based on the second-order relationship between OCP and ExPen in Ω illustrated in Theorem 3.4, we can immediately obtain their second-order relationship in $\mathbb{R}^{n\times p}$ by utilizing Proposition 2.4. We omit the proof, since it is quite straightforward.

Corollary 3.2. *Suppose Assumptions* 2.1 *and* 2.2 *hold, and $\beta \geq \hat{\beta}$, then OCP and ExPen share the same second-order stationary points.*

3.3. Estimating stationarity

When we implement an infeasible approach to ExPen, the returned solution is usually infeasible since the iterates are not necessarily restricted on $\mathcal{S}_{n,p}$. Sometimes we pursue high accuracy for the feasibility at the same time. To this end, we impose an orthonormalization as a postprocess after obtaining a solution X with mild accuracy by applying an unconstrained optimization approach to solve ExPen. Namely,

$$X \;\to\; \mathcal{P}_{\mathcal{S}_{n,p}}(X), \tag{3.5}$$

where $\mathcal{P}_{\mathcal{S}_{n,p}} : \mathbb{R}^{n\times p} \to \mathcal{S}_{n,p}$ is the projection on Stiefel manifold defined in Section 1.3.

In this subsection, we study the relationship between the stationarity of $\mathcal{P}_{\mathcal{S}_{n,p}}(X)$ with respect to OCP and the stationarity X with respect to ExPen. More specifically, we aim to estimate an upper-bound for $\|\mathrm{grad}\, f(\mathcal{P}_{\mathcal{S}_{n,p}}(X))\|_{\mathrm{F}}$ from $\|\nabla h(X)\|_{\mathrm{F}}$. Therefore, we could explicitly setting the stopping criteria for ExPen to achieve a desired accuracy in solving OCP. Moreover, the iteration complexity of various unconstrained optimization approaches directly follows from existing rich results when applied to solve ExPen.

The following lemma guarantees that the postprocess (3.5) can further reduce the function value if the current iterate is sufficiently close to the Stiefel manifold.

Proposition 3.1. *Suppose $X \in \overline{\Omega}_{1/6}$, then it holds that*

$$h\big(\mathcal{P}_{\mathcal{S}_{n,p}}(X)\big) \leq h(X) - \left(\frac{\beta}{4} - \frac{M_1}{2}\right)\left\|X^\top X - I_p\right\|_{\mathrm{F}}^2.$$

Proof. By the SVD of X, we first conclude that

$$\begin{aligned}
&\left\|X\left(\frac{3}{2}I_p - \frac{1}{2}X^\top X\right) - \mathcal{P}_{\mathcal{S}_{n,p}}(X)\right\|_{\mathrm{F}} \\
={}& \left\|U\Sigma\left(\frac{3}{2}I_p - \frac{1}{2}\Sigma^2\right)V^\top - UV^\top\right\|_{\mathrm{F}} \\
={}& \left\|\Sigma\left(\frac{3}{2}I_p - \frac{1}{2}\Sigma^2\right) - I_p\right\|_{\mathrm{F}} = \left\|\left(\frac{1}{2}\Sigma + I_p\right)(\Sigma - I_p)^2\right\|_{\mathrm{F}} \\
\overset{(i)}{\leq}{}& \frac{1}{2}\left\|\Sigma^2 - I_p\right\|_{\mathrm{F}}^2 = \frac{1}{2}\left\|X^\top X - I_p\right\|_{\mathrm{F}}^2.
\end{aligned} \tag{3.6}$$

Here (i) directly follows from $(\Sigma + 2I_p)(\Sigma - I_p)^2 \preceq (\Sigma + I_p)^2(\Sigma - I_p)^2$ when $\Sigma \succeq 5I_p/6$, and the fact that $\|A^2\|_{\mathrm{F}} \leq \|A\|_{\mathrm{F}}^2$ holds for any symmetric matrix A.

Then we can conclude that

$$\begin{aligned}
h(X) - h\big(\mathcal{P}_{\mathcal{S}_{n,p}}(X)\big) &= f\left(X\left(\frac{3}{2}I_p - \frac{1}{2}X^\top X\right)\right) - f\big(\mathcal{P}_{\mathcal{S}_{n,p}}(X)\big) + \frac{\beta}{4}\left\|X^\top X - I_p\right\|_{\mathrm{F}}^2 \\
&\overset{(ii)}{\geq} -M_1\left\|X\left(\frac{3}{2}I_p - \frac{1}{2}X^\top X\right) - \mathcal{P}_{\mathcal{S}_{n,p}}(X)\right\|_{\mathrm{F}} + \frac{\beta}{4}\left\|X^\top X - I_p\right\|_{\mathrm{F}}^2 \\
&\overset{(iii)}{\geq} \left(\frac{\beta}{4} - \frac{M_1}{2}\right)\left\|X^\top X - I_p\right\|_{\mathrm{F}}^2.
\end{aligned}$$

Here (ii) follows the Lipschitz continuity of f, and (iii) is directly from (3.6). $\square$

Lemma 3.3. *For any $X \in \overline{\Omega}_{1/6}$, we have*

$$\|\nabla h(X)\|_{\mathrm{F}}^2 \geq \|\nabla g(X)\|_{\mathrm{F}}^2 + \left(\frac{2}{3}\beta^2 - 4\beta M_1\right)\left\|X^\top X - I_p\right\|_{\mathrm{F}}^2.$$

Proof. Since

$$h(X) = g(X) + \frac{\beta}{4}\left\|X^\top X - I_p\right\|_{\mathrm{F}}^2,$$

we have

$$\begin{aligned}
\langle \nabla h(X), \nabla h(X)\rangle &= \langle \nabla g(X), \nabla g(X)\rangle + 2\beta\left\langle \nabla g(X), X\big(X^\top X - I_p\big)\right\rangle + \beta^2\left\|X\big(X^\top X - I_p\big)\right\|_{\mathrm{F}}^2 \\
&= \|\nabla g(X)\|_{\mathrm{F}}^2 - 3\beta\left\langle \Phi\big(X^\top G(X)\big), \big(X^\top X - I_p\big)^2\right\rangle + \beta^2\left\|X\big(X^\top X - I_p\big)\right\|_{\mathrm{F}}^2 \\
&\geq \|\nabla g(X)\|_{\mathrm{F}}^2 + \left(\frac{2}{3}\beta^2 - 4\beta M_1\right)\left\|X^\top X - I_p\right\|_{\mathrm{F}}^2.
\end{aligned}$$

Here the second equality directly follows Lemma 2.1. □

Lemma 3.4 illustrates the relationship between $\|\nabla h(X)\|_{\mathrm{F}}$ and $\|\mathrm{grad}\, f(\mathcal{P}_{\mathcal{S}_{n,p}}(X))\|_{\mathrm{F}}$.

Lemma 3.4. *Suppose $\beta \geq \bar{\beta}, X \in \overline{\Omega}_{1/6}$, then it holds that*

$$\|\nabla h(X)\|_{\mathrm{F}} \geq \frac{1}{2}\left\|\mathrm{grad}\, f\big(\mathcal{P}_{\mathcal{S}_{n,p}}(X)\big)\right\|_{\mathrm{F}} + \frac{\beta}{4}\left\|X^\top X - I_p\right\|_{\mathrm{F}}.$$

Proof. Suppose X has singular value decomposition as $X = U\Sigma V^\top$, we can conclude that

$$\left\|X - \mathcal{P}_{\mathcal{S}_{n,p}}(X)\right\|_{\mathrm{F}} = \|\Sigma - I_p\|_{\mathrm{F}} \leq \frac{6}{11}\|\Sigma^2 - I_p\|_{\mathrm{F}} = \frac{6}{11}\left\|X^\top X - I_p\right\|_{\mathrm{F}}. \tag{3.7}$$

Therefore, the results in Lemma 3.3 illustrates that

$$\begin{aligned}\|\nabla h(X)\|_{\mathrm{F}} &\geq \frac{1}{2}\|\nabla g(X)\|_{\mathrm{F}} + \frac{1}{3}\sqrt{6\beta^2 - 36\beta M_1}\left\|X^\top X - I_p\right\|_{\mathrm{F}} \\ &\geq \frac{1}{2}\left\|\nabla g\big(\mathcal{P}_{\mathcal{S}_{n,p}}(X)\big)\right\|_{\mathrm{F}} - \frac{3M_2}{11}\left\|X^\top X - I_p\right\|_{\mathrm{F}} \\ &\quad + \frac{1}{3}\sqrt{6\beta^2 - 36\beta M_1}\left\|X^\top X - I_p\right\|_{\mathrm{F}} \\ &\geq \frac{1}{2}\left\|\mathrm{grad}\, f\big(\mathcal{P}_{\mathcal{S}_{n,p}}(X)\big)\right\|_{\mathrm{F}} + \frac{\beta}{4}\left\|X^\top X - I_p\right\|_{\mathrm{F}}.\end{aligned}$$

The proof is complete. □

Proposition 3.2. *Suppose Assumption* 2.2 *holds, given any $X \in \overline{\Omega}_{1/6}$, then it holds that*

$$\begin{aligned}\sigma_{\min}\big(\mathrm{hess}\, f\big(\mathcal{P}_{\mathcal{S}_{n,p}}(X)\big)\big) \geq{}& \sigma_{\min}\big(\nabla^2 h(X)\big) - \left\|\nabla^2 g(X) - \nabla^2 g\big(\mathcal{P}_{\mathcal{S}_{n,p}}(X)\big)\right\|_{\mathrm{F}} \\ &- \frac{9}{2}\|\nabla h(X)\|_{\mathrm{F}}.\end{aligned} \tag{3.8}$$

Proof. Let $Y := \mathcal{P}_{\mathcal{S}_{n,p}}(X)$, i.e. let $X = U\Sigma V^\top$ be the singular value decomposition of X, then $Y = UV^\top$. Then from Lemma 3.1, for any $D_1 \in \mathcal{T}_Y$, it holds that

$$\left\langle D_1, \nabla^2 h(Y)[D_1]\right\rangle = \left\langle D_1, \nabla^2 f(X)[D_1] - D_1\Phi\big(X^\top \nabla f(X)\big)\right\rangle.$$

Therefore, the $\sigma_{\min}(\mathrm{hess}\, f(Y))$ can be expressed by

$$\sigma_{\min}\big(\mathrm{hess}\, f(Y)\big) = \min_{\substack{D_1 \in \mathcal{T}_Y, \\ \|D_1\|_{\mathrm{F}} = 1}} \langle D_1, \nabla g(Y)[D_1]\rangle.$$

Let

$$\tilde{D} := \arg\min_{\substack{D_1 \in \mathcal{T}_Y, \\ \|D_1\|_{\mathrm{F}} = 1}} \langle D_1, \nabla g(Y)[D_1]\rangle,$$

then it holds that

$$\left|\left\langle \tilde{D}, \nabla^2 g(X)[\tilde{D}]\right\rangle - \left\langle \tilde{D}, \nabla^2 g(X)[\tilde{D}]\right\rangle\right| \leq \|\nabla^2 g(X) - \nabla^2 g(Y)\|_{\mathrm{F}}.$$

Besides, from the expression of the Hessian of $\|X^\top X - I_p\|_{\mathrm{F}}^2$, we achieve

$$\begin{aligned}&\left\langle \tilde{D}, \tilde{D}\big(X^\top X - I_p\big) + 2X\Phi(\tilde{D}^\top X)\right\rangle \\ \leq{}& \left\langle \tilde{D}, \tilde{D}\big(X^\top X - I_p\big)\right\rangle + 2\left\langle \tilde{D}, X\Phi(\tilde{D}^\top X)\right\rangle\end{aligned}$$

$$
\begin{aligned}
&\leq \|D\|_{\mathrm{F}}^2 \left\|X^\top X - I_p\right\|_{\mathrm{F}} + 2\left\|\Phi(X^\top \tilde{D})\right\|_{\mathrm{F}}^2 \\
&= \|D\|_{\mathrm{F}}^2 \left\|X^\top X - I_p\right\|_{\mathrm{F}} + 2\left\|\Phi\big((X-Y)^\top \tilde{D}\big)\right\|_{\mathrm{F}}^2 \\
&\leq \left\|X^\top X - I_p\right\|_{\mathrm{F}} + 2\left\|X - Y\right\|_{\mathrm{F}}^2 \|D\|_{\mathrm{F}}^2 \\
&\leq \left\|X^\top X - I_p\right\|_{\mathrm{F}} + \frac{72}{121}\left\|X^\top X - I_p\right\|_{\mathrm{F}}^2 \\
&\leq \frac{11}{10}\left\|X^\top X - I_p\right\|_{\mathrm{F}}.
\end{aligned}
$$

Therefore, we conclude that

$$
\begin{aligned}
\sigma_{\min}\big(\nabla h(X)\big) &= \min_{\substack{D\in\mathbb{R}^n,\\ \|D\|_{\mathrm{F}}=1}} \left\langle D, \nabla^2 h(X)[D]\right\rangle \\
&\leq \min_{\substack{D\in\mathcal{T}_Y,\\ \|D\|_{\mathrm{F}}=1}} \left\langle D, \nabla^2 h(X)[D]\right\rangle \leq \left\langle \tilde{D}, \nabla^2 h(X)\tilde{D}\right\rangle \\
&= \left\langle \tilde{D}, \nabla^2 g(X)\tilde{D}\right\rangle + \beta\left\langle \tilde{D}, \tilde{D}\big(X^\top X - I_p\big) + 2X\Phi(\tilde{D}^\top X)\right\rangle \\
&\leq \sigma_{\min}\big(\mathrm{hess}\, f(Y)\big) + \|\nabla^2 g(X) - \nabla^2 g(Y)\|_{\mathrm{F}} + \frac{11\beta}{10}\left\|X^\top X - I_p\right\|_{\mathrm{F}} \\
&\leq \sigma_{\min}\big(\mathrm{hess}\, f(Y)\big) + \|\nabla^2 g(X) - \nabla^2 g(Y)\|_{\mathrm{F}} + \frac{9}{2}\left\|\nabla h(X)\right\|_{\mathrm{F}},
\end{aligned}
$$

and complete the proof. □

3.4. Łojasiewicz gradient inequality

In this section, we study the relationship between the Riemannian Łojasiewicz gradient inequality for $f(X)$ and the Euclidean Łojasiewicz gradient inequality for $h(X)$.

Proposition 3.3. *Suppose $f(X)$ satisfies the Riemannian Łojasiewicz gradient inequality at $X \in \mathcal{S}_{n,p}$ with Łojasiewicz exponent $\theta \in (0, 1/2]$, i.e there exists a neighborhood $U \subset \mathcal{S}_{n,p}$ and a constant $C > 0$ such that*

$$
\left\|\mathrm{grad}\, f(Y)\right\|_{\mathrm{F}} \geq C|f(Y) - f(X)|^{1-\theta}
$$

holds for any $Y \in U$ and the penalty parameter of ExPen satisfies $\beta > \max\{8CM_1, 1, \bar{\beta}\}$. Then $h(X)$ satisfies the Łojasiewicz gradient inequality at $X \in \mathcal{S}_{n,p}$ with Łojasiewicz exponent $\theta \in (0, 1/2]$.

Proof. For any $Y \in \Omega$, we denote $Z := Y(Y^\top Y)^{-1/2}$. It is clear that $Z \in \mathcal{S}_{n,p}$. By Lemma 2.2 and the Riemannian Łojasiewicz gradient inequality of f, we have

$$
\|\nabla g(Z)\|_{\mathrm{F}} = \left\|\mathrm{grad}\, f(Z)\right\|_{\mathrm{F}} \geq C|f(Z) - f(X)|^{1-\theta} = C|g(Z) - g(X)|^{1-\theta}.
$$

Besides, since

$$
\begin{aligned}
&\left\|\left(\frac{3}{2}I_p - \frac{1}{2}X^\top X\right)^2 X^\top X - I_p\right\|_{\mathrm{F}}^2 \\
={}& \left\|X^\top X - I_p + \big(I_p - X^\top X\big)X^\top X + \frac{1}{4}X^\top X\big(X^\top X - I_p\big)^2\right\|_{\mathrm{F}} \\
={}& \left\|\left(\frac{1}{4}X^\top X - I_p\right)\big(X^\top X - I_p\big)^2\right\|_{\mathrm{F}},
\end{aligned}
$$

we obtain

$$|g(Y) - g(Z)| \leq M_1 \left\|Y^\top Y - I_p\right\|_{\mathrm{F}}^2. \tag{3.9}$$

Together with Lemma 3.4, we can conclude that

$$\|\nabla h(Y)\|_{\mathrm{F}} \geq \frac{1}{2} \|\nabla g(Z)\|_{\mathrm{F}} + \frac{\beta}{4} \left\|Y^\top Y - I_p\right\|_{\mathrm{F}}.$$

In addition, since $\theta \in (0, 1/2]$, and $Y \in \Omega$, we have

$$\begin{aligned}
|h(Y) - h(X)|^{1-\theta} &= \left|g(Y) - g(X) + \frac{\beta}{4} \left\|Y^\top Y - I_p\right\|_{\mathrm{F}}^2\right|^{1-\theta} \\
&\leq |g(Y) - g(X)|^{1-\theta} + \left(\frac{\beta}{4} \left\|Y^\top Y - I_p\right\|_{\mathrm{F}}^2\right)^{1-\theta} \\
&\leq |g(Y) - g(X)|^{1-\theta} + \frac{\beta}{4} \left\|Y^\top Y - I_p\right\|_{\mathrm{F}}.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\|\nabla h(Y)\|_{\mathrm{F}} &\geq \|\nabla g(Z)\|_{\mathrm{F}} + \frac{\beta}{4} \left\|Y^\top Y - I_p\right\|_{\mathrm{F}} \\
&\geq C|g(Z) - g(X)|^{1-\theta} + \frac{\beta}{4} \left\|Y^\top Y - I_p\right\|_{\mathrm{F}} \\
&\overset{(i)}{\geq} C|g(Y) - g(X)|^{1-\theta} - C|g(Z) - g(Y)|^{1-\theta} + \frac{\beta}{4} \left\|Y^\top Y - I_p\right\|_{\mathrm{F}} \\
&\overset{(ii)}{\geq} C|g(Y) - g(X)|^{1-\theta} - CM_1^{1-\theta} \left\|Y^\top Y - I_p\right\|_{\mathrm{F}}^{2-2\theta} + \frac{\beta}{4} \left\|Y^\top Y - I_p\right\|_{\mathrm{F}} \\
&\geq C|g(Y) - g(X)|^{1-\theta} + \frac{\beta}{8} \left\|Y^\top Y - I_p\right\|_{\mathrm{F}} \\
&\geq \min\{C, 1/2\}|h(Y) - h(X)|^{1-\theta}.
\end{aligned}$$

Here inequality (i) uses the fact that

$$|a|^{1-\theta} + |b|^{1-\theta} \geq (|a| + |b|)^{1-\theta} \geq |a + b|^{1-\theta} \quad \text{for any} \quad a, b \in \mathbb{R}, \quad \theta \in (0, 1/2].$$

Besides, inequality (ii) directly follows from (3.9). As a result, we obtain

$$\|\nabla h(Y)\|_{\mathrm{F}} \geq \min\{C, 1/2\}|h(Y) - h(X)|^{1-\theta},$$

which concludes the proof. □

Besides, we can even show that OCP and ExPen share the same local minimizers.

Theorem 3.5. *Suppose Assumptions* 2.1 *and* 2.2 *hold, and* $\beta \geq \hat{\beta}$, *then ExPen and OCP share the same local minimizers.*

Proof. By Corollaries 3.1 and 3.2, we can conclude that any local minimizers of ExPen are on Stiefel manifold. Since $h(X) = f(X)$ holds for any $X \in \mathcal{S}_{n,p}$, then any local minimizers of ExPen are local minimizers of OCP.

On the other hand, let $X^* \in \mathcal{S}_{n,p}$ be a local minimizer of OCP, then there exists $\gamma \in (0, 1/12)$ such that $f(Z) \geq f(X)$ holds for any $Z \in \mathcal{S}_{n,p}$, $\|Z - X^*\|_{\mathrm{F}} \leq \gamma$. Then for any $Y \in \mathbb{R}^{n\times p}$,

$$\|Y - X^*\|_{\mathrm{F}} \leq \frac{\gamma}{2}\gamma, \quad Y \in \overline{\Omega}_{\frac{11}{12}\gamma},$$

we can obtain that

$$\left\|\mathcal{P}_{\mathcal{S}_{n,p}}(Y) - X^*\right\|_{\mathrm{F}} \leq \left\|\mathcal{P}_{\mathcal{S}_{n,p}}(Y) - Y\right\|_{\mathrm{F}} + \|Y - X^*\|_{\mathrm{F}} \leq \frac{\gamma}{2} + \frac{\gamma}{2} \leq \gamma.$$

Here the second inequality recalls the relationship (3.7). Then it follows from Proposition 3.1 that

$$\begin{aligned} h(Y) - h(X^*) &= h(Y) - h\big(\mathcal{P}_{\mathcal{S}_{n,p}}(Y)\big) + h\big(\mathcal{P}_{\mathcal{S}_{n,p}}(Y)\big) - h(X^*) \\ &\geq \left(\frac{\beta}{4} - \frac{M_1}{2}\right) \left\|Y^\top Y - I_p\right\|_{\mathrm{F}}^2 \geq 0, \end{aligned}$$

which concludes the proof. □

4. Application

4.1. Theoretical analysis for nonlinear conjugate gradient method

Nonlinear conjugate gradient (CG) methods are a class of important methods for solving unconstrained optimization problems. The first CG method, called Fletcher-Reeves CG (FR-CG), was proposed in [19]. Then various nonlinear CG methods were developed [14, 15, 18, 19, 22, 29, 44, 50]. Interested readers can refer to the survey [23] for details. Recently, a number of works extend CG methods to optimization problems over the Stiefel manifold [3, 52, 53, 70]. These works are developed within the frameworks provided by [4], and thus extensively involve retractions and parallel or vector transports. Hence, as discussed before, we are forced to comply with the low efficiency if using the parallel transports or the lack of convergence if choosing the vector transport instead.

Contrarily, applying nonlinear CG methods to minimizing ExPen over $\mathbb{R}^{n\times p}$ can inherit both of the efficiency and convergence properties directly. The exact penalty model ExPen provides a bridge between the unconstrained optimization approaches and the original model OCP. In this section, we demonstrate the power of this bridge through directly applying the FR-CG method to solve OCP through ExPen. First of all, we present an ExPen version of the FR-CG in Algorithm 4.1.

To prove the convergence of Algorithm 4.1, we first illustrate a nice property of ExPen through the following lemma.

Lemma 4.1. *Suppose $\beta \geq 384M_0$. For any sequence $\{X_k\}$ that satisfies $X_0 \in \overline{\Omega}_{1/24}$, $h(X_k) \leq h(X_0)$ and $\|X_{k+1} - X_k\|_{\mathrm{F}} \leq 1/24$ for any $k \geq 0$. Then it holds that $\{X_k\} \subset \overline{\Omega}_{1/12}$.*

Proof. Firstly, for any $Y \in \overline{\Omega}_{1/24}$ and $Z \in \Omega \setminus \overline{\Omega}_{1/12}$, we have

$$h(Y) - h(Z) < \sup_{W\in\Omega} h(W) - \inf_{W\in\Omega} h(W) + \frac{\beta}{1152} - \frac{\beta}{288} \leq M_0 - \frac{\beta}{384} \leq 0. \tag{4.1}$$

Then we prove the lemma by induction. Suppose $\{X_0, \ldots, X_k\} \subset \overline{\Omega}_{1/12}$. Then notice that $\|X_{k+1} - X_k\|_{\mathrm{F}} \leq 1/24$, it holds that $X_{k+1} \in \overline{\Omega}_{1/6}$. Together with the fact that $h(X_{k+1}) \leq h(X_0)$, it directly follows from (4.1) that $X_{k+1} \in \overline{\Omega}_{1/12}$. Therefore, the induction illustrates that $\{X_k\} \subset \overline{\Omega}_{1/12}$, thus we complete the proof. □

Lemma 4.1 guarantees that the iterates generated by any monotonic algorithm starting from an initial point $X_0 \in \overline{\Omega}_{1/12}$ are restricted in the region $\overline{\Omega}_{1/6}$ under mild conditions. Then

Algorithm 4.1: Nonlinear FR-CG Method for Solving ExPen.

Require: Input data: functions f.

1 Choose initial guess X_0 and parameters $0 < \delta \leq \sigma \leq 1/2$, set $k := 0$, $D_0 = -\nabla h(X_0)$.

2 **while** not terminate **do**

3   Compute the stepsize η_k that satisfies $\eta_k \left\| D_k \right\|_{\mathrm{F}} \leq 1/24$ by strong Wolfe line search [23],

$$
\begin{aligned}
h(X_k + \eta_k D_k) - h(X_k) &\leq \delta \eta_k \left\langle \nabla h(X_k), D_k \right\rangle, \\
\left| \left\langle \nabla h(X_k + \eta_k D_k), D_k \right\rangle \right| &\leq -\sigma \left\langle \nabla h(X_k), D_k \right\rangle.
\end{aligned}
$$

4     $X_{k+1} = X_k + \eta_k D_k$.

5   Compute the CG update parameter $\tau_k = \|\nabla h(X_{k+1})\|_{\mathrm{F}}^2 / \|\nabla h(X_k)\|_{\mathrm{F}}^2$.

6   Compute the search direction $D_{k+1} = -\nabla h(X_{k+1}) + \tau_k D_k$.

7   Set $k := k + 1$.

8 **end**

9 Return X_k.

the Step 3 in Algorithm 4.1 indicates that $\{h(X_k)\}$ is monotone decreasing. Then combining Lemma 4.1 and Assumption 1.1, we conclude that the objective function f satisfies the Lipschitz conditions and boundness conditions in [23]. Furthermore, since $\langle D_k, \nabla h(X_k) \rangle < 0$ holds for any $k \geq 0$ and the step sizes are generated using the strong Wolfe condition, the validity of the Zoutendijk condition [71] is guaranteed by [23, Theorem 2.1]. Therefore, based on [23, Theorem 4.2], we can directly establish the global convergence result for Algorithm 4.1 and omit its proof for simplicity.

Theorem 4.1. *Suppose $\beta \geq \max\{384M_0, \hat{\beta}\}$. Let $\{(X_k, D_k)\}$ be the sequence generated by Algorithm 4.1 initiated from $X_0 \in \overline{\Omega}_{1/24}$. If $\langle \nabla h(X_k), D_k \rangle < 0$ holds for any $k \geq 0$, then any accumulation point of $\{X_k\}$ is a first-order stationary point of OCP.*

Remark 4.1. Algorithm 4.1 and Theorem 4.1 take FR-CG as example. In fact, if we update the parameter sequence $\{\tau_k\}$ by the PRP [50], DY [14], or HS [29] formulas, we can obtain similar global convergence properties as well. Interested readers are referred to the survey paper [23] for details.

Remark 4.2. It is worth mentioning that a small penalty parameter β may lead to the failure of convergence, while a large penalty parameter may result in a large condition number, thus lead to slow convergence rate. Several existing works on developing penalty methods for OCP have suggested some practically useful choice of the penalty parameter β, interested readers can refer to [21, 61, 62] for details.

4.2. Numerical experiments

In this section, we numerically demonstrate the power of the bridge, provided by ExPen, between the unconstrained optimization approaches and the original model OCP. All the numerical experiments in this section are run in serial in a platform with AMD Ryzen 5800 H CPU and 16 GB RAM under Ubuntu 18.10 running Python 3.7.0 and Numpy 1.20.0 [26].

We choose nonlinear eigenvalue problem and the Brockett function minimization as the test problems. The details of how to construct the test instances are described in the following two subsections, respectively.

In the presented experiments, we set the penalty parameter β in ExPen as suggested in [21, 61, 62], i.e. $\beta = \|\nabla f(X_0)\|_{\mathrm{F}}/10$, where X_0 is the initial point. Besides, we choose the nonlinear conjugate gradient solver [19,48] provided in the package `SciPy` 1.6.3 [57] to minimize ExPen in $\mathbb{R}^{n\times p}$. This optimization approach is referred as ExPen-CG. We terminate ExPen-CG when $\|\nabla h(X_k)\|_{\mathrm{F}} \leq 10^{-3}$, or the maximum number of iterations exceeds 10000, while keeping all the other parameters as their default values in the package.

For comparison, we first select the Riemannian conjugate gradient (RCG) solver from the package `PyManopt` (version 0.2.5) [56]. RCG is one of the state-of-the-art Riemannian solvers in Python platform. Furthermore, we also choose some state-of-the-art infeasible optimization solvers into the comparisons. These solvers include PCAL [21], PenC [61] and SLPG [63] from the `STOP` package [64]. We terminate these solvers when the maximum number of iterations exceeds 10000 and set the tolerance for gradient as 10^{-3}. Meanwhile, we set the other parameters in these solvers by default. Furthermore, for the final solution $\tilde{X}$ generated by all the compared algorithms, we project $\tilde{X}$ onto the Stiefel manifold as the post-processing step employed in [21, 61–63].

4.2.1. Nonlinear eigenvalue problems

In this subsection, we test the performance of all the compared solves in solving a class of nonlinear eigenvalue problems arisen from electronic structure calculation [11, 42, 65]

$$\begin{aligned} &\min_{X\in\mathbb{R}^{n\times p}} f(X) = \frac{1}{2}\mathrm{tr}(X^\top LX) + \frac{\alpha}{4}\rho_X^\top L^\dagger \rho_X \\ &\text{s.t.} \quad X^\top X = I_p, \end{aligned} \tag{4.2}$$

where $\rho_X := \mathrm{diag}(XX^\top)$, L is a tridiagonal matrix with 2 as diagonal entries and -1 as subdiagonal entries. Besides, $L^\dagger$ refers to the pseudo-inverse of L. We initiate all the compared solvers at the same initial point, which is randomly generated over $\mathcal{S}_{n,p}$. Tables 4.1 and 4.2 illustrate the performance of all the compared algorithms in solving problem 4.2 with different combinations of problem parameters n, p. Here, we run each instance for 10 times and present the averaged results. We can learn from Tables 4.1 and 4.2 that all the compared solvers reach similar function values while ExPen-CG is comparable with the state-of-the-art solvers. Remarkably, ExPen-CG outperforms RCG in terms of CPU time and iterations.

Furthermore, we exhibit the convergence curves of ExPen-CG in the aspect of function value gap evaluated by $f(X_k) - f(X^*)$, the stationarity $\|\nabla f(X_k)\|$ and the feasibility $\|{X_k}^\top X_k - I_p\|_{\mathrm{F}}$. Here $f(X^*)$ is computed by RCG solver from `PyManopt` package that satisfies $\|\mathrm{grad}\, f(X^*)\|_F \leq 10^{12}$. In Fig. 4.1, we present the curves of ExPen-CG under different combination of the parameters. From Figs. 4.1(a), 4.1(d) and 4.1(g), we can observe that the sequence $\{X_k\}$ generated by ExPen-CG achieves almost the same function values as RCG. Moreover, Figs. 4.1(c), 4.1(f) and 4.1(i) illustrate that the sequence generated by ExPen-CG converges towards $\mathcal{S}_{n,p}$.

4.2.2. Brockett function minimization problems

In this subsection, we test the numerical performance of all the tested algorithms on minimizing the Brockett function over the Stiefel manifold [58],

Table 4.1: The results of the nonlinear eigenvalue problems with varying n.

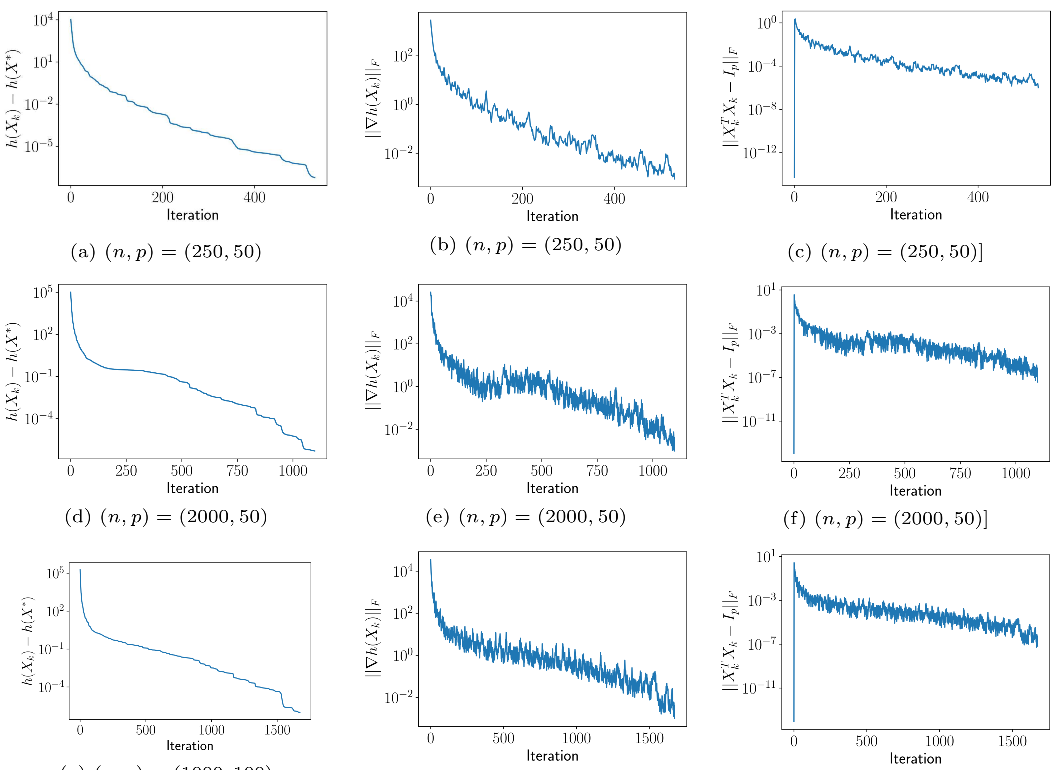
Solver	Fval	Iteration	Stationarity	Feasibility	CPU time(s)
$n = 250, p = 50$					
ExPen-CG	2.810709e+03	567.5	5.63e-04	1.79e-14	1.16
PCAL	2.810709e+03	1528.3	8.44e-04	9.83e-15	1.54
PenC	2.810709e+03	1211.4	8.00e-04	1.01e-14	1.21
SLPG	2.810709e+03	1275.3	8.40e-04	9.97e-15	1.25
RCG	2.810709e+03	1103.1	9.78e-04	5.49e-15	4.28
$n = 500, p = 50$					
ExPen-CG	2.810709e+03	632.6	6.83e-04	1.92e-14	3.52
PCAL	2.810709e+03	1739.5	8.67e-04	1.00e-14	4.64
PenC	2.810709e+03	1486.1	9.04e-04	9.99e-15	3.82
SLPG	2.810709e+03	1274.7	8.16e-04	9.77e-15	3.67
RCG	2.810709e+03	1111.7	9.78e-04	5.85e-15	9.96
$n = 1000, p = 50$					
ExPen-CG	2.810709e+03	715.6	9.35e-04	2.06e-14	5.14
PCAL	2.810709e+03	1849.6	9.72e-04	1.07e-14	7.61
PenC	2.810709e+03	1580.9	9.05e-04	1.09e-14	6.18
SLPG	2.810709e+03	1303.6	7.69e-04	1.11e-14	6.12
RCG	2.810709e+03	1492.2	9.82e-04	7.80e-15	18.53
$n = 1500, p = 50$					
ExPen-CG	2.810709e+03	787.0	7.88e-04	2.22e-14	7.29
PCAL	2.810709e+03	1680.7	9.24e-04	1.13e-14	9.70
PenC	2.810709e+03	1651.8	9.29e-04	1.10e-14	9.00
SLPG	2.810709e+03	1281.3	8.54e-04	1.10e-14	8.51
RCG	2.810709e+03	1206.9	9.83e-04	8.32e-15	20.17
$n = 2000, p = 50$					
ExPen-CG	2.810709e+03	866.4	7.52e-04	2.35e-14	10.14
PCAL	2.810709e+03	1863.0	8.45e-04	1.16e-14	14.17
PenC	2.810709e+03	1619.0	9.18e-04	1.15e-14	11.78
SLPG	2.810709e+03	1583.1	8.96e-04	1.14e-14	10.93
RCG	2.810709e+03	1140.2	9.85e-04	8.43e-15	24.49

Table 4.2: The results of the nonlinear eigenvalue problems with varying p.

Solver	Fval	Iteration	Stationarity	Feasibility	CPU time(s)
$n = 1000, p = 10$					
ExPen-CG	3.570857e+01	165.1	6.52e-04	5.74e-15	0.23
PCAL	3.570857e+01	331.9	8.36e-04	3.03e-15	0.22
PenC	3.570857e+01	295.9	7.34e-04	3.42e-15	0.19
SLPG	3.570857e+01	266.8	8.64e-04	3.30e-15	0.13
RCG	3.570857e+01	179.7	9.68e-04	2.43e-15	0.43
$n = 1000, p = 30$					
ExPen-CG	6.482086e+02	434.7	7.39e-04	1.07e-14	2.11
PCAL	6.482086e+02	860.0	7.21e-04	7.09e-15	1.74
PenC	6.482086e+02	846.7	8.77e-04	7.19e-15	1.61
SLPG	6.482086e+02	773.3	7.44e-04	7.11e-15	1.30
RCG	6.482086e+02	596.3	9.90e-04	5.31e-15	5.26

Table 4.2: The results of the nonlinear eigenvalue problems with varying p (cont'd).

Solver	Fval	Iteration	Stationarity	Feasibility	CPU time(s)
$n = 1000, p = 50$					
ExPen-CG	2.810709e+03	752.3	6.95e-04	2.06e-14	5.51
PCAL	2.810709e+03	1845.1	9.58e-04	1.09e-14	7.56
PenC	2.810709e+03	1481.5	8.77e-04	1.08e-14	5.78
SLPG	2.810709e+03	1394.7	7.66e-04	1.10e-14	5.07
RCG	2.810709e+03	1019.1	9.77e-04	7.85e-15	17.62
$n = 1000, p = 70$					
ExPen-CG	7.523209e+03	1111.8	7.15e-04	2.34e-14	14.15
PCAL	7.523209e+03	3612.8	1.61e-03	1.35e-14	19.82
PenC	7.523209e+03	2928.3	8.91e-04	1.36e-14	15.10
SLPG	7.523209e+03	2439.5	9.46e-04	1.37e-14	15.11
RCG	7.523209e+03	2248.9	9.84e-04	1.06e-14	52.19
$n = 1000, p = 100$					
ExPen-CG	2.156071e+04	1639.7	7.49e-04	3.50e-14	29.80
PCAL	2.156071e+04	4775.8	2.77e-02	1.74e-14	39.30
PenC	2.156071e+04	4499.7	5.82e-03	1.76e-14	35.53
SLPG	2.156071e+04	4642.2	8.88e-04	1.72e-14	34.73
RCG	2.156071e+04	2871.3	1.20e-03	1.50e-14	102.80



(a) $(n, p) = (250, 50)$ (b) $(n, p) = (250, 50)$ (c) $(n, p) = (250, 50)]$

(d) $(n, p) = (2000, 50)$ (e) $(n, p) = (2000, 50)$ (f) $(n, p) = (2000, 50)]$

(g) $(n, p) = (1000, 100)$ (h) $(n, p) = (1000, 100)$ (i) $(n, p) = (1000, 100)$

Fig. 4.1. The convergence curves of ExPen-CG on nonlinear eigenvalue problems.

$$\begin{aligned} \min_{X \in \mathbb{R}^{n\times p}} \; & f(X) = \frac{1}{2}\mathrm{tr}(X^\top BXC) \\ \text{s.t.} \quad & X^\top X = I_p, \end{aligned} \tag{4.3}$$

where $B \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{p\times p}$ are two randomly generated symmetric matrices. We initiate ExPen-CG and RCG at the same point, which is randomly generated over $\mathcal{S}_{n,p}$. Tables 4.3 and 4.4 illustrate the performance of these compared algorithms with different combinations of problem parameters n and p. Here, we run each instance for 10 times and present the averaged results. We observe that all the compared algorithms achieve almost the same function values, and ExPen-CG achieves comparable performance as all the compared algorithms. In particular, ExPen-CG outperforms RCG in all the test instances.

Table 4.3: The results of the Brockett function minimization problems with varying n.

Solver	Fval	Iteration	Stationarity	Feasibility	CPU time(s)
$n = 250, p = 50$					
ExPen-CG	-9.611694e+00	962.7	7.97e-04	1.64e-14	2.99
PCAL	-9.610215e+00	2209.5	9.96e-04	8.46e-15	3.07
PenC	-9.610213e+00	2326.1	9.93e-04	8.43e-15	3.26
SLPG	-9.610306e+00	1848.4	9.46e-04	8.27e-15	2.45
RCG	-9.613055e+00	2417.4	9.92e-04	3.84e-15	12.39
$n = 500, p = 50$					
ExPen-CG	-9.570277e+00	859.4	7.45e-04	1.69e-14	6.97
PCAL	-9.568621e+00	1917.1	9.90e-04	8.20e-15	6.81
PenC	-9.568620e+00	2176.8	9.90e-04	8.29e-15	7.61
SLPG	-9.568844e+00	1844.6	9.44e-04	8.11e-15	6.15
RCG	-9.571521e+00	2172.6	9.96e-04	3.15e-15	24.76
$n = 1000, p = 50$					
ExPen-CG	-1.056750e+01	816.8	7.55e-04	1.69e-14	9.17
PCAL	-1.056576e+01	1789.7	9.91e-04	8.22e-15	10.03
PenC	-1.056577e+01	2190.5	9.90e-04	8.20e-15	11.79
SLPG	-1.056600e+01	1637.0	9.48e-04	7.97e-15	10.81
RCG	-1.056893e+01	3078.1	9.94e-04	3.07e-15	49.66
$n = 1500, p = 50$					
ExPen-CG	-1.059979e+01	723.9	7.62e-04	1.63e-14	12.14
PCAL	-1.059830e+01	1632.3	9.87e-04	7.80e-15	14.00
PenC	-1.059828e+01	2004.8	9.96e-04	7.87e-15	16.55
SLPG	-1.059850e+01	1518.6	9.28e-04	7.90e-15	12.49
RCG	-1.060131e+01	2845.3	9.96e-04	2.79e-15	66.11
$n = 2000, p = 50$					
ExPen-CG	-1.098884e+01	712.2	7.63e-04	1.67e-14	16.49
PCAL	-1.098736e+01	1584.2	9.81e-04	7.97e-15	19.29
PenC	-1.098741e+01	1837.1	9.67e-04	7.93e-15	21.78
SLPG	-1.098743e+01	1514.2	9.61e-04	7.80e-15	15.59
RCG	-1.099041e+01	3106.6	9.94e-04	2.61e-15	101.91

Table 4.4: The results of the Brockett function minimization problems with varying p.

Solver	Fval	Iteration	Stationarity	Feasibility	CPU time(s)
$n = 1000, p = 10$					
ExPen-CG	-2.551708e+00	343.1	6.20e-04	4.62e-15	0.86
PCAL	-2.551070e+00	623.4	9.50e-04	2.19e-15	0.98
PenC	-2.551053e+00	675.0	9.78e-04	1.90e-15	1.01
SLPG	-2.551101e+00	582.3	9.41e-04	2.21e-15	0.87
RCG	-2.552008e+00	1278.9	9.91e-04	1.09e-15	5.69
$n = 1000, p = 30$					
ExPen-CG	-6.479578e+00	576.9	6.95e-04	9.92e-15	3.16
PCAL	-6.478547e+00	1345.6	9.89e-04	4.87e-15	3.99
PenC	-6.478539e+00	1553.4	9.94e-04	5.10e-15	4.42
SLPG	-6.478541e+00	1289.7	9.72e-04	4.86e-15	4.81
RCG	-6.480546e+00	2474.2	9.95e-04	2.11e-15	22.56
$n = 1000, p = 50$					
ExPen-CG	-1.032617e+01	838.2	7.45e-04	1.64e-14	9.00
PCAL	-1.032438e+01	1698.3	9.94e-04	8.25e-15	9.66
PenC	-1.032442e+01	2127.4	9.88e-04	8.22e-15	11.58
SLPG	-1.032457e+01	1522.5	9.39e-04	8.13e-15	9.71
RCG	-1.032752e+01	2895.4	9.96e-04	3.10e-15	49.96
$n = 1000, p = 70$					
ExPen-CG	-1.312272e+01	850.7	7.84e-04	2.02e-14	12.65
PCAL	-1.312096e+01	2131.2	9.96e-04	1.03e-14	14.91
PenC	-1.312099e+01	2492.5	9.97e-04	1.03e-14	16.79
SLPG	-1.312106e+01	1848.2	9.54e-04	1.01e-14	16.20
RCG	-1.312455e+01	3298.1	9.96e-04	3.74e-15	69.18
$n = 1000, p = 100$					
ExPen-CG	-2.039827e+01	1169.3	7.91e-04	2.69e-14	22.14
PCAL	-2.039602e+01	2413.2	9.97e-04	1.31e-14	24.27
PenC	-2.039603e+01	2838.3	9.94e-04	1.31e-14	27.86
SLPG	-2.039626e+01	2289.9	9.43e-04	1.29e-14	24.60
RCG	-2.040059e+01	4261.6	9.96e-04	4.85e-15	128.97

5. Conclusion

The optimization over the Stiefel manifold has a close connection with unconstrained optimization. To efficiently extend existing unconstrained optimization approaches to their Stiefel versions and establish the corresponding theoretical analysis, most existing approaches are mainly based on the frameworks summarized in [4]. These approaches always involve computing the retractions and parallel/vector transports. However, computing retractions or parallel transport on the Stiefel manifold lack efficiency or scalability, while computing the vector transport can hardly inherit nice techniques in theoretical analysis.

In this paper, we present a novel exact smooth penalty function and its corresponding penalty model ExPen for OCP. We show that ExPen is well-defined under mild assumptions and study its theoretical properties. As illustrated in Figs. 3.1 and 3.2, we have proved the first-order and second-order relationships between OCP and ExPen, respectively. These properties guarantee that ExPen and OCP share first-order or second-order stationary points or local minimizers with a sufficiently large given penalty parameter.

In conclusion, we can directly adopt unconstrained optimization approaches to solve OCP through the bridge built by ExPen. Meanwhile, we can easily inherit the nice convergence properties of those approaches. We use the nonlinear conjugate gradient method as an instance. We present its ExPen version and establish its global convergence. It is worth mentioning that this ExPen version is performed in Euclidean space and hence avoids computing the retractions or parallel transports on the Stiefel manifold. Our present example highlights that those progress in nonconvex unconstrained optimization will immediately benefit optimization over the Stiefel manifold through ExPen. Moreover, the presented numerical examples further address the great potential of ExPen.

Acknowledgments. Research is supported in part by the National Natural Science Foundation of China (Grant Nos. 12125108, 11971466, 12288201, 12021001, 11991021), and by the Key Research Program of Frontier Sciences, Chinese Academy of Sciences (Grant No. ZDBS-LY-7022).

References

[1] P. Ablin and G. Peyré, Fast and accurate optimization on the orthogonal manifold without retraction, in: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, PMLR, (2022), 5636–5657.

[2] T.E. Abrudan, J. Eriksson, and V. Koivunen, Steepest descent algorithms for optimization under unitary matrix constraint, *IEEE Trans. Signal Process.*, **56**:3 (2008), 1134–1147.

[3] T. Abrudan, J. Eriksson, and V. Koivunen, Conjugate gradient algorithm for optimization under unitary matrix constraint, *Signal Process.*, **89**:9 (2009), 1704–1714.

[4] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, 2009.

[5] K. Ahn and S. Sra, From Nesterov's estimate sequence to Riemannian acceleration, in: *Proceedings of Thirty Third Conference on Learning Theory*, PMLR, (2020), 84–118.

[6] N. Bansal, X. Chen, and Z. Wang, Can we gain more from orthogonality regularizations in training deep networks? in: *Adv. Neural Inf. Process. Syst.*, (2008), 4261–4271.

[7] G. Bécigneul and O.-E. Ganea, Riemannian adaptive optimization methods, *arXiv:1810.00760*, 2018.

[8] D.P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, 2014.

[9] J. Bolte, S. Sabach, and M. Teboulle, Proximal alternating linearized minimization for nonconvex and nonsmooth problems, *Math. Program.*, **146**:1 (2014), 459–494.

[10] N. Boumal, *An Introduction to Optimization on Smooth Manifolds*, Cambridge University Press, 2023.

[11] Y.F. Cai, L.H. Zhang, Z.J. Bai, and R.C. Li, On an eigenvector-dependent nonlinear eigenvalue problem, *SIAM J. Matrix Anal. Appl.*, **39**:3 (2018), 1360–1382.

[12] C. Criscitiello and N. Boumal, Efficiently escaping saddle points on manifolds, in: *Adv. Neural Inf. Process. Syst.*, **32** (2019).

[13] C. Criscitiello and N. Boumal, An accelerated first-order method for non-convex optimization on manifolds, *Found. Comput. Math.*, **23**:4 (2023), 1433–1509.

[14] Y.H. Dai and Y.X. Yuan, A nonlinear conjugate gradient method with a strong global convergence property, *SIAM J. Optim.*, **10**:1 (1999), 177–182.

[15] J.W. Daniel, The conjugate gradient method for linear and nonlinear operator equations, *SIAM J. Numer. Anal.*, **4**:1 (1967), 10–26.

[16] E.F. Figueroa and O. Dalmau, Transportless conjugate gradient for optimization on Stiefel manifold, *Comput. Appl. Math.*, **39** (2020), 1–18.

[17] M.A. Fischler and R.C. Bolles, Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography, *Commun. ACM*, **24**:6 (1981), 381–395.

[18] R. Fletcher, *Practical Methods of Optimization*, John Wiley & Sons, 2013.

[19] R. Fletcher and C.M. Reeves, Function minimization by conjugate gradients, *Comput. J.*, **7**:2 (1964), 149–154.

[20] B. Gao, X. Liu, X.J. Chen, and Y.X. Yuan, A new first-order algorithmic framework for optimization problems with orthogonality constraints, *SIAM J. Optim.*, **28**:1 (2018), 302–332.

[21] B. Gao, X. Liu, and Y.X. Yuan, Parallelizable algorithms for optimization problems with orthogonality constraints, *SIAM J. Sci. Comput.*, **41**3 (2019), A1949–A1983.

[22] W.W. Hager and H.C. Zhang, A new conjugate gradient method with guaranteed descent and an efficient line search, *SIAM J. Optim.*, **16**:1 (2005), 170–192.

[23] W.W Hager and H.C. Zhang, A survey of nonlinear conjugate gradient methods, *Pacific J. Optim.*, **2**:1 (2006), 35–58.

[24] A. Han and J. Gao, Escape saddle points faster on manifolds via perturbed Riemannian stochastic recursive gradient, *arXiv:2010.12191*, 2020.

[25] D. Han and J. Kim, Unsupervised simultaneous orthogonal basis clustering feature selection, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2015), 5016–5023.

[26] C.R. Harris et al., Array programming with NumPy, *Nature*, **585**:7825 (2020), 357–362.

[27] K. Helfrich, D. Willmott, and Q. Ye, Orthogonal recurrent neural networks with scaled Cayley transform, in: *Proceedings of the 35th International Conference on Machine Learning*, PMLR, (2018), 1969–1978.

[28] M.R. Hestenes, Multiplier and gradient methods, *J. Optim. Theory Appl.*, **4**:5 (1969), 303–320.

[29] M.R. Hestenes and E. Stiefel, Methods of conjugate gradients for solving linear systems, *Journal of Research of the National Bureau of Standards* **49** (1952), 409–435.

[30] N.J. Higham and P. Papadimitriou, A parallel algorithm for computing the polar decomposition, *Parallel Comput.*, **20** (1994), 1161–1173.

[31] S. Hosseini, Convergence of nonsmooth descent methods via Kurdyka-Lojasiewicz inequality on Riemannian manifolds, *Hausdorff Center for Mathematics and Institute for Numerical Simulation, University of Bonn, INS Preprint No. 1523*, 2015.

[32] T.Y. Hou, Z.Z. Li, and Z.Y. Zhang, Analysis of asymptotic escape of strict saddle sets in manifold optimization, *SIAM J. Math. Data Sci.*, **2**:3 (2020), 840–871.

[33] J. Hu, X. Liu, Z.W. Wen, and Y.X. Yuan, A brief introduction to manifold optimization, *J. Oper. Res. Soc. China*, **8**:2 (2020), 199–248.

[34] X.Y. Hu and X. Liu, An efficient orthonormalization-free approach for sparse dictionary learning and dual principal component pursuit, *Sensors*, **20** (2020), 3041.

[35] W. Huang and K. Wei, Riemannian proximal gradient methods, *Math. Program.*, **194**, (2022), 371–413.

[36] B. Jiang and Y.H. Dai, A framework of constraint preserving update schemes for optimization on Stiefel manifold, *Math. Program.*, **153**:2 (2015), 535–575.

[37] W. Kohn and L.J. Sham, Self-consistent equations including exchange and correlation effects, *Phys. Rev.*, **140** (1965), A1133.

[38] M. Lezcano-Casado, Trivializations for gradient-based optimization on manifolds, in: *Proceedings*

of the 33rd International Conference on Neural Information Processing Systems, (2019), 9157–9168.

[39] M. Lezcano-Casado and D. Martınez-Rubio, Cheap orthogonal constraints in neural networks: A simple parametrization of the orthogonal and unitary group, in: *Proceedings of the 36th International Conference on Machine Learning*, PMLR, (2019), 3794–3803.

[40] J. Li, L. Fuxin, and S. Todorovic, Efficient Riemannian optimization on the Stiefel manifold via the Cayley transform, *arXiv:2002.01113*, 2020.

[41] X.L. Li, H. Zhang, R. Zhang, Y. Liu, and F.P. Nie, Generalized uncorrelated regression with adaptive graph for unsupervised feature selection, *IEEE Trans. Neural Netw. Learn. Syst.*, **30**:5 (2018), 1587–1595.

[42] X. Liu, X. Wang, Z.W. Wen, and Y.X. Yuan, On the convergence of the self-consistent field iteration in Kohn-Sham density functional theory, *SIAM J. Matrix Anal. Appl.*, **35**:2 (2014), 546–558.

[43] X. Liu, Z.W. Wen, X. Wang, M. Ulbrich, and Y. Yuan, On the analysis of the discretized Kohn-Sham density functional theory, *SIAM J. Numer. Anal.*, **53**:4 (2015), 1758–1785.

[44] Y. Liu and C. Storey, Efficient generalized conjugate gradient algorithms, Part 1: Theory, *J. Optim. Theory Appl.*, **69**:1 (1991), 129–137.

[45] S. Łojasiewicz, *Sur le Probleme de la Division*, Instytut Matematyczny Polskiej Akademi Nauk, 1961.

[46] S. Lojasiewicz, Une propriété topologique des sous-ensembles analytiques réels, *Les équations aux dérivées partielles*, **117** (1963), 87–89.

[47] K.D.G. Maduranga, K.E. Helfrich, and Q. Ye, Complex unitary recurrent neural networks using scaled Cayley transform, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, **33** (2019), 4528–4535.

[48] J. Nocedal and S. Wright, *Numerical Optimization*, Springer Science and Business Media, 2006.

[49] K. Pearson, LIII. On lines and planes of closest fit to systems of points in space, *Lond. Edinb. Dublin Philos. Mag. J. Sci.*, **2**:11 (1901), 559–572.

[50] B.T. Polyak, The conjugate gradient method in extremal problems, *USSR Comput. Math. Math. Phys.*, **9**:4 (1969), 94–112.

[51] M.J.D. Powell, A method for nonlinear constraints in minimization problems, *Optimization*, (1969), 283–298.

[52] H. Sato, A Dai–Yuan-type Riemannian conjugate gradient method with the weak Wolfe conditions, *Comput. Optim. Appl.*, **64**:1 (2016), 101–118.

[53] H. Sato and T. Iwai, A new, globally convergent Riemannian conjugate gradient method, *Optimization*, **64**:4 (2015), 1011–1031.

[54] J.W. Siegel, Accelerated optimization with orthogonality constraints, *J. Comput. Math.*, **39**:2 (2021), 207–226.

[55] Y. Sun, N. Flammarion, and M. Fazel, Escaping from saddle points on Riemannian manifolds, *arXiv:1906.07355*, 2019.

[56] J. Townsend, N. Koep, and S. Weichwald, Pymanopt: A Python toolbox for optimization on manifolds using automatic differentiation, *J. Mach. Learn. Res.*, **17** (2016), 4755–4759.

[57] P. Virtanen et al., SciPy 1.0: Fundamental algorithms for scientific computing in Python, *Nat. methods*, **17**:3 (2020), 261–272.

[58] L. Wang, B. Gao, and X. Liu, Multipliers correction methods for optimization problems over the Stiefel manifold, *CSIAM Trans. Appl. Math.*, **2**:3 (2021), 508–531.

[59] L. Wang, X. Liu, and Y. Zhang, A distributed and secure algorithm for computing dominant SVD based on projection splitting, *arXiv:2012.03461*, 2020.

[60] Z.W. Wen and W.T. Yin, A feasible method for optimization with orthogonality constraints, *Math. Program.*, **142**:1-2 (2013), 397–434.

[61] N.C. Xiao, X. Liu, and Y.X. Yuan, A class of smooth exact penalty function methods for optimiza-

tion problems with orthogonality constraints, *Optim. Methods Softw.*, **37**:4 (2022), 1205–1241.

[62] N.C. Xiao, X. Liu, and Y.X. Yuan, Exact penalty function for $\ell_{2,1}$ norm minimization over the Stiefel Manifold, *SIAM J. Optim.*, **31** (2021), 3097–3126.

[63] N.C. Xiao, X. Liu, and Y.X. Yuan, A penalty-free infeasible approach for a class of nonsmooth opimtization problems over the Stiefel manifold, *arXiv:2103.03514*, 2021.

[64] N.C. Xiao, L. Wang, B. Gao, X. Liu, and Y.X. Yuan, STOP: A toolbox for Stiefel manifold optimization, 2022. `https://stmopt.gitee.io/`

[65] C. Yang, W.G. Gao, and J.C. Meza, On the convergence of the self-consistent field iteration for a class of nonlinear eigenvalue problems, *SIAM J. Matrix Anal. Appl.*, **30**:4 (2009), 1773–1788.

[66] H. Zhang and S. Sra, Towards Riemannian accelerated gradient methods, *arXiv:1806.02812*, 2018.

[67] J. Zhang, H. Zhang, and S. Sra, R-SPIDER: A fast Riemannian stochastic optimization algorithm with curvature independent rate, *arXiv:1811.04194*, 2018.

[68] R. Zhang, F.P. Nie, and X.L. Li, Feature selection under regularized orthogonal least square regression with optimal scaling, *Neurocomputing*, **273** (2018), 547–553.

[69] P. Zhou, X.-T. Yuan, S. Yan, and J. Feng, Faster first-order methods for stochastic non-convex optimization on Riemannian manifolds, *IEEE Trans. Pattern. Anal. Mach. Intell.*, **43**:2 (2021), 459–472.

[70] X.J. Zhu, A Riemannian conjugate gradient method for optimization on the Stiefel manifold, *Comput. Optim. Appl.*, **67**:1 (2017), 73–110.

[71] G. Zoutendijk, Nonlinear programming, computational methods, in: *Integer and Nonlinear Programming*, North-Holllad, (1970), 37–86.