Noname manuscript No.
(will be inserted by the editor)

A Class of Smooth Exact Penalty Function Methods for Optimization Problems with Orthogonality Constraints

Nachuan Xiao · Xin Liu · Ya-xiang Yuan

Received: date / Accepted: date

Abstract

Updating the augmented Lagrangian multiplier by closed-form expression yields efficient first-order infeasible approach for optimization problems with orthogonality constraints. Hence, parallelization becomes tractable in solving this type of problems. Inspired by this closed-form updating scheme, we propose an exact penalty function model with compact convex constraints (PenC). We show that PenC can act as an exact penalty model in some senses. Based on PenC, we first propose a first-order algorithm called PenCF and establish its global convergence and local linear convergence rate under some mild assumptions. For the case that the computation and storage of Hessian is achievable, and we pursue high precision solution and fast local convergence rate, a second-order approach called PenCS is proposed for solving PenC. To avoid expensive calculation or solving a hard subproblem in computing the Newton step, we propose a new strategy to do it approximately which still leads to quadratic convergence locally. Moreover, the main iterations of both PenCF and PenCS are orthonormalization-free and hence parallelizable. Numerical experiments illustrate that PenCF is comparable with the existing first-order methods. Furthermore, PenCS shows its stability and high efficiency in obtaining high precision solution comparing with the existing second-order methods.

Keywords orthogonality constraint, Stiefel manifold, augmented Lagrangian method

Mathematics Subject Classification (2010) 15A18, 65F15, 65K05, 90C06

1 Introduction

In this paper, we consider the following matrix optimization problem with orthogonality constraints:

$$\begin{aligned} \min_{X\in\mathbb{R}^{n\times p}} \quad & f(X) \\ \text{s.t.} \quad & X^\top X = I_p, \end{aligned} \tag{1.1}$$

where I_p is any $p \times p$ identity matrix, and $f : \mathbb{R}^{n\times p} \mapsto \mathbb{R}$ satisfying the following assumption throughout this paper.

Assumption 1 (blanket assumption) *$f(X)$ is differentiable and $\nabla f(X)$ is locally Lipschitz continuous.*

For brevity, the orthogonality constraints $X^\top X = I_p$ in problem (1.1) can be expressed as $X \in \mathcal{S}_{n,p} := \{X \in \mathbb{R}^{n\times p} | X^\top X = I_p\}$. Here $\mathcal{S}_{n,p}$ denotes the Stiefel manifold in real matrix space, and we call it as Stiefel manifold for simplicity in the rest of our paper.

Problem (1.1) plays an important role in data science and engineering. We mention a few examples in the following.

N. Xiao
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, E-mail: xnc@lsec.cc.ac.cn

X. Liu
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, E-mail: liuxin@lsec.cc.ac.cn

Y. Yuan
State Key Laboratory of Scientific and Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China, E-mail: yyx@lsec.cc.ac.cn

Example 1 (Discretized Kohn-Sham Energy Minimization) Kohn-Sham density functional theory (KSDFT) is an important tool in electronic structure calculation [22]. In the last step of KSDFT, we need to minimize the so-called discretized Kohn-Shan energy function subjected to orthogonality constraints.

$$\begin{aligned} \min_{X\in\mathbb{R}^{n\times p}} \quad & E(X) := \frac{1}{4}\mathrm{tr}\left(X^\top L X\right) + \frac{1}{2}\mathrm{tr}\left(X^\top V_{ion} X\right) + \frac{1}{4}\rho^\top L^\dagger \rho + \frac{1}{2}\rho^\top \epsilon_{\mathrm{xc}}(\rho) \\ \text{s.t.} \quad & X^\top X = I_p, \end{aligned} \tag{1.2}$$

where $L \in \mathbb{R}^{n\times n}$ and diagonal matrix $V_{\mathrm{ion}} \in \mathbb{R}^{n\times n}$ refer to the Laplace operator in the planewave basis and discretized local ionic potential, respectively, $\rho := \mathrm{diag}(XX^\top)$ denotes the charge density, and $\epsilon_{\mathrm{xc}} : \mathbb{R}^n \mapsto \mathbb{R}^n$ stands for the exchange correlation function.

Example 2 (Unsupervised feature selection) In solving unsupervised feature selection problem in [16, 24, 49], to compute the indicator matrix embedded in the input data space, the corresponding subproblem involves minimizing a quadratic function over Stiefel manifold:

$$\begin{aligned} \min_{X\in\mathbb{R}^{n\times p}} \quad & \frac{1}{2}\mathrm{tr}(X^\top W^\top W X) - \mathrm{tr}(G^\top X) \\ \text{s.t.} \quad & X^\top X = I_p, \end{aligned} \tag{1.3}$$

where $W \in \mathbb{R}^{m\times n}$ and $G \in \mathbb{R}^{n\times p}$ are constructed from input data, and the orthogonality constraints are imposed to keep each rowllj of X uncorrelated.

In fact, if we neglect the restriction on the objective value from Assumption 1. There are more applications such as [5] in machine learning area, [21] in sparse principle component analysis, [47, 39] for unsupervised feature selection, etc. But these problems are beyond the scope of this paper.

1.1 Existing methods

By exploring the geometric structure of the Stiefel manifold $\mathcal{S}_{n,p}$, quite a few first-order methods have been proposed to solve (1.1) in recent years, such as gradient-based method [28, 29, 2], conjugate gradient methods [13, 1], projection-based methods [4, 11], constraint preserving updating scheme [41, 20], multipliers correction framework [14], etc. Interested readers are referred to the references in the book [4].

Besides, there are also second-order methods, see [3, 4, 19], for instance, proposed for optimization problems with orthogonality constraints. The main ideas of these existing methods can be concluded as trust-region strategy with quadratic approximation. The difference among those methods lies in the sets where the quadratic approximation models are made.

For example, in Riemannian trust-region method (RTR) [3, 8], the quadratic model is chosen as

$$m_k(X) := \mathrm{tr}\left((X - X_k)^\top \mathrm{grad} f(X_k) + \frac{1}{2}(X - X_k)^\top \mathrm{Hess} f(X_k)[X - X_k]\right),$$

where $\mathrm{grad} f(X_k)$ and $\mathrm{Hess} f(X_k)$ stand for the Riemannian gradient and Hessian, respectively. The latter one can be expressed as a linear mapping from $\mathbb{R}^{n\times p}$ to $\mathbb{R}^{n\times p}$. Hence, the trust-region subproblem can be expressed as

$$\begin{aligned} \min_{X\in\mathcal{T}(X_k)} \quad & m_k(X) \\ \text{s.t.} \quad & \|X - X_k\|_{\mathrm{F}} \le \Delta_k. \end{aligned} \tag{1.4}$$

(1.4) is a trust-region subproblem with linear equation. To solve it efficiently, Boumal et al. [8] suggests to use projected conjugate method, in which calculating projections to $\mathcal{T}(X_k)$ is intensively invoked.

On contrast, the adaptive regularized Newton method for Riemannian optimization (ARNT) of Hu et al. [19] builds up the following quadratic model

$$m_k(X) := \mathrm{tr}\left((X - X_k)^\top \nabla f(X_k) + \frac{1}{2}(X - X_k)^\top \nabla^2 f(X_k)[X - X_k]\right) + \frac{\sigma_k}{2}\|X - X_k\|_{\mathrm{F}}^2, \tag{1.5}$$

where $\sigma_k > 0$ is a regularization parameter, $\nabla f(X)$ and $\nabla^2 f(X)$ refer to the gradient and Hessian, respectively, in the Euclidean space. The corresponding trust-region subproblem can be formulated as in the following

$$\begin{aligned} \min \quad & m_k(X) \\ \text{s.t.} \quad & X^\top X = I_p, \end{aligned} \tag{1.6}$$

which is an optimization problem on the manifold as well. Therefore, to compute the global minimizer of the subproblem (1.6) is theoretically intractable. In Hu et al. [19], the authors suggest to solve the subproblem inexactly by projection-based Riemannian optimization approach. Convergence to the first-order stationary points of the subproblem can be guaranteed.

All the above-mentioned approaches require feasibility in each iteration. To keep the orthogonality, either explicit orthonormalization, such as Gram-Schmidt or QR orthonormalization, or implicit orthonormalization, such as constraint preserving updating scheme, is intensively involved. The orthonormalization process lacks of concurrency which leads to low scalability in column-wise parallel computing, particularly when the number of columns is large. Refer to [9, 36, 32, 25, 33], parallelizable optimization methods attain lots of attentions in various application scenarios. Particularly, for optimization problems with orthogonality constraints with application in electronic structure calculation based on Kohn-Sham density functional theory, there is urgent demand to develop parallelizable approaches, for example see Gao et al. [15]. Classical constrained optimization methods, such as sequential quadratic programming, augmented Lagrangian method, are much less efficient than the existing feasible optimization methods when applied to solve (1.1). Therefore, we can not expect to gain efficiency by parallelizing them directly. In Lai and Osher [23], the authors propose an approach called "Splitting Orthogonality Constraints (SOC)" whose main steps include splitting the variables in the objective and contraints, introducing linear equality constraints to balance the two groups of variables, and adopting the alternating direct method of multipliers to solve the split problem. SOC works quite well in solving problems arisen from image processing, but performs not satisfactory in other scenarios like quadratic objective minimization with orthogonality constraints, and discretized Kohn-Sham energy minimization (see details in Gao et al. [15]). Moreover, the global convergence of SOC has not been established.

Consequently, we focus on investigating efficient infeasible approaches by penalty function methods, which preserve the structure of the orthogonality constraints. The L_1 penalization [31, 5] can provide exact penalty functions. However, minimizing an unconstrained nonsmooth nonconvex penalty function is not an easy task. For dominant eigenpairs computation, Liu et al. [27], and Wen et al. [42] propose an "exact" penalty function method by Courant penalty function [10, 30, 38]. However, such idea only works for homogeneous quadratic objective and can not be extended to general cases. Very recently, based on augmented Lagrangian method (ALM) [17, 35, 30, 7], the authors of [15] propose the proximal linearized augmented Lagrangian method (PLAM) and its column-wise normalization version (PCAL). Both PLAM and PCAL update the multipliers corresponding to the orthogonality constraints by a closed-form expression introduced in Gao et al. [14] which hold at any first-order stationary point. Numerical experiments in Gao et al. [15] illustrate the great potential of PLAM and PCAL, particularly, in parallel computing. However, there is little understanding on the merit function used in the theoretical analysis, and to extend to the second-order methods is intractable.

1.2 Contributions

In this paper, we begin with revisiting the merit function

$$h(X) = f(X) - \frac{1}{2}\left\langle \Phi(\nabla f(X)^\top X), X^\top X - I_p \right\rangle + \frac{\beta}{4}||X^\top X - I_p||_{\mathrm{F}}^2, \tag{1.7}$$

where $\Phi : \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}, \Phi(M) = \frac{M+M^\top}{2}$ denotes the linear operator that symmetrizes M. This merit function is first defined in Gao et al. [15, Equation (4.2)] to evaluate the function value reduction of PLAM. We propose a new penalty model which minimizes $h(X)$ subject to a convex compact constraint (PenC) as follows.

$$\min_{X\in\mathcal{M}} \quad h(X), \tag{1.8}$$

where $\mathcal{M}$ is a compact convex set $\mathcal{M}$ that contains $\mathcal{S}_{n,p}$.

We establish the equivalence between the original problem (1.1) and PenC in the sense that they share the same second-order stationary points, and hence, the global minimizers under some mild conditions. In addition, PCAL can be viewed as an algorithm for solving PenC with particular constraint, but PLAM can not.

Moreover, when choosing $\mathcal{M}$ as the ball $\mathcal{B}_K$ with radius $K > \sqrt{p}$, we propose an approximate gradient method (PenCF) to solve the corresponding PenC model. The orthonormalization process is completely waived in PenCF, except that a one time orthonormalization is invoked after the last iteration of PenCF if high feasibility accuracy is required. Since the iterates are restricted in a compact region, the performance of PenCF is not sensitive with the choice of parameter β. We prove the global convergence of PenCF and establish the local linear convergence rate in a neighborhood of any isolated

local minimizer of (1.1). The projection to a ball from outside is nothing but normalization, and hence is very cheap. On the other hand, numerical experiments show that such projection is only required in very few iterations, namely, the ball B is often inactive in the iterative procedure.

Furthermore, we propose a second-order method (PenCS) to solve PenC. In each iteration of PenCS, we construct an approximate Hessian of the merit function $h(X)$ and calculate the truncated Newton direction by solving a trust-region subproblem. We also provide the local quadratic convergence rate for PenCS in a neighborhood of any isolated local minimizer of (1.1).

The primarily numerical experiments demonstrate that PenCF is comparable with the existing feasible methods and PCAL in solving discretized Kohn-Sham total energy minimizing problems. In addition, PenCS is superior to the well-established second-order algorithms ARNT and RTR in pursuing second-order stationary points with high precision, particularly in solving problems with large column size.

1.3 Organization

The rest of this paper is organized as follows. We put all the preliminaries including the optimality condition and the analysis on our new model in Section 2. In Section 3, we study the relationship between the original problem (1.1) and PenC. We present a first-order algorithm PenCF and establish its global convergence and local linear convergence rate in Section 4. In section 5, we propose a second-order algorithm PenCS as well as its local quadratic convergence rate. Numerical experiments are reported in section 6. In the last section, we draw a brief conclusion.

1.4 Notation

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 a matrix $A \in \mathbb{R}^{p\times p}$. $\|\cdot\|_2$ 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. $X^\dagger$ refers to the pseudo-inverse of X. We denote the smallest eigenvalue of A by $\lambda_{\min}(A)$. The i-th column of matrix $X \in \mathbb{R}^{n\times p}$ is denoted by X_i $(i = 1, ..., p)$. $\mathbf{qr}(X)$ is the Q matrix of the reduced QR decomposition[1] of X. $\mathcal{P}_{\mathcal{S}_{n,p}}(X)$ denotes the projection[2] of X to the Stiefel manifold $\mathcal{S}_{n,p}$. conv Ω is denoted as the convex hull of the set Ω. Finally, $\mathbf{0}_{n,p}$ stands for the $n \times p$ matrix with all entries being equal to zero.

2 Preliminaries

In this section, we begin with the optimality conditions of the optimization problems with orthogonality constraints. Then we discuss the main thoughts of the algorithms PLAM and PCAL proposed in [15]. Finally, we study the properties of the merit function (1.7) and propose our new penalty model.

2.1 Optimality conditions

The first-order optimality condition of problem (1.1) can be written as the following.

Definition 1 *Given a point $X \in \mathbb{R}^{n\times p}$, if the relationship*

$$\begin{cases} \mathrm{tr}(Y^\top \nabla f(X)) \geq 0; \\ X^\top X = I_p \end{cases}$$

holds for any $Y \in \mathcal{T}(X)$, we call X a first-order stationary point of (1.1). *Here, $\mathcal{T}(X) := \{Y \mid Y^\top X + X^\top Y = 0\}$ is the tangent space of the orthogonality constraints at X.*

According to Lemma 2.2 in [14], a point X is a first-order stationary point if and only if

$$\begin{cases} (I_n - XX^\top)\nabla f(X) = 0; \\ X^\top \nabla f(X) = \nabla f(X)^\top X; \\ X^\top X = I_p. \end{cases} \tag{2.1}$$

[1] $Q \in \mathbb{R}^{n\times p}$ is the Q matrix of the reduced QR decomposition of $X \in \mathbb{R}^{n\times p}$, if $X = QR$, $Q \in \mathbb{R}^{n\times p}$ is orthogonal and $R \in \mathbb{R}^{p\times p}$ is an upper triangle matrix.

[2] $\mathcal{P}_{\mathcal{S}_{n,p}}(X) = \tilde{U}\tilde{V}^\top$ where $\tilde{U}\Sigma\tilde{V}^\top$ is the reduced singular value decomposition of X.

It can be easily derived from (2.1) that $\Lambda := \nabla f(X)^\top X \in \mathbb{S}^p$ can be viewed as the Lagrangian multipliers of the orthogonality constraints at any first-order stationary point.

In some context of this paper, we also assume second-order differentiability of $f(X)$ than what is assumed in Assumption 1.

Assumption 2 *$f(X)$ is twice continuously differentiable on $\mathbb{R}^{n\times p}$, and $\nabla^2 f(X)$ is locally Lipschitz continuous in $\mathbb{R}^{n\times p}$.*

Definition 2 *We call X a first-order stationary point of problem* (1.1)*, if condition* (2.1) *holds. Suppose Assumption 2 holds, we call X a second-order stationary point of problem* (1.1)*, if it is a first-order stationary point and satisfies*

$$\mathrm{tr}(Y^\top \nabla^2 f(X)[Y] - \Lambda Y^\top Y) \geq 0, \quad \forall\, Y \in \mathcal{T}(X). \tag{2.2}$$

Next, we state the optimality conditions of PenC. Since the differentiability of $h(X)$ often involves higher order differentiability of $f(X)$ which will be strictly discussed later. Here, we try to impose the least requirements on the differentiability of $h(X)$.

Definition 3 *1. The sequential feasible direction of $\mathcal{M}$ at X is defined as*

$$\mathrm{SFD}(X) := \{d \in \mathbb{R}^{n\times p} \mid \mathcal{M} \ni X_k \to X, \quad d = \lim_{t_k\to+\infty} t_k(X_k - X)\}.$$

2. X^* is a first-order stationary point of PenC if and only if for any $d \in \mathrm{SFD}(X^*)$, it holds

$$\liminf_{t\to 0^+} \quad \frac{1}{t}\left(h(X^* + td) - h(X^*)\right) \geq 0.$$

3. Suppose Assumption 2 holds, X^* is a second-order stationary point of PenC if and only if X^* is a first-order stationary point of PenC, it holds

$$\liminf_{X_k\to X^*, X_k\in\mathcal{M}} \quad \frac{h(X_k) - h(X^*)}{\|X_k - X^*\|_{\mathrm{F}}^2} \geq 0.$$

Remark 3 *The above definition of the first-order and second-order stationary points are equivalent to those standard ones, if $\nabla h(X^*)$ and $\nabla^2 h(X^*)$ exist, respectively.*

2.2 Augmented Lagrangian methods for solving optimization problems with orthogonality constraints

Augmented Lagrangian method (ALM) [34] is widely used in solving constrained optimization problems. The augmented Lagrangian penalty function for (1.1) can be written as,

$$\mathcal{L}(X,\Lambda) := f(X) - \frac{1}{2}\mathrm{tr}\left(\Lambda(X^\top X - I_p)\right) + \frac{\beta}{4}\left\|X^\top X - I_p\right\|_{\mathrm{F}}^2, \tag{2.3}$$

where $\Lambda \in \mathbb{R}^{p\times p}$ is called the Lagrangian multiplier corresponding to the orthogonality constraints.

In each step of ALM, X_k is first updated by solving the following unconstrained subproblem to certain precision when the multiplier Λ_k is fixed,

$$X_{k+1} = \operatorname*{arg\,min}_{X\in\mathbb{R}^{n\times p}} \quad \mathcal{L}(X,\Lambda_k).$$

Then, the multiplier Λ_k is updated by the following dual-ascend step,

$$\Lambda_{k+1} = \Lambda_k - \beta(X_{k+1}^\top X_{k+1} - I_p) \tag{2.4}$$

with the primal variable X^{k+1} fixed.

However, as shown by the numerical experiments in [15], the original ALM does not perform efficiently in solving optimization problems with orthogonality constraints. It is not competitive with most existing manifold based approaches. Besides, its convergence highly depends on the choice of the penalty parameter β.

PLAM, proposed by Gao et al. [15], updates the Lagrangian multiplier by the closed-form expression $\Lambda_k = \Lambda(X_k)$, where

$$\Lambda(X) := \Phi(\nabla f(X)^\top X), \tag{2.5}$$

and the operator Φ is defined by (1.7). The primal variable is then updated by a gradient step, which can be viewed as the closed-form solution of the following proximal linearized approximation of the augmented Lagrangian penalty function (2.3),

$$X_{k+1} = \arg\min_{X \in \mathbb{R}^{n\times p}} \quad (X - X_k)^\top \nabla_X \mathcal{L}(X_k, \Lambda_k) + \frac{1}{2\eta_k} \|X - X_k\|_{\mathrm{F}}^2, \tag{2.6}$$

where $\nabla_X \mathcal{L}(X, \Lambda) = \frac{\partial \mathcal{L}(X,\Lambda)}{\partial X}$ is the gradient of the augmented Lagrangian function with respect to X.

In practice, the performance of PLAM is sensitive to the penalty parameter β. Small β leads to divergence while large β results in slow convergence. To solve the sensitivity difficulty, Gao et al. [15] proposes an upgraded version PCAL, in which the primal variable is normalized column-wisely after the gradient step,

$$(X_{k+1})_i = (X_{k+1})_i / \|(X_{k+1})_i\|_2 .$$

The iterates of PCAL is restricted to the Oblique manifold $\mathcal{OB}_{n,p} := \{X \in \mathbb{R}^{n\times p} | \, \|X(:,i)\|_2 = 1, i = 1, ..., p\}$.

2.3 Motivation

In this paper, we revisit the merit function (1.7) and try to understand whether it can act as a penalty function or not. For convenience, we assume Assumption 2 holds throughout this subsection. We can verify that any first-order stationary point $\bar{X}$ of problem (1.1) satisfies $\nabla h(\bar{X}) = 0$. Moreover, $h(X_k) = \mathcal{L}(X_k, \Lambda_k)$ holds for any $k > 0$. Therefore, $h(X)$ inherits lots of good properties from $\mathcal{L}(X, \Lambda)$. On the other hand, $h(X)$ is not bounded below for a large variety of objective functions f. For example, let $f(X) = \frac{1}{4}\mathrm{tr}(X^\top X X^\top X)$, then

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

and for any orthogonal matrix $\tilde{U}$,

$$h(t\tilde{U}) = -\frac{p}{2}t^6 + \frac{(3+\beta)p}{4}t^4 - \frac{\beta p}{2}t^2 + \frac{\beta p}{4}.$$

Clearly, $h(X) \to -\infty$ when $t \to \infty$, namely, $||X|| \to \infty$.

Therefore, if we want to utilize $h(X)$ as a penalty function, we need first to restrict it into a compact region to keep it bounded from below. Consequently, we consider the PenC model defined in (1.8). The possible choices of $\mathcal{M}$ include, but not limited to, any ball with radius not smaller than $\sqrt{p}$, the convex hull of the Stiefel manifold

$$\mathrm{conv}\,(\mathcal{S}_{n,p}) = \{X \in \mathbb{R}^{n\times p} \mid ||X||_2 \le 1\}$$

and the convex hull of the Oblique manifold

$$\mathrm{conv}\,(\mathcal{OB}_{n,p}) = \{X \in \mathbb{R}^{n\times p} \mid ||X_i||_2 \le 1\}.$$

By direct calculation, $\nabla h(X)$ can be expressed as

$$\begin{aligned}\nabla h(X) = \nabla f(X) - X\Lambda(X) - \frac{1}{2}\nabla f(X)(X^\top X - I_p) \\ -\frac{1}{2}\nabla^2 f(X)[X(X^\top X - I_p)] + \beta X(X^\top X - I_p).\end{aligned} \tag{2.7}$$

Proposition 1 *Suppose $f(X)$ is three times continuously differentiable at X, the Hessian of $h(X)$ can be expressed in the following formula.*

$$\begin{aligned}\nabla^2 h(X)[D] =& \nabla^2 f(X)[D] - D\Lambda(X) - X\Phi(D^\top \nabla f(X)) - X\Phi(X^\top \nabla^2 f(X)[D]) - \nabla f(X)\Phi(D^\top X) \\ &-\nabla^2 f(X)[X\Phi(X^\top D)] - \frac{1}{2}\nabla^2 f(X)[D](X^\top X - I_p) - \frac{1}{2}\nabla^2 f(X)[D(X^\top X - I_p)] \\ &-\frac{1}{2}\nabla^3 f(X)[D, X(X^\top X - I_p)] + 2\beta X\Phi(X^\top D) + \beta D(X^\top X - I_p).\end{aligned} \tag{2.8}$$

Here,

$$\nabla^2 f(X) : D \in \mathbb{R}^{n\times p} \mapsto \nabla^2 f(X)[D] \in \mathbb{R}^{n\times p},$$

and

$$\nabla^3 f(X) : (D, D) \in \mathbb{R}^{n\times p} \otimes \mathbb{R}^{n\times p} \mapsto \nabla^3 f(X)[D, D] \in \mathbb{R}^{n\times p}$$

are the expressions of the Hessian and the third-order derivative of $f(X)$ in the linear mapping form, respectively.

Proof Since

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

we denote $f_1(X) := f(X), f_2(X) := \frac{1}{2}\mathrm{tr}\left(\Lambda(X)(X^\top X - I_p)\right), f_3 := \frac{\beta}{4}\left\|X^\top X - I_p\right\|_{\mathrm{F}}^2$.

For $f_1(X)$, by the definition of Hessian, we have $\nabla^2 f_1(X)[D] = \nabla^2 f(X)[D]$.

For $f_2(X)$, first we have that

$$\nabla f_2(X) = X\Lambda(X) + \frac{1}{2}\nabla f(X)(X^\top X - I_p) + \frac{1}{2}\nabla^2 f(X)[X(X^\top X - I_p)].$$

Then we define $g_{21}(X) := X\Lambda(X), g_{22}(X) := \frac{1}{2}\nabla f(X)(X^\top X - I_p)$ and $g_{23}(X) := \frac{1}{2}\nabla^2 f[X](X(X^\top X - I_p))$.

For any $D \in \mathbb{R}^{n\times p}$, we have

$$\begin{aligned}
&g_{21}(X+D) - g_{21}(X)\\
=&D\Lambda(X) + X(\Lambda(X+D) - \Lambda(X)) + \mathcal{O}(\|D\|_{\mathrm{F}}^2)\\
=&D\Lambda(X) + X\left[\Phi((X+D)^\top \nabla f(X+D)) - \Phi(X^\top \nabla f(X))\right] + \mathcal{O}(\|D\|_{\mathrm{F}}^2)\\
=&D\Lambda(X) + X\left[\Phi((X+D-X)^\top \nabla f(X)) + \Phi\left(X^\top(\nabla f(X+D) - \nabla f(X))\right)\right] + \mathcal{O}(\|D\|_{\mathrm{F}}^2)\\
=&D\Lambda(X) + X\Phi(D^\top \nabla f(X)) + X^\top \Phi(X^\top \nabla^2 f(X)[D]) + \mathcal{O}(\|D\|_{\mathrm{F}}^2).
\end{aligned}$$

For g_{22}, we have

$$\begin{aligned}
&g_{22}(X+D) - g_{22}(X)\\
=&\frac{1}{2}(\nabla f(X+D) - \nabla f(X))(X^\top X - I_p) + \nabla f(X)\Phi(X^\top D) + \mathcal{O}(\|D\|_{\mathrm{F}}^2)\\
=&\frac{1}{2}\nabla^2 f(X)[D](X^\top X - I_p) + \nabla f(X)\Phi(X^\top D) + \mathcal{O}(\|D\|_{\mathrm{F}}^2)
\end{aligned}$$

For $g_{23}(X)$, we have

$$\begin{aligned}
&g_{23}(X+D) - g_{23}(X) = \frac{1}{2}\left(\nabla^2 f(X+D) - \nabla^2 f(X)\right)[X(X^\top X - I_p)]\\
&+ \frac{1}{2}\nabla^2 f(X)[(X+D)((X+D)^\top (X+D) - I_p) - X(X^\top X - I_p)] + \mathcal{O}(\|D\|_{\mathrm{F}}^2)\\
=&\frac{1}{2}\nabla^3 f(X)[D, X(X^\top X - I_p)] + \frac{1}{2}\nabla^2 f(X)[D(X^\top X - I_p)] + \nabla^2 f(X)[X\Phi(X^\top D)] + \mathcal{O}(\|D\|_{\mathrm{F}}^2)
\end{aligned}$$

Then we can conclude that

$$\begin{aligned}
\nabla f_2(X)[D] =&\nabla g_{21}(X)[D] + \nabla g_{22}(X)[D] + \nabla g_{23}(X)[D]\\
=&D\Lambda(X) + X\Phi(D^\top \nabla f(X)) + X^\top \Phi(X^\top \nabla^2 f(X)[D]) + \frac{1}{2}\nabla^2 f(X)[D](X^\top X - I_p)\\
&+ \nabla f(X)\Phi(X^\top D) + \frac{1}{2}\nabla^3 f(X)[D, X(X^\top X - I_p)]\\
&+ \frac{1}{2}\nabla^2 f(X)[D(X^\top X - I_p)] + \nabla^2 f(X)[X\Phi(X^\top D)]
\end{aligned}$$

Moreover, for any $D \in \mathbb{R}^{n\times p}$, we have

$$\nabla f_3(X) = \beta X(X^\top X - I_p),$$

and

$$\nabla f_3(X+D) - \nabla f_3(X) = \beta D(X^\top X - I_p) + 2\beta X\Phi(X^\top D). \tag{2.9}$$

Since $h(X) = f_1(X) - f_2(X) + f_3(X)$, we can conclude that

$$\begin{aligned}
\nabla^2 h(X)[D] =&\nabla^2 f(X)[D] - D\Lambda(X) - X\Phi(D^\top \nabla f(X)) - X\Phi(X^\top \nabla^2 f(X)[D])\\
&- \nabla f(X)\Phi(D^\top X) - \nabla^2 f(X)[X\Phi(X^\top D)] - \frac{1}{2}\nabla^2 f(X)[D](X^\top X - I_p)\\
&- \frac{1}{2}\nabla^2 f(X)[D(X^\top X - I_p)] - \frac{1}{2}\nabla^3 f(X)[D, X(X^\top X - I_p)]\\
&+ 2\beta X\Phi(X^\top D) + \beta D(X^\top X - I_p).
\end{aligned}$$

3 Model Analysis

In this section, we explore the equivalence between problem (1.1) and PenC in some senses. We denote $\sigma_{\min}(A)$ as the smallest singular value of a given matrix A, and let

- $M_0 := \sup_{X\in\mathcal{M}} \max\{1, ||\nabla f(X)||_{\mathrm{F}}\}$;
- $M_1 := \sup_{X\in\mathcal{M}} \max\{1, \|\Lambda(X)\|_{\mathrm{F}}\}$;
- $C_1 := \sup_{X\in\mathcal{M}} \tilde{h}(X) - \inf_{X\in\mathcal{M}} \tilde{h}(X)$;
- $L_0 := \sup_{X,Y\in\mathcal{M}} \frac{\|\nabla f(X)-\nabla f(Y)\|_2}{\|X-Y\|_2}$
- $L_1 := \sup_{X,Y\in\mathcal{M}} \max\left\{1, \frac{\|\Lambda(X)-\Lambda(Y)\|_2}{\|X-Y\|_2}\right\}$;
- $L_2 := \sup_{X\in\mathcal{M},Y\in\mathbb{R}^{n\times p}} \limsup_{t\to 0} \frac{\left\|\nabla\tilde{h}(X+tY)-\nabla\tilde{h}(X)\right\|_{\mathrm{F}}}{t\|Y\|_{\mathrm{F}}}$;
- $M_2 := \sup_{X\in\mathcal{M}} \max\{1, \nabla\tilde{h}(X)\}$;
- $M_3 := \sup_{X\in\mathcal{M}} \max\{1, ||\nabla^2 f(X)||_{\mathrm{F}}\}$.

Here, $\tilde{h}(X) := f(X) - \frac{1}{2}\mathrm{tr}\left(\Lambda(X)(X^\top X - I_p)\right)$. The first four constants are defined under Assumption 1 and the last three ones are defined under Assumption 2. Please notice that the constants defined above are independent of β.

Lemma 1 *Suppose Assumption 1 holds, $\beta \geq \max\{2(M_0+M_1), 2pL_1\}$ and $\tilde{X}$ is a first order stationary point of PenC. Then the inequality $\left\|\tilde{X}\right\|_2^2 \leq 1 + \frac{2pL_1}{\beta} \leq 2$ holds.*

Proof Let $\tilde{X} = U\Sigma V^\top$ be the singular value decomposition of $\tilde{X}$, and $\sigma_1 \geq ... \geq \sigma_p$ are the diagonal entries of Σ, namely, the singular values of $\tilde{X}$. If $\sigma_1 \leq 1$, we have $\left\|\tilde{X}\right\|_2^2 \leq 1$ which concludes the proof. In the rest of the proof, we assume that $\sigma_1 > 1$. If $\sigma_p > 1$, we denote $X^+ = \tilde{X}$ and $X^- = \mathbf{0}_{n,p}$. Otherwise, there exists $1 \leq l < p$ satisfying $\sigma_l > 1 \geq \sigma_{l+1}$. Let $\Sigma_1 = \mathrm{Diag}(\sigma_1, ..., \sigma_l)$ and $\Sigma_2 = \mathrm{Diag}(\sigma_{l+1}, ..., \sigma_p)$, we denote

$$X^+ := U\begin{bmatrix}\Sigma_1 & \mathbf{0}_{l,p-l}\\ \mathbf{0}_{p-l,l} & \mathbf{0}_{p-l,p-l}\end{bmatrix}V^\top \quad \text{and} \quad X^- := U\begin{bmatrix}\mathbf{0}_{l,l} & \mathbf{0}_{l,p-l}\\ \mathbf{0}_{p-l,l} & \Sigma_2\end{bmatrix}V^\top,$$

respectively. By the definition of X^+, we can conclude that $\|X^+\|_2 = \left\|\tilde{X}\right\|_2$. Besides, since $\|X^-\|_2 \leq 1$, we have that $X^- \in \mathrm{conv}\,(\mathcal{S}_{n,p})$. As a result, for any $t \in \left[0, \frac{1}{2}\right]$, by the convexity of $\mathcal{M}$, $\tilde{X} - tX^+ = (1-t)\tilde{X} + tX^- \in \mathcal{M}$ and hence $-X^+ \in \mathrm{SFD}(\tilde{X})$. In addition, we denote $Q := V\begin{bmatrix}I_l & \mathbf{0}_{l,p-l}\\ \mathbf{0}_{p-l,l} & \mathbf{0}_{p-l,p-l}\end{bmatrix}V^\top$, then we can conclude that $X^+ = \tilde{X}Q$.

We now assume that the desired inequality does not hold. Namely, $\left\|\tilde{X}\right\|_2^2 > 1 + \frac{2pL_1}{\beta}$. By simple calculation, we obtain that

$$\mathcal{L}(\tilde{X}, \Lambda(\tilde{X})) - \mathcal{L}(\tilde{X} - tX^+, \Lambda(\tilde{X})) = t\left\langle X^+, \nabla f(\tilde{X}) - \tilde{X}\Lambda(\tilde{X}) + \beta\tilde{X}(\tilde{X}^\top\tilde{X} - I_p)\right\rangle \tag{3.1}$$

holds for any $t \in (0, 1]$. By using the facts that $\mathrm{tr}\,(AB) = \mathrm{tr}(BA)$ holds for any two square matrices A and B with same size, ${X^+}^\top\tilde{X} = {X^+}^\top X^+$, and $X^+ = X^+Q$, we have the following statements hold

$$\left\langle X^+, \nabla f(\tilde{X}) - \tilde{X}\Lambda(\tilde{X})\right\rangle = -\mathrm{tr}\left(\left({X^+}^\top X^+ - Q\right){X^+}^\top\nabla f(\tilde{X})\right),$$
$$\left\langle X^+, \beta\tilde{X}(\tilde{X}^\top\tilde{X} - I_p)\right\rangle = \beta\mathrm{tr}\left({X^+}^\top X^+\left({X^+}^\top X^+ - Q\right)\right).$$

Substituting the above equalities into (3.1), we obtain

$$\begin{aligned}&\mathcal{L}(\tilde{X}, \Lambda(\tilde{X})) - \mathcal{L}(\tilde{X} - tX^+, \Lambda(\tilde{X})) = t\mathrm{tr}\left(\left({X^+}^\top X^+ - Q\right)\left(\beta{X^+}^\top X^+ - {X^+}^\top\nabla f(\tilde{X})\right)\right)\\ \geq\ & t\mathrm{tr}\left(\left({X^+}^\top X^+ - Q\right)\left(\frac{\beta}{2}{X^+}^\top X^+\right)\right) \geq \frac{t\beta}{2}\left(\left\|\tilde{X}\right\|_2^2 - 1\right)\left\|\tilde{X}\right\|_2^2.\end{aligned} \tag{3.2}$$

Here the first inequality uses the fact that $\frac{\beta}{2}{X^+}^\top X^+ - {X^+}^\top\nabla f(\tilde{X}) \succeq \frac{\beta}{2}{X^+}^\top X^+ - \Lambda(\tilde{X}) - {X^-}^\top\nabla f(\tilde{X}) \succeq 0$, which is implied by the facts that $\beta > 2(M_0+M_1)$ and $||{X^-}^\top\nabla f(\tilde{X})||_2 \leq ||X^-||_2||\nabla f(\tilde{X})||_{\mathrm{F}} \leq ||\nabla f(\tilde{X})||_{\mathrm{F}}$. The second inequality uses the definition of X^+.

On the other hand, for any $t \in (0, \bar{t}]$ where $\bar{t} = \min\left\{1, (\sigma_1^2 - 1)\sigma_1^2 \Big/ ||X^{+\top} X^+||_{\mathrm{F}}^2\right\}$, it holds that

$$\begin{aligned}
&\left\|(\tilde{X} - tX^+)^\top(\tilde{X} - tX^+) - I_p\right\|_{\mathrm{F}}^2 = \left\|\tilde{X}^\top\tilde{X} - I_p + (t^2 - 2t)X^{+\top}X^+\right\|_{\mathrm{F}}^2 \\
&= \left\|\tilde{X}^\top\tilde{X} - I_p\right\|_{\mathrm{F}}^2 + 2(t^2 - 2t)\left\langle \tilde{X}^\top\tilde{X} - I_p, X^{+\top}X^+\right\rangle + (t^2 - 2t)^2||X^{+\top}X^+||_{\mathrm{F}}^2 \\
&\leq \left\|\tilde{X}^\top\tilde{X} - I_p\right\|_{\mathrm{F}}^2 - (2t - t^2)\left[2(\sigma_1^2 - 1)\sigma_1^2 - (2t - t^2)||X^{+\top}X^+||_{\mathrm{F}}^2\right] \\
&\leq \left\|\tilde{X}^\top\tilde{X} - I_p\right\|_{\mathrm{F}}^2 - 2(2t - t^2)\left[(\sigma_1^2 - 1)\sigma_1^2 - t||X^{+\top}X^+||_{\mathrm{F}}^2\right] \leq \left\|\tilde{X}^\top\tilde{X} - I_p\right\|_{\mathrm{F}}^2.
\end{aligned}$$

Combining the above inequality with the Lipschitz continuity of $\Lambda(X)$, we have

$$\begin{aligned}
&|\mathcal{L}(\tilde{X} - tX^+, \Lambda(\tilde{X})) - \mathcal{L}(\tilde{X} - tX^+, \Lambda(\tilde{X} - tX^+))| \\
&= \left|\frac{1}{2}\left\langle \Lambda(\tilde{X}), (\tilde{X} - tX^+)^\top(\tilde{X} - tX^+) - I_p\right\rangle - \frac{1}{2}\left\langle \Lambda(\tilde{X} - tX^+), (\tilde{X} - tX^+)^\top(\tilde{X} - tX^+) - I_p\right\rangle\right| \\
&\leq \frac{1}{2}\left\|\Lambda(\tilde{X}) - \Lambda(\tilde{X} - tX^+)\right\|_{\mathrm{F}}\left\|(\tilde{X} - tX^+)^\top(\tilde{X} - tX^+) - I_p\right\|_{\mathrm{F}} \qquad (3.3) \\
&\leq \frac{tL_1}{2}\left\|X^+\right\|_{\mathrm{F}}\left\|(\tilde{X} - tX^+)^\top(\tilde{X} - tX^+) - I_p\right\|_{\mathrm{F}} \leq \frac{t\sqrt{p}L_1}{2}\left\|X^+\right\|_2\left\|\tilde{X}^\top\tilde{X} - I_p\right\|_{\mathrm{F}} \\
&\leq \frac{tpL_1}{2}\max\left\{1, \left\|\tilde{X}\right\|_2^2 - 1\right\}\left\|\tilde{X}\right\|_2^2.
\end{aligned}$$

Combining (3.2) with (3.3), we immediately obtain

$$\begin{aligned}
&h(\tilde{X}) - h(\tilde{X} - tX^+) \\
&\geq \left(\mathcal{L}(\tilde{X}, \Lambda(\tilde{X})) - \mathcal{L}(\tilde{X} - tX^+, \Lambda(\tilde{X}))\right) - \left|\left(\mathcal{L}(\tilde{X} - tX^+, \Lambda(\tilde{X})) - \mathcal{L}(\tilde{X} - tX^+, \Lambda(\tilde{X} - tX^+))\right)\right| \\
&\geq \frac{t\beta}{2}\left(\left\|\tilde{X}\right\|_2^2 - 1\right)\left\|\tilde{X}\right\|_2^2 - \frac{tpL_1}{2}\max\left\{1, \left\|\tilde{X}\right\|_2^2 - 1\right\}\left\|\tilde{X}\right\|_2^2 \\
&\geq \frac{tpL_1}{2}\max\left\{1, \left\|\tilde{X}\right\|_2^2 - 1\right\}\left\|\tilde{X}\right\|_2^2
\end{aligned}$$

holds for any $t \in (0, \bar{t}]$. Here, The third inequality follows the fact that $\beta \geq 2pL_1$ and $\left\|\tilde{X}\right\|_2^2 > 1 + \frac{2pL_1}{\beta}$. Therefore, we can conclude that $\liminf_{t\to 0^+} \frac{h(\tilde{X} - tX^+) - h(\tilde{X})}{t} < 0$, which reveals the contradictory to the optimality condition. Hence, we complete the proof.

Lemma 1 shows that the first-order stationary points of PenC have a uniform bound once $\mathcal{M}$ and β are given.

Lemma 2 *Suppose Assumption 1 holds. Let $\tilde{X}$ be an first-order stationary point of* (1.8) *with $\beta \geq \max\{2(M_0 + M_1), 2pL_1\}$, then it holds that $\tilde{X} - \frac{1}{2}\tilde{X}(\tilde{X}^\top\tilde{X} - I_p) \in \mathrm{conv}\,(\mathcal{S}_{n,p})$, which implies $-\tilde{X}(\tilde{X}^\top\tilde{X} - I_p) \in \mathrm{SFD}(\tilde{X})$.*

Proof It follows from Lemma 1 that $||\tilde{X}||_2 \leq \sqrt{2}$. Let $\tilde{X} = U\Sigma V^\top$ be the SVD of $\tilde{X}$, namely, $U \in \mathbb{R}^{n\times p}$ and $V \in \mathbb{R}^{p\times p}$ are the orthogonal matrices and Σ is a diagonal matrix with singular values of $\tilde{X}$ on its diagonal. Let σ_i $(i = 1, ..., p)$ be the diagonal entries of Σ, it holds that $\sigma_i \in [0, \sqrt{2}]$. Denote $W = \tilde{X} - \frac{1}{2}\tilde{X}(\tilde{X}^\top\tilde{X} - I_p)$, and we have

$$W = U\left(\Sigma - \frac{1}{2}\Sigma(\Sigma^2 - I_p)\right)V^\top.$$

Let $\nu(t) := t - \frac{1}{2}t(t^2 - 1)$, then $|\nu(\sigma_i(X))|$ $(i = 1, ..., p)$ are the singular values of W. By easy calculation, we can obtain all the critical points of $\nu(t)$ in $[0, \sqrt{2}]$. They are $t = 0$, $t = 1$ and $t = \sqrt{2}$ with function values

$$\nu(0) = 0, \quad \nu(1) = 1, \quad \nu(\sqrt{2}) = \frac{\sqrt{2}}{2},$$

respectively. Hence, we can conclude that $|\nu(t)| \leq 1$ holds for any $t \in [0, \sqrt{2}]$. Namely, $\sigma_i(W) \leq 1$ holds for any $i = 1, ..., p$. Therefore, $\tilde{X} - \frac{1}{2}\tilde{X}(\tilde{X}^\top\tilde{X} - I_p) = W \in \mathrm{conv}\,(\mathcal{S}_{n,p}) \subset \mathcal{M}$. Together with the convexity of $\mathcal{M}$, we arrive at $-\frac{1}{2}\tilde{X}(\tilde{X}^\top\tilde{X} - I_p) = W - \tilde{X} \in \mathrm{SFD}(\tilde{X})$ and complete the proof.

Next we establish the relationships between the first-order or second-order stationary points of PenC and problem (1.1).

Theorem 1 *Suppose Assumption 1 holds. Let $\tilde{X}$ be an first-order stationary point of PenC with $\beta \geq \max\{2(M_0 + M_1), 2pL_1\}$, then either $\tilde{X}$ is a first-order stationary point of* (1.1)*, or $\sigma_{\min}(\tilde{X}^\top \tilde{X}) \leq \frac{2M_1+\sqrt{2}L_1}{2\beta}$.*

Proof We assume that the desired statement does not hold. Namely, $\tilde{X} \notin \mathcal{S}_{n,p}$ and $\tilde{X}^\top \tilde{X} \succ \frac{2M_1+\sqrt{2}L_1}{2\beta} I_p$, which implies $W := \tilde{X}(\tilde{X}^\top \tilde{X} - I_p) \neq 0$. By Lemmas 1 and 2, we have $\left\|\tilde{X}\right\|_2 \leq \sqrt{2}$ and $-W \in \mathrm{SFD}(\tilde{X})$, respectively. Let $Q = -\tilde{X}^\top \tilde{X} + I_p$, we then obtain $W = -\tilde{X}Q$. By simple calculation, it holds that for any $t \in (0, 1]$

$$\begin{aligned}
&\mathcal{L}(\tilde{X}, \Lambda(\tilde{X})) - \mathcal{L}(\tilde{X} - tW, \Lambda(\tilde{X})) = t\left\langle W, \nabla f(\tilde{X}) - \tilde{X}\Lambda(\tilde{X}) + \beta\tilde{X}(\tilde{X}^\top \tilde{X} - I_p)\right\rangle \\
&= t\left\langle W, \nabla f(\tilde{X}) - \tilde{X}\Lambda(\tilde{X}) + \beta\tilde{X}(\tilde{X}^\top \tilde{X} - I_p)\right\rangle = t\left\langle W, \beta\tilde{X}(\tilde{X}^\top \tilde{X} - I_p)\right\rangle + t\left\langle W, \nabla f(\tilde{X}) - \tilde{X}\Lambda(\tilde{X})\right\rangle \\
&= t\beta||W||_{\mathrm{F}}^2 - t\mathrm{tr}\left((\tilde{X}^\top \tilde{X} - I_p)^2\Lambda(\tilde{X})\right) = t\mathrm{tr}\left((\tilde{X}^\top \tilde{X} - I_p)^2(\beta\tilde{X}^\top \tilde{X} - \Lambda(\tilde{X}))\right). \qquad (3.4)
\end{aligned}$$

On the other hand, $||\tilde{X}||_2 \leq \sqrt{2}$ results in the fact that $I_p \succ I_p - 2t\tilde{X}^\top \tilde{X} + t^2\tilde{X}^\top \tilde{X}(\tilde{X}^\top \tilde{X} - I_p) \succ -I_p$ holds for any $t \in \left(0, \frac{1}{3}\right]$, which further implies

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

Together with the fact that $\|W\|_{\mathrm{F}}^2 = \mathrm{tr}\left(\tilde{X}^\top \tilde{X}(\tilde{X}^\top \tilde{X} - I_p)^2\right) \leq \left\|\tilde{X}\right\|_2^2 \left\|\tilde{X}^\top \tilde{X} - I_p\right\|_{\mathrm{F}}^2$ and the Lipschitz continuity of $\Lambda(X)$, we obtain

$$\begin{aligned}
&|\mathcal{L}(\tilde{X} - tW, \Lambda(\tilde{X})) - \mathcal{L}(\tilde{X} - tW, \Lambda(\tilde{X} - tW))| \\
&= \frac{1}{2}\left|\left\langle \Lambda(\tilde{X}) - \Lambda(\tilde{X} - tW), (\tilde{X} - tW)^\top(\tilde{X} - tW) - I_p\right\rangle\right| \leq \frac{tL_1}{2}\|W\|_{\mathrm{F}}\left\|(\tilde{X} - tW)^\top(\tilde{X} - tW) - I_p\right\|_{\mathrm{F}} \\
&\leq \frac{tL_1}{2}\left\|\tilde{X}\right\|_2\left\|\tilde{X}^\top \tilde{X} - I_p\right\|_{\mathrm{F}}^2 \leq \frac{\sqrt{2}tL_1}{2}\left\|\tilde{X}^\top \tilde{X} - I_p\right\|_{\mathrm{F}}^2. \qquad (3.5)
\end{aligned}$$

Combining (3.4) with (3.5), we have

$$\begin{aligned}
&h(\tilde{X}) - h(\tilde{X} - tW) \\
&= \mathcal{L}(\tilde{X}, \Lambda(\tilde{X})) - \mathcal{L}(\tilde{X} - tW, \Lambda(\tilde{X})) + \mathcal{L}(\tilde{X} - tW, \Lambda(\tilde{X})) - \mathcal{L}(\tilde{X} - tW, \Lambda(\tilde{X} - tW)) \\
&\geq \mathcal{L}(\tilde{X}, \Lambda(\tilde{X})) - \mathcal{L}(\tilde{X} - tW, \Lambda(\tilde{X})) - |\mathcal{L}(\tilde{X} - tW, \Lambda(\tilde{X})) - \mathcal{L}(\tilde{X} - tW, \Lambda(\tilde{X} - tW))| \\
&\geq t \cdot \mathrm{tr}\left((\tilde{X}^\top \tilde{X} - I_p)^2(\beta\tilde{X}^\top \tilde{X} - \Lambda(\tilde{X}))\right) - \frac{\sqrt{2}tL_1}{2}\left\|\tilde{X}^\top \tilde{X} - I_p\right\|_{\mathrm{F}}^2 \\
&= t \cdot \left((\tilde{X}^\top \tilde{X} - I_p)^2\left(\beta\tilde{X}^\top \tilde{X} - \Lambda(\tilde{X}) - \frac{\sqrt{2}L_1}{2}I_p\right)\right) > 0,
\end{aligned}$$

which contradicts to the first-order stationarity of $\tilde{X}$. Here, the last equality holds because the facts that $\tilde{X}^\top \tilde{X} \succ \frac{2M_1+\sqrt{2}L_1}{2\beta} I_p$, $\left\|\Lambda(\tilde{X}) + \frac{\sqrt{2}L_1}{2}I_p\right\|_2 < M_1 + \frac{\sqrt{2}L_1}{2}$, and $\tilde{X} \notin \mathcal{S}_{n,p}$. We complete the proof.

Lemma 3 *Suppose Assumption 1 holds. Let $0 < \delta \leq \frac{1}{3}$ and $\beta \geq \max\left\{2(M_0 + M_1), 2pL_1, \left(3M_1 + \frac{3\sqrt{2}}{2}L_1\right), \frac{2C_1}{\delta^2}\right\}$. For any $X \in \mathcal{M}$, it holds that*

$$\sup_{\left\|X^\top X - I_p\right\|_{\mathrm{F}} \leq \delta} h(X) < \inf_{\left\|X^\top X - I_p\right\|_{\mathrm{F}} \geq 2\delta} h(X). \qquad (3.6)$$

Moreover, any global minimizer X^* of PenC satisfies ${X^*}^\top X^* = I_p$ which further implies that it is a global minimizer of problem (1.1)*.*

Proof For any Y and Z satisfy $\left\|Y^\top Y - I_p\right\|_{\mathrm{F}} \leq \delta \leq \frac{1}{3}$ and $\left\|Z^\top Z - I_p\right\|_{\mathrm{F}} \geq 2\delta$, respectively, it holds that

$$\begin{aligned}
h(Y) &\leq \sup_{X \in \mathcal{M}}\left(f(X) - \frac{1}{2}\mathrm{tr}\left(\Lambda(X)(X^\top X - I_p)\right)\right) + \frac{\delta^2\beta}{4}, \\
h(Z) &\geq \inf_{X \in \mathcal{M}}\left(f(X) - \frac{1}{2}\mathrm{tr}\left(\Lambda(X)(X^\top X - I_p)\right)\right) + \delta^2\beta.
\end{aligned}$$

As a result, we have

$$h(Y) - h(Z) \leq C_1 - \frac{3\beta\delta^2}{4} \leq \min\left\{-\frac{C_1}{2}, -3L_1\delta^2\right\} < 0.$$

Hence, inequality (3.6) holds.

Let X^* be a global minimizer of PenC. Suppose X^* does not satisfy the orthogonal constraints, according to Theorem 1, we have $\sigma_{\min}(X^{*\top}X^*) \leq \frac{2M_1+\sqrt{2}L_1}{2\beta} \leq \frac{1}{3} \leq 1-2\delta$, which implies $\left\|X^\top X - I_p\right\|_{\mathrm{F}} \geq 2\delta$. Then, the global optimality of X^* contradicts to the inequality (3.6).

Hence $X^{*\top}X^* = I_p$ holds. For any X satisfying the feasibility, we have $f(X) = h(X)$. Thus X^* is a global minimizer of the original Problem (1.1).

The proof of Lemma 3 directly gives the following corollary.

Corollary 1 *Suppose Assumption 1 holds, $\delta \in \left(0, \frac{1}{3}\right]$, $\beta \geq \max\left\{2(M_0+M_1), 2pL_1, \left(3M_1 + \frac{3\sqrt{2}}{2}L_1\right), \frac{2C_1}{\delta^2}\right\}$, and $X^0 \in \mathcal{M}$ satisfying $\left\|X^{0\top}X^0 - I_p\right\|_{\mathrm{F}} \leq \delta$. Let $\tilde{X}$ be a first-order stationary point of PenC and $h(\tilde{X}) \leq h(X^0)$ holds, then we can conclude $\tilde{X}^\top\tilde{X} = I_p$ and $\tilde{X}$ is a first-order stationary point of* (1.1).

Although PenC may bring first-order stationary points which does not satisfy the orthogonality constraints, Lemma 3 guarantees that those stationary points can not be global minimizer of PenC. If we generate a sequence of $\{X_k\}$ from an initial point X_0 satisfying $\left\|X_0^\top X_0 - I_p\right\|_{\mathrm{F}} \leq \delta$, and the sequence $\{h(X_k)\}$ is monotonically decreasing, Corollary 1 tells us that any limit point of $\{X_k\}$ is a first-order stationary point of (1.1) if it is a first-order stationary point of PenC.

In the following, we try to investigate the equivalence between the second-order stationary points of Problem (1.1) and PenC under some mild assumptions.

Theorem 2 *Suppose Assumption 2 holds, and $\mathcal{M}$ is chosen as $\mathcal{B}_K := \{X \in \mathbb{R}^{n\times p} \mid \|X\|_{\mathrm{F}} \leq K, K > \sqrt{p}\}$ and $\beta \geq \max\left\{2(M_0+M_1), 2pL_1, 6M_1+3\sqrt{2}L_1, 2L_2+1, \frac{6L_2+12KM_2+1}{5}\right\}$. Then, any second-order stationary point $\tilde{X}$ of PenC satisfies $\tilde{X}^\top\tilde{X} = I_p$. Moreover, $\tilde{X}$ is a second-order stationary point of problem* (1.1).

Proof Let $\tilde{X}$ be a second-order stationary point of PenC. Suppose it satisfies the orthogonality constraints, i.e. $\tilde{X}^\top\tilde{X} = I_p$, recalling Definition 2 and Proposition 1, we can easily verify that $\tilde{X}$ is also a second-order stationary point of problem (1.1).

In the following, we only need to consider the case that $\tilde{X}^\top\tilde{X} \neq I_p$. By Theorem 1, it holds that $\sigma_{\min}(\tilde{X}^\top\tilde{X}) \leq \frac{1}{6}$. Let $\tilde{X} = U\Sigma V^\top$ be the singular value decomposition of $\tilde{X}$, namely, $U \in \mathbb{R}^{n\times p}$ and $V \in \mathbb{R}^{p\times p}$ 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 $\sigma_1 \leq \frac{1}{\sqrt{6}}$ which is the first entry of the diagonal matrix Σ.

First we consider the case that $\left\|\tilde{X}\right\|_{\mathrm{F}} < K$. We denote $D = -\sigma_1 u_1 v_1^\top$, where u_1 and v_1 are the first columns of U and V, respectively. It holds that

$$(\tilde{X}+tD)^\top(\tilde{X}+tD) = \tilde{X}^\top\tilde{X} + 2tD^\top\tilde{X} + t^2D^\top D = V^\top\Sigma^2V - 2t\sigma_1^2 v_1v_1^\top + t^2\sigma_1^2 v_1v_1^\top. \tag{3.7}$$

Clearly, $\tilde{X}+tD \in \mathcal{B}_K$ holds if $t \in [0,2]$ and $\tilde{X} \in \mathcal{B}_K$. Namely, we obtain $D \in \mathrm{SFD}(\tilde{X})$. Due to the first-order stationarity of $\tilde{X}$, it holds $\nabla h(\tilde{X}) = 0$ which implies $D^\top\nabla h(\tilde{X}) = 0$. We notice that $h(X) = \tilde{h}(X) + f_3(X)$, where f_3 is defined in Proposition 1. Thus, we obtain

$$\begin{aligned}
&h(\tilde{X}+tD) \leq h(\tilde{X}) + t\cdot\mathrm{tr}\left(D^\top\nabla h(\tilde{X})\right) + \frac{t^2}{2}L_2\|D\|_{\mathrm{F}}^2 + \frac{t^2}{2}\mathrm{tr}\left(D^\top\nabla^2 f_3(\tilde{X})[D]\right) + \mathcal{O}(t^3)\\
=&h(\tilde{X}) + \frac{t^2}{2}L_2\|D\|_{\mathrm{F}}^2 + \frac{t^2}{2}(3\beta\sigma_1^4 - \beta\sigma_1^2) + \mathcal{O}(t^3)\\
\leq&h(\tilde{X}) + \frac{t^2}{12}L_2 - \frac{\beta t^2}{24} + \mathcal{O}(t^3) \leq h(\tilde{X}) - \frac{t^2}{24} + \mathcal{O}(t^3),
\end{aligned} \tag{3.8}$$

where the first equality uses the relationship (2.9). Obviously, the inequality (3.8) contradicts to the second-order stationarity of $\tilde{X}$.

Hence $\left\|\tilde{X}\right\|_{\mathrm{F}} = K$, we let $D := [\hat{u}, 0, ..., 0]V^\top$ where $\hat{u}^\top\tilde{X} = 0$. Besides, for each t, we define $\tilde{X}^t := q^t(\tilde{X}+tD)$ and $q^t := \frac{K}{\left\|\tilde{X}+tD\right\|_{\mathrm{F}}}$. It can be directly verified that

$$\mathrm{tr}\left((\tilde{X}^t)^\top\tilde{X}^t\right) = (q^t)^2\cdot||\tilde{X}+tD||_{\mathrm{F}}^2 = K^2$$

which implies $\tilde{X}^t \in \mathcal{B}_K$ for any t, and both D and $-D$ belongs to $\mathrm{SFD}(\tilde{X})$. By the first-order optimality condition, we have both $\langle D, \nabla h(\tilde{X})\rangle \geq 0$ and $\langle -D, \nabla h(\tilde{X})\rangle \geq 0$, which further gives us that $\langle \nabla h(\tilde{X}), D\rangle = 0$.

In addition, it holds that $||D||_{\mathrm{F}}^2 = 1$ and

$$q^t = \frac{K}{\sqrt{||\tilde{X}||_{\mathrm{F}}^2 + t^2||D||_{\mathrm{F}}^2}} = \frac{K}{\sqrt{K^2+t^2}} \leq 1,$$
$$1-(q^t)^2 = 1 - \frac{K^2}{K^2+t^2} = \frac{t^2}{K^2+t^2} \leq t^2. \tag{3.9}$$

We first evaluate the reduction of the constraint violation.

$$\begin{aligned}\left\|(\tilde{X}^t)^\top \tilde{X}^t - I_p\right\|_{\mathrm{F}}^2 &= \left\|\left((\tilde{X}+tD)^\top(\tilde{X}+tD) - I_p\right) - \left(1-(q^t)^2\right)(\tilde{X}+tD)^\top(\tilde{X}+tD)\right\|_{\mathrm{F}}^2 \\ &= \left\|(\tilde{X}+tD)^\top(\tilde{X}+tD) - I_p\right\|_{\mathrm{F}}^2 \\ &\quad - \left(1-(q^t)^2\right)\mathrm{tr}\left(\left((\tilde{X}+tD)^\top(\tilde{X}+tD) - I_p\right)(\tilde{X}+tD)^\top(\tilde{X}+tD)\right) \\ &\quad + \left(1-(q^t)^2\right)^2\left\|(\tilde{X}+tD)^\top(\tilde{X}+tD)\right\|_{\mathrm{F}}^2.\end{aligned}$$

We notice that

$$\begin{aligned}&\mathrm{tr}\left((A-I_p)A\right) = \sum_{i=1}^p \lambda_i(A)(\lambda_i(A)-1) \\ &= \sum_{i=1}^p\left((\lambda_i(A)-1)^2 + \lambda_i(A) - 1\right) \geq \sum_{i=1}^p(\lambda_i(A)-1) = \mathrm{tr}(A) - p\end{aligned}$$

holds for any symmetric and positive semi-definite matrix $A \in \mathbb{R}^{p,p}$, $\lambda_i(A)$ $(i = 1, ..., p)$ are the eigenvalues of A. Hence, the second term of the right hand side of the above equality satisfies

$$\mathrm{tr}\left(\left((\tilde{X}+tD)^\top(\tilde{X}+tD) - I_p\right)(\tilde{X}+tD)^\top(\tilde{X}+tD)\right) \geq \mathrm{tr}\left((\tilde{X}+tD)^\top(\tilde{X}+tD)\right) - p \geq K^2 - p \geq 0.$$

Hence, we further obtain

$$\begin{aligned}\left\|(\tilde{X}^t)^\top\tilde{X}^t - I_p\right\|_{\mathrm{F}}^2 &\leq \left\|(\tilde{X}+tD)^\top(\tilde{X}+tD) - I_p\right\|_{\mathrm{F}}^2 + \mathcal{O}(t^4) \\ &= \left\|\tilde{X}^\top\tilde{X} + t^2D^\top D - I_p\right\|_{\mathrm{F}}^2 + \mathcal{O}(t^4) \\ &= \mathrm{tr}\left((\tilde{X}^\top\tilde{X} - I_p)^2 + 2t^2D^\top D(\tilde{X}^\top\tilde{X} - I_p) + t^4(D^\top D)^2\right) + \mathcal{O}(t^4) \\ &\leq \left\|\tilde{X}^\top\tilde{X} - I_p\right\|_{\mathrm{F}}^2 - \frac{5}{3}t^2\left\|D\right\|_{\mathrm{F}}^2 + \mathcal{O}(t^4).\end{aligned}$$

Here, the last inequality uses the fact that $\sigma_{\min}(\tilde{X}^\top\tilde{X}) \leq \frac{1}{6}$.

According to the Taylor's expansion and the definition of L_2, we have

$$\tilde{h}(\tilde{X}+tD) - \tilde{h}(\tilde{X}) \leq t\left\langle D, \nabla\tilde{h}(\tilde{X})\right\rangle + \frac{L_2}{2}t^2\left\|D\right\|_{\mathrm{F}}^2 + \mathcal{O}(t^3) \leq \frac{L_2}{2}t^2\left\|D\right\|_{\mathrm{F}}^2 + \mathcal{O}(t^3) \tag{3.10}$$

$$\begin{aligned}\tilde{h}(\tilde{X}^t) - \tilde{h}(\tilde{X}+tD) &\leq \left\langle (q^t-1)(\tilde{X}+tD), \nabla\tilde{h}(\tilde{X}+tD)\right\rangle + \frac{L_2}{2}t^2(1-q^t)^2\left\|D\right\|_{\mathrm{F}}^2 + \mathcal{O}(t^3) \\ &\leq t^2|\left\langle\tilde{X}, \nabla\tilde{h}(\tilde{X})\right\rangle| + \mathcal{O}(t^3) \leq t^2KM_2\left\|D\right\|_{\mathrm{F}}^2 + \mathcal{O}(t^3).\end{aligned} \tag{3.11}$$

Here the second inequality of (3.10) uses the fact that $D^\top\tilde{X} = 0$ which leads to

$$0 = \left\langle\nabla h(\tilde{X}), D\right\rangle = \left\langle\nabla\tilde{h}(\tilde{X}) + \beta\tilde{X}(\tilde{X}^\top\tilde{X} - I_p), D\right\rangle = \left\langle D, \nabla\tilde{h}(\tilde{X})\right\rangle.$$

Combining the inequalities (3.10) and (3.11), we arrive at

$$\tilde{h}(\tilde{X}^t) - \tilde{h}(\tilde{X}) \leq \frac{L_2 + 2KM_2}{2}t^2\left\|D\right\|_{\mathrm{F}}^2 + \mathcal{O}(t^3).$$

Together with the fact that $h(X) = \tilde{h}(X) + f_3(X)$, we obtain

$$\begin{aligned} h(\tilde{X}^t) - h(\tilde{X}) &\leq \frac{L_2 + 2KM_2}{2} t^2 \|D\|_{\mathrm{F}}^2 + (f_3(\tilde{X}^t) - f_3(\tilde{X})) + \mathcal{O}(t^3) \\ &= \frac{L_2 + 2KM_2}{2} t^2 \|D\|_{\mathrm{F}}^2 - \frac{5\beta}{12} t^2 \|D\|_{\mathrm{F}}^2 + \mathcal{O}(t^3) \leq -\frac{t^2}{12} + \mathcal{O}(t^3), \end{aligned}$$

which contradicts to the second-order stationarity of $\tilde{X}$. Hence $\tilde{X}$ must satisfy the orthogonal constraints and we complete the proof.

4 The First-Order Method

In this section, we consider to design a first-order method for solving PenC with $\mathcal{M}$ chosen as a ball with radius K in F-norm, i.e. $\mathcal{B}_K := \{X \in \mathbb{R}^{n\times p} | \, \|X\|_{\mathrm{F}} \leq K\}$, where $K > \sqrt{p}$.

4.1 Algorithm Framework

Without specific mentioning, we suppose Assumption 2 holds in this subsection. According to Proposition 1, we notice that the Hessian of $f(X)$ involves in computing a gradient of $h(X)$. Therefore, any gradient method for solving PenC is already a second-order method for solving the optimization problem with orthogonality constraints. Therefore, we consider to design an inexact gradient method for solving PenC. A natural idea is to use $\nabla_X \mathcal{L}(X_k, \Lambda)\Big|_{\Lambda=\Lambda(X_k)} = \nabla f(X_k) - X_k\Lambda(X_k) + \beta X_k(X_k^\top X_k - I_p)$ to approximate $\nabla h(X_k)$. We present the projected inexact gradient algorithm to solve PenC in the following, and for convenience, we call it a first-order method.

Algorithm 1 First-order method for solving PenC (PenCF)

Require: $f : \mathbb{R}^{n\times p} \mapsto \mathbb{R}$, $\beta > 0$;
1: Randomly choose X_0 on Stiefel manifold, set $k = 0$;
2: **while** not terminate **do**
3: Compute inexact gradient: $D_k := \nabla_X \mathcal{L}(X_k, \Lambda)\Big|_{\Lambda=\Lambda(X_k)}$;
4: $\tilde{X}_{k+1} = X_k - \eta_k D_k$;
5: **if** $\|X_{k+1}\|_{\mathrm{F}} > K$ **then**
6: $X_{k+1} = \frac{K}{\|\tilde{X}_{k+1}\|_{\mathrm{F}}} \tilde{X}_{k+1}$;
7: **else**
8: $X_{k+1} = \tilde{X}_{k+1}$;
9: **end if**
10: $k++$;
11: **end while**
12: Return X_k.

4.2 Global Convergence

The main steps of establishing the global convergence of PenCF are showing that $\left\|X_k^\top X_k - I_p\right\|_{\mathrm{F}} \leq \delta$ always holds for some $\delta > 0$ and $\left\|X_0^\top X_0 - I_p\right\|_{\mathrm{F}} \leq \delta$, which implies the projection step is waived. Then we demonstrate the strictly monotonic decreasing of $\{h(X_k)\}$, which results in the fact that any clustering point is a stationary point of PenC. Recalling the theorems in the previous session, we can show that any such clustering point must to be a stationary point of Problem (1.1).

Next we present the details of our theoretical analysis.

Lemma 4 *Suppose Assumption 1 holds, $\delta \in \left(0, \frac{1}{3}\right]$, $K \geq \sqrt{p + \delta\sqrt{p}}$, and $\beta \geq 3M_1$. Let $\{X_k\}$ be the iterate sequence generated by Algorithm 1, starting from any initial point X_0 satisfying $||X_0^\top X_0 - I_p||_{\mathrm{F}} \leq \delta$, and the stepsize $\eta_k \leq \bar{\eta}$, where $\bar{\eta} = \min\left\{\frac{\delta}{8KM_4}, \frac{\beta\delta^2}{9K^2M_4^2}\right\}$, $M_4 = M_0 + M_1K + \beta\delta K$. Then $\left\|X_k^\top X_k - I_p\right\|_{\mathrm{F}} \leq \delta$ and $\|X_k\|_{\mathrm{F}} \leq K$ holds for any $k = 1, ...$*

Proof Suppose the argument to be proved holds for X_k ($k = 0, 1, ...$), and in the following, we show it also holds for X_{k+1}. We use the intermediate variable $\tilde{X}_{k+1} := X_k - \eta_k D_k$ introduced in Algorithm 1, where $D_k = \nabla f(X_k) - X_k\Lambda(X_k) + \beta X_k(X_k^\top X_k - I_p)$.

We first estimate the upper-bound for D_k by using the fact that $\left\|X_k^\top X_k - I_p\right\|_{\mathrm{F}} \le \delta$,

$$\begin{aligned} \|D_k\|_{\mathrm{F}} &\le \|\nabla f(X_k)\|_{\mathrm{F}} + \|X_k\Lambda(X_k)\|_{\mathrm{F}} + \beta\left\|X_k(X_k^\top X_k - I_p)\right\|_{\mathrm{F}} \\ &\le \|\nabla f(X_k)\|_{\mathrm{F}} + \|X_k\|_{\mathrm{F}}\|\Lambda(X_k)\|_2 + \beta\sqrt{\mathrm{tr}\left(X_k^\top X_k(X_k^\top X_k - I_p)^2\right)} \\ &\le \|\nabla f(X_k)\|_{\mathrm{F}} + \|X_k\|_{\mathrm{F}}\|\Lambda(X_k)\|_2 + \beta\sqrt{\left\|X_k^\top X_k\right\|_2}\left\|X_k^\top X_k - I_p\right\|_{\mathrm{F}} \\ &\le M_0 + M_1K + \beta\delta K = M_4. \end{aligned} \tag{4.1}$$

Then, we have

$$\begin{aligned} &\left\|X_{k+1}^\top X_{k+1} - I_p\right\|_{\mathrm{F}} = \left\|X_k^\top X_k - I_p - 2\eta_k\Phi(X_k^\top D_k) + \eta_k^2 D_k^\top D_k\right\|_{\mathrm{F}} \\ =\ &\left\|\Phi\left(\left(X_k^\top X_k - I_p\right)\left(I_p + 2\eta_k(\Lambda(X_k) - \beta X_k^\top X_k)\right)\right) + \eta_k^2 D_k^\top D_k\right\|_{\mathrm{F}} \\ \le\ &\left\|\Phi\left(\left(X_k^\top X_k - I_p\right)\left(I_p + 2\eta_k(\Lambda(X_k) - \beta X_k^\top X_k)\right)\right)\right\|_{\mathrm{F}} + \eta_k^2\|D_k\|_{\mathrm{F}}^2 \\ \le\ &\left\|X_k^\top X_k - I_p\right\|_{\mathrm{F}} + \eta_k^2\|D_k\|_{\mathrm{F}}^2. \end{aligned} \tag{4.2}$$

Here the last inequality uses the fact that $\sigma_{\min}(X_k^\top X_k) \ge 1 - \delta \ge \frac{2}{3}$ which further implies $\beta X_k^\top X_k \succeq M_1 \cdot I_p \succeq \Lambda(X_k)$.

Next, we consider two different situations of $\left\|X_k^\top X_k - I_p\right\|$. First we assume $\left\|X_k^\top X_k - I_p\right\|_{\mathrm{F}} \le \frac{\delta}{2}$,

$$\begin{aligned} \left\|\tilde{X}_{k+1}^\top\tilde{X}_{k+1} - I_p\right\|_{\mathrm{F}} &\le \left\|(X_k - \eta_k D_k)^\top(X_k - \eta_k D_k) - I_p\right\|_{\mathrm{F}} \\ &\le \left\|X_k^\top X_k - I_p\right\|_{\mathrm{F}} + \eta_k^2||D||_{\mathrm{F}}^2 \le \frac{\delta}{2} + \eta_k^2 M_4^2 \le \frac{\delta}{2} + \frac{\delta^2}{64K^2} < \delta. \end{aligned}$$

Then we assume $\delta \ge \left\|X_k^\top X_k - I_p\right\|_{\mathrm{F}} > \frac{\delta}{2}$, which implies $\sigma_{\min}(X_k^\top X_k) \ge 1 - \delta \ge \frac{2}{3}$. Hence,

$$\begin{aligned} &\mathrm{tr}\left(\left(X_k(X_k^\top X_k - I_p)\right)^\top D_k\right) \\ =\ &\mathrm{tr}\left((X_k^\top X_k - I_p)X_k^\top\nabla f(X_k) - (X_k^\top X_k - I_p)X_k^\top X_k\Lambda(X_k) + (X_k^\top X_k - I_p)(\beta X_k^\top X_k)(X_k^\top X_k - I_p)\right) \\ =\ &\mathrm{tr}\left((X_k^\top X_k - I_p)^2(\beta X_k^\top X_k - \Lambda(X_k))\right) \ge \left(\frac{2\beta}{3} - M_1\right)||X_k^\top X_k - I_p||_{\mathrm{F}}^2 > \frac{\beta}{3}||X_k^\top X_k - I_p||_{\mathrm{F}}^2. \end{aligned} \tag{4.3}$$

Let $c(X) = ||X^\top X - I_p||_{\mathrm{F}}^2$. According to the Taylor expansion, we have

$$c(\tilde{X}_{k+1}) \le c(X_k) + \mathrm{tr}\left(\nabla c(X_k)^\top(\tilde{X}_{k+1} - X_k)\right) + \frac{C_2}{2}||\tilde{X}_{k+1} - X_k||_{\mathrm{F}}^2,$$

where $C_2 = \max_{X\in\mathcal{M}}||\nabla^2 c(X)||_2 \le 12K^2$. Therefore, we can obtain

$$c(\tilde{X}_{k+1}) \le c(X_k) - 2\eta_k\mathrm{tr}\left(\left(X_k(X_k^\top X_k - I_p)\right)^\top D_k\right) + 6K^2\eta_k^2||D_k||_{\mathrm{F}}^2.$$

Together with (4.3) and the fact that $||D_k|| \le M_4$, it holds that

$$\begin{aligned} &c(\tilde{X}_{k+1}) \le c(X_k) - 2\eta_k\mathrm{tr}\left(\left(X_k(X_k^\top X_k - I_p)\right)^\top D_k\right) + 6K^2\eta_k^2||D_k||_{\mathrm{F}}^2 \\ &\le c(X_k) - \frac{2\beta\delta^2\eta_k}{3} + 6K^2M_4^2\eta_k^2 \le c(X_k), \end{aligned}$$

which implies $\left\|\tilde{X}_{k+1}^\top\tilde{X}_{k+1} - I_p\right\|_{\mathrm{F}}^2 \le \left\|X_k^\top X_k - I_p\right\|_{\mathrm{F}}^2$.

On the other hand, $\left\|X_k^\top X_k - I_p\right\|_{\mathrm{F}} \le \delta$ implies

$$\sqrt{p} + \delta \ge ||\tilde{X}_{k+1}^\top\tilde{X}_{k+1}||_{\mathrm{F}} \ge \sqrt{\frac{(||\tilde{X}_{k+1}||_{\mathrm{F}}^2)^2}{p}}.$$

Then we have $||\tilde{X}_{k+1}||_{\mathrm{F}}^2 \le p + \sqrt{p}\delta \le K^2$, which implies that the projection step can be waived, namely, $X_{k+1} = \tilde{X}_{k+1}$. By mathematical induction, we have proved $\left\|X_k^\top X_k - I_p\right\|_{\mathrm{F}} \le \delta$ and $\|X_k\|_{\mathrm{F}} \le K$ hold for any $k = 0, 1, ...$, which completes the proof.

Lemma 5 *Suppose Assumption 1 holds, $||X^\top X - I_p||_{\mathrm{F}} \le \delta \le \frac{1}{3}$, and $\beta \ge 3M_1$. Then*

$$\|D(X)\|_{\mathrm{F}} \ge \frac{\sqrt{3}\beta}{6} \cdot \left\|X^\top X - I_p\right\|_{\mathrm{F}}, \tag{4.4}$$

where $D(X) := \nabla_X \mathcal{L}(X, \Lambda)\Big|_{\Lambda=\Lambda(X)}$.

Proof First, we present two linear algebra relationships. The first is the inequality $||A||_{\mathrm{F}} \ge \left\|\frac{A+A^\top}{2}\right\|_{\mathrm{F}}$ holds for any square matrix A, which is quite obvious and the proof is omitted. The second is the equality $||AB + BA||_{\mathrm{F}} = 2||AB||_{\mathrm{F}}$ holds for any symmetric matrices A and B, which results from the fact $||AB + BA||_{\mathrm{F}}^2 = 2||AB||_{\mathrm{F}}^2 + 2\mathrm{tr}(ABAB) = 2||AB||_{\mathrm{F}}^2 + 2\mathrm{tr}(A^{\frac{1}{2}}BA^{\frac{1}{2}}A^{\frac{1}{2}}BA^{\frac{1}{2}}) = 4||AB||_{\mathrm{F}}^2$.

It follows from the above facts that

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

where the last equality uses the fact that $\sigma_{\min}(X^\top X) \ge 1 - \delta \ge \frac{2}{3}$.

Together with the facts that $\left\|X^\top D(X)\right\|_{\mathrm{F}} \le \|X\|_2 \|D(X)\|_{\mathrm{F}}$ and $\sigma_{\max}(X^\top X) \le 1 + \delta \le \frac{4}{3}$, we have

$$\begin{aligned}
\|D(X)\|_{\mathrm{F}} &\ge \frac{1}{\|X\|_{\mathrm{F}}}\left\|X^\top D(X)\right\|_{\mathrm{F}} \ge \frac{\sqrt{3}}{2}\left\|X^\top D(X)\right\|_{\mathrm{F}} \\
&\ge \frac{\sqrt{3}\beta}{6}\left\|X^\top X - I_p\right\|_{\mathrm{F}}.
\end{aligned}$$

Lemma 6 *Suppose Assumption 1 holds, $\delta \in \left(0, \frac{1}{3}\right]$, $K \ge \sqrt{p + \delta\sqrt{p}}$, and $\beta \ge 3M_1$. Let $\{X_k\}$ be the iterate sequence generated by Algorithm 1, starting from any initial point X_0 satisfying $||X_0^\top X_0 - I_p||_{\mathrm{F}} \le \delta$, and the stepsize $\eta_k \in \left[\frac{1}{2}\bar{\eta}, \bar{\eta}\right]$, where $\eta = \min\left\{\frac{\delta}{8KM_4}, \frac{\beta\delta^2}{9K^2L_1M_4^2}, \frac{1}{45(L_0+M_1)+137\beta}\right\}$, $M_4 = M_0 + M_1K + \beta\delta K$. Then it holds that*

$$h(X_{k+1}) \le h(X_k) - \frac{\bar{\eta}}{5}\|D_k\|_{\mathrm{F}}^2 \tag{4.5}$$

for any $k = 0, 1, ...$

Proof Recalling Lemma 4, we have $X_{k+1} := X_k - \eta_k D_k$, where $D_k = \nabla f(X_k) - X_k\Lambda(X_k) + \beta X_k(X_k^\top X_k - I_p)$. For any $X \in \mathcal{B}_K$, and $Y \in \mathbb{R}^{n\times p}$, it holds that

$$\begin{aligned}
&\frac{\|[\nabla f(X) - X\Lambda(X)] - [\nabla f(X+tY) - (X+tY)\Lambda(X)]\|_{\mathrm{F}}}{t\|Y\|_{\mathrm{F}}} = \frac{\|[\nabla f(X) - \nabla f(X+tY)] - tY\Lambda(X)\|_{\mathrm{F}}}{t\|Y\|_{\mathrm{F}}} \\
&\le \frac{\|\nabla f(X) - \nabla f(X+tY)\|_{\mathrm{F}} + \|tY\Lambda(X)\|_{\mathrm{F}}}{t\|Y\|_{\mathrm{F}}} \le L_0 + M_1.
\end{aligned}$$

Here, the last inequality uses the fact that $\|Y\Lambda(X+tY)\|_{\mathrm{F}} \le \|\Lambda(X+tY)\|_2\|Y\|_{\mathrm{F}}$. As a result,

$$\begin{aligned}
&\left(f(X_k) - \frac{1}{2}\left\langle X_k^\top X_k - I_p, \Lambda(X_k)\right\rangle\right) - \left(f(X_{k+1}) - \frac{1}{2}\left\langle X_{k+1}^\top X_{k+1} - I_p, \Lambda(X_k)\right\rangle\right) \\
&\ge \eta_k\langle \nabla f(X_k) - X_k\Lambda(X_k), D_k\rangle - \frac{M_1 + L_0}{2}\eta_k^2\|D_k\|_{\mathrm{F}}^2.
\end{aligned} \tag{4.6}$$

Recalling the expression for $\nabla^2 f_3$ in proposition 2.8, we obtain that

$$\begin{aligned}
&\left|\left\langle D, \nabla^2 f_3(Y)[D]\right\rangle\right| \le 2\beta\left|\left\langle D, Y\Phi(Y^\top D)\right\rangle\right| + \beta\left|\left\langle D, D(Y^\top Y - I_p)\right\rangle\right| \\
&\le 2\beta\left|\mathrm{tr}\left(D^\top YY^\top D\right)\right| + \beta\left|\mathrm{tr}(D^\top D(Y^\top Y - I_p))\right| \le 2\beta||Y||_2^2||D||_{\mathrm{F}}^2 + \beta||Y^\top Y - I_p||_2||D||_{\mathrm{F}}^2 \\
&\le 2(1+\delta)\beta||D||_{\mathrm{F}}^2 + \delta\beta||D||_{\mathrm{F}}^2 \le 3\beta\|D\|_{\mathrm{F}}^2
\end{aligned}$$

holds for any Y satisfying $\left\|Y^\top Y - I_p\right\| \le \delta$, and the last inequality uses the fact taht $\delta \le \frac{1}{3}$.

Together with (4.6) and the differential mean value theorem, there exists $t \in [0,1]$ such that

$$\begin{aligned}
&\mathcal{L}(X_k, \Lambda(X_k)) - \mathcal{L}(X_{k+1}, \Lambda(X_k)) \\
&\geq \eta_k \cdot \mathrm{tr}\left(D_k^\top \nabla_X \mathcal{L}(X_k, \Lambda(X_k))\right) - \frac{L_0 + M_1}{2}\eta_k^2 \|D_k\|_{\mathrm{F}}^2 - \frac{\eta_k^2}{2}\left\langle D_k, \nabla^2 f_3(tX_{k+1} + (1-t)X_k)[D_k]\right\rangle \\
&\geq \eta_k \cdot \mathrm{tr}\left(D_k^\top D_k\right) - \frac{L_0 + M_1 + 3\beta}{2}\eta_k^2 \|D_k\|_{\mathrm{F}}^2.
\end{aligned}$$

Here the last inequality holds resulting from the Lemma 4 and the relationship $tX_{k+1} + (1-t)X_k = X_k - t\eta_k D_k$. Following the deduction in Theorem 1, we can conclude that

$$\begin{aligned}
&h(X_k) - h(X_{k+1}) = \mathcal{L}(X_k, \Lambda(X_k)) - \mathcal{L}(X_{k+1}, \Lambda(X_k)) + \mathcal{L}(X_{k+1}, \Lambda(X_k)) - \mathcal{L}(X_{k+1}, \Lambda(X_{k+1})) \\
&\geq \eta_k \cdot \mathrm{tr}\left(D_k^\top D_k\right) - \frac{L_0 + M_1 + 3\beta}{2}\eta^2 \|D_k\|_{\mathrm{F}}^2 - \eta_k L_1 ||D_k||_{\mathrm{F}} ||X_{k+1}^\top X_{k+1} - I_p||_{\mathrm{F}} - \eta_k^3 L_1 \|D_k\|_{\mathrm{F}}^3 \\
&\geq \eta_k ||D_k||_{\mathrm{F}}^2 - \frac{L_0 + M_1 + 3\beta}{2}\eta_k^2 \|D_k\|_{\mathrm{F}}^2 - \eta_k L_1 \cdot \frac{2\sqrt{3}}{\beta} ||D_k||_{\mathrm{F}}^2 - \eta_k^3 L_1 \|D_k\|_{\mathrm{F}}^3 \\
&\geq \eta_k \|D_k\|_{\mathrm{F}}^2 - \frac{L_0 + M_1 + 3\beta}{2}\eta_k^2 \|D_k\|_{\mathrm{F}}^2 - \frac{\sqrt{3}}{3}\eta_k \|D_k\|_{\mathrm{F}}^2 - \frac{\beta}{50}\eta_k^2 \|D_k\|_{\mathrm{F}}^2 \\
&> \frac{3-\sqrt{3}}{6}\bar{\eta} \cdot ||D_k||_{\mathrm{F}}^2 - \frac{L_0 + M_1 + 4.96 \cdot \beta}{2}\bar{\eta}^2 \|D_k\|_{\mathrm{F}}^2 \\
&> \frac{1}{5}\bar{\eta} ||D_k||_{\mathrm{F}}^2 + \left(\frac{1}{90}\bar{\eta} - \frac{L_0 + M_1 + 3.03 \cdot \beta}{2}\bar{\eta}^2\right)\|D_k\|_{\mathrm{F}}^2 \geq \frac{\bar{\eta}}{5}\|D_k\|_{\mathrm{F}}^2
\end{aligned}$$

holds for any $k = 0, 1, ...$ Here, the second inequality uses (4.2) and Lemma 5. Besides, the third inequality uses the estimation in (4.1), which illustrates that

$$\eta_k L_1 \|D_k\|_{\mathrm{F}} \leq \frac{\delta^2\beta}{9K^2 L_1 M_4^2} \cdot L_1 \cdot M_4 \leq \frac{\delta^2}{9K^2 M_4}\beta \leq \frac{\delta^2}{9K^2}\beta < \frac{3}{200}\beta.$$

The fifth inequality follows the fact that $\frac{3-\sqrt{3}}{6} \geq \frac{19}{90}$. And the last inequality holds with the fact that $\bar{\eta} \leq \bar{\eta} \leq \frac{1}{45(L_0+M_1)+137\beta}$.

Theorem 3 *Suppose Assumption 1 holds, $\delta \in \left(0, \frac{1}{3}\right]$, $K \geq \sqrt{p + \delta\sqrt{p}}$, and $\beta \geq \max\{2(M_0+M_1), 2pL_1, 3M_1 + \frac{3\sqrt{2}}{2}L_1\}$. Let $\{X_k\}$ be the iterate sequence generated by Algorithm 1 initiated from $X_0 \in \mathcal{M}$ satisfying $||X_0^\top X_0 - I_p||_{\mathrm{F}} \leq \delta$, and the stepsize $\eta_k \in \left[\frac{1}{2}\bar{\eta}, \bar{\eta}\right]$, where $\bar{\eta} = \min\left\{\frac{\delta}{8KM_4}, \frac{\beta\delta^2}{9K^2L_1M_4^2}, \frac{1}{45(L_0+M_1)+137\beta}\right\}$, $M_4 = M_0 + M_1K + \beta\delta K$. Then, the iterate sequence $\{X^k\}$ has at least one cluster point, and each cluster point of $\{X^k\}$ is a stationary point of problem* (1.1). *More precisely, for any $k \geq 1$, it holds that*

$$\sum_{0 \leq i \leq N-1} \|D_i\|_{\mathrm{F}}^2 \leq \frac{20C_1 + 5\beta\delta^2}{4N\bar{\eta}}.$$

Proof Using Lemma 4, we first obtain that $\left\|X_k^\top X_k - I_p\right\|_{\mathrm{F}} \leq \delta$ and $\|X_k\| \leq K$ hold. Therefore, $\{X_k\}$ exists cluster point.

By Lemma 6, it holds that

$$h(X_k) - h(X_{k+1}) \geq \frac{\bar{\eta}}{5}\|D_k\|_{\mathrm{F}}^2.$$

If X^* is a cluster point of $\{X_k\}$, we have $\nabla_X \mathcal{L}(X^*, \Lambda)\Big|_{\Lambda=\Lambda(X^*)} = 0$. Together with ${X^*}^\top X^* = I_p$ implied by Corollary 1, we can conclude that X^* is a first-order stationary point of problem (1.1).

Calculating the summation of the above inequalities from $k = 0$ to $N-1$, we have

$$\begin{aligned}
&\sum_{k=0}^{N-1}\frac{\bar{\eta}}{5}\|D_i\|_{\mathrm{F}}^2 \leq h(X_0) - h(X_N) < h(X_0) - \inf_{||X^\top X - I_p||_{\mathrm{F}} \leq \delta} h(X) \\
&< \sup_{X \in \mathcal{M}} \tilde{h}(X) - \inf_{X \in \mathcal{M}} \tilde{h}X + \frac{\beta}{4}\left(||X_0^\top X_0 - I_p||_{\mathrm{F}}^2 - ||X_N^\top X_N - I_p||_{\mathrm{F}}^2\right) \leq C_1 + \frac{\beta\delta^2}{4},
\end{aligned}$$

which completes the proof.

Corollary 2 *Suppose all the assumptions of Theorem 3 hold. It holds that*

$$\min_{0\le i\le N-1} \max\left\{||X_i^\top X_i - I_p||_{\mathrm{F}}, \|D_i\|_{\mathrm{F}}\right\} \le \max\left\{\frac{2\sqrt{3}}{3M_1}, 1\right\}\cdot\sqrt{\frac{5C_1+\frac{5}{4}\beta\delta^2}{N\bar{\eta}}}.$$

Proof This is a direct corollary of Lemma 5 and Theorem 3.

Remark 4 *The sublinear convergence rate of Corollary 2 actually tells us that PenCF terminates after $O(1/\epsilon^2)$ iterations, if the stopping criterion is set as* $\max\left\{||X_k{}^\top X_k - I_p||_{\mathrm{F}}, \|D_k\|_{\mathrm{F}}\right\} < \epsilon$.

4.3 Local Convergence

In this section, we deliver the local convergence rate of the PenCF. According to the Lemma 4, when the iterate is close enough to the Stiefel manifold, the projection process will never be invoked again. In fact, PenCF reduces to PLAM proposed in [15]. Therefore, the local convergence rate of PenCF can be established in the same manner.

Theorem 4 *Suppose Assumption 2 holds. Let X^* be an isolated local minimizer of* (1.1)*, namely,*

$$\tau := \inf_{Y^\top X^* + X^{*\top} Y = 0} \frac{\nabla^2 f(X^*)[Y,Y] - \mathrm{tr}\left(Y^\top Y \Lambda(X^*)\right)}{\|Y\|_{\mathrm{F}}^2} > 0. \tag{4.7}$$

Assume that the parameters $\beta \ge \frac{1}{2}M_3 + \frac{\sqrt{3}M_0}{3} + \frac{1}{2}\tau$ and $\eta_k \in [\frac{\bar{\eta}}{2}, \bar{\eta}]$, where $\bar{\eta} \ge M_3 + \frac{2\sqrt{3}M_0}{3} + 2\beta$. Then, there exists $\varepsilon > 0$ such that starting from any X_0 satisfying $||X_0 - X^*||_{\mathrm{F}} < \varepsilon$, and the iterate sequence $\{X_k\}$ generated by Algorithm 1 converges to X^* Q-linearly.

The proof of Theorem 4 can follow the same steps of proving Theorem 4.12 of [15].

Remark 5 *PLAM dose not have any constraint for the generated sequence, which can leads to an unbounded sequence. This fact further explains the fact that PLAM requires large β to guarantee its convergence in practical, and its convergence rate is sensitive to β [15]. Moreover, the sequence $\{X^k\}$ generated by PCAL is restricted on the Oblique manifold, which is an compact but nonconvex subset in $\mathbb{R}^{n\times p}$. As a result, PCAL computes the retraction to Oblique manifold in each iteration. As illustrated in Algorithm 1, the sequence is restricted in a convex compact set $\mathcal{M}$ in PenCF. As illustrated in Algorithm 1, the sequence is restricted in a convex compact set $\mathcal{B}_K$ in PenCF. The generated sequence by PenCF is restricted in $\mathcal{B}_K$, implying that PenCFis more stable than PLAM. Moreover, $\mathcal{S}_{n,p}$ lies in the interior of $\mathcal{B}_K$. As a result, when $\nabla h(X_k)$ the constraint in PenC is inactive and PenCF avoids the projection to $\mathcal{B}_K$.*

5 The Second-Order Method

In this section, we consider to design a second-order method for solving PenC with $\mathcal{M}$ chosen as a ball with radius K in F-norm, i.e. $\mathcal{B}_K := \{X \in \mathbb{R}^{n\times p} |\ \|X\|_{\mathrm{F}} \le K\}$, where $K > \sqrt{p}$.

5.1 Algorithm Framework

Without specific mentioning, we suppose Assumption 2 holds in this subsection. According to Proposition 1, we notice that the third-order derivative of $f(X)$ involves in computing the Hessian of $h(X)$. Therefore, we consider the approximate $\nabla^2 f(X)$ a matrix $W(X)$ defined by

$$W(X)[D] = \nabla^2 f(X)[D] - D\Lambda(X) \tag{5.1}$$

$$-X\left[\Phi(D^\top \nabla f(X)) + \Phi(X^\top \nabla^2 f(X)[D]) + 2\beta\Phi(X^\top D)\right] \tag{5.2}$$

$$-\left[\nabla f(X)\Phi(D^\top X) + \nabla^2 f(X)[X\Phi(X^\top D)]\right]. \tag{5.3}$$

Clearly, we have $W(X^*) = \nabla^2 h(X^*)$.

An inexact Newton direction D can be calculated by solving the following

$$\min\ \ \mathrm{tr}\left(D^\top \nabla h(X_k) + \frac{1}{2}D^\top W(X_k)D\right) \tag{5.4}$$

$$\text{s.t.}\ \ \|X_k + D\|_{\mathrm{F}} \le K. \tag{5.5}$$

We present the inexact projected Newton method to solve PenC as the following. For convenience, we call Algorithm 2 a second-order method.

Algorithm 2 Second-order method for solving PenC (PenCS)

Require: $f : \mathbb{R}^{n\times p} \mapsto \mathbb{R}$, $\beta > 0$;
Choose an initial point X_0 close enough to a local minimizer, set $k = 0$;
while not terminate **do**
  Compute the inexact Newton direction D_k by solving (5.4);
  Select a stepsize η_k;
  $\tilde{X}_{k+1} = X_k - \eta_k D_k$;
  if $\left\|\tilde{X}_{k+1}\right\|_{\mathrm{F}} > K$ **then**
    $X_{k+1} = \frac{K}{\left\|\tilde{X}_{k+1}\right\|}\tilde{X}_{k+1}$;
  else
    $X_{k+1} = \tilde{X}_{k+1}$;
  end if
  $k++$;
end while
Return X_k.

5.2 Local Convergence

In this subsection, we establish the local quadratic convergence of PenCS.

Theorem 5 *Suppose Assumption 2 holds. Let X^* be an isolated local minimizer of* (1.1) *with τ defined by* (4.7). *Assume that the parameters $K \geq \sqrt{p}+1$, $\beta \geq \max\left\{\frac{2}{3}L_2, \frac{4L_2^2}{\tau}+\tau\right\}$. Then, there exists a sufficiently small $\varepsilon \in (0,1)$ such that the iterate sequence $\{X_k\}$ generated by Algorithm 2 initiated from X_0 satisfying $\|X_0 - X^*\|_{\mathrm{F}} \leq \varepsilon$, and with unit stepsize $\eta_k = 1$ converges to X^* quadratically.*

Proof The orthogonality of X^* directly results in the relationship $W(X^*) = \nabla^2 h(X^*)$. We first prove the positive definiteness of $W(X^*)$. For any $D \in \mathbb{R}^{n\times p}$, we decompose it into two parts $D = D_1 + D_2$, where $D_1 = D - X^*\Phi(D^\top X^*)$ in the null space of X^*, namely, $D_1^\top X^* = 0$, and $D_2 = X^*\Phi(D^\top X^*)$ in the range space of X^*. Then we can obtain

$$\begin{aligned}
&\langle W(X^*)[D_1], D_1\rangle = \nabla^2 h(X^*)[D_1, D_1] = \left\langle D_1, \nabla^2 f(X^*)[D_1] - D_1\Lambda(X^*)\right\rangle \\
&- \left\langle D_1, X^*\left[\Phi(D_1^\top \nabla f(X^*)) + \Phi(X^{*\top}\nabla^2 f(X^*)[D_1]) + 2\beta\Phi(X^{*\top}D_1)\right]\right\rangle \\
&- \left\langle D_1, \left[\nabla f(X)\Phi(D_1^\top X^*) + \nabla^2 f(X^*)[X^*\Phi(X^{*\top}D_1)]\right]\right\rangle \\
&= \left\langle D_1, \nabla^2 f(X^*)[D_1] - D_1\Lambda(X^*)\right\rangle = \nabla^2 f(X^*)[D_1, D_1] - \mathrm{tr}\left(D_1^\top D_1\Lambda(X^*)\right) \geq \tau||D_1||_{\mathrm{F}}^2. \quad (5.6)
\end{aligned}$$

On the other hand, we also have

$$\begin{aligned}
\langle W(X^*)[D_1], D_2\rangle &= \nabla^2 h(X^*)[D_1, D_2] = \nabla^2\tilde{h}(X^*)[D_1, D_2] \geq -L_2\left\|D_1\right\|_{\mathrm{F}}\left\|D_2\right\|_{\mathrm{F}}; \quad (5.7)\\
\langle W(X^*)[D_2], D_2\rangle &= \nabla^2 h(X^*)[D_2, D_2] = \nabla^2\tilde{h}(X^*)[D_2, D_2] + 2\beta\left\|D_2\right\|_{\mathrm{F}}^2 \\
&\geq -L_2\left\|D_2\right\|_{\mathrm{F}}^2 + 2\beta\left\|D_2\right\|_{\mathrm{F}}^2 \geq \frac{\beta}{2}\left\|D_2\right\|_{\mathrm{F}}^2. \quad (5.8)
\end{aligned}$$

Here, the last inequality uses the fact that $\beta \geq \frac{2L_2}{3}$.

Combining (5.6)-(5.8) and the inequality $\beta \geq \frac{4L_2^2}{\tau} + \tau$, we can further obtain

$$\begin{aligned}
\nabla^2 h(X^*)[D, D] = \nabla^2 h(X^*)[D_1 + D_2, D_1 + D_2] &\geq \frac{\tau}{2}\left[\|D_1\|_{\mathrm{F}}^2 + \|D_2\|_{\mathrm{F}}^2\right] \\
&+ \left[\frac{\tau}{2}\left\|D_1\right\|_{\mathrm{F}}^2 + \frac{\beta-\tau}{2}\left\|D_2\right\|_{\mathrm{F}}^2 - 2L_2\left\|D_1\right\|_{\mathrm{F}}\left\|D_2\right\|_{\mathrm{F}}\right] \geq \frac{\tau}{2}\left[\|D_1\|_{\mathrm{F}}^2 + \|D_2\|_{\mathrm{F}}^2\right],
\end{aligned}$$

which implies the positive definiteness of both $\nabla^2 h(X^*)$ and $W(X^*)$. Recalling the Gershgorin Circle Theorem, there exists $\varepsilon_1 > 0$ such that

$$\lambda_{\min}(W(X)) \geq \frac{\tau}{4} \quad (5.9)$$

holds for any $||X - X^*||_{\mathrm{F}} \leq \varepsilon_1$.

Assumption 2 implies the Lipschitz continuity of $W(X)$ in $\mathcal{B}_K$. Hence, we can denote

$$L_W := \sup_{X,Y\in\mathcal{B}_K} \frac{\|W(X)-W(Y)\|_{\mathrm{F}}}{\|X-Y\|_{\mathrm{F}}}.$$

On the other hand, we define

$$g_Y(X) := \nabla f(X) - X\Lambda(X) - \frac{1}{2}\nabla f(Y)(X^\top X - I_p) - \frac{1}{2}\nabla^2 f(Y)[Y(X^\top X - I_p)] + \beta X(X^\top X - I_p).$$

It can be verified that $g_X(X) = \nabla h(X)$, $\nabla g_X(X)[D] = W(X)[D]$ holds for any $X \in \mathbb{R}^{n\times p}$, and

$$\nabla g_X(X^*) = \nabla f(X^*) - X^*\Lambda(X^*) = \nabla h(X^*).$$

By using the Lipschitz continuity of $W(X)$, the mean value theorem, and the convexity of $\mathcal{B}_K$, there exists $\varepsilon_2 > 0$ such that it holds that

$$\begin{aligned}
&\|\nabla h(X) - \nabla h(X^*) - W(X)[X-X^*]\|_{\mathrm{F}} = \|\nabla h(X^*) - \nabla h(X) - W(X)[X^*-X]\|_{\mathrm{F}} \\
&= \|g_X(X^*) - g_X(X) - \nabla g_X(X)[X^*-X]\|_{\mathrm{F}} \\
&= \|\nabla g_X(\xi X^* + (1-\xi)X)[X^*-X] - \nabla g_X(X)[X^*-X]\|_{\mathrm{F}} \le L_W \|X-X^*\|_{\mathrm{F}}^2
\end{aligned} \tag{5.10}$$

for any X satisfying $\|X-X^*\|_{\mathrm{F}} \le \varepsilon_2$, where $\xi \in [0,1]$.

By the assumption on K, we know that $X_0 \in \mathcal{M}$. Denote $\varepsilon := \min\{1, \varepsilon_1, \varepsilon_2, \frac{\tau}{5L_W}\}$ Now, we assume that $X_k \in \mathcal{M}$ and $\|X_k - X^*\|_{\mathrm{F}} \le \varepsilon$ hold. Combining the inequality (5.9) and (5.10), we arrive at

$$\begin{aligned}
&\|X_{k+1}-X^*\|_{\mathrm{F}} = \left\|X_k - X^* - W(X_k)^{-1}[\nabla h(X_k)]\right\|_{\mathrm{F}} \\
&= \left\|W(X_k)^{-1}\left[W(X_k)[X_k-X^*] - \nabla h(X_k) + \nabla h(X^*)\right]\right\|_{\mathrm{F}} \\
&\le \left\|W(X_k)^{-1}\right\|_{\mathrm{F}} \|W(X_k)[X_k-X^*] - \nabla h(X_k) + \nabla h(X^*)\|_{\mathrm{F}} \\
&\le \frac{4L_W}{\tau}\|X_k-X^*\|_{\mathrm{F}}^2 \le \frac{4}{5}\|X_k-X^*\|_{\mathrm{F}} \le \varepsilon.
\end{aligned}$$

Therefore, we have $\|X_{k+1}-X^*\|_{\mathrm{F}} \le \varepsilon$ and hence $X_{k+1} \in \mathcal{M}$. By mathematical induction, we conclude that $\{X_k\}$ converges quadratically to X^*, which completes the proof.

5.3 Computational Cost

In this section, we compare the computational costs among ARNT, RTR and PenCS. We analyze the computational cost of the basic linear algebra operations of solving the correlated subproblem in ARNT, RTR and PenCS by conjugate gradient method. The comparison is listed in Table 1.

As mentioned in section 1, most of Newton methods for optimization problems on Stiefel manifold involves constructing a quadratic approximation $m_k(X)$ to $f(X)$ at X_k, and minimizing the quadratic function. For ARNT, the corresponding subproblem (1.5) is an quadratic optimization problem on $\mathcal{S}_{n,p}$, which is nonconvex and it is only possible to compute its first-order stationary point. As illustrated in [19], (1.5) can be approximately solved by minimizing m_k in the tangent subspace, which is solved by conjugate gradient in their code[3]. For RTR, the corresponding subproblem is illustrated in (1.4). Due to the nonconvexity of f and $\mathcal{S}_{n,p}$, subproblem (1.4) can be nonconvex and RTR uses the truncated conjugate gradient method to compute an approximated solution[4] [40, 37, 48]. And the subproblem for PenCS is minimizing (5.4) in $\mathbb{R}^n$, which is a trust-region subproblem in $\mathbb{R}^{n\times p}$. Since both ARNT and RTR use conjugate gradient as the solver for their subproblems, in PenCSwe compute an approximated solution of (5.4) by truncated gradient method. And in this section we only compare the computational costs of solving subproblems by conjugate gradient method.

In addition, since ARNT and RTR are retraction-based algorithm, orthonormalization process is involved to project the approximated solution for their subproblem to Stiefel manifold. The computational costs for Gram-Schmidt orthonormalization process is $2np^2$ [15]. Since this process lacks scalability and is hard for column-wise parallelism, we add its computational costs to the total costs of each algorithm in Table 1.

[3] The code can be downloaded from https://github.com/wenstone/ARNT

[4] The code can be downloaded from https://www.manopt.org/

The result is presented in Table 1. In Table 1, the iterations that conjugate gradient method takes in subproblems for ART, RTR and PenCS is denoted as Iter_A, Iter_R and Iter_P, respectively. Besides, the hessian-matrix multiplication $\nabla^2 f(X_k)[D]$ needs to be computed in each iteration in conjugate gradient method. The computational costs for $\nabla^2 f(X_k)[D]$ depends on f and varies from case to case. As a result, in Table 1 we denote its computational complexity as Hmm, an abbreviation for Hessian-matrix multiplication.

ARNT		
Riemannian Hessian of $m_k(D)$	$D\Phi(X_k^\top \nabla f(X_k))$	$2np^2$
	$\nabla^2 f(X_k)[D] - D\Phi(X_k^\top \nabla f(X_k))$	1 Hessian-matrix production
Projection to tangent space	$D \to D - X_k\Phi(D^\top X_k)$	$2np^2$
Total	$\left[1 \cdot \text{Hmm} + 4np^2\right] \cdot \text{Iter}_A + 2np^2$	
RTR		
Riemannian Hessian of $m_k(D)$	$D\Phi(X_k^\top \nabla f(X_k))$	$2np^2$
	$\nabla^2 f(X_k)[D] - D\Phi(X_k^\top \nabla f(X_k))$	1 Hessian-matrix production
Projection to tangent space	$D \to D - X_k\Phi(D^\top X_k)$	$2np^2$
total	$\left[1 \cdot \text{Hmm} + 4np^2\right] \cdot \text{Iter}_R + 2np^2$	
PenCS		
Projection	None	0
Inherent from outer iteration	$X_k^\top X_k$	0
	$\Lambda(X_k)$	0
Basic calculation	$\Phi(X_k^\top D)$	$2np^2$
	$X_k\Phi(X_k^\top D)$	$2np^2$
Hessian-matrix production	$\nabla^2 f(X_k)[X_k\Phi(X_k^\top D)]$	1 Hessian-matrix production
	$\nabla^2 f(X_k)[D]$	1 Hessian-matrix production
(5.1)	$\nabla^2 f(X)[D] - D\Lambda(X)$	$2np^2$
(5.2)	$\Phi(D^\top \nabla f(X))$	$2np^2$
	$\Phi(X^\top \nabla^2 f(X)[D])$	$2np^2$
	$\Phi(X^\top \nabla^2 f(X_k)[X_k\Phi(X_k^\top D)])$	$2np^2$
	$-X_k\left[\Phi(D^\top \nabla f(X_k)) + \Phi(X_k^\top \nabla^2 f(X_k)[D]) + 2\beta\Phi(X_k^\top D)\right]$	$2np^2$
(5.3)	$-\left[\nabla f(X_k)\Phi(D^\top X_k) + \nabla^2 f(X_k)[X_k\Phi(X_k^\top D)]\right]$	$2np^2$
Total	$\left[2 \cdot \text{Hmm} + 16np^2\right] \cdot \text{Iter}_P$	

Table 1: Complexity in each iteration in solving the corresponding subproblem of ARNT, RTR and PenCS by conjugate gradient method.

6 Numerical Experiments

In this section, we present the numerical experiments to illustrate the efficiency of PenCF and PenCS in practice. In each part, we introduce the test problems, investigate how to choose the default settings of the algorithm parameters, and compare the proposed algorithms with some of the state-of-the-art ones. All the numerical experiments in this section are run in serial in a desktop with Intel(R) Core(TM) i7-9700K @ 3.60 GHz×8 and 16GB RAM running under Ubuntu 18.04 and MATLAB R2018b.

6.1 The First-order Algorithms

In this subsection, we demonstrate the efficiency of our first-order algorithm PenCF.

First we investigate how to choose the default values for the penalty parameter β, the stepsize η_k, and the radius K through the comparison on the following two test problems.

Problem 3 Kohn-Sham total energy minimization including the non-classical and quantum interaction between electrons:

$$\begin{aligned} \min_{X \in \mathbb{R}^{n\times p}} \quad & \frac{1}{2}\text{tr}\left(X^\top L X\right) + \frac{1}{4}\text{tr}\left(\rho^\top L^\dagger \rho\right) + \frac{3\gamma}{4}\rho^\top \rho^{\frac{1}{3}} \\ \text{s.t.} \quad & X^\top X = I_p, \end{aligned} \tag{6.1}$$

where $\rho = \text{Diag}\left(XX^\top\right)$ and γ is an constant. In our numerical experiment, we choose L as some graph Laplician matrix and A as a randomly generated matrix with Gaussian distribution. In the numerical experiments, if without any specific setting, we set $\gamma = \gamma_0 := -2\left(\frac{3}{\pi}\right)^{\frac{1}{3}}$ and $s = \|\nabla f(X_0)\|_{\text{F}}$.

Problem 4 Minimizing quadratic function over Stiefel manifold:

$$\begin{aligned} \min_{X\in\mathbb{R}^{n\times p}} \quad & \mathrm{tr}\left(X^\top AX\right) + \mathrm{tr}(G^\top X) \\ \text{s.t.} \quad & X^\top X = I_p. \end{aligned} \tag{6.2}$$

In this experiment, both A and G is as a randomly generated symmetric matrix with Gaussian distribution and $s = \|\nabla f(X_0)\|_{\mathrm{F}}$.

In Theorem 3, to guarantee the convergence, a sufficient condition of β is to be sufficiently large. Although we can estimate a suitable β satisfying the conditions in previous theorems and propositions, such β is too huge to be practically useful. In our numerical experiments, we set β no bigger than $s = \|\nabla f(X_0)\|_{\mathrm{F}}$ in PenCF.

The second parameter to tune is the stepsize η_k. Similarly, the upper bound of η_k adopted in Theorem 3 is too restrictive in practice. Here, we suggest to use the Barzilai-Borwein (BB) stepsize [6],

$$\eta_{BB1,k} = \frac{\langle S_k, Y_k\rangle}{\langle Y_k, Y_k\rangle}, \quad \eta_{BB2,k} = \frac{\langle S_k, S_k\rangle}{\langle S_k, Y_k\rangle}, \tag{6.3}$$

and alternating Barzilai-Borwein (ABB) stepsize [12],

$$\eta_{ABB,k} = \begin{cases} \eta_{BB1,k} & mod(k,2) = 1 \\ \eta_{BB2,k} & mod(k,2) = 0, \end{cases} \tag{6.4}$$

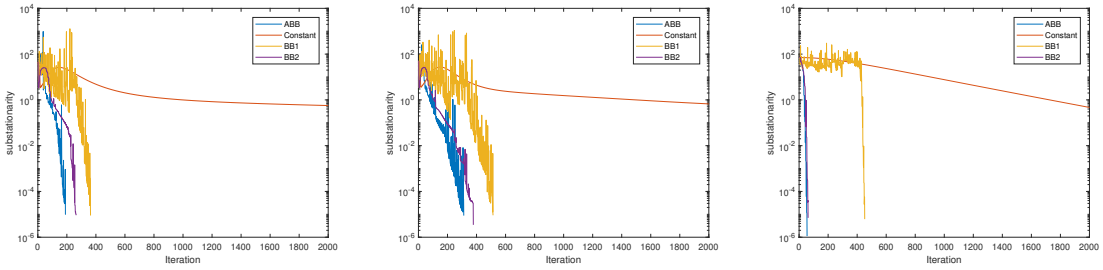
where and $S_k = X_k - X_{k-1}$, $Y_k = \nabla_X \mathcal{L}(X_k, \Lambda)\left|_{\Lambda=\Lambda(X_k)}\right. - \nabla_X \mathcal{L}(X_{k-1}, \Lambda)\left|_{\Lambda=\Lambda(X_{k-1})}\right.$.

Without specific mentioning, the default stopping criteria in this subsection is set as the following,

$$\left\|\nabla f(X_k) - X_k \nabla f(X_k)^\top X_k\right\|_{\mathrm{F}} \le 10^{-8}.$$

Besides, we evaluate the substationarity by the value of $\left\|\nabla f(X_k) - X_k \nabla f(X_k)^\top X_k\right\|_{\mathrm{F}}$.

We first run PenCF under different stepsizes mentioned above. The penalty parameter β is set to be $0.1s$. The numerical results are illustrated in Figure 1, from which we can learn that PenCF with η_{ABB} always outperforms those with other choices. Therefore, we set η_{ABB} as the default stepsize for PenCFin the rest of this section.

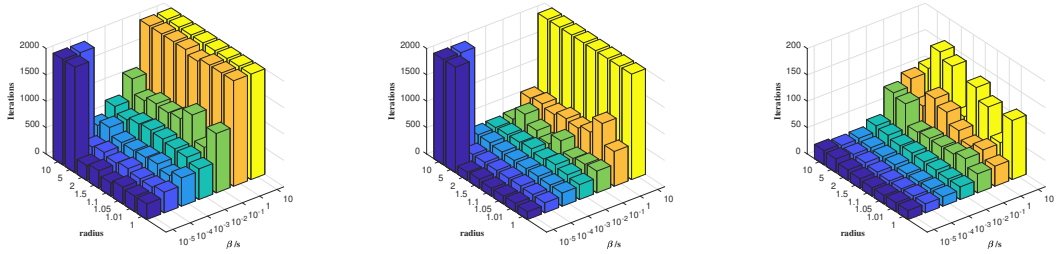


(a) PenCF for Problem 3 with $\gamma = 0$ (b) PenCF for Problem 3 with $\gamma = \gamma_0$ (c) PenCF for Problem 4

Fig. 1: A comparison of substationarity for PenCF with different stepsize η_k.

Moreover, we compare PenCF with different pairs of β and K that vary among $\beta = 10^{-5}s, 10^{-4}s, 10^{-3}s, 10^{-2}s, 10^{-1}s, s, 10s$ and $K = 1 * \sqrt{p}, 1.01 * \sqrt{p}, 1.05 * \sqrt{p}, 1.1 * \sqrt{p}, 1.5 * \sqrt{p}, 2 * \sqrt{p}, 5 * \sqrt{p}, 10 * \sqrt{p}$, respectively. We test Problem 3 and 4 with $n = 1000, p = 30$, and the stopping criteria is changed to $\|\nabla f(X) - X\Lambda(X)\|_{\mathrm{F}} \le 10^{-5}$. In Figure 2, "iter" denotes the iteration number of our algorithms, and we record the number of iterations required by running PenCF under different combinations of parameters β and K.

The results are shown in Figure 2. From these figures, we can conclude that PenCF is not sensitive to the choice of K when $K \le 2$. Furthermore, a small β can lead to fast convergence rate. For test Problem 3, PenCF may fail to converge with small β and large K which are the very extreme cases. For test Problem 4, PenCF even converges even with sufficiently small β and large K. As a result, we suggest to choose $\beta = 0.1s$ while $K = 1.1\sqrt{p}$.

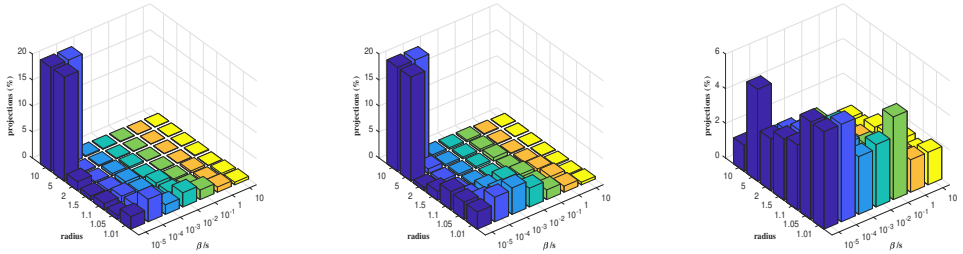


(a) PenCF for Problem 3 with $\gamma = 0$ (b) PenCF for Problem 3 with $\gamma = \gamma_0$ (c) PenCF for Problem 4

Fig. 2: A comparison of substationarity for PenCF with different β and K.

Next, we observe how the percentage of $\mathcal{B}_K$ being active with all iterations varies with different pairs of parameters β and K. The choices of β and K are set in the same manner as the previous experiment. The results are presented in Figure 3.

All the tests, except very extreme cases, in which β is chosen too small and K is set too large for Problem 3 and consequently PenCF diverges, the percentage of B being active with all iterations is always under 5%. This observation reveals the fact that the additional projection procedure in PenCF can be waived in most iterations.



(a) Problem 3, $\gamma = 0$ (b) Problem 3, $\gamma = \gamma_0$ (c) Problem 4

Fig. 3: A comparison of projections with different β and K. $s = \|\nabla f(X_0)\|_{\mathrm{F}}$ is an approximation for L_1 and 'radius' denotes $K/\sqrt{p}$.

When using the first-order methods to solve constrained optimization problems, we usually expect mild accuracy for the substationarity, but pursue high accuracy for the feasibility. To this end, we impose an orthonormalization postprocess after obtaining the last iterate X^k by PenCF. Namely,

$$\mathrm{orth}(X^k) := UV^\top, \tag{6.5}$$

where $X^k = U\Sigma V^\top$ is the economic singular value decomposition of X^k with $U \in \mathbb{R}^{n\times p}$ and $V \in \mathbb{R}^{p\times p}$ are the orthogonal matrices and Σ is a diagonal matrix with singular values of X^k on its diagonal. Proposition 2 in Appendix illustrate the postprocess (6.5) can reduce the function value at the same time. Consequently, $\mathrm{orth}(X^k)$ is better point than X^k in any sense.

In the end of this subsection, we compare PenCF with some of the start-of-the-art solvers for solving optimization problems with orthogonality constraints. PenCF takes all the default settings and the postprocess.

The test problems are the Kohn-Sham total energy minimization chosen from KSSOLV [46], which is a MATLAB toolbox designed for electronic structure calculation. We choose two build-in solvers in KSSOLV. The first is the self-consistent field (SCF) iteration, which reformulates the Kohn-Sham total energy minimization as a nonlinear eigenvalue problem and solves it by a sequence of linear eigenvalue problems [45, 26]. The other one is the trust-region direct constrained minimization (TRDCM) [43, 44], which adopts the trust-region method to solve the minimization problem and uses SCF to solve the trust-region subproblem. Beside these two solvers, we select another two state-of-the-art solvers for optimization problems with orthogonality constraints. One of them is OptM [41], which adopts the Cayley transform to preserve the feasibility in each iteration. The other one is MOptQR

[4, 8], which is a projection-based feasible method. In our numerical test, we choose alternating BB stepsize with nonmonotone line search to accelerate MOptQR.

We select 9 test problems with respect to different molecules. To avoid unrepresentative CPU time caused by slow convergence, we set the max iteration number as 200 for SCF, 200 for TRDCM and 2000 for MOptQR, OptM, PCAL and PenCF. All the parameters for these algorithms take their default values, and the initial guess X_0 is generated by function "getX0" in KSSOLV. The numerical results are illustrated in Table 2 where E_{tot}, "Substationarity", "Iteration", "Feasibility violation" and "CPU time" stands for the function value, $\|\nabla f(X) - X\Lambda(X)\|_{\mathrm{F}}$, the number of iterations, $\left\|X^TX - I_p\right\|_{\mathrm{F}}^2$, and the wall-clock running time, respectively.

From Table 2, we can observe that PCAL and PenCF have better numerical behaviors than the other algorithms in comparison. Moreover, PenCF outperforms PCAL in most cases. Therefore, we can conclude that PenCF is more efficient and robust than the state-of-the-art algorithms in comparison.

Moreover, with increasing column size p, the projection to Stiefel manifold turns to be more and more expensive, hence, the advantages of orthonormalization-free approaches PCAL and PenCF becomes more obvious. Moreover, since projection back to a ball costs much less than projection back to the Oblique manifold, PenCF performs better than PCAL in most cases, particularly in cases "ptnio" and "pentacene".

Solver	E_{tot}	Substationarity	Iteration	Feasibility violation	CPU time(s)
alanine,$n = 12671, p = 18$					
SCF	-6.1161921213050e+01	3.14e-09	20	7.31e-15	21.63
TRDCM	-6.1161921213046e+01	2.28e-06	200	4.91e-15	150.53
ManOptQR	-6.1161921213050e+01	5.68e-09	185	3.89e-15	30.10
OptM	-6.1161921213050e+01	2.30e-09	105	3.79e-14	18.23
PCAL	-6.1161921213050e+01	5.94e-09	106	3.63e-15	19.47
PenCF	-6.1161921213050e+01	5.96e-09	113	3.55e-15	18.92
al,$n = 16879, p = 12$					
SCF	-1.5769678051112e+01	1.08e-01	200	5.59e-15	175.33
TRDCM	-1.5803817596149e+01	3.33e-08	200	3.68e-15	133.41
ManOptQR	-1.5630343515632e+01	9.67e-01	2000	6.39e-15	311.84
OptM	-1.5803791154679e+01	2.47e-09	1942	1.10e-14	303.16
PCAL	-1.5803817596151e+01	1.90e-08	2000	1.54e-12	310.82
PenCF	-1.5803817596151e+01	9.93e-09	1722	1.67e-14	266.62
benzene,$n = 8407, p = 15$					
SCF	-3.7225751362902e+01	3.45e-09	17	7.21e-15	9.20
TRDCM	-3.7225751362902e+01	8.83e-09	44	6.77e-15	16.68
ManOptQR	-3.7225751362902e+01	1.27e-09	135	2.62e-15	11.97
OptM	-3.7225751362902e+01	2.42e-09	99	2.17e-14	9.20
PCAL	-3.7225751362902e+01	9.13e-09	89	2.44e-15	8.72
PenCF	-3.7225751362902e+01	5.70e-09	95	3.04e-15	8.72
c12h26,$n = 5709, p = 37$					
SCF	-8.1536091936606e+01	3.19e-09	23	1.15e-14	29.92
TRDCM	-8.1536091936555e+01	6.80e-06	200	9.83e-15	149.25
ManOptQR	-8.1536091936606e+01	9.26e-09	822	5.56e-15	141.71
OptM	-8.1536091936606e+01	1.41e-09	120	9.51e-14	23.07
PCAL	-8.1536091936606e+01	9.05e-09	117	1.19e-14	24.45
PenCF	-8.1536091936606e+01	8.54e-09	108	5.95e-15	20.85
glutamine,$n = 16517, p = 29$					
SCF	-9.1839425243648e+01	3.75e-09	23	8.69e-15	71.50
TRDCM	-9.1839425243571e+01	9.55e-06	200	8.11e-15	496.17
ManOptQR	-9.1839425243648e+01	6.01e-09	119	6.61e-15	59.35
OptM	-9.1839425243648e+01	1.18e-09	143	5.67e-15	71.75
PCAL	-9.1839425243648e+01	9.28e-09	127	9.53e-15	66.50
PenCF	-9.1839425243648e+01	5.02e-09	128	5.99e-15	62.64
graphene16,$n = 3071, p = 37$					

Table 2 continued from previous page

SCF	-9.4032618962855e+01	6.28e-02	200	1.24e-14	129.52
TRDCM	-9.4046217544979e+01	8.13e-06	200	9.09e-15	104.55
ManOptQR	-9.4046217545036e+01	7.08e-09	746	5.61e-15	77.18
OptM	-9.4046217545036e+01	1.66e-09	298	5.01e-15	31.89
PCAL	-9.4046217545036e+01	8.83e-09	276	5.44e-15	32.45
PenCF	-9.4046217545036e+01	6.07e-09	270	5.44e-15	28.41
ptnio,$n = 4069$, $p = 43$					
SCF	-2.2678884272587e+02	5.39e-09	99	1.50e-14	88.09
TRDCM	-2.2678883639168e+02	2.89e-04	200	1.05e-14	136.59
ManOptQR	-2.2678884272587e+02	9.52e-09	697	5.01e-15	100.67
OptM	-2.2678884272587e+02	2.40e-09	864	4.52e-15	125.62
PCAL	-2.2678884272587e+02	9.70e-09	699	5.36e-15	110.28
PenCF	-2.2678884272587e+02	7.83e-09	693	4.38e-15	95.34
ctube661,$n = 12599$, $p = 48$					
SCF	-1.3463843176502e+02	6.79e-09	19	1.39e-14	62.98
TRDCM	-1.3463843176491e+02	1.05e-05	200	1.04e-14	487.15
ManOptQR	-1.2304610718869e+02	7.04e+00	2000	6.70e-15	967.80
OptM	-1.3463843176501e+02	1.97e-09	120	5.91e-15	64.29
PCAL	-1.3463843176502e+02	8.39e-09	112	5.75e-15	61.94
PenCF	-1.3463843176502e+02	3.17e-09	120	7.61e-15	59.55
graphene30,$n = 12279$, $p = 67$					
SCF	-1.7358463196091e+02	7.24e-02	200	1.86e-14	1039.02
TRDCM	-1.7359510505877e+02	6.39e-06	200	1.27e-14	757.31
ManOptQR	-1.7359509080081e+02	1.77e-03	2000	7.54e-15	1459.20
OptM	-1.7359510505880e+02	1.78e-09	892	6.46e-15	1260.97
PCAL	-1.7359510505880e+02	5.08e-09	332	7.07e-15	269.51
PenCF	-1.7359510505880e+02	8.55e-09	351	6.32e-15	266.32
pentacene,$n = 44791$, $p = 51$					
SCF	-1.3189029495352e+02	7.36e-09	22	1.45e-14	310.38
TRDCM	-1.3189029495346e+02	7.39e-06	200	1.17e-14	1763.47
ManOptQR	-1.3189029495352e+02	7.60e-09	215	8.75e-15	417.12
OptM	-1.3189029495352e+02	2.44e-09	168	1.01e-14	335.30
PCAL	-1.3189029495352e+02	8.80e-09	150	1.04e-14	333.78
PenCF	-1.3189029495352e+02	9.87e-09	156	1.00e-14	297.68

Table 2: The results in Kohn-Sham total energy minimization

6.2 The Second-order Algorithms

In this subsection, we compare the numerical behaviors among PenCS and two existing algorithms RTR and ARNT introduced in Subsection 1.1. The test problem is the following simplified Kohn-Sham energy minimization (Problem 3 without exchange correlation energy).

Problem 5 Simple nonlinear eigenvalue problem

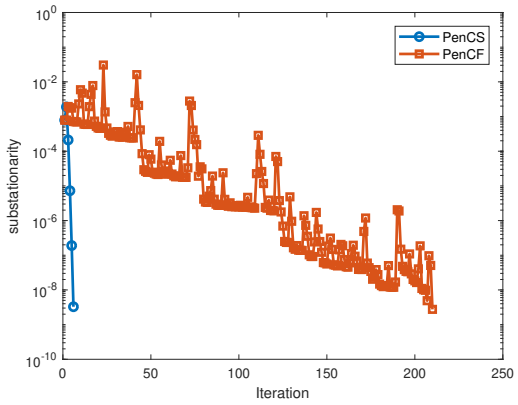
$$\begin{aligned} \min_{X\in\mathbb{R}^{n\times p}} \quad & \frac{1}{2}\mathrm{tr}\left(X^\top L X\right) + \frac{\alpha}{4}\rho^\top L^\dagger \rho \\ \text{s.t.} \quad & X^\top X = I_p. \end{aligned} \tag{6.6}$$

Since the second-order methods are usually not global convergent. We often use it combined with a first-order method. Namely, we first adopt a first-order algorithm to reach certain precision and then swift to the second-order algorithm, see Hu et al. [19] for instance. The default setting for parameter η_k is set to be 1.

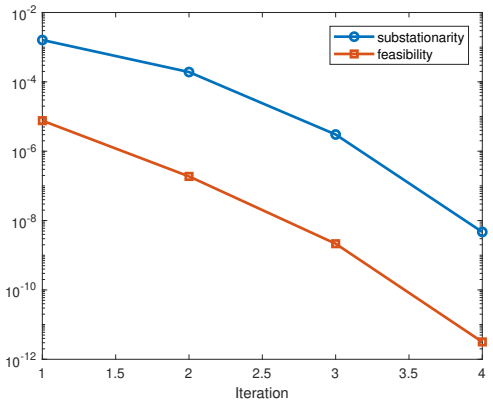
We first perform a simple test to compare the PenCS with PenCF locally. In this numerical test, $(n, p, \alpha) = (500, 10, 1)$ and A is generated by the following MATLAB code,

$$L = \mathrm{randn}(n, p); \qquad L = 0.5 * (L + L').$$

We call PenCF with default setting to obtain an initial point satisfying $\|\nabla f(X_0) - X_0\Phi(X_0)\|_F \leq 10^{-3}$ for both algorithms in comparison. The numerical results are presented in Figure 4. Figure 4(a) illustrates the quadratic convergence rate of PenCS, and figure 4(b) shows that feasibility violation reduces in the same rate as the substationarity.



(a) Comparison between PenCF and PenCS.



(b) Substationarity and constraint violations in PenCS.

Fig. 4: A preliminary numerical example for PenCS.

Then we compare PenCS with ARNT and RTR in solving Problems 5 with different triples of n, p, α and the matrix L is always set as tridiagonal with all diagonal entries and sub-diagonal entries equal to 2 and -1, respectively. All the algorithms PenCS , RTR and ARNT start from the same initial point X_0 satisfying $\|\nabla f(X_0) - X_0\Phi(X_0)\|_F \leq 10^{-4}$. The stopping criteria is $\|\nabla f(X_k) - X_k\Lambda(X_k)\|_F \leq 10^{-12}$. More specifically, we first fix $p = 30$, $\alpha = 10$ and n varies from 2000 to 12000. Then we fix $n = 5000$, $\alpha = 10$ and choose p from 10 to 40. Besides, we fix $n = 5000$ and $p = 30$ while choosing α from 0.5 to 20. The detailed results are demonstrated in Table 3-4, respectively. Table 3 shows that PenCS has good scalability as n grows. As n grows, PenCS is comparable with ARNT and RTR. Table 4 indicates that when the column size p increases, PenCS has better scalability than ARNT and RTR. From these two tables, we can conclude that PenCS takes far less outer iterations than ARNT and RTR. As a result, PenCS significantly reduces the number of Newton directions to be evaluated, which may save CPU time a lot.

Moreover, we also plot out how the substationarity decreases with the CPU time elapses in these three algorithms. The results are illustrated in Figure 5. From these figures we can further conclude that PenCS is comparable with ARNT and RTR in most cases, especially when p is relative large. In the early stage of the iteration, we can find that both ARNT and RTR are slightly faster. We explain this phenomenon as the following.

In PenCS, since a good initial point is obtained by the first-order method, D_k can be solved by directly solving the linear system $W(X_k)[D] = -\nabla h(X_k)$, which can be realized by conjugate gradient method (CG) [18]. In each CG iteration, both $\nabla^2 f(X_k)[D_k]$ and $\nabla^2 f(X_k)[X_k\Phi(X_k^\top D_k)]$ should be computed. For testing problem 5, since L is sparse, the computational cost for computing $\nabla f(X)$ is $\mathcal{O}(np^2)$, and computing $\nabla^2 f(X)[D]$ is also $\mathcal{O}(np^2)$. But in each CG iteration for solving subproblems of RTR and ARNT, only $\nabla^2 f(X_k)[D]$ is required to be computed. Therefore, in PenCS requires more arithmetic operations in the early stage.

However, with the decreasing of the substationarity, we can scale the Riemannian gradient or $\nabla h(X)$ to reduce the round-off error. However, such technique fails in ARNT due to the nonconvexity of the subproblems. As a result, from these figures we can find that when the substationarity reaches 10^{-6} to 10^{-9}, ARNT fails to make any progress. Besides, since the updating scheme in RTR is based on trust-region updating scheme, the solution generated by trust-region subproblem in RTR may be rejected by trust-region updating formula, especially when the Riemannian Hessian is ill-conditioned. This fact leads to more outer iterations in RTR and helps to explain that PenCS has better numerical performance when p is large.

Solver	fval	iter.	sub-iter.	substationarity	feasibility	CPU time(s)
$(n, p, \alpha) = (2000, 30, 10.0)$						
ARNT	1.160612e+03	10	192	9.86e-13	2.72e-15	0.58
RTR	1.160612e+03	7	244	8.73e-13	1.44e-15	0.80
PenCS	1.160612e+03	4	223	7.77e-13	3.10e-15	0.45
$(n, p, \alpha) = (5000, 30, 10.0)$						
ARNT	1.160612e+03	12	202	7.77e-13	2.59e-15	1.42
RTR	1.160612e+03	7	242	9.27e-13	1.19e-15	1.71
PenCS	1.160612e+03	4	230	7.40e-13	2.47e-15	0.97
$(n, p, \alpha) = (8000, 30, 10.0)$						
ARNT	1.160612e+03	12	204	8.82e-13	2.77e-15	2.51
RTR	1.160612e+03	8	247	8.23e-13	1.56e-15	2.97
PenCS	1.160612e+03	4	237	9.28e-13	2.49e-15	1.74
$(n, p, \alpha) = (10000, 30, 10.0)$						
ARNT	1.160612e+03	11	202	7.79e-13	2.94e-15	3.19
RTR	1.160612e+03	7	249	9.55e-13	1.58e-15	4.00
PenCS	1.160612e+03	4	232	7.58e-13	2.40e-15	2.32
$(n, p, \alpha) = (12000, 30, 10.0)$						
ARNT	1.160612e+03	11	206	9.29e-13	2.27e-15	4.06
RTR	1.160612e+03	8	247	6.85e-13	1.45e-15	4.94
PenCS	1.160612e+03	4	229	8.74e-13	2.60e-15	2.81

Table 3: Comparison with fixed p and α.

Solver	fval	iter.	inner iter.	substationarity	feasibility	CPU time(s)
$(n, p, \alpha) = (5000, 10, 10.0)$						
ARNT	7.576019e+01	8	65	9.88e-13	1.70e-15	0.24
RTR	7.576019e+01	3	77	6.79e-13	7.25e-16	0.23
PenCS	7.576019e+01	4	63	8.23e-13	9.98e-16	0.14
$(n, p, \alpha) = (5000, 20, 10.0)$						
ARNT	3.950328e+02	8	128	9.82e-13	1.70e-15	0.55
RTR	3.950328e+02	4	166	6.50e-13	1.57e-15	0.75
PenCS	3.950328e+02	4	144	9.10e-13	1.96e-15	0.40
$(n, p, \alpha) = (5000, 30, 10.0)$						
ARNT	1.160612e+03	12	202	7.77e-13	2.59e-15	1.32
RTR	1.160612e+03	7	242	9.27e-13	1.19e-15	1.60
PenCS	1.160612e+03	4	230	7.40e-13	2.47e-15	0.90
$(n, p, \alpha) = (5000, 40, 10.0)$						
ARNT	2.580831e+03	16	269	9.97e-13	3.38e-15	2.65
RTR	2.580831e+03	31	556	6.09e-12	1.40e-15	5.60
PenCS	2.580831e+03	4	309	8.56e-13	3.85e-15	1.85

Table 4: Comparison with fixed n and α.

Solver	fval	iter.	inner iter.	substationarity	feasibility	CPU time(s)
$(n, p, \alpha) = (5000, 30, 0.5)$						
ARNT	1.201359e+02	8	174	9.31e-13	2.63e-15	1.22
RTR	1.201359e+02	4	183	9.80e-13	1.23e-15	1.17
PenCS	1.201359e+02	4	177	9.30e-13	2.41e-15	0.80
$(n, p, \alpha) = (5000, 30, 1.0)$						
ARNT	1.808320e+02	8	162	9.25e-13	2.30e-15	1.11
RTR	1.808320e+02	4	224	2.84e-13	1.31e-15	1.54
PenCS	1.808320e+02	4	181	7.55e-13	2.59e-15	0.77
$(n, p, \alpha) = (5000, 30, 5.0)$						
ARNT	6.204195e+02	9	187	7.54e-13	2.83e-15	1.23
RTR	6.204195e+02	4	204	9.91e-13	1.80e-15	1.43
PenCS	6.204195e+02	4	208	7.90e-13	2.86e-15	0.87
$(n, p, \alpha) = (5000, 30, 10.0)$						
ARNT	1.160612e+03	12	202	7.77e-13	2.59e-15	1.42
RTR	1.160612e+03	7	242	9.27e-13	1.19e-15	1.67
PenCS	1.160612e+03	4	230	7.40e-13	2.47e-15	0.99
$(n, p, \alpha) = (5000, 30, 20.0)$						
ARNT	2.238241e+03	14	213	8.07e-13	2.68e-15	1.51
RTR	2.238241e+03	31	318	1.20e-12	1.79e-15	2.57
PenCS	2.238241e+03	4	247	7.66e-13	3.38e-15	1.04

Table 5: Comparison with fixed n and p.

7 Conclusion

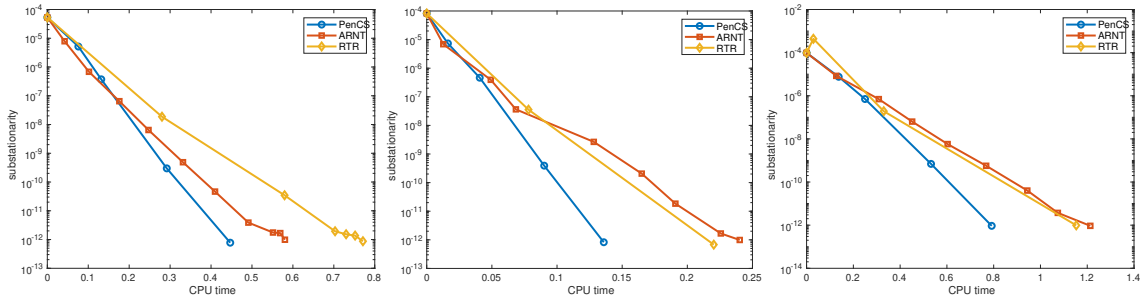
Optimization problems with orthogonality constraints play an important role in many application areas such as material science, machine learning, image processing and so on. The existing efficient approaches for this type of problem are retraction-based. Namely, explicit or implicit orthonormalization is inevitable. When the number of columns of the variable increases, orthonormalization becomes costly and lacks concurrency, and becomes the bottleneck of parallelization. Different with Gao et al. [15], in which the authors propose a technique to update the augmented multiplier by a closed-form formula in using the augmented Lagrangian penalty function, we deeply investigate the merit function $h(X)$ and propose a new model PenC which is to minimize $h(X)$ over a compact convex set. We show that PenC can be exact penalty under some mild conditions. Based on PenC with a ball constraint, we propose an inexact projected gradient method, called PenCF, and an inexact projected Newton method, called PenCS. The global convergence of PenCF and local convergence rate of PenCF and PenCS are established. Numerical experiments show that PenCF outperforms the existing retraction-based approaches and PCAL proposed in Gao et al. [15] in solving the test problems. In pursuing high accuracy solution, PenCS shows its robustness and efficiency comparing with ARNT and RTR proposed in [19, 3], respectively.

Acknowledgements. Xin Liu was supported by NSFC grants 11971466, Key Research Program of Frontier Sciences, CAS, Grant NO. ZDBS-LY-7022, the National Center for Mathematics and Interdisciplinary Sciences, CAS and the Youth Innovation Promotion Association, CAS. Ya-xiang Yuan was supported by NSFC grants 11688101.

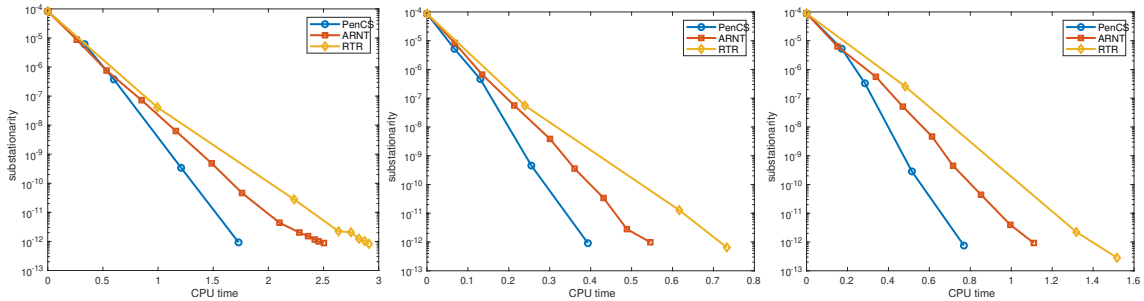
Appendix A

Lemma 7 *For any $X, Y \in \mathcal{M}$, it holds that*

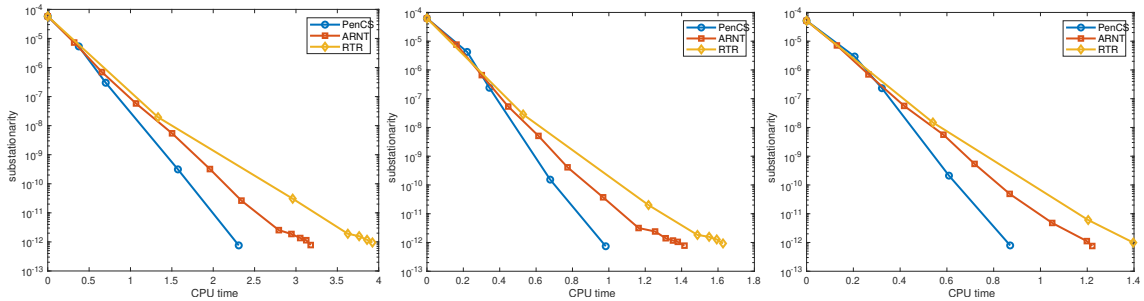
$$\begin{aligned} &\left|\left\langle \Lambda(X), X^\top X - I_p \right\rangle - 2\left\langle X - Y, Y\Lambda(Y)\right\rangle - \left\langle \Lambda(Y), Y^\top Y - I_p \right\rangle\right| \\ &\le L_1 \left\|X - Y\right\|_{\mathrm{F}} \left\|X^\top X - I_p\right\|_{\mathrm{F}} + M_1 \left\|X - Y\right\|_{\mathrm{F}}^2. \end{aligned} \tag{7.1}$$



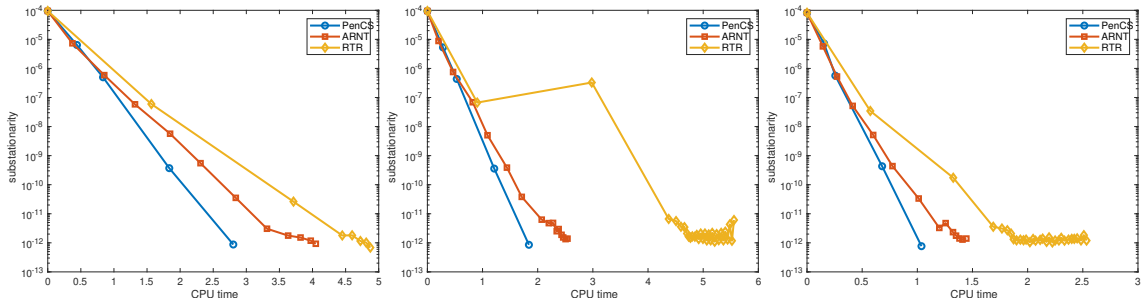
(a) CPU time with $(2000, 30, 10)$ (b) CPU time with $(5000, 10, 10)$ (c) CPU time with $(5000, 30, 0.5)$



(d) CPU time with $(8000, 30, 10)$ (e) CPU time with $(5000, 20, 10)$ (f) CPU time with $(5000, 30, 1)$



(g) CPU time with $(10000, 30, 10)$ (h) CPU time with $(5000, 30, 10)$ (i) CPU time with $(10000, 30, 5)$



(j) CPU time with $(12000, 30, 10)$ (k) CPU time with $(5000, 40, 10)$ (l) CPU time with $(5000, 30, 20)$

Fig. 5: A preliminary numerical example for comparing CPU time for PenCS and ARNT with different (n, p, α).

Proof

$$\begin{aligned}
&\left|\left\langle \Lambda(X), X^\top X - I_p \right\rangle - 2\left\langle X - Y, Y\Lambda(Y)\right\rangle - \left\langle \Lambda(Y), Y^\top Y - I_p \right\rangle\right| \\
=&\left|\left\langle \Lambda(X), X^\top X - I_p \right\rangle - \left\langle \Lambda(Y), X^\top X - I_p \right\rangle + \left\langle \Lambda(Y), X^\top X - I_p \right\rangle \right. \\
&\left. - \left\langle \Lambda(Y), Y^\top Y - I_p \right\rangle - 2\left\langle X - Y, Y\Lambda(Y)\right\rangle\right|
\end{aligned}$$

$$
\begin{aligned}
&\leq \left|\left\langle \Lambda(X), X^\top X - I_p \right\rangle - \left\langle \Lambda(Y), X^\top X - I_p \right\rangle\right| \\
&\quad + \left|\left\langle \Lambda(Y), X^\top X - I_p \right\rangle - \left\langle \Lambda(Y), Y^\top Y - I_p \right\rangle - 2\left\langle X - Y, Y\Lambda(Y) \right\rangle\right| \\
&\leq L_1 \left\| X - Y \right\|_{\mathrm{F}} \left\| X^\top X - I_p \right\|_{\mathrm{F}} + \left|\left\langle \Lambda(Y), (X-Y)^\top (X-Y) \right\rangle\right| \\
&\leq L_1 \left\| X - Y \right\|_{\mathrm{F}} \left\| X^\top X - I_p \right\|_{\mathrm{F}} + M_1 \left\| Y - X \right\|_{\mathrm{F}}^2,
\end{aligned}
$$

which completes the proof.

Proposition 2 *Suppose Assumption 1 holds, $\beta \geq 1 + 2L_0 + 2L_1 + 2M_1$ and $X \in \mathcal{M}$. Let $X = U\Sigma V^\top$ be the economic SVD for X and $\mathrm{orth}(X) = UV^\top$. Then, it holds that*

$$
h(\mathrm{orth}(X)) \leq h(X) - \frac{1}{4}\left\| X^\top X - I_p \right\|_{\mathrm{F}}^2.
$$

Proof First we define $T := X - \mathrm{orth}(X)$. It can be easily verified that $T = \mathrm{orth}(X)(V^\top(\Sigma - I_p)V)$. Then $\|T\|_{\mathrm{F}} = \left\|\Sigma - I_p\right\|_{\mathrm{F}} \leq \left\|(\Sigma - I_p)(\Sigma + I_p)\right\|_{\mathrm{F}} = \left\|X^\top X - I_p\right\|_{\mathrm{F}}$.

Putting $Y := \mathrm{orth}(X)$ into Lemma 7, we have

$$
\begin{aligned}
&\left|\left\langle \Lambda(X), X^\top X - I_p \right\rangle - \left\langle \Lambda(\mathrm{orth}(X)), \mathrm{orth}(X)^\top \mathrm{orth}(X) - I_p \right\rangle - 2\left\langle T, \mathrm{orth}(X)\Lambda(\mathrm{orth}(X)) \right\rangle\right| \\
&\leq L_1 \|T\|_{\mathrm{F}} \left\| X^\top X - I_p \right\|_{\mathrm{F}} + M_1 \|T\|_{\mathrm{F}}^2 \leq (L_1 + M_1) \left\| X^\top X - I_p \right\|_{\mathrm{F}}^2. \qquad (7.2)
\end{aligned}
$$

Moreover, by the Lipschitz continuity of ∇f,

$$
|f(X) - f(\mathrm{orth}(X)) - \langle T, \nabla f(\mathrm{orth}(X)) \rangle| \leq \frac{L_0}{2}\|T\|_{\mathrm{F}}^2 \leq \frac{L_0}{2}\left\| X^\top X - I_p \right\|_{\mathrm{F}}^2. \qquad (7.3)
$$

Substituting relationships (7.2) and (7.3) into $h(X) = f(X) - \frac{1}{2}\left\langle \Lambda(X), X^\top X - I_p \right\rangle + \frac{\beta}{4}\left\| X^\top X - I_p \right\|_{\mathrm{F}}$, we have

$$
\begin{aligned}
&h(X) - h(\mathrm{orth}(X)) = \left(f(X) - \frac{1}{2}\left\langle \Lambda(X), X^\top X - I_p \right\rangle + \frac{\beta}{4}\left\| X^\top X - I_p \right\|_{\mathrm{F}}^2 \right) \\
&- \left(f(\mathrm{orth}(X)) - \frac{1}{2}\left\langle \Lambda(\mathrm{orth}(X)), \mathrm{orth}(X)^\top \mathrm{orth}(X) - I_p \right\rangle \right) \\
&\geq (f(X) - f(\mathrm{orth}(X)) - \langle T, \nabla f(\mathrm{orth}(X)) \rangle) + \langle T, \nabla f(\mathrm{orth}(X)) - \mathrm{orth}(X)\Lambda(\mathrm{orth}(X)) \rangle \\
&- \left(\frac{1}{2}\left\langle \Lambda(X), X^\top X - I_p \right\rangle - \frac{1}{2}\left\langle \Lambda(\mathrm{orth}(X)), \mathrm{orth}(X)^\top \mathrm{orth}(X) - I_p \right\rangle - \left\langle T, \mathrm{orth}(X)\Lambda(\mathrm{orth}(X)) \right\rangle \right) \\
&+ \frac{\beta}{4}\left\| X^\top X - I_p \right\|_{\mathrm{F}}^2 \geq -\frac{L_0}{2}\left\| X^\top X - I_p \right\|_{\mathrm{F}}^2 + 0 - \frac{L_1 + M_1}{2}\left\| X^\top X - I_p \right\|_{\mathrm{F}}^2 + \frac{\beta}{4}\left\| X^\top X - I_p \right\|_{\mathrm{F}}^2 \\
&= \left(\frac{\beta}{4} - \left(\frac{L_0 + L_1 + 2M_1}{2} \right) \right) \left\| X^\top X - I_p \right\|_{\mathrm{F}}^2 \geq \frac{1}{4}\left\| X^\top X - I_p \right\|_{\mathrm{F}}^2,
\end{aligned}
$$

This completes the proof.

References

1. Traian Abrudan, Jan Eriksson, and Visa Koivunen. Conjugate gradient algorithm for optimization under unitary matrix constraint. *Signal Processing*, 89(9):1704–1714, 2009.
2. Traian E Abrudan, Jan Eriksson, and Visa Koivunen. Steepest descent algorithms for optimization under unitary matrix constraint. *IEEE Transactions on Signal Processing*, 56(3):1134–1147, 2008.
3. P-A Absil, Christopher G Baker, and Kyle A Gallivan. Trust-region methods on riemannian manifolds. *Foundations of Computational Mathematics*, 7(3):303–330, 2007.
4. P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.
5. Nitin Bansal, Xiaohan Chen, and Zhangyang Wang. Can we gain more from orthogonality regularizations in training deep networks? In *Advances in Neural Information Processing Systems*, pages 4261–4271, 2018.
6. Jonathan Barzilai and Jonathan M Borwein. Two-point step size gradient methods. *IMA journal of numerical analysis*, 8(1):141–148, 1988.
7. Dimitri P Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

8. Nicolas Boumal, Bamdev Mishra, P-A Absil, and Rodolphe Sepulchre. Manopt, a matlab toolbox for optimization on manifolds. *The Journal of Machine Learning Research*, 15(1):1455–1459, 2014.
9. Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.
10. Richard Courant et al. *Variational methods for the solution of problems of equilibrium and vibrations*. Verlag nicht ermittelbar, 1943.
11. Xiaoying Dai, Liwei Zhang, and Aihui Zhou. Adaptive step size strategy for orthogonality constrained line search methods. *arXiv preprint arXiv:1906.02883*, 2019.
12. Yu-Hong Dai and Roger Fletcher. Projected barzilai-borwein methods for large-scale box-constrained quadratic programming. *Numerische Mathematik*, 100(1):21–47, 2005.
13. Alan Edelman, Tomás A Arias, and Steven T Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.
14. Bin Gao, Xin Liu, Xiaojun Chen, and Ya-xiang Yuan. A new first-order algorithmic framework for optimization problems with orthogonality constraints. *SIAM Journal on Optimization*, 28(1): 302–332, 2018.
15. Bin Gao, Xin Liu, and Ya-xiang Yuan. Parallelizable algorithms for optimization problems with orthogonality constraints. *SIAM Journal on Scientific Computing*, 41(3):A1949–A1983, 2019.
16. Dongyoon Han and Junmo Kim. Unsupervised simultaneous orthogonal basis clustering feature selection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5016–5023, 2015.
17. Magnus R Hestenes. Multiplier and gradient methods. *Journal of optimization theory and applications*, 4(5):303–320, 1969.
18. Magnus Rudolph Hestenes and Eduard Stiefel. *Methods of conjugate gradients for solving linear systems*, volume 49. NBS Washington, DC, 1952.
19. Jiang Hu, Andre Milzarek, Zaiwen Wen, and Yaxiang Yuan. Adaptive quadratically regularized newton method for riemannian optimization. *SIAM Journal on Matrix Analysis and Applications*, 39 (3):1181–1207, 2018.
20. Bo Jiang and Yu-Hong Dai. A framework of constraint preserving update schemes for optimization on stiefel manifold. *Mathematical Programming*, 153(2):535–575, 2015.
21. Ian T Jolliffe, Nickolay T Trendafilov, and Mudassir Uddin. A modified principal component technique based on the lasso. *Journal of computational and Graphical Statistics*, 12(3):531–547, 2003.
22. Walter Kohn and Lu Jeu Sham. Self-consistent equations including exchange and correlation effects. *Physical review*, 140(4A):A1133, 1965.
23. Rongjie Lai and Stanley Osher. A splitting method for orthogonality constrained problems. *Journal of Scientific Computing*, 58(2):431–449, 2014.
24. Xuelong Li, Han Zhang, Rui Zhang, Yun Liu, and Feiping Nie. Generalized uncorrelated regression with adaptive graph for unsupervised feature selection. *IEEE transactions on neural networks and learning systems*, 30(5):1587–1595, 2018.
25. Ji Liu, Stephen J Wright, Christopher Ré, Victor Bittorf, and Srikrishna Sridhar. An asynchronous parallel stochastic coordinate descent algorithm. *The Journal of Machine Learning Research*, 16(1): 285–322, 2015.
26. Xin Liu, Xiao Wang, Zaiwen Wen, and Yaxiang Yuan. On the convergence of the self-consistent field iteration in kohn–sham density functional theory. *SIAM Journal on Matrix Analysis and Applications*, 35(2):546–558, 2014.
27. Xin Liu, Zaiwen Wen, and Yin Zhang. An efficient gauss–newton algorithm for symmetric low-rank product matrix approximations. *SIAM Journal on Optimization*, 25(3):1571–1608, 2015.
28. Jonathan H Manton. Optimization algorithms exploiting unitary constraints. *IEEE Transactions on Signal Processing*, 50(3):635–650, 2002.
29. Yasunori Nishimori and Shotaro Akaho. Learning algorithms utilizing quasi-geodesic flows on the stiefel manifold. *Neurocomputing*, 67:106–135, 2005.
30. Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.
31. Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. 1999.
32. Zhimin Peng, Ming Yan, and Wotao Yin. Parallel and distributed sparse optimization. In *2013 Asilomar conference on signals, systems and computers*, pages 659–646. IEEE, 2013.
33. Zhimin Peng, Yangyang Xu, Ming Yan, and Wotao Yin. Arock: an algorithmic framework for asynchronous parallel coordinate updates. *SIAM Journal on Scientific Computing*, 38(5):A2851–A2879, 2016.
34. M. J. D Powell. A method for nonlinear constraints in minimization problems. *Optimization*, 5(6): 283–298, 1969.

35. Michael JD Powell. A method for nonlinear constraints in minimization problems. *Optimization*, pages 283–298, 1969.
36. Benjamin Recht, Christopher Re, Stephen Wright, and Feng Niu. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in neural information processing systems*, pages 693–701, 2011.
37. Trond Steihaug. The conjugate gradient method and trust regions in large scale optimization. *SIAM Journal on Numerical Analysis*, 20(3):626–637, 1983.
38. Wenyu Sun and Ya-Xiang Yuan. *Optimization theory and methods: nonlinear programming*, volume 1. Springer Science & Business Media, 2006.
39. Jiliang Tang and Huan Liu. Unsupervised feature selection for linked social media data. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 904–912. ACM, 2012.
40. Philippe Toint. Towards an efficient sparsity exploiting newton method for minimization. In *Sparse matrices and their uses*, pages 57–88. Academic Press, 1981.
41. Zaiwen Wen and Wotao Yin. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1-2):397–434, 2013.
42. Zaiwen Wen, Chao Yang, Xin Liu, and Yin Zhang. Trace-penalty minimization for large-scale eigenspace computation. *Journal of Scientific Computing*, 66(3):1175–1203, 2016.
43. Chao Yang, Juan C Meza, and Lin-Wang Wang. A constrained optimization algorithm for total energy minimization in electronic structure calculations. *Journal of Computational Physics*, 217(2): 709–721, 2006.
44. Chao Yang, Juan C Meza, and Lin-Wang Wang. A trust region direct constrained minimization algorithm for the kohn–sham equation. *SIAM Journal on Scientific Computing*, 29(5):1854–1875, 2007.
45. Chao Yang, Weiguo Gao, and Juan C Meza. On the convergence of the self-consistent field iteration for a class of nonlinear eigenvalue problems. *SIAM Journal on Matrix Analysis and Applications*, 30 (4):1773–1788, 2009.
46. Chao Yang, Juan C Meza, Byounghak Lee, and Lin-Wang Wang. Kssolva matlab toolbox for solving the kohn-sham equations. *ACM Transactions on Mathematical Software (TOMS)*, 36(2):10, 2009.
47. Yi Yang, Heng Tao Shen, Zhigang Ma, Zi Huang, and Xiaofang Zhou. $\ell_{2,1}$-norm regularized discriminative feature selection for unsupervised. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.
48. Yaxiang Yuan. On the truncated conjugate gradient method. *Mathematical Programming*, 87(3): 561–573, 2000.
49. Rui Zhang, Feiping Nie, and Xuelong Li. Feature selection under regularized orthogonal least square regression with optimal scaling. *Neurocomputing*, 273:547–553, 2018.