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

Proximal Point Algorithm on the Stiefel Manifold

Harry Oviedo

Received: date / Accepted: date

Abstract In this paper, we consider the problem of minimizing a continuously differentiable function on the Stiefel manifold. To solve this problem, we develop a geodesic–free proximal point algorithm, which does not require to use the Riemannian distance. The proposed method can be regarded as an iterative fixed–point method, which repeatedly applies a proximal operator to an initial point. In addition, we establish the global convergence of the new approach without any restrictive assumption. Numerical experiments on linear eigenvalues problems and the minimization of sums of heterogeneous quadratic functions, show that the developed algorithm is competitive with some procedures existing in the literature.

Keywords Proximal point method · Stiefel manifold · Orthogonality constraint · Riemannian optimization.

Mathematics Subject Classification (2000) 90C30 · 90C48 · 90C51.

1 Introduction.

In this paper, we are interested in designing a proximal point procedure to solve the following optimization problem

$$\min_{X\in\mathbb{R}^{n\times p}} \mathcal{F}(X), \quad \text{s.t.} \quad X^\top X = I_p, \tag{1}$$

where $\mathcal{F}:\mathbb{R}^{n\times p}\to\mathbb{R}$ is a continuously differentiable matrix function, and $I_p\in\mathbb{R}^{p\times p}$ denotes the identity matrix. The feasible set of problem (1) $St(n,p)=\{X\in\mathbb{R}^{n\times p}: X^\top X=I_p\}$, is known as the *Stiefel manifold* [1]. Actually, this set constitutes an embedded Riemannian sub–manifold of the Euclidean space $\mathbb{R}^{n\times p}$, with dimension equals to $np-\frac{1}{2}p(p+1)$, see [1]. Notice that (1) is a well–defined optimization problem, because $St(n,p)$ is a compact set and $\mathcal{F}$ is a continuous function, therefore the Weierstrass theorem ensures the existence of at least one global minimizer (and even a global maximizer) for $\mathcal{F}$ on the Stiefel manifold.

The orthogonality constrained minimization problem (1) is widely applicable in many fields such as nearest low–rank correlation matrix problem [9,22], linear eigenvalue problem [17,18,28], sparse principal component analysis [5,18], Kohn–Sham total energy minimization [17,20,28,30], low–rank matrix completion [12,13], orthogonal procrustes problem [19,20], maximization of sums of heterogeneous quadratic functions from statistics [3,17,31], joint diagonalization problem [31], image segmentation [14], dimension reduction techniques in pattern recognition [11], among others.

In the Euclidean setting, given $f:\mathbb{R}^n\to\mathbb{R}\cup\{\infty\}$ a closed proper convex function, the proximal operator $\texttt{prox}_f:\mathbb{R}^n\to\mathbb{R}^n$ is defined by

$$\texttt{prox}_f(x)=\arg\min_{y\in\mathbb{R}^n} f(y)+\frac{1}{2\alpha}||y-x||_2^2, \tag{2}$$

H. Oviedo
Escola de Matemática Aplicada, Fundação Getulio Vargas (FGV/EMAp). Rio de Janeiro, RJ, Brazil.
E-mail: harry.leon@fgv.br

where $||\cdot||_2$ is the standard norm of $\mathbb{R}^n$ and $\alpha > 0$ is the proximal parameter [21].

The scalar α plays an important role controlling the magnitude by which the proximal operator sends the points towards the optimum values of f. In particular, larger values of α are related to mapped points near the optimum, while smaller values of this parameter promote a smaller movement to the minimum. It can design iterative optimization procedures that use the proximal operator to define the recursive update scheme. For example, the *proximal minimization algorithm* [21], minimizes the cost function f by consecutively applying the proximal operator $\texttt{prox}_f(\cdot)$, similar to fixed–point methods, to some given initial vector $x_0 \in \mathbb{R}^n$.

Several researchers have proposed generalizations of the proximal minimization algorithm in the Riemannian context. This kind of method was first considered in the Riemannian context by Ferreira and Oliveira [7], in the particular case of Hadamard manifolds. Papa Quiroz and Oliveira [23] adapted the proximal point method for quasiconvex function and proved full convergence of the sequence $\{x_k\}$ to a minimizer point over Hadamard manifolds. In addition, Souza and Oliveira [24] introduced a proximal point algorithm for minimize DC functions on Hadamard manifolds. Wang et. al. [27] established linear convergence and finite termination of this type of algorithms on Hadamard manifolds. For this same type of manifold, in [26] the authors proved global convergence of inexact proximal point methods. Recently Almeida et. al. [2] developed a modified version of the proximal point procedure for minimization over Hadamard manifolds. For the specific case of optimization on the Stiefel manifold, in [4], the authors proposed an proximal gradient method for minimize sum of two function $f(X) + g(X)$ over $St(n, p)$, where f is smooth and its gradient is Lipschitz continuous, and g is convex and Lipschitz continuous.

To minimize a function $\mathcal{F} : \mathcal{M} \to \mathbb{R}$ defined on a Riemannian manifold $\mathcal{M}$, all the approaches presented in [2,7,23,24,27] consider the following generalization of (2)

$$\texttt{prox}_{\mathcal{F}}(X) = \arg\min_{Y \in \mathcal{M}} \mathcal{F}(Y) + \frac{1}{2\alpha}\text{dist}^2(X, Y), \tag{3}$$

where $\text{dist}(X, Y)$ is the Riemannian distance [1]. The main disadvantage of all these works is that the authors proposed methods and theoretical analysis based on the exponential mapping, which requires the construction of geodesics on $\mathcal{M}$. However, it is not always possible to find closed expressions for geodesics over a given Riemannian manifold, since geodesics are defined through a differential equation, [1,6]. Even in the case when we have available a closed formula for the corresponding geodesics on $\mathcal{M}$, the computational cost of calculating the exponential mapping over a matrix space is too high, which is an obstacle to solve large–scale problems.

In this paper, we introduce a very simple proximal point algorithm to tackle Stiefel manifold constrained optimization problems. The proposed approach replaces the term $\text{dist}(X, Y)$ in (3), by the usual matrix distance $||X - Y||_F$, in order to avoid purely Riemannian concepts and techniques such as Riemannian distance, and geodesics. The proposed iterative method tries to solve the optimization problem (1) by repeatedly applying our modified proximal operator to a given starting point. We prove (without imposing Lipschitz continuity hypothesis) that our method converges to critical points of the restriction of the cost function to the Stiefel manifold. Our preliminary computational results suggest that our proposal presents a competitive numerical performance against several feasible methods existing in the literature.

The rest of this manuscript is organized as follows. Section 2 summarizes a few well–known notations and concepts of linear algebra and Riemannian geometry, that will be exploited in this paper. Afterwards, Sections 3 introduces a new proximal point algorithm to deal with optimization problems with orthogonality constraints. Section 4 provides a concise convergence analysis for the proposed algorithm. Section 5 presents some illustrative numerical results, where we compare our approach with several state–of–the–art methods on the solution of linear eigenvalue problems and the minimization of sums of heterogeneous quadratic functions. Finally, the paper ends with a conclusion in Section 6.

2 Preliminaries.

Throughout this paper, we say that $W \in \mathbb{R}^{n\times n}$ is skew–symmetric if $W = -W^\top$. Given a square matrix $A \in \mathbb{R}^{m\times m}$, $\text{skew}(A)$ will denotes the skew–symmetric part of A, that is, $\text{skew}(A) = 0.5(A - A^\top)$. The trace of X is defined as the sum of the diagonal elements which we denote by $Tr[X]$. The standard inner product between two matrices $A, B \in \mathbb{R}^{m\times n}$ is given by $\langle A, B\rangle \equiv \sum_{i,j} a_{ij}b_{ij} = Tr[A^\top B]$. The Frobenius norm is

defined by $||A||_F = \sqrt{\langle A, A\rangle}$. Let $X \in St(n,p)$ an arbitrary matrix in the Stiefel manifold, the tangent space of the Stiefel manifold at X is given by [1]

$$T_X St(n,p) = \{Z \in \mathbb{R}^{n\times p} : Z^\top X + X^\top Z = 0\}.$$

Let $X \in St(n,p)$, the canonical metric [28] associated with the tangent space of the Stiefel manifold is defined by

$$\langle \xi_X, \eta_X \rangle_c \equiv Tr\left[\xi_X^\top \left(I - \frac{1}{2}XX^\top\right)\eta_X\right], \quad \forall \xi_X, \eta_X \in T_X St(n,p).$$

Let $\mathcal{F} : \mathbb{R}^{n\times p} \to \mathbb{R}$ be a differentiable function, we denote by $\mathcal{DF}(X) \equiv (\frac{\partial \mathcal{F}(X)}{\partial X_{ij}})$ the matrix of partial derivatives of $\mathcal{F}$ (the Euclidean gradient of $\mathcal{F}$). Let $\Phi : St(n,p) \to \mathbb{R}$ be a smooth function defined on the Stiefel manifold, then the Riemannian gradient of Φ at $X \in St(n,p)$, denote by $\nabla\Phi(X)$, is the unique vector in $T_X St(n,p)$ satisfying

$$\mathcal{D}\Phi(X)[\xi_X] := \lim_{\tau\to 0} \frac{\Phi(\gamma(\tau)) - \Phi(\gamma(0))}{\tau} = \langle \nabla\Phi(X), \xi_X\rangle, \quad \forall \xi_X \in T_X St(n,p), \tag{4}$$

where $\gamma : [0, \tau_{\max}] \to St(n,p)$ is any curve that verifies $\gamma(0) = X$ and $\dot\gamma(0) = \xi_X$.

The Riemannian gradient of $\Phi : St(n,p) \to \mathbb{R}$ under the canonical metric has the following closed expression [17,28]

$$\nabla\Phi(X) = \mathcal{D}\Phi(X) - X\mathcal{D}\Phi(X)^\top X, \quad \forall X \in St(n,p). \tag{5}$$

In addition, based on the formula (5), we can define the following projection operator over $T_X St(n,p)$

$$P_X^c[Z] = Z - XZ^\top X, \quad \text{with} \quad Z \in \mathbb{R}^{n\times p}. \tag{6}$$

It can be shown easily that the operator (6) effectively projects matrices from $\mathbb{R}^{n\times p}$ to the tangent space $T_X St(n,p)$. This projection operator was also considered in [25].

Similarly to the case of smooth unconstrained optimization, $X \in St(n,p)$ is a critical point of $\Phi : St(n,p) \to \mathbb{R}$ if it satisfies [1]

$$\nabla\Phi(X) = 0.$$

Therefore, the critical points of the restriction of the cost function $\mathcal{F}$ to the Stiefel manifold are candidates to be local minimizers of problem (1). Here, we clarify that the objective function $\mathcal{F}$ that appears in (1) has both a Euclidean gradient and a Riemannian gradient. In the rest of this paper, we will denote by $\nabla\mathcal{F}(\cdot)$, the Riemannian gradient of the restriction of $\mathcal{F}$ to the set $St(n,p)$ under the canonical metric.

3 Proximal point algorithm on $St(n,p)$.

In this section, we propose an implicit defined curve on the Stiefel manifold. We also validate that the proposed curve verifies analogous properties that Riemannian gradient methods has. Using this curve, we further present in detail the proposed proximal point algorithm.

As we mentioned in the introduction, we consider an adaptation of the exact proximal point method to the framework of minimization over the Stiefel manifold: given a feasible point $X \in St(n,p)$, we compute the new iterated $X(\bar\alpha)$ as a point on the curve $X(\cdot) : [0, \alpha_{\max}] \to St(n,p)$ defined by

$$X(\alpha) \equiv \arg\min_{Y\in\mathbb{R}^{n\times p}} \alpha\mathcal{F}(Y) + \frac{1}{2}||Y - X||_F^2, \quad \text{s.t.} \quad Y^\top Y = I_p. \tag{7}$$

Note that this proximal optimization problem is obtained from (3) by substituting the Riemannian distance with the standard metric associated with the matrix space $\mathbb{R}^{n\times p}$. Additionally, $X(\alpha)$ satisfies that $X(0) = X$. Thus $X(\alpha)$ is a curve on $St(n,p)$ that connects the consecutive iterates X and $X(\bar\alpha)$. This is an analogous property to that the retraction–based line–search curves verify, see [1,10].

The following lemma shows that (7) defines a descent curve at $\alpha = 0$ for the objective function on $St(n,p)$.

Lemma 1 *Given $X \in St(n,p)$ and consider the curve (7). Then,*

$$\mathcal{D}\mathcal{F}(X)[\dot{X}(0)] = -Tr[\nabla\mathcal{F}(X)^\top (I + XX^\top)^{-1}\nabla\mathcal{F}(X)] < 0. \tag{8}$$

Proof. From the optimality condition related to the minimization problem (7), we have that $X(\alpha)$ must satisfy

$$\alpha\nabla\mathcal{F}(X(\alpha)) + P_X^c[X(\alpha) - X] = 0,$$

or equivalently,

$$X(\alpha) - X = X(\alpha)(X(\alpha) - X)^\top X(\alpha) - \alpha\nabla\mathcal{F}(X(\alpha)) \tag{9}$$

$$= X(\alpha) - X(\alpha)X^\top X(\alpha) - \alpha\nabla\mathcal{F}(X(\alpha)). \tag{10}$$

By differentiating the curve $X(\alpha)$ presented in (10) with respect to α, we obtain

$$\dot{X}(\alpha) = \dot{X}(\alpha) - \dot{X}(\alpha)X^\top X(\alpha) - X(\alpha)X^\top\dot{X}(\alpha) - \nabla\mathcal{F}(X(\alpha)) - \alpha\frac{\partial\left(\nabla\mathcal{F}(X(\alpha))\right)}{\partial\alpha}. \tag{11}$$

Substituting $\alpha = 0$ in (12), we obtain

$$\dot{X}(0) = -(I + XX^\top)^{-1}\nabla\mathcal{F}(X). \tag{12}$$

Let's denotes by $W_X := \text{skew}(X^\top\mathcal{D}\mathcal{F}(X))$ and $P_X = I - XX^\top$. Now, we need to proof that $\dot{X}(0)$ belongs to $T_X St(n,p)$. To demonstrate this fact, first note that $\nabla\mathcal{F}(X) = (I + XX^\top)(XW_X + P_X\mathcal{D}\mathcal{F}(X))$. Using this last relation, we can rewrite $\dot{X}(0)$ by

$$\dot{X}(0) = -XW_X - P_X\mathcal{D}\mathcal{F}(X). \tag{13}$$

Then, we have

$$\dot{X}(0)^\top X + X^\top\dot{X}(0) = W_X^\top + W_X = 0.$$

Hence $X(\alpha)$ is a curve on the Stiefel manifold that satisfies $X(0) = X$ and $\dot{X}(0) \in T_X St(n,p)$. So, form (4), the directional derivative of $\mathcal{F}(X(\alpha))$ at $\alpha = 0$ verifies

$$\mathcal{D}\mathcal{F}(X)[\dot{X}(0)] = -Tr[\nabla\mathcal{F}(X)^\top (I + XX^\top)^{-1}\nabla\mathcal{F}(X)] < 0.$$

The strict inequality above is obtained from the positive definiteness of $I + XX^\top$. □

Lemma 1 guarantees that the proposed approach is a descent iterative process. Therefore, by executing an iterative process based on the proximal curve (7), we can minimize the objective function and at the same time move towards stationarity. Taking into account this fact, we propose our proximal point algorithm on $St(n,p)$, whose steps are described in Algorithm 1 (PPA–St).

Algorithm 1 PPA–St

Require: $X_0 \in St(n,p)$, $0 < \alpha_{\min} \leq \alpha_{\max} < \infty$, $\{\alpha_k\}$ be a sequence such that $\alpha_k \in [\alpha_{\min}, \alpha_{\max}]$ for all $k \in \mathbb{N}$, $k \leftarrow 0$.

1: **while** $||\nabla\mathcal{F}(X_k)||_F \neq 0$ **do**

2:

$$X_{k+1} = \arg\min_{Y \in St(n,p)} \alpha_k\mathcal{F}(Y) + \frac{1}{2}||Y - X_k||_F^2, \tag{14}$$

3: **If** $||X_{k+1} - X_k||_F = 0$ **then** stop the algorithm.

4: $k \leftarrow k + 1$,

5: **end while**

In the implementation of PPA–St, we will use an efficient Riemannian gradient method based on the QR–retraction mapping (see Example 4.1.3 in [1]), in order to solve the optimization sub–problem (14). By introducing the notation $\phi_k(Y) \equiv \alpha_k\mathcal{F}(Y) + \frac{1}{2}||Y - X_k||_F^2$, we propose to use the following feasible line–search method, starting at $Y_0^k = X_k$ and $\tau_0 = \alpha_k$,

$$Y_{i+1}^k = (\, Y_i^k - \tau_i\nabla\phi_k(Y_i^k)\,)\texttt{chol}(\, I_p + \tau_i^2\nabla\phi_k(Y_i^k)^\top\nabla\phi_k(Y_i^k)\,)^{-1}, \tag{15}$$

where $\nabla\phi_k(Y)$ is the Riemannian gradient under the canonical metric of $\phi_k(\cdot)$ evaluated at Y, that is, $\nabla\phi_k(Y) = P_Y^c[\mathcal{D}\phi_k(Y)]$, where $\mathcal{D}\phi_k(Y)$ is the Euclidean gradient of $\phi_k(\cdot)$, i.e. $\mathcal{D}\phi_k(Y) = \alpha_k\mathcal{D}\mathcal{F}(Y) + Y - X_k$.

In addition, $\texttt{chol}(A)$ denotes the Cholesky factor obtained form the Cholesky factorization of $A \in \mathbb{R}^{p\times p}$, i.e., let $A \in \mathbb{R}^{p\times p}$ be a symmetric and positive definite matrix (PSD), and suppose that $A = L_A^\top L_A$ is its Cholesky decomposition then $\texttt{chol}(A) \equiv L_A$. Observe that this function is well–defined due to the uniqueness of Cholesky factorization for PSD matrices. Additionally, notice that in the recursive scheme (15), we are projecting $Y_i^k - \tau_i \nabla\phi_k(Y_i^k)$ over the Stiefel manifold using its QR factorization obtained from the Cholesky decomposition, see equation (1.3) in [8].

It is well–known that if we endow the iterative method (15) with a globalization strategy to determine the step–size $\tau_i > 0$, such as the Armijo's rule [15,1] or a non–monotone Zhang–Hager type condition [29,16], then the Riemannian line–search method (15) is globally convergent, please see [1,16].

4 Convergence results.

In this section, we analyze the convergence properties of Algorithm 1 by revealing the relationships between the residuals $||\nabla\mathcal{F}(X_k)||_F$, $|\mathcal{F}(X_{k+1}) - \mathcal{F}(X_k)|$ and $||X_{k+1} - X_k||_F$.

Since $\mathcal{F}$ is continuously differentiable over $\mathbb{R}^{n\times p}$, then its derivative $\mathcal{DF}(X)$ is continuous. Hence $\mathcal{DF}(X)$ is bounded on $St(n,p)$, due to the compactness of the Stiefel manifold. Then, there exists a constant $\kappa > 0$ such that

$$||\mathcal{DF}(X)||_F < \kappa, \quad \forall X \in St(n,p). \tag{16}$$

Consequently, the Riemannian gradient of $\mathcal{F}$ satisfies

$$||\nabla\mathcal{F}(X)||_F = ||\mathcal{DF}(X) - X\mathcal{DF}(X)^\top X||_F \leq 2\kappa, \tag{17}$$

for all $X \in St(n,p)$.

The following proposition states that Algorithm 1 stops at Riemannian critical points of $\mathcal{F}$.

Proposition 1 *Let $\{X_k\}$ be a sequence generated by the Algorithm 1. Suppose that Algorithm terminates at iteration $k \in \mathbb{N}$, then $\nabla\mathcal{F}(X_k) = 0$.*

Proof. The first–order necessary optimality condition associated with the sub–problem (14) leads to

$$\alpha_k \nabla\mathcal{F}(X_{k+1}) + P^c_{X_{k+1}}[X_{k+1} - X_k] = 0, \tag{18}$$

but, since Algorithm 1 terminates at the k–th iteration, we have that $X_{k+1} = X_k$, which directly implies that $P^c_{X_{k+1}}[X_{k+1} - X_k] = P^c_{X_{k+1}}[0] = 0$. By substituting this fact in (18), we obtain the desired result. □

The rest of this section is devoted to study the asymptotic behavior of Algorithm 1 for infinite sequences $\{X_k\}$ generated by our approach, since otherwise, Proposition 1 says that Algorithm 1 returns a stationary point for problem (1). The lemma below provides us two key theoretical results.

Lemma 2 *Let $\{X_k\}$ be an infinite sequence generated by the Algorithm 1. Then, we have*

(a) $\{\mathcal{F}(X_k)\}$ is a convergent sequence.
(b) The residual sequence $\{||X_{k+1} - X_k||_F\}$ converges to zero.

Proof. In view of the optimization sub–problem (14), we have

$$\mathcal{F}(X_k) \leq \mathcal{F}(X_{k-1}) - \frac{1}{2\alpha_{k-1}}||X_k - X_{k-1}||_F^2. \tag{19}$$

Therefore, $\{\mathcal{F}(X_k)\}$ is a monotonically decreasing sequence. Now, since Stiefel manifold is a compact set and $\mathcal{F}$ is a continuous function, we obtain that $\mathcal{F}$ has maximum and minimum on $St(n,p)$. Therefore, $\{\mathcal{F}(X_k)\}$ is bounded, and then $\{\mathcal{F}(X_k)\}$ is a convergent sequence, which proves the first part of the lemma.

On the other hand, by rearranging inequality (19), we arrive at

$$\begin{aligned} ||X_k - X_{k-1}||_F^2 &\leq 2\alpha_{k-1}\left(\mathcal{F}(X_{k-1}) - \mathcal{F}(X_k)\right) \\ &\leq 2\alpha_{\max}\left(\mathcal{F}(X_{k-1}) - \mathcal{F}(X_k)\right). \end{aligned} \tag{20}$$

Applying limits in (20) and using part (a) of this lemma, we obtain

$$\lim_{k\to\infty} ||X_{k+1} - X_k||_F = 0. \ \square$$

Now we are ready to prove the global convergence of Algorithm 1, which is established in the theorem below.

Theorem 1 *Let $\{X_k\}$ be an infinite sequence generated by the Algorithm 1. Then*

$$\lim_{k\to\infty} ||\nabla \mathcal{F}(X_{k+1})||_F = 0.$$

Proof. Firstly, let us denote by $P_k := X_{k+1}(X_{k+1} - X_k)^\top X_{k+1}$. Now, notice that

$$||P_k||_F = ||X_{k+1}(X_{k+1} - X_k)^\top X_{k+1}||_F \leq ||X_{k+1} - X_k||_F. \tag{21}$$

By applying Lemma 2 in (21), we obtain

$$\lim_{k\to\infty} ||P_k||_F = 0. \tag{22}$$

From the minimization property (14), we have the following relation

$$\mathcal{F}(X_{k+1}) \leq \mathcal{F}(X_k) - \frac{1}{2\alpha_k}||X_{k+1} - X_k||_F^2. \tag{23}$$

It follows form (9), (23), the Cauchy–Schwarz inequality and (17) that

$$\begin{aligned}
\mathcal{F}(X_{k+1}) &\leq \mathcal{F}(X_k) - \frac{1}{2\alpha_k}||X_{k+1} - X_k||_F^2 \\
&= \mathcal{F}(X_k) - \frac{1}{2\alpha_k}||P_k - \alpha_k \nabla \mathcal{F}(X_{k+1})||_F^2 \\
&= \mathcal{F}(X_k) - \frac{1}{2\alpha_k}||P_k||_F^2 + Tr[P_k^\top \nabla \mathcal{F}(X_{k+1})] - \frac{\alpha_k}{2}||\nabla \mathcal{F}(X_{k+1})||_F^2 \\
&< \mathcal{F}(X_k) + Tr[P_k^\top \nabla \mathcal{F}(X_{k+1})] - \frac{\alpha_k}{2}||\nabla \mathcal{F}(X_{k+1})||_F^2 \\
&\leq \mathcal{F}(X_k) + Tr[P_k^\top \nabla \mathcal{F}(X_{k+1})] - \frac{\alpha_{\min}}{2}||\nabla \mathcal{F}(X_{k+1})||_F^2 \\
&\leq \mathcal{F}(X_k) + 2\kappa||P_k||_F - \frac{\alpha_{\min}}{2}||\nabla \mathcal{F}(X_{k+1})||_F^2,
\end{aligned} \tag{24}$$

which implies that,

$$||\nabla \mathcal{F}(X_{k+1})||_F^2 \leq \frac{2}{\alpha_{\min}}\left(\mathcal{F}(X_k) - \mathcal{F}(X_{k+1})\right) + \frac{4\kappa}{\alpha_{\min}}||P_k||_F. \tag{25}$$

Finally, taking limits in (25) and, considering (22) and Lemma 2, we obtain

$$\lim_{k\to\infty} ||\nabla \mathcal{F}(X_{k+1})||_F = 0,$$

which completes the proof. □

5 Computational experiments

In this section, we present some numerical results to verify the practical performance of the proposed algorithm. We test our Algorithm 1 on academic problems, considering linear eigenvalue problems and minimization of sums of heterogeneous quadratic functions. We coded our simulations in Matlab (version 2017b) with double precision on a machine intel(R) CORE(TM) i7–4770, CPU 3.40 GHz with 500GB HD and 16GB RAM. We compare our approach with the Riemannian gradient method based on the Cayley transform [28][1] (OptStiefel), and with three Riemannian conjugate gradient methods RCG1a, RCG1b and RCG1b+ZH developed in [31][2]. In addition, we stop all the methods when the algorithms find a matrix $\hat{X} \in St(n,p)$ such that $||\nabla \mathcal{F}(\hat{X})||_F < 1e\text{-}4$. In all the tests, we consider Algorithm 1 with $\alpha = p$. The implementation of our algorithm is available in `http://www.optimization-online.org/DB_HTML/2021/05/8401.html`, (see the compressed postscript).

In the rest of this section, we use the following notation: *Time*, *Iter*, *Grad*, *Feasi* denote the averaged total computing time in seconds, the averaged number of iterations, the averaged residual $||\nabla \mathcal{F}(\hat{X})||_F$ and the averaged feasibility error $||\hat{X}^\top \hat{X} - I_p||_F$, respectively. In all experiment presented below, we solve ten independent instances for each pair (n,p) and then we report all these mean values. For all the computational test, we randomly generate the starting point X_0 using the Matlab command $[X_0, \sim] = \texttt{qr}(\texttt{randn}(n,p), 0)$.

[1] The OptSt Matlab code is available in `https://github.com/wenstone/OptM`

[2] The Riemannian conjugate gradient methods RCG1a, RCG1b and RCG1b+ZH can be downloaded from `http://www.optimization-online.org/DB_HTML/2016/09/5617.html`

Table 1 Eigenvalues on randomly generated dense matrices for fixed $n = 1000$.

p	**1**	**50**	**100**	**300**	**500**
			PPASt		
Nitr	9.9	18.3	19.7	21.1	26.1
Time	0.04	0.89	1.75	9.51	21.64
Grad	3.79e-5	5.61e-5	6.41e-5	6.31e-5	7.51e-5
Feasi	1.67e-16	2.13e-15	3.53e-15	7.89e-15	1.19e-14
			OptStiefel		
Nitr	141.3	273.4	293.3	295.1	336.7
Time	0.04	0.93	2.53	14.58	31.12
Grad	7.98e-5	7.86e-5	8.25e-5	8.25e-5	8.19e-5
Feasi	8.88e-17	2.59e-14	3.05e-14	1.65e-14	1.59e-14
			RCG1a		
Nitr	143.9	298.0	303.5	339.6	358.9
Time	0.06	1.59	4.39	30.64	78.21
Grad	8.48e-5	8.34e-5	8.02e-5	8.78e-5	8.56e-5
Feasi	5.00e-16	5.75e-15	9.62e-15	2.55e-14	1.57e-14
			RCG1b		
Nitr	137.2	288.7	296.1	309.9	360.8
Time	0.05	1.48	4.01	25.34	68.74
Grad	8.19e-5	8.61e-5	8.25e-5	8.51e-5	8.37e-5
Feasi	3.55e-16	6.22e-15	1.01e-14	2.55e-14	1.57e-14
			RCG1b+ZH		
Nitr	197.0	272.2	319.0	355.7	384.8
Time	0.06	1.25	3.99	27.24	67.52
Grad	7.84e-5	8.10e-5	8.53e-5	8.60e-5	8.29e-5
Feasi	3.55e-16	7.23e-15	1.04e-14	2.85e-14	1.57e-14

5.1 Linear eigenvalues problem.

In order to illustrate the numerical behavior of our method computing some eigenvalues of a given symmetric matrix $A \in \mathbb{R}^{n\times n}$, we present a numerical experiment taken from [20]. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ be the eigenvalues of A. The p–largest eigenvalue problem can be mathematically formulated as

$$\sum_{i=1}^{p} \lambda_i = \max_{X\in\mathbb{R}^{n\times p}} Tr[X^\top AX] \quad \text{s.t.} \quad X^\top X = I_p. \tag{26}$$

In this subsection, we consider the following experiment design: given (n, p), we randomly generated dense matrices assembled as $A = 0.5(\bar{A}^\top + \bar{A})$, where $\bar{A} \in \mathbb{R}^{n\times n}$ is a matrix whose entries are sampled from the standard Gaussian distribution. Table 1 contains the computational results associated to varying $p \in \{1, 50, 100, 300, 500\}$ but fixed $n = 1000$. As shown in Tables 1 all the methods obtained estimates of a solution of problem (27) with the required precision. Furthermore, we clearly observe that as p approaches n our proposal converges more quickly, even in terms of computational time, than the rest of the methods.

5.2 Heterogeneous quadratic minimization.

In this subsection, we consider the minimization of sums of heterogeneous quadratic functions over the Stiefel manifold, this problem is formulated as

$$\min_{X\in\mathbb{R}^{n\times p}} \sum_{i=1}^{p} Tr[X_{[i]}^\top A_i X_{[i]}] \quad \text{s.t.} \quad X^\top X = I_p, \tag{27}$$

where $A_i s$ are n–by–n symmetric matrices and $X_{[i]}$ denotes the i–th column of X. For benchmarking, we consider two structures for the data matrices $A_i s$ obtained by using the following Matlab's commands:

- Structure I: $A_i = \texttt{diag}\left(\frac{(i-1)n+1}{p} : \frac{1}{p} : \frac{in}{p}\right)$, for all $i \in \{1, 2, \ldots, p\}$.
- Structure II: $A_i = \texttt{diag}\left(\frac{(i-1)n+1}{p} : \frac{1}{p} : \frac{in}{p}\right) + B_i^\top + B_i$, for all $i \in \{1, 2, \ldots, p\}$,

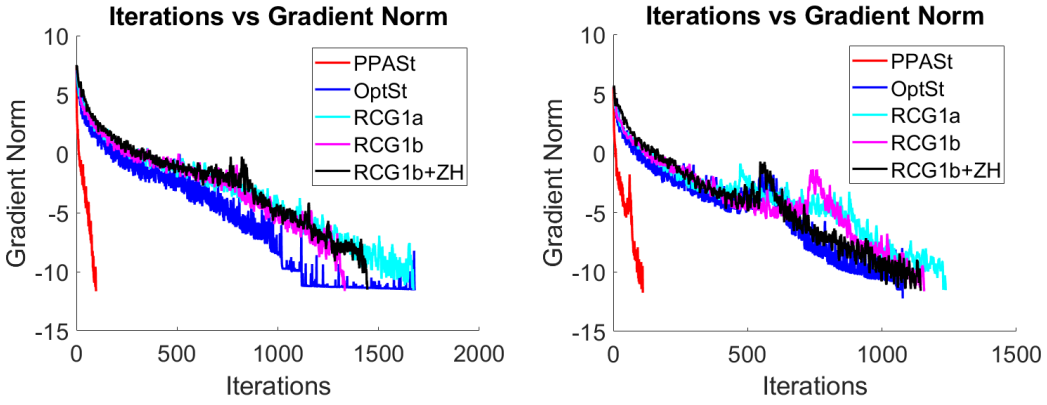
where $B_i s$ are random matrices generated by $B_i = 0.1\texttt{randn}(n)$. This experiment design was taken from [31]. The numerical results concerning the structures I and II are contained in tables 2 and 3 respectively. From Table 2, we see that the most efficient method both in terms of the number of iterations and in total computational

Table 2 Numerical results on the heterogeneous quadratics minimization, considering the Structure I.

Method	Nitr	Time	Grad	Feasi
	Structure I: $n = 10000$, $p = 10$			
PPASt	76.7	3.25	8.14e-5	8.17e-16
OptStiefel	1012.6	4.74	8.39e-5	3.98e-15
RCG1a	1216.4	9.13	8.90e-5	6.72e-14
RCG1b	1140.1	8.47	8.58e-5	5.77e-14
RCG1b+ZH	1191.9	8.45	8.36e-5	5.72e-14

Table 3 Numerical results on the heterogeneous quadratics minimization, considering the Structure II.

Method	Nitr	Time	Grad	Feasi
	Structure II: $n = 1000$, $p = 5$			
PPASt	57.6	16.09	8.21e-5	7.66e-16
OptStiefel	592.1	14.27	8.18e-5	2.86e-15
RCG1a	672.7	23.23	9.01e-5	4.44e-15
RCG1b	643.6	21.77	8.50e-5	4.05e-15
RCG1b+ZH	638.7	18.55	8.79e-5	4.98e-15



(a) Structure I with $(n, p) = (10000, 10)$ (b) Structure II with $(n, p) = (1000, 5)$

Fig. 1 Convergence behaviour of all the methods, from the same initial point, for the minimization of the heterogeneous quadratics function.

time was our procedure. The second most efficient method was OptStiefel. However, the numerical performance of PPASt and OptStiefel is very similar.

On the other hand, the results related to the Structure II show that OptStiefel is slightly superior to PPASt in terms of computational time. However our PPASt approach was much more efficient than the three Riemannian conjugate gradient methods, and took fewer number of iterations to reach the desired tolerance than the rest of the methods. In addition, in Figure 1 we plot the average convergence history of the five methods for each of the structures. From this figure, we clearly see that our proximal point method converges faster (in terms of iterations) to a local minimizer than the rest of the methods.

6 Concluding remarks.

In this article, we have introduced a new feasible method, free of exponential mapping, for solving smooth optimization problems on the Stiefel manifold. In particular, the proposal is an exact proximal point algorithm constructed with the standard distance of the matrix space $\mathbb{R}^{n\times p}$, which exploits the geometric properties of the Stiefel manifold. The proposed algorithm constructs a sequence $\{X_k\} \subset St(n, p)$ through an iterative process that successively applies a simple proximal operator to the initial point X_0. The global convergence to stationary points of the new approach is guaranteed without imposing any restrictive hypotheses on the objective function. Our computational studies show that the developed method can be a good tool to solve trace maximization problems and heterogeneous quadratic minimization problems on the Stiefel manifold.

This is a first approach that can be used as a basis for future methods, such as proximal gradient methods on the Stiefel manifold or inertial accelerated proximal gradient methods. These ideas will remain as future work.

Acknowledgements The research was financially supported in part by FGV (Fundação Getulio Vargas).

References

1. P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.
2. Yldenilson Torres Almeida, João Xavier da Cruz Neto, Paulo Roberto Oliveira, and João Carlos de Oliveira Souza. A modified proximal point method for dc functions on hadamard manifolds. *Computational Optimization and Applications*, pages 1–25, 2020.
3. Marianna Bolla, György Michaletzky, Gábor Tusnády, and Margit Ziermann. Extrema of sums of heterogeneous quadratic forms. *Linear Algebra and its Applications*, 269(1-3):331–365, 1998.
4. Shixiang Chen, Shiqian Ma, Anthony Man-Cho So, and Tong Zhang. Proximal gradient method for nonsmooth optimization over the stiefel manifold. *SIAM Journal on Optimization*, 30(1):210–239, 2020.
5. Shixiang Chen, Shiqian Ma, Anthony Man-Cho So, and Tong Zhang. Proximal gradient method for manifold optimization. *arXiv preprint arXiv:1811.00980*, 5(6):8, 2018.
6. David W Dreisigmeyer. Equality constraints, riemannian manifolds and direct search methods. *Submitted to SIOPT*, 2006.
7. OP Ferreira and PR Oliveira. Proximal point algorithm on riemannian manifolds. *Optimization*, 51(2):257–270, 2002.
8. Takeshi Fukaya, Ramaseshan Kannan, Yuji Nakatsukasa, Yusaku Yamamoto, and Yuka Yanagisawa. Shifted cholesky qr for computing the qr factorization of ill-conditioned matrices. *SIAM Journal on Scientific Computing*, 42(1):A477–A503, 2020.
9. Igor Grubišić and Raoul Pietersz. Efficient rank reduction of correlation matrices. *Linear algebra and its applications*, 422(2-3):629–653, 2007.
10. Jiang Hu, Xin Liu, Zai-Wen Wen, and Ya-Xiang Yuan. A brief introduction to manifold optimization. *Journal of the Operations Research Society of China*, 8(2):199–248, 2020.
11. Effrosini Kokiopoulou, Jie Chen, and Yousef Saad. Trace optimization and eigenproblems in dimension reduction methods. *Numerical Linear Algebra with Applications*, 18(3):565–602, 2011.
12. Hugo Lara, Oviedo Harry, and Jiyun Yuan. Matrix completion via a low rank factorization model and an augmented lagrangean succesive overrelaxation algorithm. *Bulletin of Computational Applied Mathematics*, 2(2), 2014.
13. Zhizhong Li, Deli Zhao, Zhouchen Lin, and Edward Y Chang. A new retraction for accelerating the riemannian three-factor low-rank matrix completion algorithm. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4530–4538, 2015.
14. Renee T Meinhold, Tyler L Hayes, and Nathan D Cahill. Efficiently computing piecewise flat embeddings for data clustering and image segmentation. In *2016 IEEE MIT Undergraduate Research Technology Conference (URTC)*, pages 1–4. IEEE, 2016.
15. Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.
16. Harry Oviedo. Global convergence of riemannian line search methods with a zhang–hager–type condition. *Technical report in* `http://www.optimization-online.org/DB_HTML/2021/03/8297.html`.
17. Harry Oviedo. Implicit steepest descent algorithm for optimization with orthogonality constraints. *Technical report in* `http://www.optimization-online.org/DB_HTML/2020/03/7682.html`.
18. Harry Oviedo and Oscar Dalmau. A scaled gradient projection method for minimization over the stiefel manifold. In *Mexican International Conference on Artificial Intelligence*, pages 239–250. Springer, 2019.
19. Harry Oviedo and Shaday Guerrero. Solving weighted orthogonal procrustes problems via a projected gradient method. *Technical report in* `http://www.optimization-online.org/DB_HTML/2021/05/8375.html`.
20. Harry Oviedo, Hugo Lara, and Oscar Dalmau. A non-monotone linear search algorithm with mixed direction on stiefel manifold. *Optimization Methods and Software*, 34(2):437–457, 2019.
21. Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239, 2014.
22. Raoul Pietersz 4 and Patrick JF Groenen. Rank reduction of correlation matrices by majorization. *Quantitative Finance*, 4(6):649–662, 2004.
23. EA Papa Quiroz and P Roberto Oliveira. Proximal point methods for quasiconvex and convex functions with bregman distances on hadamard manifolds. *J. Convex Anal*, 16(1):49–69, 2009.
24. JCO Souza and PR Oliveira. A proximal point algorithm for dc fuctions on hadamard manifolds. *Journal of Global Optimization*, 63(4):797–810, 2015.
25. Hugo Lara Urdaneta and Harry Fernando Oviedo Leon. Solving joint diagonalization problems via a riemannian conjugate gradient method in stiefel manifold. *Proceeding Series of the Brazilian Society of Computational and Applied Mathematics*, 6(2), 2018.
26. Jinhua Wang, Chong Li, Genaro Lopez, and Jen-Chih Yao. Convergence analysis of inexact proximal point algorithms on hadamard manifolds. *Journal of Global Optimization*, 61(3):553–573, 2015.
27. Jinhua Wang, Chong Li, Genaro Lopez, and Jen-Chih Yao. Proximal point algorithms on hadamard manifolds: linear convergence and finite termination. *SIAM Journal on Optimization*, 26(4):2696–2729, 2016.
28. Zaiwen Wen and Wotao Yin. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1):397–434, 2013.
29. Hongchao Zhang and William W Hager. A nonmonotone line search technique and its application to unconstrained optimization. *SIAM journal on Optimization*, 14(4):1043–1056, 2004.
30. Xin Zhang, Jinwei Zhu, Zaiwen Wen, and Aihui Zhou. Gradient type optimization methods for electronic structure calculations. *SIAM Journal on Scientific Computing*, 36(3):C265–C289, 2014.
31. Xiaojing Zhu. A riemannian conjugate gradient method for optimization on the stiefel manifold. *Computational optimization and Applications*, 67(1):73–110, 2017.