SIAM J. OPTIM.
Vol. 28, No. 1, pp. 302–332
© 2018 Society for Industrial and Applied Mathematics

Downloaded 04/19/20 to 124.16.148.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

A NEW FIRST-ORDER ALGORITHMIC FRAMEWORK FOR OPTIMIZATION PROBLEMS WITH ORTHOGONALITY CONSTRAINTS*

BIN GAO†, XIN LIU†, XIAOJUN CHEN‡, AND YA-XIANG YUAN§

Abstract. In this paper, we consider a class of optimization problems with orthogonality constraints, the feasible region of which is called the Stiefel manifold. Our new framework combines a function value reduction step with a correction step. Different from the existing approaches, the function value reduction step of our algorithmic framework searches along the standard Euclidean descent directions instead of the vectors in the tangent space of the Stiefel manifold, and the correction step further reduces the function value and guarantees a symmetric dual variable at the same time. We construct two types of algorithms based on this new framework. The first type is based on gradient reduction including the gradient reflection (GR) and the gradient projection (GP) algorithms. The other one adopts a columnwise block coordinate descent (CBCD) scheme with a novel idea for solving the corresponding CBCD subproblem inexactly. We prove that both GR/GP with a fixed step size and CBCD belong to our algorithmic framework, and any clustering point of the iterates generated by the proposed framework is a first-order stationary point. Preliminary experiments illustrate that our new framework is of great potential.

Key words. orthogonality constraint, Stiefel manifold, Householder transformation, gradient projection, block coordinate descent

AMS subject classifications. 15A18, 65F15, 65K05, 90C06

DOI. 10.1137/16M1098759

1. Introduction.

We consider numerical methods for solving the following matrix variable 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 stands for the p-by-p identity matrix, $f : \mathbb{R}^{n\times p} \longrightarrow \mathbb{R}$, with $p \le n$, satisfying the following assumption.

Assumption 1.1 (blanket assumption).

(i) f is twice differentiable. We define ρ as

$$\rho := \sup_{X\in\tilde{\mathcal{S}}} \left\|\nabla^2 f(X)\right\|_2,$$

where $\tilde{\mathcal{S}} := \{Y \mid \|Y\|_{\mathrm{F}}^2 < p+1\}$.[1]

*Received by the editors October 13, 2016; accepted for publication (in revised form) October 6, 2017; published electronically February 1, 2018.

http://www.siam.org/journals/siopt/28-1/M109875.html

Funding: The second author's research supported in part by NSFC grants 11622112, 11471325, 91530204, and 11688101, the National Center for Mathematics and Interdisciplinary Sciences, CAS, and Key Research Program of Frontier Sciences, CAS. The third author's research supported in part by Hong Kong Research Council Grant N_PolyU504/14. The fourth authors research supported in part by NSFC grants 11331012 and 11461161005.

†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 (gaobin@lsec.cc.ac.cn, liuxin@lsec.cc.ac.cn).

‡Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China (xiaojun.chen@polyu.edu.hk).

§State Key Laboratory of Scientific and Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China (yyx@lsec.cc.ac.cn).

[1] In fact, $\tilde{\mathcal{S}}$ can be defined as any given bounded open set containing $\mathcal{S}_{n,p} := \{Y \in \mathbb{R}^{n\times p} |\ Y^\top Y = I_p\}$.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

(ii) $f(X)$ can be represented as $h(X) + \mathrm{tr}(G^\top X)$, where $G \in \mathbb{R}^{n\times p}$, and $h(X)$ is orthogonal invariant, namely, $h(XQ) = h(X)$ holds for any $Q \in \mathcal{S}_{p,p}$, and $\nabla h(X) = H(X)X$, where $H : \mathbb{R}^{n\times p} \longrightarrow \mathbb{S}^n$ is a matrix function.

Here, $\mathbb{S}^n$ refers to the set of n-by-n symmetric matrices. The feasible region of problem (1.1) can be consequently denoted as $\mathcal{S}_{n,p}$. In practice, the value of ρ is often not known and difficult to estimate. Fortunately, we can overcome this difficulty in computation as shown in section 5.1.

Optimization problems of the above type with orthogonality constraints have many applications in scientific engineering computing and data science. More specifically, they play an important role in electronic structure calculations [35, 36, 34], linear eigenvalue problems [6], low-rank correlation matrix problems [14], sparse principal component analysis [39, 8], the orthogonal Procrustes problem [27, 11], etc. For other applications, we refer the interested readers to [10, 33, 17].

Remark 1.2. If $\rho = 0$, the objective function $f(X)$ reduces to a linear function $\mathrm{tr}(G^\top X)$. In this case, the solution of (1.1) has the closed form $X = -RQ^\top$, where $RSQ^\top$ is the reduced singular value decomposition[2] of G. In this paper, this special situation will not be discussed.

Assumption 1.1 is sufficient for the global convergence of our algorithmic framework. In this paper, we will not investigate how to weaken this sufficient condition. Fortunately, many interesting problems satisfy this assumption. Here are two simple examples.

Example 1.1.

$$f(X) := \frac{1}{2}\mathrm{tr}\left(X^\top AX\right) + \mathrm{tr}\left(G^\top X\right),$$

where $A \in \mathbb{S}^n$. In this case

$$\nabla f(X) = AX + G.$$

We notice that if the objective function defined in Example 1.1 takes $G = 0$, the corresponding optimization problem with orthogonality constraints (1.1) reduces to the Rayleigh–Ritz minimization which is exactly the optimization model for the eigenvalue problem. However, the problem with $G \neq 0$ is difficult to solve, even if A is positive definite. Example 1.1 is a key subproblem in the trust-region method for solving optimization problems with orthogonality constraints (see [36, (4.5)]). Therefore, it is challenging and interesting to explore efficient solvers for this problem.

Example 1.2.

$$f(X) := \frac{1}{2}\mathrm{tr}\left(X^\top AX\right) + \frac{1}{2}\sum_{i=1}^{m} q_i(z),$$

where $z = \mathrm{diag}(XX^\top)$, $q_i : \mathbb{R}^n \to \mathbb{R}$ $(i = 1, \ldots, m)$, and $A \in \mathbb{S}^n$. In this case,

$$\nabla f(X) = \left(A + \sum_{i=1}^{m} \mathrm{Diag}(\nabla q_i(z))\right) X.$$

This example often appears in electronic structure calculations [20], which is one of the most important topics in materials science.

[2]For $G \in \mathbb{R}^{n\times p}$ with $p < n$, the reduced singular value decomposition refers to $RSQ^\top$, where $R \in \mathbb{R}^{n\times p}$ and $Q \in \mathbb{R}^{p\times p}$ are orthogonal matrices and $S \in \mathbb{R}^{p\times p}$ is diagonal.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 04/19/20 to 124.16.148.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Downloaded 04/19/20 to 124.16.148.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

1.1. Overview of existing methods. In general, it is difficult to find a global solution of problem (1.1) due to the nonconvexity. In fact, finding a stationary point or a feasible point is not an easy task because it can be numerically expensive to maintain the orthogonality for large p. There are some existing infeasible methods such as the splitting method [19] or the penalty method for large-scale eigenspace computation [32]. However, the former does not guarantee global convergence, and the latter only works for a very special case. Exploring practically useful infeasible methods for optimization problems with orthogonality constraints is beyond the discussion of this paper.

Recently, some algorithms have been developed for special cases of (1.1), such as electronic structure calculations [38, 31], dominant eigenpair calculation [21, 22], computing the coupling between matrices [12]. Usually, these approaches utilize the special structures of the problems and can hardly be extended to the generic optimization problems with orthogonality constraints.

The feasible region of problem (1.1), $\mathcal{S}_{n,p}$, is usually called the Stiefel manifold [28]. Various optimization methods designed for solving optimization problems restricted on a matrix manifold can be applied to problem (1.1). For instance, gradient-based methods [23, 24, 1], conjugate gradient methods [10, 2], trust-region methods [36], Newton methods [10], quasi-Newton methods [26, 16, 15], etc. The key principle of these methods is to find a feasible point with a lower function value than that at the current iterate. In [10, 3], the authors study the geometric structure of the Stiefel manifold from the optimization point of view, and bring up a new concept, which is called "retraction," to connect previously unrelated algorithms. A map $\mathcal{R}_X : \mathcal{T}_X\mathcal{S}_{n,p} \to \mathcal{S}_{n,p}$ is called a retraction if the following properties hold:

(1) $\mathcal{R}_X(0_X) = X$, where 0_X is the origin of $\mathcal{T}_X\mathcal{S}_{n,p}$;

(2) $\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{R}_X(tZ)|_{t=0} = Z$ for all $Z \in \mathcal{T}_X\mathcal{S}_{n,p}$,

where $\mathcal{T}_X\mathcal{S}_{n,p} := \{Y \in \mathbb{R}^{n\times p} \mid Y^\top X + X^\top Y = 0\}$ is the tangent space of the Stiefel manifold $\mathcal{S}_{n,p}$ at point X. The retraction $\mathcal{R}_X$ maps a tangent vector into the manifold, so it defines an update rule to preserve the orthogonality.

There are two major classes of retractions for optimization problems with orthogonality constraints. The first one searches along the geodesic of a manifold to find a suitable trial point. Methods in this class are called geodesic-like retractions [10, 1, 3]. Calculating geodesics involves solving ordinary differential equations which often causes computational difficulties. The authors of [24] propose a quasi-geodesic updating formula based on the Cayley transformation whose main computation is to solve an n-by-n linear system. The methods in the other major class consist of two steps, line search in the tangent space and projection back to the Stiefel manifold. Thus, they are called projection-like methods [23, 3, 4]. The orthogonal projection can be calculated by QR factorization or polar decomposition. The projection-like methods coincide with the geodesic-like methods, in the special case of $p = 1$. The above mentioned retraction-based approaches, including both geodesic-like and projection-like methods, should work with a certain line search strategy, such as the Armijo inexact line search [25, 29] or a nonmonotonic line search strategy. The line search procedure is to guarantee the global convergence, but in the meantime, it induces additional function value evaluations.

Recently, Wen and Yin [33] proposed a feasible method for optimization with orthogonality constraints. In their work, an efficient way to calculate the Cayley transformation is introduced. In each iteration, it only requires one to solve a $2p \times 2p$ linear system instead of an $n \times n$ one. Combining a curvilinear search algorithm [13] with Barzilai–Borwein (BB) [5] nonmonotonic line search [37], it achieves much

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

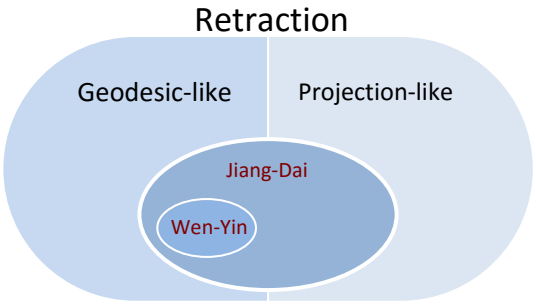


FIG. 1. *Relationship among retraction-based methods.*

lower computational cost than the other existing retraction-based algorithms and is illustrated to have robust numerical performance in solving a bunch of optimization problems with orthogonality constraints. Later on, Jiang and Dai [17] significantly extended the idea of [33], and found out that a large group of retraction-based methods enjoy such a reducible iterative formulation. It can be proved that all the algorithms under their framework with BB nonmonotonic line search are globally convergent to a stationary point.

In order to clarify the difference among the aforementioned retraction-based algorithms, we demonstrate their relationship through Figure 1.

It is worth mentioning that the retraction-based algorithms highly depend on the geometry of the Stiefel manifold and hence have very low compatibility with additional constraints such as nonnegative constraints or linear inequality constraints.

1.2. Contributions. In this paper, we revisit the first-order optimality condition of problem (1.1), and find that it is of the following form,

$$\text{(1.2)} \qquad \begin{cases} (I_n - XX^\top)\nabla f(X) = 0, & \textit{substationarity,} \\ X^\top \nabla f(X) = \nabla f(X)^\top X, & \textit{symmetry,} \\ X^\top X = I_p, & \textit{feasibility.} \end{cases}$$

For convenience, we call the three equalities of (1.2) substationarity,[3] symmetry, and feasibility, respectively. Based on the first-order optimality condition, we propose a new algorithmic framework consisting of two main steps.

The first step is function value reduction. Namely, we find a feasible point which reduces the objective function value to a certain amount in proportion to the norm square of the projected gradient. We then propose two types of algorithms which can achieve such a requirement. Gradient reflection (GR) and gradient projection (GP) are the representatives of the first type of algorithm which uses different strategies to pull a gradient descent point back to the Stiefel manifold. The second type of algorithm employs a columnwise block coordinate descent (CBCD) iteration. A novel idea for solving the corresponding subproblem efficiently is proposed.

The second step is to find a feasible point satisfying the symmetry property. This correction step, whose main calculation is a $p \times p$ singular value decomposition, is highly dependent on Assumption 1.1. The correction step can be viewed as a rotation of the trial point obtained in the first step. In the special cases when $p = 1$ or $G = 0$, the symmetry of (1.2) always holds and hence this step can be waived.

[3]Here, substationarity stands for the stationarity of the gradient of the objective function in the null space of $X^\top$.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 04/19/20 to 124.16.148.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Downloaded 04/19/20 to 124.16.148.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

It is worth mentioning that GR and GP iterations belong to particular retractions if the correction step is not necessary, but either CBCD or GR/GP with correction step is not a retraction-based iteration. According to the construction way, the proposed algorithmic framework is expected to be compatible with additional nonmanifold constraints. Our framework exposes the essential mechanism of the gradient methods for optimization problems with orthogonality constraints, with which the global convergence of gradient-based algorithms with fixed step sizes can be established. Moreover, the numerical experiments for solving a class of generic quadratic minimization problems and the instances arising from electronic structure calculations show that our new algorithmic framework performs robustly and more efficiently than the existing algorithms.

Finally, the global convergence of CBCD is of great potential itself, as this is the first convergence result for the BCD method for nonconvex optimization problems with coupled constraints.

1.3. Organization. The rest of this paper is organized as follows. In section 2, we study the first-order optimality condition of problem (1.1), and provide a new first-order framework. The main step of our new framework is only required to meet a condition for sufficient function value reduction. We then develop two types of algorithms, in section 3, to achieve this requirement and form three concrete algorithms under the scheme of the new framework, namely, GR, GP, and CBCD, respectively. Global convergence of our new algorithmic framework is established in section 4. In section 5, we demonstrate the efficiency of our algorithmic framework in solving a class of general quadratic minimization problems and the energy minimization problem arising from the electronic structure calculations. We show the great potential of our proposed approach in solving large-scale problems. Finally, concluding remarks are given in the last section.

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 pseudoinverse of X. We denote the smallest positive eigenvalue and the smallest eigenvalue in magnitude of A by $\lambda^+_{\min}(A)$ and $\lambda_{|\min|}(A)$, respectively. The ith column of matrix $X \in \mathbb{R}^{n\times p}$ is denoted by X_i. $X_{\bar{i}} \in \mathbb{R}^{n\times(p-1)}$ denotes the matrix X with its ith column removed, i.e., $X_{\bar{i}} = [X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_p]$. We use $X_{i,v} \in \mathbb{R}^{n\times p}$ to denote X with its ith column replaced by a given vector v, i.e., $X_{i,v} = [X_1, \ldots, X_{i-1}, v, X_{i+1}, \ldots, X_p]$. Finally, $\mathcal{B}(C, r)$ is the ball defined as $\{X \in \mathbb{R}^{m_1\times m_2} \mid \|X - C\|_{\mathrm{F}} \le r\}$, where $C \in \mathbb{R}^{m_1\times m_2}$ is the center and r is the radius. $\mathbf{qr}(X)$ is the Q matrix of the reduced QR decomposition[4] of X. $\mathcal{P}_{\mathcal{S}_{n,p}}(X)$ denotes the projection[5] of X to the Stiefel manifold $\mathcal{S}_{n,p}$. Finally, $\mathbf{rand}(n, p)$ and $\mathbf{randn}(n, p)$ represent $n \times p$ randomly generated matrices under independently and identically distributed (i.i.d.) uniformed distribution in $[0, 1]$ and i.i.d. standard Gaussian distribution, respectively.

2. A new first-order framework. In this section, we first give a new presentation of the first-order optimality condition of the optimization problem with

[4] $Q \in \mathbb{R}^{n\times p}$ 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.

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

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

orthogonality constraints (1.1), which motivates our new framework. The details of the new framework will also be presented.

2.1. Optimality condition. The first-order optimality condition of problem (1.1) can be interpreted as follows.

DEFINITION 2.1. *Given a point $X \in \mathbb{R}^{n\times p}$, if the relationship*

$$\begin{cases} \operatorname{tr}(Y^\top \nabla f(X)) \geq 0, \\ X^\top X = I_p, \end{cases} \tag{2.1}$$

holds for any $Y \in \mathcal{T}_X\mathcal{S}_{n,p}$, we call X a first-order stationary point of (1.1). *The set containing all the first-order stationary points is denoted as Ω_{FON}.*

Since condition (2.1) cannot be verified numerically, we show the following equivalent result.

LEMMA 2.2. *A point X is a first-order stationary point if and only if equalities* (1.2) *hold.*

Proof. We notice that any $Y \in \mathcal{T}_X\mathcal{S}_{n,p}$ can be uniquely decomposed as $Y = XS + K$,[6] where $S \in \mathbb{R}^{p\times p}$ is a skew matrix (i.e., $S^\top + S = 0$) and $K \in \mathbb{R}^{n\times p}$ satisfies $K^\top X = 0$, which is equivalent to $K = (I_n - XX^\top)K$. Likewise, any matrix of the form $XS + K$ lies in $\mathcal{T}_X\mathcal{S}_{n,p}$.

Since S and K are arbitrary, condition (2.1) is equivalent to the following relationships,

$$\operatorname{tr}\left(S^\top X^\top \nabla f(X)\right) \geq 0 \quad \forall\, S \in \mathbb{R}^{p\times p} \text{ and } S^\top + S = I_p, \tag{2.2}$$

$$\operatorname{tr}\left(K^\top \nabla f(X)\right) \geq 0 \quad \forall\, K \in \mathbb{R}^{n\times p} \text{ and } K^\top X = 0, \tag{2.3}$$

$$X^\top X = I_p.$$

By using (2.2) and the skew symmetry of $Q^\top - Q$, where $Q := X^\top \nabla f(X)$, we obtain

$$\operatorname{tr}\left(\left(Q - Q^\top\right) Q\right) \geq 0. \tag{2.4}$$

It then follows from (2.4) that

$$\begin{aligned} 0 &\leq \operatorname{tr}\left(\left(Q - Q^\top\right) Q\right) + \operatorname{tr}\left(\left(Q - Q^\top\right) Q\right) = \operatorname{tr}\left(QQ - Q^\top Q\right) + \operatorname{tr}\left(Q^\top\left(Q^\top - Q\right)\right) \\ &= \operatorname{tr}\left(QQ - Q^\top Q\right) + \operatorname{tr}\left(\left(Q^\top - Q\right) Q^\top\right) = \operatorname{tr}\left(QQ - Q^\top Q + Q^\top Q^\top - QQ^\top\right) \\ &= \operatorname{tr}\left(\left(Q - Q^\top\right)\left(Q - Q^\top\right)\right) = -\operatorname{tr}\left(\left(Q - Q^\top\right)^\top \left(Q - Q^\top\right)\right) \leq 0. \end{aligned}$$

This implies $Q = Q^\top$. On the other hand, if $X^\top \nabla f(X)$ is symmetric, the equality $\operatorname{tr}(S^\top X^\top \nabla f(X)) = 0$ holds for any skew symmetric matrix S. Hence, (2.2) is equivalent to the symmetry of $X^\top \nabla f(X)$.

Following from the property $K = (I_n - XX^\top)K$ and the arbitrariness of K, we can easily obtain the equivalence between (2.3) and $(I_n - XX^\top)\nabla f(X) = 0$. This completes the proof. ∎

[6] This is because $Y = XX^\top Y + (I_n - XX^\top)Y$ for any Y. It is not difficult to verify that $(I_n - XX^\top)Y$ satisfies $Y^\top(I_n - XX^\top)^\top X = 0$. $X^\top Y$ is a skew matrix due to the fact that $Y \in \mathcal{T}_X\mathcal{S}_{n,p} := \{Y \in \mathbb{R}^{n\times p} \mid Y^\top X + X^\top Y = 0\}$.

Downloaded 04/19/20 to 124.16.148.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 04/19/20 to 124.16.148.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Remark 2.3. It is very easy to check that our first-order optimality condition (1.2) in the Euclidean space is exactly the same as the one in the tangent space:

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

which is stated in [33]. Moreover, it actually holds that

$$\begin{aligned} \left\|\nabla f(X) - X\nabla f(X)^\top X\right\|_{\mathrm{F}}^2 = {} & \left\|\nabla f(X) - XX^\top \nabla f(X)\right\|_{\mathrm{F}}^2 \\ & + \left\|X^\top \nabla f(X) - \nabla f(X)^\top X\right\|_{\mathrm{F}}^2. \end{aligned} \tag{2.5}$$

2.2. Correction step and algorithm framework. We notice that there are three properties, substationarity, symmetry, and feasibility in our first-order optimality condition (1.2) of problem (1.1). Motivated by the relationship (2.5), to make the gradient in the tangent space equal to zero, we can adopt the following two step procedure. From the current iterate, we first find a trial point which reduces the function value in proportion to the norm square of the projected gradient. Based on this trial point, we then find the next iterate which makes the symmetry property hold without increasing the function value. Then we repeat the procedure until convergence. In these two steps, the feasibility holds all the time. The details of these two steps are described in the following.

Suppose the current iteration point is X^k. In the first step, we find an intermediate point $\bar{X} \in \mathcal{S}_{n,p}$, which satisfies sufficient function value reduction, i.e.,

$$f\left(X^k\right) - f(\bar{X}) \geq C_1 \cdot \left\|\left(I_n - X^k {X^k}^\top\right) \nabla f\left(X^k\right)\right\|_{\mathrm{F}}^2, \tag{2.6}$$

where $C_1 > 0$ is a positive constant. The right-hand side of (2.6) measures the square of the Frobenius norm of the projected gradient at X^k in the Euclidean space.

Although the intermediate point $\bar{X} \in \mathcal{S}_{n,p}$ satisfies (2.6), it does not satisfy the symmetry property in (1.2). In the second part, we consider constructing a correction step which makes the symmetry property hold without increasing the function value.

Resulting from Assumption 1.1, it holds that

$$\bar{X}^\top \nabla f(\bar{X}) = \bar{X}^\top H(\bar{X})\bar{X} + \bar{X}^\top G. \tag{2.7}$$

The term $\bar{X}^\top H(\bar{X})\bar{X}$ is symmetric. Hence, the next iterate X^{k+1} can take $\bar{X}$, if $\bar{X}^\top G$ is symmetric. Otherwise, it suffices to find a point X^{k+1} satisfying the symmetry property ${X^{k+1}}^\top G = G^\top X^{k+1}$. To achieve this, we use the rotation correction, namely, $X^{k+1} = -\bar{X}UT^\top$, where U and T come from the singular value decomposition of a $p \times p$ matrix

$$\bar{X}^\top G = U\Lambda T^\top. \tag{2.8}$$

The motivation of this correction step is to find a $p \times p$ orthogonal matrix Q^* which minimizes $f(\bar{X}Q^*)$, and then set $X^{k+1} = \bar{X}Q^*$. By recalling Assumption 1.1, we obtain $Q^* = -UT^\top$, which is the global minimizer of

$$\min_{Q \in \mathcal{S}_{p,p}} \operatorname{tr}\left(\left(\bar{X}Q\right)^\top G\right).$$

Therefore, we set the next iterate as

$$X^{k+1} = \begin{cases} \bar{X} & \text{if } \bar{X}^\top G = G^\top \bar{X}, \\ -\bar{X}UT^\top & \text{otherwise.} \end{cases} \tag{2.9}$$

We can then establish the following properties of such a correction step X^{k+1}.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 04/19/20 to 124.16.148.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

LEMMA 2.4. *Suppose $\bar{X} \in \mathcal{S}_{n,p}$. Let $\{X^{k+1}\}$ be calculated by (2.9), where U and T are determined by (2.8). Then, it holds that $X^{k+1} \in \mathcal{S}_{n,p}$ and ${X^{k+1}}^\top \nabla f(X^{k+1})$ is symmetric. Furthermore, we have*

$$8\theta \left(f(\bar{X}) - f\left(X^{k+1}\right)\right) \geq \left\|\bar{X}^\top \nabla f(\bar{X}) - \nabla f(\bar{X})^\top \bar{X}\right\|_{\mathrm{F}}^2, \tag{2.10}$$

where

$$\theta := \|G\|_2. \tag{2.11}$$

Proof. The orthogonality of X^{k+1} and the symmetry of ${X^{k+1}}^\top \nabla f(X^{k+1})$ can be directly derived by formula (2.9). Next, we prove inequality (2.10). If $\theta = 0$, which means $\nabla f(X) = H(X)X$, then the symmetry of $\bar{X}^\top \nabla f(\bar{X})$ implies (2.10) immediately. On the other hand, according to Assumption 1.1, we have

$$\begin{aligned} f(\bar{X}) - f\left(X^{k+1}\right) &= h(\bar{X}) + \mathrm{tr}\left(G^\top \bar{X}\right) - h\left(X^{k+1}\right) - \mathrm{tr}\left(G^\top X^{k+1}\right) \\ &= \mathrm{tr}\left(G^\top \bar{X} - G^\top X^{k+1}\right) \\ &= \mathrm{tr}\left(U\Lambda T^\top + \Lambda\right) = \mathrm{tr}(B + \Lambda), \end{aligned} \tag{2.12}$$

where $B = (\Lambda T^\top U + U^\top T \Lambda)/2$.

On the other hand,

$$\begin{aligned} \left\|\bar{X}^\top \nabla f(\bar{X}) - \nabla f(\bar{X})^\top \bar{X}\right\|_{\mathrm{F}}^2 &= \left\|\bar{X}^\top G - G^\top \bar{X}\right\|_{\mathrm{F}}^2 \\ &= \left\|U\Lambda T^\top - T\Lambda U^\top\right\|_{\mathrm{F}}^2 = 2\mathrm{tr}\left(\Lambda^2\right) - 2\mathrm{tr}\left(\Lambda T^\top U \Lambda T^\top U\right) \\ &= 4\mathrm{tr}\left(\Lambda^2 - B^2\right), \end{aligned} \tag{2.13}$$

where the last equality uses the fact that

$$\mathrm{tr}\left(B^2\right) = \frac{1}{2}\mathrm{tr}\left(\Lambda^2\right) + \frac{1}{2}\mathrm{tr}\left(\Lambda T^\top U \Lambda T^\top U\right).$$

Moreover, we have

$$\begin{aligned} \mathrm{tr}\left(\Lambda^2 - B^2\right) &\leq \sum_{i=1}^{p} \left(\Lambda_{ii}^2 - B_i^\top B_i\right) \leq \sum_{i=1}^{p} \left(\Lambda_{ii}^2 - B_{ii}^2\right) = \sum_{i=1}^{p} (\Lambda_{ii} - B_{ii})(\Lambda_{ii} + B_{ii}) \\ &\leq \sum_{i=1}^{p} 2\Lambda_{ii}(\Lambda_{ii} + B_{ii}) \leq 2\|\Lambda\|_2 \cdot \sum_{i=1}^{p} (\Lambda_{ii} + B_{ii}) = 2\|\Lambda\|_2 \cdot \mathrm{tr}(\Lambda + B) \\ &\leq 2\|G\|_2 \cdot \mathrm{tr}(\Lambda + B) = 2\theta \cdot \mathrm{tr}(\Lambda + B). \end{aligned} \tag{2.14}$$

Here the third inequality uses the fact that

$$|B_{ii}| = \Lambda_{ii} \cdot \left|T_i^\top U_i\right| \leq \Lambda_{ii}.$$

Combining (2.12)–(2.14), we complete the proof. ∎

We can adopt

$$c(X) := \left(I_n - XX^\top\right) \nabla f(X), \tag{2.15}$$

because the symmetry and the feasibility of (1.2) hold at each iteration. The complete framework can be described as the following.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Algorithm 1: First-order framework for optimization problems with orthogonality constraints.

1 Set tolerance $\epsilon > 0$; Initialize: $X^0 \in \mathcal{S}_{n,p}$; Set $k := 0$
2 **while** $\|c(X^k)\|_{\mathrm{F}} > \epsilon$ **do**
3 Based on X^k, find a feasible point $\bar{X}$ satisfying (2.6);
4 Based on $\bar{X}$, calculate a feasible point X^{k+1} by (2.9);
5 Set $k := k + 1$.
6 Return X^k.

3. Algorithms for finding $\bar{X}$ from X^k. In section 2, we propose a new algorithmic framework; however, how to find a point $\bar{X}$ satisfying sufficient function value reduction (2.6) is still open. In this section, we introduce two types of algorithms to achieve step 3 in Algorithm 1. The first type of algorithm is based on gradient descent in the Euclidean space which will be introduced in the first two subsections. The second type of algorithm adopts a columnwise coordinate descent idea, and it will be introduced in the third subsection. In the last subsection, we list the computational cost per iteration of some existing algorithms and our new proposed algorithms.

3.1. Gradient-type methods. An intuitive idea to reduce the function value in the Euclidean space is to take the gradient descent direction. Unfortunately, a trial point obtained by a gradient descent step from the current iterate may violate the orthogonality constraint. Therefore, in this section we discuss two concrete strategies to pull the trial point back to the Stiefel manifold. Each of them can be used in step 3 in Algorithm 1.

Both strategies are based on the following observation.

LEMMA 3.1. *For any* $Y \in \mathcal{B}_{X,\tau} := \mathcal{B}(X - \tau\nabla f(X), \tau\|\nabla f(X)\|_{\mathrm{F}})$, *where* $\tau \in (0, \rho^{-1})$, *it holds that*

$$f(X) - f(Y) \geq \frac{1 - \rho\tau}{2\tau} \cdot \|X - Y\|_{\mathrm{F}}^2. \tag{3.1}$$

Proof. For any $Y \in \mathcal{B}_{X,\tau}$, we can derive

$$\langle Y - X, Y - X + 2\tau\nabla f(X)\rangle \leq 0,$$

which implies

$$\begin{aligned} f(Y) &\leq f(X) + \langle Y - X, \nabla f(X)\rangle + \frac{\rho}{2}\|Y - X\|_{\mathrm{F}}^2 \\ &= f(X) + \frac{1}{2\tau} \cdot \langle Y - X, Y - X + 2\tau\nabla f(X)\rangle - \frac{\tau^{-1} - \rho}{2} \cdot \|Y - X\|_{\mathrm{F}}^2 \\ &\leq f(X) - \frac{\tau^{-1} - \rho}{2} \cdot \|Y - X\|_{\mathrm{F}}^2. \end{aligned}$$

This completes the proof. ∎

We illustrate the relationship among the feasible region, current iterate, gradient step, and the auxiliary ball $\mathcal{B}_{X,\tau}$ in Figure 2.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 04/19/20 to 124.16.148.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

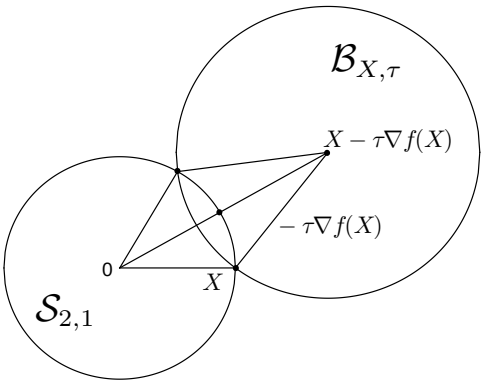


FIG. 2. *Gradient-type method.*

3.1.1. Gradient reflection. The first possible choice of a feasible trial point is to take the reflection point of the current iterate X^k reflecting on the null space of $X^k - \tau\nabla f(X^k)$. This point can be actually calculated by the Householder transformation

$$\textbf{GR}: \quad \begin{cases} V = X^k - \tau\nabla f(X^k) \quad \text{for a fixed chosen } \tau \in (0, \rho^{-1}), \\ \bar{X}_{\mathrm{GR}} = (-I_n + 2V(V^\top V)^\dagger V^\top)X^k. \end{cases} \tag{3.2}$$

Since $\bar{X}_{\mathrm{GR}}$ is the reflection point, we call Algorithm 1, using (3.2), to get $\bar{X} := \bar{X}_{\mathrm{GR}}$ in step 3, the GR.

Next, we show the intermediate point $\bar{X}_{\mathrm{GR}}$ defined in (3.2) is feasible and achieves sufficient function value reduction (2.6).

LEMMA 3.2. *Let $X^k \in \mathcal{S}_{n,p}$ and $\bar{X}_{\mathrm{GR}}$ be defined by* (3.2). *Then it holds that $\bar{X}_{\mathrm{GR}} \in \mathcal{S}_{n,p}$ and*

$$f\left(X^k\right) - f\left(\bar{X}_{\mathrm{GR}}\right) \geq \frac{2(\tau^{-1} - \rho)}{(\tau^{-1} + \rho + \theta)^2} \cdot \left\| \left(I_n - X^k {X^k}^\top\right) \nabla f\left(X^k\right) \right\|_{\mathrm{F}}^2, \tag{3.3}$$

where $\tau \in (0, \rho^{-1})$, ρ, and θ are defined in Assumption 1.1 *and equality* (2.11), *respectively.*

Proof. With a slight abuse of notation, we omit the superscript k and use X to denote X^k in this proof.

First, by simple calculation, we have

$$\bar{X}_{\mathrm{GR}}^\top \bar{X}_{\mathrm{GR}} = X^\top(-I_n + 2V(V^\top V)^\dagger V^\top)^\top(-I_n + 2V(V^\top V)^\dagger V^\top)X = I_p,$$

which implies $\bar{X}_{\mathrm{GR}} \in \mathcal{S}_{n,p}$.

Let $RSQ^\top$ be the reduced singular value decomposition of V. If $S = 0$, we have $X = \tau\nabla f(X)$, which implies that $(I_n - XX^\top)\nabla f(X) = 0$, and inequality (3.3) holds immediately. Now we consider the case that $S \neq 0$. We have

(3.4)

$$\begin{aligned} &\|\bar{X}_{\mathrm{GR}} - X\|_{\mathrm{F}} \\ &= 2\left\|\left(I_n - V(V^\top V)^\dagger V^\top\right)X\right\|_{\mathrm{F}} = 2\left\|\left(I_n - RR^\top\right)X\right\|_{\mathrm{F}} = 2\sqrt{p - \|R^\top X\|_{\mathrm{F}}^2} \\ &= 2\left\|\left(I_n - XX^\top\right)R\right\|_{\mathrm{F}} \geq 2\left\|\left(I_n - XX^\top\right)VQS^\dagger\right\|_{\mathrm{F}} \geq 2\left\|\left(I_n - XX^\top\right)VQ\right\|_{\mathrm{F}} \cdot \lambda_{\min}^+\left(S^\dagger\right) \\ &= 2\left\|\left(I_n - XX^\top\right)V\right\|_{\mathrm{F}} / \|S\|_2 = 2\tau\left\|c(X)\right\|_{\mathrm{F}} / \|V\|_2 \geq \frac{2}{\tau^{-1} + \rho + \theta}\left\|c(X)\right\|_{\mathrm{F}}, \end{aligned}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 04/19/20 to 124.16.148.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Downloaded 04/19/20 to 124.16.148.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

where $\lambda_{\min}^{+}$ and $c(X)$ are defined in section 1.4 and equality (2.15), respectively. Here, the second inequality uses the fact that all the entries of the jth column of VQ are zero for any j satisfying $S_{jj} = 0$ which is implied by the equality $RS = VQ$. The last inequality of (3.4) results from

$$\begin{aligned} \|\nabla f(X)\|_2 &\le \|H(X)X\|_2 + \|G\|_2 \le \rho + \theta, \\ \|V\|_2 &\le \|X\|_2 + \tau\|\nabla f(X)\|_2 \le 1 + \tau(\rho+\theta). \end{aligned} \tag{3.5}$$

Substituting inequality (3.4) into (3.1) in Lemma 3.1 with $Y = \bar{X}_{\mathrm{GR}}$, we arrive at

$$\begin{aligned} f(X) - f\left(\bar{X}_{\mathrm{GR}}\right) &\ge \frac{4}{(\tau^{-1}+\rho+\theta)^2} \cdot \frac{\tau^{-1}-\rho}{2} \cdot \left\|\left(I_n - XX^\top\right)\nabla f(X)\right\|_{\mathrm{F}}^2 \\ &= \frac{2(\tau^{-1}-\rho)}{(\tau^{-1}+\rho+\theta)^2} \cdot \left\|\left(I_n - XX^\top\right)\nabla f(X)\right\|_{\mathrm{F}}^2, \end{aligned} \tag{3.6}$$

which completes the proof. ∎

3.1.2. Gradient projection. Another possible choice of a feasible trial point is to directly take the projection of $X^k - \tau\nabla f(X^k)$ onto the Stiefel manifold, which can be calculated by,

$$\mathbf{GP}: \quad \begin{cases} V = X^k - \tau\nabla f(X^k) & \text{for a fixed chosen } \tau \in (0, \rho^{-1}), \\ \bar{X}_{\mathrm{GP}} = \mathcal{P}_{\mathcal{S}_{n,p}}(V). \end{cases} \tag{3.7}$$

We call Algorithm 1 using (3.7) to get $\bar{X} := \bar{X}_{\mathrm{GP}}$ in step 3 the GP. We can similarly prove the feasibility of $\bar{X}_{\mathrm{GP}}$ and show sufficient function value reduction can be achieved.

LEMMA 3.3. *Let $X^k \in \mathcal{S}_{n,p}$ and $\bar{X}_{\mathrm{GP}}$ be defined by* (3.7). *Then it holds that $\bar{X}_{\mathrm{GP}} \in \mathcal{S}_{n,p}$ and*

$$f(X^k) - f\left(\bar{X}_{\mathrm{GP}}\right) \ge \frac{\tau^{-1}-\rho}{2(\tau^{-1}+\rho+\theta)^2} \cdot \left\|\left(I_n - X^k {X^k}^\top\right)\nabla f\left(X^k\right)\right\|_{\mathrm{F}}^2, \tag{3.8}$$

where $\tau \in (0, \rho^{-1})$, ρ, and θ are defined in Assumption 1.1 *and equality* (2.11), *respectively.*

Proof. The first part of the argument can be derived in the same manner as Lemma 3.2. Here, we just focus on the proof of (3.8).

By using the singular value decomposition $V = RSQ^\top$ and the first two equalities of (3.4), we arrive at

$$\begin{aligned} &\left\|\bar{X}_{\mathrm{GP}} - X^k\right\|_{\mathrm{F}}^2 - \frac{1}{4}\left\|\bar{X}_{\mathrm{GR}} - X\right\|_{\mathrm{F}}^2 \\ &= \left\|RQ^\top - X^k\right\|_{\mathrm{F}}^2 - \left\|\left(I_n - RR^\top\right)X^k\right\|_{\mathrm{F}}^2 \\ &= \mathrm{tr}(I_p) - 2\mathrm{tr}\left(QR^\top X^k\right) + \mathrm{tr}(I_p) - \mathrm{tr}(I_p) + 2\mathrm{tr}\left({X^k}^\top RR^\top X^k\right) - \left\|R^\top X^k\right\|_{\mathrm{F}}^2 \\ &= p - 2\mathrm{tr}\left(QR^\top X^k\right) + \left\|R^\top X^k\right\|_{\mathrm{F}}^2 = \left\|Q^\top - R^\top X^k\right\|_{\mathrm{F}}^2 \ge 0, \end{aligned}$$

which implies

$$\left\|\bar{X}_{\mathrm{GP}} - X^k\right\|_{\mathrm{F}} \ge \frac{1}{2}\left\|\bar{X}_{\mathrm{GR}} - X^k\right\|_{\mathrm{F}} \ge \frac{1}{\tau^{-1}+\rho+\theta}\|c(X)\|_{\mathrm{F}}.$$

Then, we can obtain (3.8) along the lines of the proof of inequality (3.6), and then complete the proof. ∎

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 04/19/20 to 124.16.148.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

3.2. CBCD method. Another popular first-order method is block coordinate descent. For optimization problems with orthogonality constraints, a natural way to build up blocks is to partition the variables by the columns. On the other hand, the convergence of block coordinate descent with blocks coupled in nonconvex constraints cannot be guaranteed by existing results. Therefore, it is worthwhile studying the CBCD for optimization problems with orthogonality constraints. In this subsection, we consider Algorithm 1 using CBCD in step 3, discuss the way to solve the subproblem efficiently, and prove that such an approach belongs to Algorithm 1.

Once we fix the values of $p-1$ columns of X and only leave the ith column as variable, we arrive at the following subproblem:

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & f_{i,X}(x) \\ \text{s.t.} \quad & \|x\|_2 = 1, \\ & X_{\bar{i}}^\top x = 0, \end{aligned} \tag{3.9}$$

where $f_{i,X}(x) := f(X_{i,x})$, $X_{i,x}$, and $X_{\bar{i}}$ are defined in subsection 1.4.

Suppose we can obtain the solution of the above subproblem or find a feasible point x^+ with sufficient function value reduction compared to $f_{i,X}(X_i)$. Then we can use this feasible point to update our iterate columnwisely in a Gauss–Seidel manner. More specifically, if X is the current iterate, the trial point $\bar{X}$ can be calculated by the following CBCD scheme.

Algorithm 2: Columnwise block coordinate descent.

1 Set $W^0 = X$, $i := 1$;
2 **while** $i \le p$ **do**
3 Solve the subproblem (3.9) with X replaced by W^{i-1}, and obtain feaible point x^+ satisfying the following sufficient function value reduction and asymptotic small step size safeguard

$$f_{i,W^{i-1}}(X_i) - f_{i,W^{i-1}}(x^+) \ge k_1\|X_i - x^+\|_2^2, \tag{3.10}$$

$$\|X_i - x^+\|_2 \ge k_2\|(I_n - W^{i-1}{W^{i-1}}^\top)\nabla f_{i,W^{i-1}}(X_i)\|_2; \tag{3.11}$$

Set $W^i = W^{i-1}_{i,x^+}$, $i := i+1$;
4 **Return** $\bar{X} = W^p$.

Remark 3.4. Algorithm 2 actually provides a cyclic CBCD scheme, i.e., the columns are updated in a cyclic order. We can similarly implement the greedy order, stochastic order (sampling with replacement), or randomly permuted order (sampling without replacement) which often appear in classical block coordinate descent algorithms. However, as we will show in section 5, these strategies will not help to improve the performance of the cyclic CBCD. Therefore, we omit the detailed descriptions and analyses of these strategies.

Before claiming that Algorithm 2 can find $\bar{X}$ in step 3 of Algorithm 1, we need to answer two questions: can we cheaply calculate a solution or feasible point achieving sufficient function value reduction and asymptotic small step size safeguard (3.10)–(3.11)? Does Algorithm 2 provide a feasible point of problem (1.1) satisfying (2.6)? We answer these two questions in the following two subsections.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 04/19/20 to 124.16.148.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

3.2.1. Solving the CBCD subproblem. In this subsection, we discuss how to obtain a feasible trial point of subproblem (3.9) efficiently. We notice that the second constraint of (3.9) restricts the variable x lying in the null space of $X_{\bar{i}}$. Hence, we can use the variable change $x = (I_n - X_{\bar{i}}X_{\bar{i}}^\top)x$ to reduce this constraint.

First, the fact $X_{\bar{i}}^\top x = 0$ holds if and only if $x = (I_n - X_{\bar{i}}X_{\bar{i}}^\top)x$. Hence, subproblem (3.9) is equivalent to the following problem,

$$\begin{array}{ll} \min\limits_{x\in\mathbb{R}^n} & f_{i,X}\left(\left(I_n - X_{\bar{i}}X_{\bar{i}}^\top\right)x\right) \\ \text{s.t.} & \left\|(I_n - X_{\bar{i}}X_{\bar{i}}^\top)x\right\|_2 = 1. \end{array} \tag{3.12}$$

Furthermore, problem (3.12) is equivalent to a well-posed problem if it is restricted to the null space of $X_{\bar{i}}$. More specifically, we have the following proposition.

PROPOSITION 3.5. *The equivalence between problem* (3.9) *restricted to a subspace* $\mathcal{D}$ *and the following sphere constrained problem*

$$\begin{array}{ll} \min\limits_{x\in\mathbb{R}^n} & q_i(x) := f_{i,X}\left(\left(I_n - X_{\bar{i}}X_{\bar{i}}^\top\right)x\right) \\ \text{s.t.} & \|x\|_2 = 1, \\ & x \in \mathcal{D} \end{array} \tag{3.13}$$

holds, if $X_{\bar{i}}^\top x = 0$ *holds for any* $x \in \mathcal{D}$.

Proof. For any $x \in \mathcal{D}$, it holds that $x = (I_n - X_{\bar{i}}X_{\bar{i}}^\top)x$ which implies the equivalence of problems (3.13) and (3.12) restricted to the subspace $\mathcal{D}$. By using the equivalence between problems (3.12) and (3.9), we complete the proof. ∎

Proposition 3.5 tells us that we can calculate a feasible point of subproblem (3.9) with sufficient function value reduction through solving problem (3.13) if we can find a suitable subspace $\mathcal{D}$.

We notice that both X_i and $\nabla q_i(X_i) = (I_n - X_{\bar{i}}X_{\bar{i}}^\top)\nabla f_{i,X}((I_n - X_{\bar{i}}X_{\bar{i}}^\top)X_i)$ lie in the null space of $X_{\bar{i}}$. Therefore, any point in the subspace $\operatorname{span}\{X_i, \nabla q_i(X_i)\}$ satisfies the orthogonality. In other words, $\operatorname{span}\{X_i, \nabla q_i(X_i)\}$ is a qualified choice of orthogonal subspace $\mathcal{D}$ in Proposition 3.5. Considering that subproblem (3.13) with $\mathcal{D} = \operatorname{span}\{X_i, \nabla q_i(X_i)\}$ is a special case of the original optimization problem with orthogonality constraints (1.1) with $n = 2$ and $p = 1$, we recommend using the GR step (3.2) or the GP step (3.7) introduced in subsection 3.1 to calculate x^+.

It can be verified that the GR as well as the GP step satisfy sufficient function value reduction (3.10) and asymptotic small step size safeguard (3.11).

LEMMA 3.6. *Let* $x^+ = (-1 + 2v(v^\top v)^{-1}v^\top)X_i$ *or* $x^+ = (v^\top v)^{-\frac{1}{2}}v$, *where* $v = X_i - \tau\cdot\nabla q_i(X_i)$, $\tau \in (0, \rho^{-1})$. *Then* x^+ *satisfies the constraints of* (3.9) *and conditions* (3.10) *and* (3.11).

The proof of Lemma 3.6 directly follows from Lemmas 3.1, 3.2, 3.3, and the fact that $I_n - XX^\top = (I_n - X_iX_i^\top)(I_n - X_{\bar{i}}X_{\bar{i}}^\top)$, and hence it is omitted here.

Remark 3.7. If $f_{i,X}$ is quadratic, subproblem (3.13) restricted to the subspace $\operatorname{span}\{X_i, \nabla q_i(X_i)\}$ is equivalent to finding the roots of a quartic equation, which can be calculated in a closed form. In this case, the global minimizer of subproblem (3.13) restricted to the subspace $\operatorname{span}\{X_i, \nabla q_i(X_i)\}$ can be an alternative option of x^+.

3.2.2. Sufficient function value reduction. In this subsection, we show that $\bar{X}$ calculated by Algorithm 2 is a feasible point of problem (1.1) and satisfies sufficient function value reduction (2.6).

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 04/19/20 to 124.16.148.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

LEMMA 3.8. *Let $X \in \mathcal{S}_{n,p}$ and $\bar{X}$ be calculated by Algorithm* 2. *Then it holds that $\bar{X} \in \mathcal{S}_{n,p}$ and*

$$f(X) - f(\bar{X}) \geq \frac{k_1 k_2^2}{\left(1 + (p-1)k_2\left((1+\sqrt{2})\rho + \sqrt{2}\theta\right)\right)^2} \cdot \left\|\left(I_n - XX^\top\right)\nabla f(X)\right\|_{\mathrm{F}}^2. \tag{3.14}$$

Proof. The feasibility of $\bar{X}$ directly follows from the cyclic Gauss–Seidel-type update and the constraints of subproblem (3.9).

Now, we prove the second part. First, we have

$$\begin{aligned} f(X) - f(\bar{X}) &= f\left(W^0\right) - f\left(W^p\right) = \sum_{i=1}^{p}\left(f\left(W^{i-1}\right) - f\left(W^i\right)\right) \\ &= \sum_{i=1}^{p}\left(f_{i,W^{i-1}}\left(W_i^{i-1}\right) - f_{i,W^{i-1}}\left(W_i^i\right)\right) \end{aligned} \tag{3.15}$$

and

$$f_{i,W^{i-1}}\left(W_i^{i-1}\right) - f_{i,W^{i-1}}\left(W_i^i\right) \geq k_1 k_2^2 \left\|\left(I_n - W^{i-1}{W^{i-1}}^\top\right)\nabla f_{i,W^{i-1}}\left(W_i^{i-1}\right)\right\|_2^2. \tag{3.16}$$

By using the Lipschitz continuity and the boundedness of gradient (3.5), we have

$$\begin{aligned} &\left\|\left(I_n - W^{i-1}{W^{i-1}}^\top\right)\nabla f_{i,W^{i-1}}\left(W_i^{i-1}\right)\right\|_2 \\ &\quad\geq \left\|\left(I_n - W^0{W^0}^\top\right)\nabla f_{i,W^{i-1}}\left(W_i^{i-1}\right)\right\|_2 \\ &\qquad - \left\|\left(W^{i-1}{W^{i-1}}^\top - W^0{W^0}^\top\right)\nabla f_{i,W^{i-1}}\left(W_i^{i-1}\right)\right\|_2 \\ &\quad\geq \left\|\left(I_n - XX^\top\right)\nabla f_{i,W^{i-1}}(X_i)\right\|_2 \\ &\qquad - \sum_{j=1}^{i-1}\left\|W^j{W^j}^\top - W^{j-1}{W^{j-1}}^\top\right\|_{\mathrm{F}} \cdot \left\|\nabla f_{i,W^{i-1}}\left(W_i^{i-1}\right)\right\|_2 \\ &\quad\geq \left\|\left(I_n - XX^\top\right)\nabla f_{i,W^0}(X_i)\right\|_2 - \left\|\left(I_n - XX^\top\right)\left(\nabla f_{i,W^0}(X_i) - \nabla f_{i,W^{i-1}}(X_i)\right)\right\|_2 \\ &\qquad - (\rho+\theta)\sum_{j=1}^{i-1}\left(\sqrt{2 - 2\left({W_j^j}^\top W_j^{j-1}\right)^2}\right) \\ &\quad\geq \left\|\left(I_n - XX^\top\right)\nabla f_{i,X}(X_i)\right\|_2 - \left\|\nabla f_{i,W^0}(X_i) - \nabla f_{i,W^{i-1}}(X_i)\right\|_2 \\ &\qquad - \sqrt{2}(\rho+\theta)\cdot\sum_{j=1}^{i-1}\left(\sqrt{2 - 2\left({W_j^j}^\top W_j^{j-1}\right)}\right) \\ &\quad\geq \left\|\left(I_n - XX^\top\right)\nabla f_{i,X}(X_i)\right\|_2 - \rho\cdot\left\|W^0 - W^{i-1}\right\|_2 - \sqrt{2}(\rho+\theta)\sum_{j=1}^{i-1}\left\|W_j^j - W_j^{j-1}\right\|_2 \\ &\quad\geq \left\|\left(I_n - XX^\top\right)\nabla f_{i,X}(X_i)\right\|_2 \\ &\qquad - \left(\left(1+\sqrt{2}\right)\rho + \sqrt{2}\theta\right)\sqrt{k_1}^{-1}\sum_{j=1}^{i-1}\sqrt{f_{j,W^{j-1}}\left(W_j^{j-1}\right) - f_{j,W^{j-1}}\left(W_j^j\right)}, \end{aligned} \tag{3.17}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

where the third last inequality uses the facts $|W_i^{i-1^\top} W_i^i| \leq 1$ and $\sqrt{2-2\delta^2} \leq 2\sqrt{1-\delta}$ $(\forall |\delta| \leq 1)$. Together with (3.16), we have

(3.18)
$$\begin{aligned}\sqrt{f_{i,W^{i-1}}(W_i^{i-1}) - f_{i,W^{i-1}}\left(W_i^i\right)} &\geq \sqrt{k_1}k_2 \left\| \left(I_n - W^{i-1}W^{i-1^\top}\right) \nabla f_{i,W^{i-1}}\left(W_i^{i-1}\right) \right\|_2 \\ &\geq \sqrt{k_1}k_2 \left\| \left(I_n - XX^\top\right) \nabla f_{i,X}(X_i) \right\|_2 \\ &\quad - k_2 \left(\left(1+\sqrt{2}\right)\rho + \sqrt{2}\theta \right) \sum_{j=1}^{i-1} \sqrt{f_{j,W^{j-1}}\left(W_j^{j-1}\right) - f_{j,W^{j-1}}\left(W_j^j\right)}.\end{aligned}$$

Let $\delta_j := \sqrt{f_{j,W^{j-1}}(W_j^{j-1}) - f_{j,W^{j-1}}(W_j^j)}$, $c := k_2((1+\sqrt{2})\rho + \sqrt{2}\theta)$, substituting relationship (3.18) into the fact that

$$\left(\delta_i + c\sum_{j=1}^{i-1} \delta_j \right)^2 \leq (1+(i-1)c)\,\delta_i^2 + \sum_{j=1}^{i-1} c\,(1+(i-1)c)\,\delta_j^2,$$

we obtain

$$(1+(i-1)c)\,\delta_i^2 + \sum_{j=1}^{i-1} c\,(1+(i-1)c)\,\delta_j^2 \geq k_1 k_2^2 \left\| \left(I_n - XX^\top\right) \nabla f_{i,X}(X_i) \right\|_2^2. \tag{3.19}$$

Summing up inequality (3.19) from $i=1$ to p, and recalling (3.15), we arrive at

(3.20)
$$\begin{aligned}&\left(1 + (p-1)k_2\left(\left(1+\sqrt{2}\right)\rho + \sqrt{2}\theta\right)\right)^2 \sum_{i=1}^{p} \left(\sqrt{f_{i,W^{i-1}}\left(W_i^{i-1}\right) - f_{i,W^{i-1}}\left(W_i^i\right)} \right)^2 \\ &\quad \geq \sum_{i=1}^{p} k_1 k_2^2 \cdot \left\| \left(I_n - XX^\top\right) \nabla f_{i,X}(X_i) \right\|_2^2 = k_1 k_2^2 \cdot \left\| \left(I_n - XX^\top\right) \nabla f(X) \right\|_{\mathrm{F}}^2,\end{aligned}$$

which implies

$$f(X) - f(\bar{X}) \geq \frac{k_1 k_2^2}{\left(1 + (p-1)k_2\left((1+\sqrt{2})\rho + \sqrt{2}\theta\right)\right)^2} \cdot \left\| \left(I_n - XX^\top\right) \nabla f(X) \right\|_{\mathrm{F}}^2.$$

This completes the proof. ∎

A byproduct of the proof of Lemma 3.8 is the following asymptotic small step size safeguard property of CBCD.

COROLLARY 3.9. *Let $X \in \mathcal{S}_{n,p}$ and $\bar{X}$ be calculated by Algorithm* 2. *Then it holds that*

$$\|X - \bar{X}\|_{\mathrm{F}} \geq \frac{k_2}{1 + (p-1)k_2((1+\sqrt{2})\rho + \sqrt{2}\theta)} \cdot \left\| \left(I_n - XX^\top\right) \nabla f(X) \right\|_{\mathrm{F}}. \tag{3.21}$$

Proof. Using condition (3.11), the second last inequality of (3.17), and following along the same lines of inequalities (3.18) and (3.20), we can immediately obtain the desired result. ∎

Downloaded 04/19/20 to 124.16.148.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 04/19/20 to 124.16.148.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

TABLE 1
Comparison on computational cost.

Update schemes	Computational cost	
	First τ	Subsequent τ
	Geodesic-like algorithms	
$Y_{\mathrm{geoc}}(\tau;X)$ [1]	$O(n^3)$	$O(n^3)$
$Y_{\mathrm{qgeo}}(\tau;X)$ [24]	$O(n^3)$	$O(n^3)$
$Y_{\mathrm{geoe}}(\tau;X)$ [10]	$10np^2+2np+O(p^3)$	$4np^2+\ O(p^3)$
$Y_{\mathrm{wy}}(\tau;X)$ [33]	$7np^2+2np+O(p^3)$	$4np^2+\ np+O(p^3)$
	Projection-like algorithms	
$Y_{\mathrm{qr}}(\tau;X)$ [3]	$6np^2+3np+O(p^3)$	$2np^2+2np$
$Y_{\mathrm{pd}}(\tau;X)$ [3]	$7np^2+4np+O(p^3)$	$2np^2+2np+O(p^3)$
$Y_{\mathrm{pj}}(\tau;X)$ [23]	$7np^2+4np+O(p^3)$	$3np^2+3np+O(p^3)$
$Y_{\mathrm{jd}}(\tau;X)$ [17]	$7np^2+3np+O(p^3)$	$2np^2+3np+O(p^3)$
	Our algorithms	
GR	$9np^2+4np+O(p^3)$	
GP	$7np^2+3np+O(p^3)$	
CBCD-GR	$4np^2+8np+O(p^3)$	
CBCD-GP	$4np^2+5np+O(p^3)$	

3.3. Computational cost. In this subsection, we compare the computational cost per iteration among the existing algorithms and our GR, GP, and CBCD algorithms. First of all, we clarify the computational cost of the basic linear algebra operations as the following. Given $A \in \mathbb{R}^{n\times n}$, B_1, $B_2 \in \mathbb{R}^{n\times p}$, S_1, $S_2 \in \mathbb{R}^{p\times p}$, and $x \in \mathbb{R}^n$, calculating matrix-matrix products $B_1^\top B_2$, $B_1^\top B_1$, B_1S_1, and S_1S_2 we need $2np^2$, np^2+np, $2np^2$, and $2p^3$ flops, respectively. Computing A^{-1} and S^{-1} we need $8n^3/3$ and $8p^3/3$ flops, respectively, and computing Ax needs $2n^2$ flops. Calculating the full singular value decomposition of $S \in \mathbb{R}^{p\times p}$ to a fixed precision costs $O(p^3)$ flops [30]. We assume $\nabla f(X)$ is already assembled and hence the computation of $\nabla f(X)$ is not counted in the computational cost per iteration. The other settings are similar to [17, Table 1]. We illustrate the comparison result as follows.

In Table 1, the two columns "first τ" and "subsequent τ" refer to the computational cost for the first trial point, and for subsequent trial points, respectively. However, the additional function evaluations have not been counted yet. For our GP, GR, and CBCD algorithms, line search is waived, as GR and GP converge with a fixed step size and the subproblem of CBCD only needs to be solved inexactly by one iteration. Hence, our computational cost per iteration is much cheaper than the retraction-based algorithms in general. Nevertheless, we have to point out that the computation time does not only depend on the flops count, but also an efficient use of the BLAS.

Moreover, CBCD-GR or CBCD-GP refer to the CBCD (using Algorithm 2 in step 3 of Algorithm 1) with the GR or GP updating formula used once in step 3 of Algorithm 2. We notice that the calculation of $\nabla f_{i,X}((I_n - W_{\bar{i}}^{i-1}W_{\bar{i}}^{i-1\top})X_i)$ is waived because it is equal to $\nabla f_{i,X}(X_i)$ which is implied by $W_{\bar{i}}^{i-1\top}X_i = 0$. If $f_{i,X}(X_i)$ $(i = 1,\ldots,p)$ are quadratic, and we solve the subproblem (3.13) restricted to the subspace span $\{X_i, \nabla q_i(X_i)\}$ to global optimality in Algorithm 2, the corresponding computational cost is $12np^2+3np+O(p^3)$.

4. Convergence analysis. In this section, we establish the global convergence of our new algorithmic framework, Algorithm 1. First, the function value convergence is shown.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 04/19/20 to 124.16.148.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

LEMMA 4.1. *Let $\{X^k\}$ be the iterate sequence generated by Algorithm* 1 *initiated from a point $X^0 \in \mathcal{S}_{n,p}$, then $\{f(X^k)\}$ converges.*

Proof. According to the construction of $\bar{X}$ in step 3 of Algorithm 1 and Lemma 2.4, we obtain

(4.1)
$$\begin{aligned} & f\left(X^k\right) - f\left(X^{k+1}\right) \\ & \quad = f\left(X^k\right) - f(\bar{X}) + f(\bar{X}) - f\left(X^{k+1}\right) \\ & \quad \geq C_1 \cdot \left\| \nabla f\left(X^k\right) - X^k {X^k}^\top \nabla f\left(X^k\right) \right\|_{\mathrm{F}}^2 + \frac{1}{8\theta + 1} \cdot \left\| \bar{X}^\top \nabla f(\bar{X}) - \nabla f(\bar{X})^\top \bar{X} \right\|_{\mathrm{F}}^2 \\ & \quad \geq C_1 \cdot \left\| \nabla f\left(X^k\right) - X^k \nabla f\left(X^k\right)^\top X^k \right\|_{\mathrm{F}}^2. \end{aligned}$$

Hence, $\{f(X^k)\}$ is monotonically decreasing. Since $\mathcal{S}_{n,p}$ is a compact set, $\{f(X^k)\}$ is bounded below so that we can conclude that $\{f(X^k)\}$ converges. ∎

Next, we show the iterate subsequence convergence.

THEOREM 4.2. *Let $\{X^k\}$ be the sequence generated by Algorithm* 1 *initiated from a point $X^0 \in \mathcal{S}_{n,p}$. Then there exists a convergent subsequence of $\{X^k\}$. Moreover, each accumulation point X^* of $\{X^k\}$ satisfies the first-order optimality condition of* (1.1) *defined by Definition* 2.1.

Proof. Notice that $\{X^k\}$ is bounded due to the feasibility of each iterate X^k, hence it has a convergent subsequence. Let X^* be an accumulation point of $\{X^k\}$. Due to the feasibility of X^k, X^* satisfies the feasibility equality of condition (1.2).

Recalling inequality (4.1) in the proof of Lemma 4.1 and the boundedness of $\{f(X^k)\}$, we can conclude that

$$\lim_{k \to +\infty} \left\| \nabla f\left(X^k\right) - X^k \nabla f\left(X^k\right)^\top X^k \right\|_{\mathrm{F}} = 0,$$

which implies $\left\| \nabla f(X^*) - X^* \nabla f(X^*)^\top X^* \right\|_{\mathrm{F}} = 0$. Hence, the first two equalities of condition (1.2) are satisfied at X^* as well. Combining with Lemma 2.2, we complete the proof. ∎

Lemma 4.1 and Theorem 4.2 guarantee that $f(X)$ is a constant on the accumulation point set of $\{X^k\}$. We denote this constant as f^*, and

$$\Omega_{FON}^{f^*} = \Omega_{FON} \cap \{X \mid f(X) = f^*\}, \tag{4.2}$$

where Ω_{FON} is defined in Definition 2.1.

Next, we show that the distance between X^k and $\Omega_{FON}^{f^*}$ goes to zero.

COROLLARY 4.3. *Let $\{X^k\}$ be the sequence generated by Algorithm* 1 *initiated from a point $X^0 \in \mathcal{S}_{n,p}$, then it holds that*

$$f\left(X^k\right) \geq f^* \qquad \forall\, k = 1, \ldots, \tag{4.3}$$

and

$$\lim_{k \to \infty} \operatorname{dist}\left(X^k, \Omega_{FON}^{f^*}\right) = 0. \tag{4.4}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 04/19/20 to 124.16.148.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Proof. Since $\{f(X^k)\}$ is nonincreasing, relationship (4.3) holds. Now, we assume statement (4.4) is not true. Then there exist $\delta > 0$ and a subsequence of $\{X^k\}$, denoted as $\{X^{k_j}\}$, such that

$$\operatorname{dist}\left(X^{k_j}, \Omega_{FON}^{f^*}\right) \geq \delta. \tag{4.5}$$

Since $\{X^{k_j}\}$ is bounded, there exists a convergent subsequence of $\{X^{k_j}\}$ and any accumulation point shall also satisfy the first-order optimality condition, which contradicts (4.5). ∎

5. Numerical experiments. In this section, we report the numerical performance of the algorithms based on Algorithm 1. Two types of testing problems are chosen based on Examples 1.1 and 1.2. All experiments are performed in MATLAB R2016a under a Windows 10 operating system on a Dell Optiplex 9020 personal computer with an Intel Core i7-4790 CPU at 3.6 GHz×2 and 8 GB of RAM.

5.1. Implementation details. In Lemmas 3.2 and 3.3, we show that GR and GP satisfy sufficient function value reduction (3.1) if the fixed step size τ is smaller than ρ^{-1}. However, to obtain a good estimation of ρ is often intractable, and ρ^{-1} can be very small, which leads to slow convergence. In practice, we can use an alternating BB step size introduced in [7], which has been already adopted in the retraction-based algorithm in [33]. More specifically, the updating rule for τ can be described as follows:

$$\tau := \begin{cases} \tau^{\mathrm{BB1}} & \text{for odd } k, \\ \tau^{\mathrm{BB2}} & \text{for even } k, \end{cases} \tag{5.1}$$

where

$$\tau^{\mathrm{BB1}} := \frac{\left\langle J^{k-1}, J^{k-1}\right\rangle}{\left|\left\langle J^{k-1}, K^{k-1}\right\rangle\right|}, \qquad \tau^{\mathrm{BB2}} := \frac{\left|\left\langle J^{k-1}, K^{k-1}\right\rangle\right|}{\left\langle K^{k-1}, K^{k-1}\right\rangle},$$
$$J^{k-1} = X^k - X^{k-1}, \qquad K^{k-1} = c\left(X^k\right) - c\left(X^{k-1}\right).$$

We call GR and GP with step size τ defined by (5.1) as GR-BB and GP-BB, respectively. In contrast, GR and GP with a fixed step size are called GR-F and GP-F, respectively.

CBCD will only be tested in solving quadratic problem (1.1). Therefore, in each inner iteration, the subproblem (3.13) restricted to the 2-dimensional subspace $\operatorname{span}\{X_i, \nabla q_i(X_i)\}$ can be solved to the global optimality. However, in each outer iteration, the column updating order $\{j_1, j_2, \ldots, j_p\}$ determines the way to classify different types of algorithms. Usually, there are four orders:

(a) *cyclic type*: $j_i = i$ for $i = 1, 2, \ldots, p$;
(b) *random* 1: $j_i = \lceil p \cdot \mathbf{rand}(1,1) \rceil$ for $i = 1, 2, \ldots, p$ (sampling with replacement);
(c) *random* 2: $\{j_1, j_2, \ldots, j_p\}$ is a random permutation of $\{1, 2, \ldots, p\}$ (sampling without replacement);
(d) *greedy type*: for $i = 1, 2, \ldots, p$,

$$j_i := \underset{j=1,\ldots,p}{\arg\max} \left\| \left(I_n - W^{i-1} W^{i-1\top}\right) \nabla f_j(X_j) \right\|_2.$$

The corresponding CBCD are denoted as CBCD-C, CBCD-R1, CBCD-R2, and CBCD-G, respectively.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 04/19/20 to 124.16.148.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

We have already shown that any iterate generated by any algorithm based on our new framework satisfies the symmetry and feasibility in (1.2). Hence, for the stopping criterion, we only need to check the projected gradient, $\|(I_n - XX^\top)\nabla f(X)\|_{\mathrm{F}}$. More specifically, the stopping criterion can be described as follows:

$$\left\|\left(I_n - XX^\top\right)\nabla f(X)\right\|_{\mathrm{F}} < \epsilon \left\|\nabla f\left(X^0\right) - X^0 \nabla f\left(X^0\right)^\top X^0\right\|_{\mathrm{F}}, \tag{5.2}$$

where $\epsilon > 0$ is a small number. The right-hand side of (5.2) is to match the scale of the initial projected gradient. On the other hand, convergence of first-order methods may slow down as the iterates approach a stationary point, so it is critical to detect the slowdown and stop properly. It is usually beneficial to have flexible stopping rules for identifying the situation that the algorithm gets trapped in a certain region. As suggested in [33], we use the following rule based on the relative error in addition:

$$\mathrm{tol}_k^x := \frac{\left\|X^k - X^{k+1}\right\|_{\mathrm{F}}}{\sqrt{n}} < \epsilon_x \quad \text{and} \quad \mathrm{tol}_k^f := \frac{\left|f(X^k) - f(X^{k+1})\right|}{|f(X^k)| + 1} < \epsilon_f, \tag{5.3}$$

$$\begin{aligned} &\mathrm{mean}\left(\left[\mathrm{tol}_{k-\min\{k,T\}+1}^x, \ldots, \mathrm{tol}_k^x\right]\right) < 10\epsilon_x \\ &\text{and} \quad \mathrm{mean}\left(\left[\mathrm{tol}_{k-\min\{k,T\}+1}^f, \ldots, \mathrm{tol}_k^f\right]\right) < 10\epsilon_f. \end{aligned} \tag{5.4}$$

We terminate the algorithm when one of the above three criteria (5.2)–(5.4) or a maximum iteration number MaxIter is reached. Unless otherwise specified, the default tolerance parameters are chosen as $\epsilon = 10^{-5}$, $\epsilon_x = 10^{-6}$, $\epsilon_f = 10^{-10}$, $T = 5$, and MaxIter = 3000.

5.2. Testing problems. In this subsection, we introduce two types of testing problems.

The first type of testing problem is based on Example 1.1. We consider the following quadratic minimization problems with orthogonality constraints,

$$\begin{array}{ll} \min\limits_{X \in \mathbb{R}^{n\times p}} & \frac{1}{2}\mathrm{tr}\left(X^\top A X\right) + \mathrm{tr}\left(G^\top X\right) \\ \text{s.t.} & X^\top X = I_p, \end{array} \tag{5.5}$$

where the matrices $A \in \mathbb{R}^{n\times n}$ and $G \in \mathbb{R}^{n\times p}$ are randomly generated by

$$A := P\Lambda P^\top, \tag{5.6}$$

$$G := \alpha \cdot QD, \tag{5.7}$$

where the matrices $P = \mathbf{qr}\left(\mathbf{rand}(n,n)\right) \in \mathbb{R}^{n\times n}$, $\tilde{Q} = \mathbf{rand}(n,p) \in \mathbb{R}^{n\times p}$, $Q \in \mathbb{R}^{n\times p}$, and $Q_i = \tilde{Q}_i / \|\tilde{Q}_i\|_2$ $(i = 1, 2, \ldots, p)$, and matrices $\Lambda \in \mathbb{R}^{n\times n}$ and $D \in \mathbb{R}^{p\times p}$ are diagonal matrices with

$$\Lambda_{ii} := \begin{cases} \beta^{1-i} & \text{if } \omega_i < \xi, \\ -\beta^{1-i} & \text{otherwise} \end{cases} \qquad \text{for all } i = 1, 2, \ldots, n, \tag{5.8}$$

$$D_{jj} := \zeta^{j-1} \quad \text{for all } j = 1, 2, \ldots, p, \tag{5.9}$$

where $\omega_i \in [0,1]$ $(i = 1, 2, \ldots, n)$ are randomly generated numbers. Here, $n \times p$ is the variable size; $\beta \geq 1$ is a parameter determining the decay of eigenvalues of A; $\zeta \geq 1$ is a parameter referring to the growth rate of the column's norm of G. The parameter $\alpha > 0$ represents the scale difference between the quadratic term and the linear term. When α is large, the linear term dominates the objective. The parameter $\xi \in [0,1]$ is

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 04/19/20 to 124.16.148.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

for determining the definiteness of A. Once $\xi = 1$, matrix A is positive definite, while $\xi = 0$ means A is negative definite. In contrast, unless specifically mentioned, the default setting of these parameters are $n = 3000, p = 60, \alpha = 1, \beta = 1.01, \zeta = 1.2, \xi = 1$. The initial point is chosen as $X^0 = \mathbf{qr}(\mathbf{rand}(n, p)) \in \mathbb{R}^{n\times p}$.

The second type of testing problem is a special case of Example 1.2. It is called Kohn–Sham total energy minimization which comes from electronic structure calculations [18]. The original Kohn–Sham equations are the Euler–Lagrange equations for the continuous total energy minimization problem. Under the planewave discretization scheme, the Kohn–Sham total energy can be transformed into a finite-dimensional approximation as follows:

$$E_{\text{total}}(X) = \text{tr}\left[X^\top \left(\frac{1}{2}L + V_{\text{ion}}\right) X\right] + \frac{1}{2}\varrho(X)^\top L^\dagger \varrho(X) + \varrho(X)^\top \epsilon_{\text{xc}}(\varrho(X)), \tag{5.10}$$

where $\varrho(X) := \text{diag}(XX^\top)$ denotes the charge density, and L is a finite-dimensional representation of the Laplacian operator in the planewave basis. The discretized local ionic potential can be represented by a diagonal matrix V_{ion}. And the matrix $L^\dagger$ which is the discrete form of the Hartree potential corresponds to the pseudoinverse of L. The exchange correlation function ϵ_{xc} is used to model the nonclassical and quantum interaction between electrons. We aim to solve the following total energy minimization problem,

$$\min_{X^\top X = I_p} E_{\text{total}}(X). \tag{5.11}$$

It is not difficult to verify that the gradient of the energy function is $H(X)X$, where $H(X) = L/2 + V_{ion} + \text{Diag}(L^\dagger \varrho(X)) + \text{Diag}(\mu_{\text{xc}}(\varrho(X)))$ is the Kohn–Sham Hamiltonian and $\mu_{\text{xc}}(\varrho(X)) = d\epsilon_{\text{xc}}/d\varrho(X)$.

5.3. Default settings of our algorithms. In this subsection, we determine the default settings for our GR, GP, and CBCD algorithms by numerical experiments.

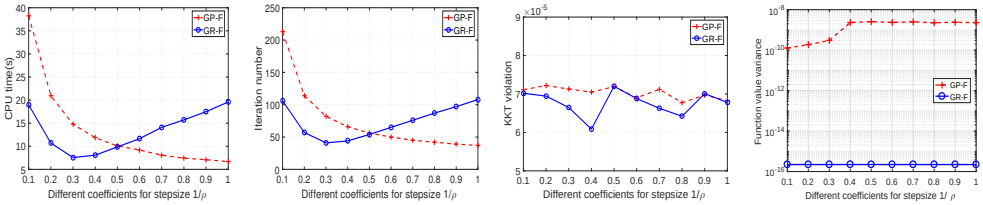
We first compare the performance of GR-F and GP-F with different fixed step sizes for choosing a proper value of the step size. The parameter p in the test is chosen as $0.1n$, the parameter ζ is 1.01, and the other parameters take their default values. We will compare four measurements: CPU time in seconds, total number of iterations, KKT violation, and function value variance, which is defined in the following. Suppose $f_{\min}$ is the smallest absolute function value of those obtained by all solvers in the comparison, then for f_s the function value returned by solver s, the function value variance is defined as

$$\frac{|f_s - f_{\min}|}{1 + |f_{\min}|} + \text{eps}, \tag{5.12}$$

where eps = 2.2204e-16 is the machine precision in MATLAB. Here, we add eps to the relative variance of function value, which is the first part of (5.12), in order to plot the variance of the function value with a logarithmic scale for the y-axis. Since both retraction-based approaches and our new framework are feasible methods, we do not report the feasibility violation $\|I - X^\top X\|_{\text{F}}$. The results with respect to the above four measurements are demonstrated in subfigures of Figure 3(a)–(d), respectively. We choose different τ's ranging from $0.1\rho^{-1}$ to ρ^{-1}.

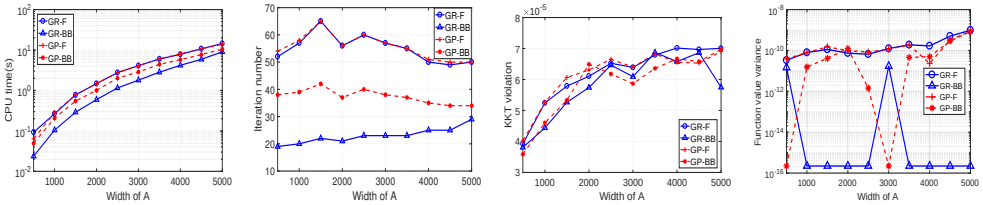
From Figure 3, we observe that $\tau = 1/3\rho$ and $1/\rho$ are the best choices for GR-F and GP-F, respectively, in this testing problem. Hence, we choose them as step sizes in the comparison with GR-BB and GP-BB.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.



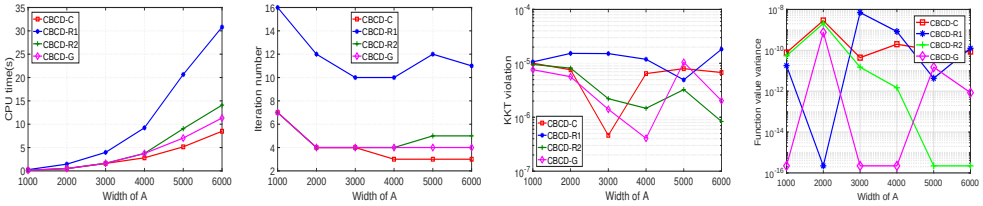
(a) CPU time (b) Iteration number (c) KKT violation (d) Function value variance

FIG. 3. *Performance of GR-F and GP-F with different step sizes.*



(a) CPU time (b) Iteration number (c) KKT violation (d) Function value variance

FIG. 4. *Performance of gradient-based algorithms.*



(a) CPU time (b) Iteration number (c) KKT violation (d) Function value variance

FIG. 5. *Performance of CBCD with different types of choosing working index.*

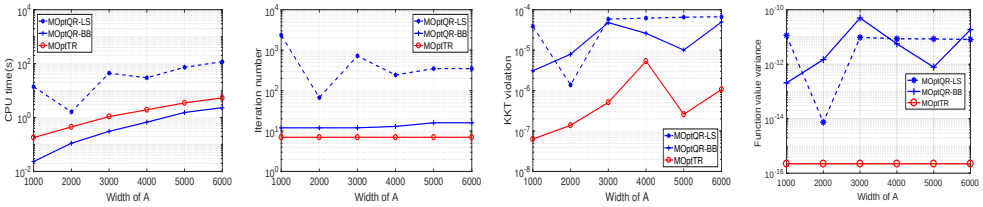
Next, we perform on a set of testing problems with ten randomly generated matrices with size n ranging from 500 to 5000, and the width of variable p is still $10\%n$. The parameter ζ is 1.01, and the other parameters take their default values. Numerical results of this test are illustrated in Figure 4.

From Figure 4, we notice that GR-BB and GP-BB require a much smaller number of iterations and less CPU time than GR-F and GP-F, and also achieve the same first-order stationary point with comparable KKT violation. Moreover, GR-BB outperforms GP-BB in terms of CPU time and iteration number in most cases. Thus, we choose GR-BB to represent the gradient-based class of algorithms in the following comparison in subsection 5.4.

We next compare the performance among CBCD variations corresponding to different updating orders. In this comparison, we run CBCD-C, CBCD-R1, CBCD-R2, and CBCD-RG to solve the testing problems with n ranging from 1000 to 6000, $p = 2\%n$, and other parameters taking their default values. The numerical results are presented in Figure 5.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 04/19/20 to 124.16.148.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php



(a) CPU time (b) Iteration number (c) KKT violation (d) Function value variance

FIG. 6. *Performance of MOptQR with different types of step sizes.*

From Figure 5, we can see that CBCD-C, CBCD-G, and CBCD-R2 have a similar performance with respect to CPU time and iteration number, and are better than CBCD-R1. Among CBCD-C, CBCD-G, and CBCD-R2, CBCD-C performs slightly better and it is easy to implement. Therefore, we will use CBCD-C to represent the CBCD class of algorithms in the following tests.

5.4. Performance comparison on random problems. In this subsection, we compare the performance of our algorithms GR-BB and CBCD-C with two state-of-the-art solvers in solving a large variety of problems (5.5). We first choose the solver OptM[7] based on the algorithm in [33]. For the other existing solver for comparison, we intend to choose one from MOptQR-LS (manifold QR method with line search[8] [3]), MOptQR-BB (for fair comparison, we implement the same alternating BB step size strategy as GR-BB for the manifold QR method), and MOptTR (manifold trust-region method [3]). We compare MOptQR-LS, MOptQR-BB, and MOptTR to solve the problem (5.5) with default settings. The result is illustrated in Figure 6.

We can learn from Figure 6 that MOptQR-BB outperforms the other two methods in the testing problems and, hence, we will choose MOptQR-BB to be the other solver to compare with our algorithms. By abuse of notation, we use MOptQR to denote MOptQR-BB hereafter.

In the following experiments, we only compare the performance between GR-BB, CBCD-C, OptM, and MOptQR. We will set the same stopping criteria as introduced in subsection 5.1, and the tolerance takes its default value. We design six groups of testing problems, in each of which there is only one parameter varying with all the others fixed. More specifically, we describe the varying parameters of each group as follows:

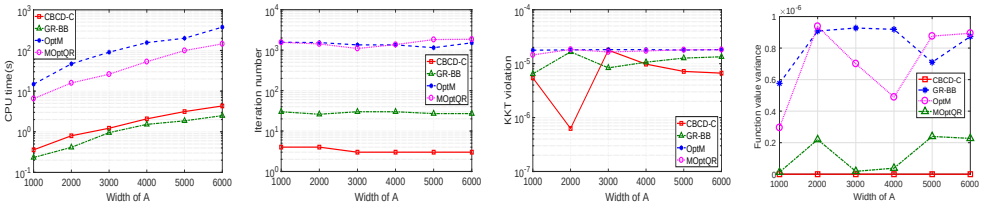
- Number of rows of the variable, $n = 1000j$ for $j = \mathbf{1}, \mathbf{2}, \mathbf{3}, \mathbf{4}, \mathbf{5}, \mathbf{6}$.
- Number of columns of the variable, $p = 20j$ for $j = \mathbf{1}, \mathbf{2}, \mathbf{3}, \mathbf{4}, \mathbf{5}, \mathbf{6}$.
- Decay of the eigenvalues of A, $\beta = 1.01 + 0.03j$ for $j = 0, \mathbf{1}, 2, 3, \mathbf{4}, 5, 6, \mathbf{7}, 8$.
- Difference between column norms of G, $\zeta = 1.01 + 0.03j$ for $j = 0, \mathbf{1}, 2, 3, \mathbf{4}, 5, 6, \mathbf{7}, 8$.
- The dominance of the linear term, $\alpha = 10^{-2}, \mathbf{10^{-1}}, \mathbf{1}, \mathbf{10}, 10^2$.
- The definiteness of A, $\xi = 0.2(j-1)$ for $j = \mathbf{1}, 2, \mathbf{3}, 4, \mathbf{5}, 6$.

The meanings of these parameters refer to equalities (5.6)–(5.9). The linear eigenvalue problem, i.e., problem (5.5) with $\alpha = 0$, is not in our testing problems. The step sizes in our new proposed GR and GP need to be tuned for different problems, and hence the algorithms do not become practically useful. On the other hand, there are a bunch

[7] Available from http://optman.blogs.rice.edu
[8] Available from http://www.manopt.org

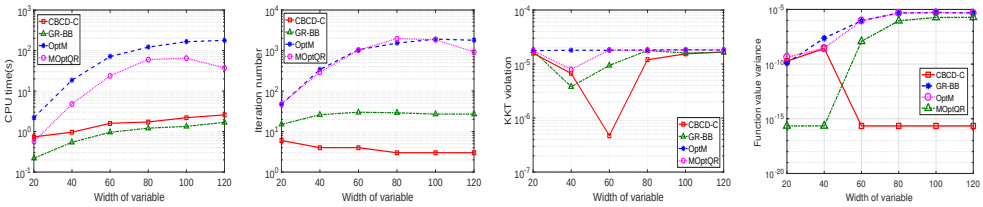
Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 04/19/20 to 124.16.148.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php



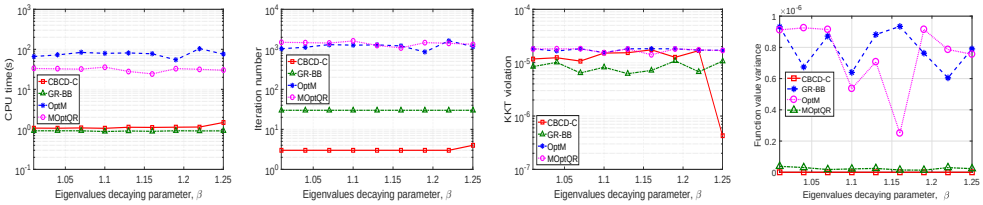
(a) CPU time (b) Iteration number (c) KKT violation (d) Function value variance

FIG. 7. *Comparison with varying matrix dimension n.*



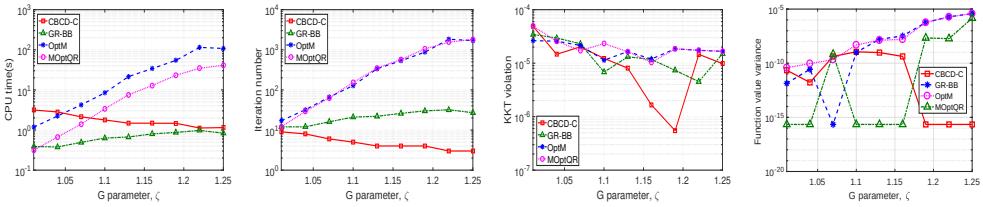
(a) CPU time (b) Iteration number (c) KKT violation (d) Function value variance

FIG. 8. *Comparison with varying width of variable p.*



(a) CPU time (b) Iteration number (c) KKT violation (d) Function value variance

FIG. 9. *Comparison with varying decay parameter β.*



(a) CPU time (b) Iteration number (c) KKT violation (d) Function value variance

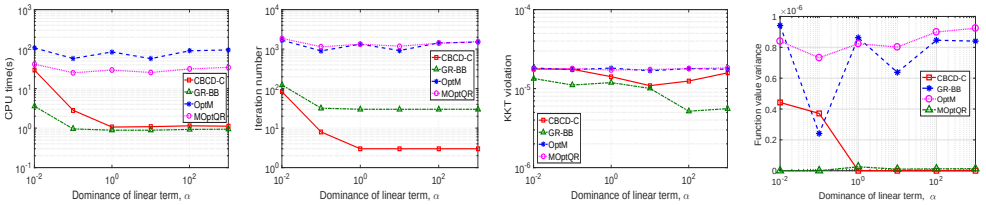
FIG. 10. *Comparison with varying G parameter ζ.*

of efficient solvers particularly designed for the linear eigenvalue problem which can hardly be beaten by general solvers for optimization problems with orthogonality constraints. The numerical results of the above six groups of testing problems are given in Figures 7 to 12, respectively.

From the above figures, we have the following observations. All solvers reach the same function value from the same initial point. They achieve comparable KKT

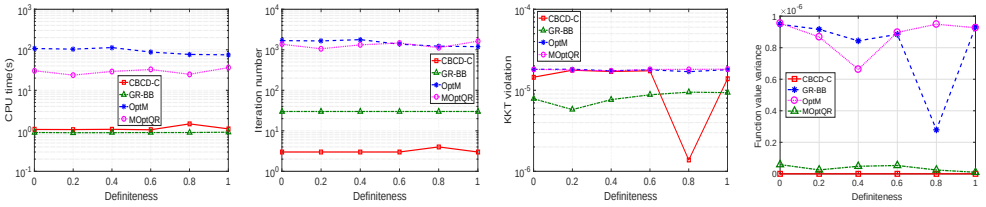
Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 04/19/20 to 124.16.148.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php



(a) CPU time (b) Iteration number (c) KKT violation (d) Function value variance

FIG. 11. *Comparison with varying dominance of linear term* α.



(a) CPU time (b) Iteration number (c) KKT violation (d) Function value variance

FIG. 12. *Comparison with varying nonnegtivity of* A, ξ.

violation with magnitude around 10^{-5}. Moreover, GR-BB and CBCD-C usually have lower KKT violation than the other two in most experiments. Among the four algorithms, CBCD-C has the lowest iteration number in all the tests, while GR-BB has the least CPU time. Except for very extreme cases, CBCD-C performs the second best in terms of CPU-time.

Finally, we select all the testing problems with options in bold in the above description, and put them into a performance profile experiment [9]. There are altogether $6 \times 6 \times 3 \times 3 \times 3 \times 3 = 2916$ randomly generated problems. The performance profile can eliminate the influence of a small number of difficult problems and the sensitivity of results associated with the different criteria, and also provide a way to visualize the expected performance difference among many solvers. We describe the key parameters of such a test as the following. For problem m and solver s, we denote $t_{m,s}$ to represent the CPU time or iteration number. Performance ratio is defined as $r_{m,s} := t_{m,s}/\min_s\{t_{m,s}\}$. If solver s fails to solve problem m, the ratio $r_{m,s}$ will be set to infinity or some sufficiently large number. Finally, the overall performance of solver s is defined by

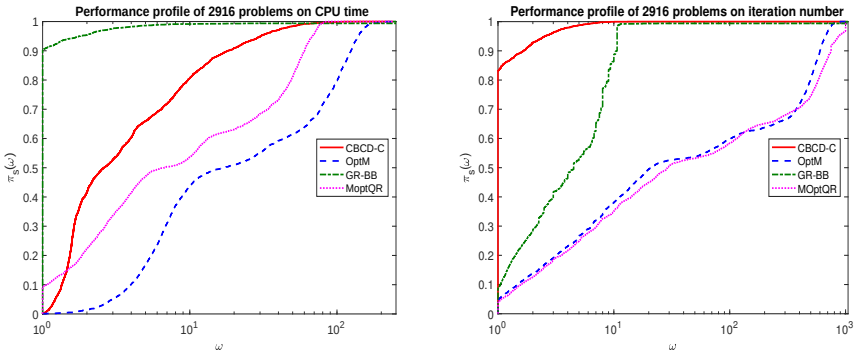
$$\pi_s(\omega) := \frac{\text{number of problems where } r_{m,s} \leq \omega}{\text{total number of problems}}.$$

It means the percentage of testing problems that can be solved in $\omega \min_s t_{m,s}$ seconds (or iterations). Of course, the closer π_s is to 1, the better performance solver s has. The performance profile results with respect to CPU time and iteration number are given in Figure 13.

We observe that GR-BB performs best and CBCD-C performs the second best among all four algorithms in solving these 2916 testing problems in CPU time, and meanwhile CBCD-C requires the smallest iteration number. In addition, we also provide the average KKT violation and feasibility over these 2916 random problems in Table 2.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 04/19/20 to 124.16.148.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php



(a) CPU time (b) Iteration number

FIG. 13. *Performance profile.*

TABLE 2
Average KKT, feasibility violation, and function value.

	CBCD-C	OptM	GR-BB	MOptQR
KKT violation	1.6075e-05	2.1730e-05	1.9501e-05	2.5072e-05
Function value variance	6.5780e-06	8.1754e-06	3.0417e-06	7.9584e-06

Table 2 shows all solvers achieve a comparable average KKT violation, feasibility, and function value variance. Here, the function value variance of solver s in solving problem m is in the same manner as (5.12). More specifically,

$$z_{m,s} := \frac{|f_{m,s} - \min_s\{f_{m,s}\}|}{1 + |\min_s\{f_{m,s}\}|}.$$

5.5. Global property of CBCD. An interesting observation of all the experiments introduced above is that all solvers reach the same function value when they converge from a randomly generated initial guess, although our problem (1.1) is nonconvex. Therefore we design a new experiment as the following. We construct the following problem

$$\begin{aligned} \min_{X\in\mathbb{R}^{3\times 2}} \ & \frac{1}{2}\mathrm{tr}\left((X - X^*)^\top A\,(X - X^*)\right) \\ \text{s. t.} \ & X^\top X = I_2, \end{aligned}$$

where

$$A = \begin{bmatrix} 13/2 & 2 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

For this special problem, we can verify that

$$X^* = \begin{bmatrix} 3/5 & 0 \\ 4/5 & 0 \\ 0 & 1 \end{bmatrix}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 04/19/20 to 124.16.148.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Downloaded 04/19/20 to 124.16.148.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

TABLE 3
Test results with initial points near X^I.

Testing methods	X^*	X^{I}	X^{II}	X^{III}	Success rate
CBCD-C	1000	0	0	0	100%
GR-BB	0	1000	0	0	0%
OptM	0	1000	0	0	0%
MOptQR	0	1000	0	0	0%

TABLE 4
Test results with initial points near X^{II}.

Testing methods	X^*	X^{I}	X^{II}	X^{III}	Success rate
CBCD-C	1000	0	0	0	100%
GR-BB	0	1000	0	0	0%
OptM	729	0	271	0	72.9%
MOptQR	78	0	922	0	7.8%

TABLE 5
Test results with initial points near X^{III}.

Testing methods	X^*	X^{I}	X^{II}	X^{III}	Success rate
CBCD-C	1000	0	0	0	100%
GR-BB	1000	0	0	0	100%
OptM	338	28	634	0	33.8%
MOptQR	5	5	990	0	0.5%

is the unique global minimizer, while

$$X^{\mathrm{I}} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad X^{\mathrm{II}} = \begin{bmatrix} 3/5 & 0 \\ 4/5 & 0 \\ 0 & -1 \end{bmatrix}, \quad X^{\mathrm{III}} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & -1 \end{bmatrix}$$

are the first-order stationary points. X^{I} is a local minimizer, while the other two are saddle points. Then we set the initial guess from the neighborhoods of the three stationary points, and run GR-BB, CBCD-C, OptM, and MOptQR to see the different performances. More specifically,

$$X^0 := \mathcal{P}_{\mathcal{S}_{n,p}}\left(X^i + \mu \cdot \mathbf{randn}(3,2)\right) \quad \text{for } i = \mathrm{I}, \mathrm{II}, \mathrm{III},$$
$$X^0 := \mathcal{P}_{\mathcal{S}_{n,p}}(\mathbf{randn}(3,2)),$$

where $\mu > 0$ controls the distance between X^0 and X^i for $i = \mathrm{I}, \mathrm{II}, \mathrm{III}$. We set $\mu = 10^{-4}$ and compare all solvers with these four types of initial points. With repeating each test 1000 times, we record the number of each solution for every solver, and the success rates are presented in Tables 3, 4, 5, and 6.

It can be observed from the above tables that the four algorithms are not necessarily convergent to same stationary points. In our tests, CBCD-C can always find the global minimizer. We are not sure whether it is a coincidence or CBCD-C has the

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

TABLE 6
Test results with random initial guesses.

Testing methods	X^*	X^{I}	X^{II}	X^{III}	Success rate
CBCD-C	1000	0	0	0	100%
GR-BB	656	344	0	0	65.6%
OptM	864	136	0	0	86.4%
MOptQR	774	226	0	0	77.4%

nice property of converging to a global minimizer with great probability. The random initialization does increase the chance to find a global minimizer for the other three algorithms.

5.6. Kohn–Sham total energy minimization. In the end of this section, we compare GR-BB with the state-of-the-art solvers in solving Kohn–Sham total energy minimization. Our test is based on the best MATLAB platform, to the best of our knowledge, for electronic structure calculations, KSSOLV [34]. KSSOLV has a friendly interface and allows researchers to investigate their own algorithms easily for different steps in electronic structure calculations. Currently, the most widely used algorithm for (5.11) is the self-consistent field (SCF) iteration, which is provided in KSSOLV. This is an iterative method for solving the nonlinear eigenvalue problem (KKT system of (5.11) briefly). Other methods focusing on discretized Kohn–Sham total energy minimization including direct constrained minimization [35] and its improved version, trust-region direct constrained minimization (TRDCM) [36] are also integrated in KSSOLV. TRDCM combines the trust-region and the subspace strategies to this special optimization problem with orthogonality constraints, and its trust-region subproblems restricted to a subspace are solved by SCF. GR-BB and MOptQR are selected in this comparison as general solvers for optimization problems with orthogonality constraints.

We select 18 testing problems with respect to different molecules, which are assembled in KSSOLV. We run methods SCF and TRDCM with $\epsilon = 10^{-5}, \mathrm{MaxIter} = 200$, and other parameters taking their default values, while GR-BB and MOptQR improve their stopping criteria with $\epsilon = 10^{-5}, \epsilon_x = 10^{-9}, \epsilon_f = 10^{-13}, \mathrm{MaxIter} = 1000$ to get a comparable solution with other methods. It is worth mentioning that here the symmetry of (1.2) is already achieved, since the total energy function is homogeneous and hence without a linear term. The stopping rule is set as $\left\|(I_n - XX^\top)H(X)X\right\|_{\mathrm{F}} < \epsilon$. For all of the testing algorithms, we set the same initial guess X^0 by using the function "getX0," which is provided by KSSOLV. The numerical results are illustrated in Tables 7 and 8.

Here, "E_{tot}," "KKT violation," "Iteration," and "CPU time(s)" represent the total energy function value, the value of $\left\|(I_n - XX^\top)H(X)X\right\|_{\mathrm{F}}$, the number of iterations and the total running time in seconds, respectively. From the tables, we observe that GR-BB outperforms the other algorithms, even the heuristic ones, in most cases, and it obtains a comparable total energy function value and a lower KKT violation. In particular, in the large size problem "ctube661," GR-BB achieves the same total energy function value and same magnitude KKT violation, but requires much less CPU time than the others.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 04/19/20 to 124.16.148.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Downloaded 04/19/20 to 124.16.148.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

TABLE 7
The results in total energy minimization.

Solver	E_{tot}	KKT violation	Iteration	CPU time(s)
al, $n = 16879$, $p = 12$				
SCF	−1.5799906179e+01	8.68e-03	200	2509.48
TRDCM	−1.5803817595e+01	8.15e-06	184	1595.83
MOptQR	−1.5802118775e+01	8.42e-03	1000	2017.61
GR-BB	−1.5802922328e+01	2.05e-03	1000	2070.80
alanine, $n = 12671$, $p = 18$				
SCF	−6.1161921213e+01	9.70e-07	15	204.20
TRDCM	−6.1161921213e+01	5.91e-06	16	147.84
MOptQR	−6.1161921213e+01	8.14e-06	65	142.70
GR-BB	−6.1161921212e+01	9.78e-06	63	142.36
benzene, $n = 8407$, $p = 15$				
SCF	−3.7225751363e+01	7.85e-07	12	85.52
TRDCM	−3.7225751363e+01	7.33e-06	14	71.13
MOptQR	−3.7225751363e+01	8.38e-06	127	154.06
GR-BB	−3.7225751362e+01	9.69e-06	50	60.38
c2h6, $n = 2103$, $p = 7$				
SCF	−1.4420491322e+01	1.12e-06	11	10.09
TRDCM	−1.4420491322e+01	5.00e-06	12	7.61
MOptQR	−1.4420491322e+01	5.56e-06	49	8.53
GR-BB	−1.4420491321e+01	9.84e-06	43	7.58
c12h26, $n = 5709$, $p = 37$				
SCF	−8.1536091936e+01	1.52e-06	16	288.09
TRDCM	−8.1536091937e+01	9.48e-06	15	171.38
MOptQR	−8.1536091935e+01	9.51e-06	442	1296.05
GR-BB	−8.1536091936e+01	8.85e-06	50	157.02
co2, $n = 2103$, $p = 8$				
SCF	−3.5124395801e+01	1.50e-06	11	11.92
TRDCM	−3.5124395801e+01	7.63e-06	13	8.72
MOptQR	−3.5124395800e+01	9.03e-06	39	7.53
GR-BB	−3.5124395801e+01	6.94e-06	39	7.52
ctube661, $n = 12599$, $p = 48$				
SCF	−1.3463843176e+02	2.80e-06	13	532.25
TRDCM	−1.3463843176e+02	5.77e-06	22	787.58
MOptQR	−1.3463843177e+02	5.06e-06	533	3817.95
GR-BB	−1.3463843176e+02	9.27e-06	68	493.53
glutamine, $n = 16517$, $p = 29$				
SCF	$-9.1839425243e+01$	2.88e-06	17	616.73
TRDCM	$-9.1839425244e+01$	8.49e-06	15	479.34
MOptQR	$-9.1839425243e+01$	7.26e-06	87	570.86
GR-BB	$-9.1839425243e+01$	9.76e-06	75	499.92
graphene16, $n = 3071$, $p = 37$				
SCF	−9.3873673630e+01	5.28e-03	200	2008.61
TRDCM	−9.4046217545e+01	6.12e-06	43	313.88
MOptQR	−9.4046217540e+01	9.56e-06	693	1110.39
GR-BB	−9.4046217543+01	8.35e-06	321	513.45

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 04/19/20 to 124.16.148.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

TABLE 8
The results in total energy minimization.

Solver	E_{tot}	KKT violation	Iteration	CPU time(s)
graphene30, $n = 12279$, $p = 67$				
SCF	−1.7358503892e+02	3.18e-03	200	15344.80
TRDCM	−1.7359510505e+02	9.77e-06	62	3768.22
MOptQR	−1.6908746446e+02	3.87e+00	1000	11930.80
GR-BB	−1.7359510453e+02	1.97e-04	1000	12027.63
h2o, $n = 2103$, $p = 4$				
SCF	−1.6440507246e+01	7.78e-07	9	5.48
TRDCM	−1.6440507246e+01	8.22e-06	11	4.55
MOptQR	−1.6440507245e+01	8.43e-06	44	5.13
GR-BB	−1.6440507245e+01	9.89e-06	42	4.53
hnco, $n = 2103$, $p = 8$				
SCF	−1.6440507246e+01	7.08e-07	9	5.52
TRDCM	−1.6440507246e+01	9.64e-06	11	4.27
MOptQR	−1.6440507245e+01	9.20e-06	82	10.41
GR-BB	−1.6440507246e+01	8.64e-06	40	5.11
nic, $n = 251$, $p = 7$				
SCF	−2.3543529955e+01	1.10e-06	12	3.13
TRDCM	−2.3543529955e+01	9.33e-06	49	5.50
MOptQR	−2.3543529955e+01	8.26e-06	100	2.84
GR-BB	−2.3543529955e+01	9.56e-06	39	0.88
pentacene, $n = 44791$, $p = 51$				
SCF	−1.3189029495e+02	9.83e-07	15	2448.72
TRDCM	−1.3189029495e+02	9.67e-06	23	2706.14
MOptQR	−1.3189029495e+02	7.02e-06	355	9145.66
GR-BB	−1.3189029495e+02	9.54e-06	100	2606.81
ptnio, $n = 4609$, $p = 43$				
SCF	−2.2678884273e+02	8.25e-07	70	1079.14
TRDCM	−2.2678882962e+02	2.93e-04	200	1957.89
MOptQR	−2.2678884235e+02	2.33e-05	1000	2281.22
GR-BB	−2.2678884272e+02	9.68e-06	512	1159.91
qdot, $n = 2103$, $p = 8$				
SCF	2.7702342351e+01	3.91e-02	200	175.16
TRDCM	2.7699896368e+01	2.72e-03	200	104.80
MOptQR	3.1736592205e+01	3.96e+00	1000	135.88
GR-BB	2.7700280932e+01	7.90e-04	1000	138.98
si2h4, $n = 2103$, $p = 6$				
SCF	−6.3009750460e+00	4.98e-07	13	12.42
TRDCM	−6.3009750459e+00	7.39e-06	16	9.09
MOptQR	−6.3009750460e+00	3.83e-06	75	11.67
GR-BB	−6.3009750457e+00	6.58e-06	58	8.97
sih4, $n = 2103$, $p = 4$				
SCF	−6.1769279851e+00	8.83e-07	10	5.80
TRDCM	−6.1769279850e+00	9.59e-06	10	4.50
MOptQR	−6.1769279851e+00	3.76e-06	42	5.14
GR-BB	−6.1769279850e+00	9.03e-06	36	4.41

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 04/19/20 to 124.16.148.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

6. Conclusion. In this paper, we propose a new first-order algorithmic framework, Algorithm 1, for optimization problems with orthogonality constraints (1.1). This algorithmic framework consists of two steps. In the first step, we choose a function value reduction approach to reduce the function value and keep the feasibility at the same time, and hence the calculation related to the tangent space of the Stiefel manifold can be waived. Second, a correction step is employed to guarantee that any accumulation point of the iterates is a first-order stationary point. Moreover, for some special cases, the correction step can be waived. We introduce two classes of approaches. The difference of them is in the first step. We first put forward a gradient-based scheme, whose global convergence can be guaranteed by a fixed step size and hence line search is no longer needed. We recommend two particular algorithms, GR and GP, from this class. The second class of algorithms is called CBCD, in which the columnwise block coordinate update is conducted in a Gauss–Seidel manner. We also propose novel ideas to solve the columnwise subproblem efficiently and guarantee the global convergence. Preliminary experiments on two large classes of testing problems including Kohn–Sham total energy minimization arising from electronic structure calculations illustrate that our new algorithms have great potential.

However, how to design second-order methods to further enhance the performance and obtain local minimizers is still under investigation. Global optimality under some random assumptions is an attractive topic for future work. How to design Jacobian-type CBCD methods is very important for the parallelization, as low scalability is an inevitable bottleneck of existing approaches for solving optimization problems with orthogonality constraints.

Acknowledgments. The authors would like to thank Zhaosong Lu, Ting Kei Pong, Zaiwen Wen, and Zaikun Zhang for the insightful discussions.

REFERENCES

[1] T. E. Abrudan, J. Eriksson, and V. Koivunen, *Steepest descent algorithms for optimization under unitary matrix constraint*, IEEE Trans. Signal Process., 56 (2008), pp. 1134–1147.
[2] T. E. Abrudan, J. Eriksson, and V. Koivunen, *Conjugate gradient algorithm for optimization under unitary matrix constraint*, Signal Process., 89 (2009), pp. 1704–1714.
[3] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, Princeton, NJ, 2009.
[4] P.-A. Absil and J. Malick, *Projection-like retractions on matrix manifolds*, SIAM J. Optim., 22 (2012), pp. 135–158.
[5] J. Barzilai and J. M. Borwein, *Two-point step size gradient methods*, IMA J. Numer. Anal., 8 (1988), pp. 141–148.
[6] A. Caboussat, R. Glowinski, and V. Pons, *An augmented Lagrangian approach to the numerical solution of a non-smooth eigenvalue problem*, J. Numer. Math., 17 (2009), pp. 3–26.
[7] Y.-H. Dai and R. Fletcher, *Projected Barzilai-Borwein methods for large-scale box-constrained quadratic programming*, Numer. Math., 100 (2005), pp. 21–47.
[8] A. d'Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. G. Lanckriet, *A direct formulation for sparse PCA using semidefinite programming*, SIAM Rev., 49 (2007), pp. 434–448.
[9] E. D. Dolan and J. J. Moré, *Benchmarking optimization software with performance profiles*, Math. Program., 91 (2002), pp. 201–213.
[10] A. Edelman, T. A. Arias, and S. T. Smith, *The geometry of algorithms with orthogonality constraints*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 303–353.
[11] L. Eldén and H. Park, *A Procrustes problem on the Stiefel manifold*, Numer. Math., 82 (1999), pp. 599–619.
[12] C. Fraikin, Y. Nesterov, and P. Van Dooren, *A gradient-type algorithm optimizing the coupling between matrices*, Linear Algebra Appl., 429 (2008), pp. 1229–1242.
[13] D. Goldfarb, Z. Wen, and W. Yin, *A curvilinear search method for p-harmonic flows on spheres*, SIAM J. Imaging Sci., 2 (2009), pp. 84–109.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

[14] I. Grubišić and R. Pietersz, *Efficient rank reduction of correlation matrices*, Linear Algebra Appl., 422 (2007), pp. 629–653.
[15] W. Huang, P.-A. Absil and K. A. Gallivan, *A Riemannian BFGS method for nonconvex optimization problems*, in Numerical Mathematics and Advanced Applications, Lecture Notes Comput. Sci. Eng. 112, Springer, Cham, Switzerland, 2016, pp. 627–634, https://doi.org/10.1007/978-3-319-39929-4_60.
[16] W. Huang, K. A. Gallivan and P.-A. Absil, *A Broyden class of quasi-Newton methods for Riemannian optimization*, SIAM J. Optim., 25 (2015), pp. 1660–1685.
[17] B. Jiang and Y.-H. Dai, *A framework of constraint preserving update schemes for optimization on Stiefel manifold*, Math. Program., 153 (2015), pp. 535–575.
[18] W. Kohn and L. J. Sham, *Self-consistent equations including exchange and correlation effects*, Phys. Rev. (2), 140 (1965), pp. A1133–1138.
[19] R. Lai and S. Osher, *A splitting method for orthogonality constrained problems*, J. Sci. Comput., 58 (2014), pp. 431–449.
[20] X. Liu, X. Wang, Z. Wen, and Y. Yuan, *On the convergence of the self-consistent field iteration in Kohn–Sham density functional theory*, SIAM J. Matrix Anal. Appl., 35 (2014), pp. 546–558.
[21] X. Liu, Z. Wen, and Y. Zhang, *Limited memory block Krylov subspace optimization for computing dominant singular value decompositions*, SIAM J. Sci. Comput., 35 (2013), pp. A1641–A1668.
[22] X. Liu, Z. Wen, and Y. Zhang, *An efficient Gauss–Newton algorithm for symmetric low-rank product matrix approximations*, SIAM J. Optim., 25 (2015), pp. 1571–1608.
[23] J. H. Manton, *Optimization algorithms exploiting unitary constraints*, IEEE Trans. Signal Process., 50 (2002), pp. 635–650.
[24] Y. Nishimori and S. Akaho, *Learning algorithms utilizing quasi-geodesic flows on the Stiefel manifold*, Neurocomputing, 67 (2005), pp. 106–135.
[25] J. Nocedal and S. Wright, *Numerical Optimization*, Springer, Berlin, 2006.
[26] B. Savas and L.-H. Lim, *Quasi-Newton methods on Grassmannians and multilinear approximations of tensors*, SIAM J. Sci. Comput., 32 (2010), pp. 3352–3393.
[27] P. H. Schönemann, *A generalized solution of the orthogonal Procrustes problem*, Psychometrika, 31 (1966), pp. 1–10.
[28] E. Stiefel, *Richtungsfelder und fernparallelismus in n-dimensionalen mannigfaltigkeiten*, Comment. Math. Helv., 8 (1935), pp. 305–353.
[29] W. Sun and Y.-X. Yuan, *Optimization Theory and Methods: Nonlinear Programming*, Vol. 1, Springer, New York, 2006.
[30] L. N. Trefethen and D. Bau, *Numerical Linear Algebra*, SIAM, Philadelphia, 1997.
[31] M. Ulbrich, Z. Wen, C. Yang, D. Klöckner, and Z. Lu, *A proximal gradient method for ensemble density functional theory*, SIAM J. Sci. Comput., 37 (2015), pp. A1975–A2002.
[32] Z. Wen, C. Yang, X. Liu, and Y. Zhang, *Trace-penalty minimization for large-scale eigenspace computation*, J. Sci. Comput., 66 (2016), pp. 1175–1203.
[33] Z. Wen and W. Yin, *A feasible method for optimization with orthogonality constraints*, Math. Program., 142 (2013), pp. 397–434.
[34] C. Yang, J. C. Meza, B. Lee, and L.-W. Wang, *KSSOLV–a MATLAB toolbox for solving the Kohn-Sham equations*, ACM Trans. Math. Software, 36 (2009), 10.
[35] C. Yang, J. C. Meza, and L.-W. Wang, *A constrained optimization algorithm for total energy minimization in electronic structure calculations*, J. Comput. Phys., 217 (2006), pp. 709–721.
[36] C. Yang, J. C. Meza, and L.-W. Wang, *A trust region direct constrained minimization algorithm for the Kohn-Sham equation*, SIAM J. Sci. Comput., 29 (2007), pp. 1854–1875.
[37] H. Zhang and W. W. Hager, *A nonmonotone line search technique and its application to unconstrained optimization*, SIAM J. Optim., 14 (2004), pp. 1043–1056.
[38] X. Zhang, J. Zhu, Z. Wen, and A. Zhou, *Gradient type optimization methods for electronic structure calculations*, SIAM J. Sci. Comput., 36 (2014), pp. C265–C289.
[39] H. Zou, T. Hastie, and R. Tibshirani, *Sparse principal component analysis*, J. Comput. Graph, Statist., 15 (2006), pp. 265–286.

Downloaded 04/19/20 to 124.16.148.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.