A Nesterov-type Acceleration with Adaptive Localized Cayley Parametrization for Optimization over the Stiefel Manifold

Keita Kume, Isao Yamada
Dept. of Information and Communications Engineering, Tokyo Institute of Technology
Email:{kume,isao}@sp.ce.titech.ac.jp

***Abstract*—Despite certain singular-point issues, the Cayley parametrization (CP) has great potential to serve as a key to import many powerful strategies, developed originally for optimization over a vector space, into the task for optimization over the Stiefel manifold. In this paper, we newly present (i) a computationally efficient CP that can circumvent the singular-point issues and (ii) a Nesterov type accelerated gradient method, based on the proposed CP, with its convergence analysis. To guarantee the convergence, we also evaluate a Lipschitz constant of the gradient of the cost function in the CP domain. Numerical experiments show excellent performance of the proposed accelerated algorithm compared with the standard algorithms, e.g., the Barzilai-Borwein method and L-BFGS method, combined with a vector transport for optimization over the Stiefel manifold as a special instance of the Riemannian manifold.**

***Index Terms*—Stiefel manifold, orthogonal group, Cayley parametrization, non-convex optimization, Nesterov acceleration**

I. Introduction

We consider orthogonal constrained optimization problem[a]:

Problem I.1. *Let* $\mathrm{St}(p,N) \coloneqq \{\boldsymbol{X} \in \mathbb{R}^{N\times p} \mid \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \boldsymbol{I}_p \in \mathbb{R}^{p\times p}\}$ *with* $p \leq N$*, where* $\boldsymbol{I}_p$ *is the identity matrix. For a given continuously differentiable function* $f\colon \mathbb{R}^{N\times p} \to \mathbb{R}$*,*

$$\text{Find } \boldsymbol{X}^{\star} \in \arg\min f(\boldsymbol{X}) \text{ s.t. } \boldsymbol{X} \in \mathrm{St}(p,N) \tag{1}$$

The feasible region $\mathrm{St}(p,N)$ of (1) is called the Stiefel manifold, and particularly $\mathrm{O}(N) \coloneqq \mathrm{St}(N,N)$ is called the orthogonal group. This optimization problem has rich applications in data sciences including signal processing and machine learning, such as joint diagonalization problem [1], orthogonal Procrustes problem [2], enhancement of the generalization performance in deep neural network [3]. However, development of numerically efficient algorithms for (1) has been challenging due to the severe non-linearity of $\mathrm{St}(p,N)$.

The standard optimization strategy [4] for the Problem I.1 are realized by two steps at nth iteration: (i) find a search direction $\boldsymbol{D}_n$ in the tangent space $T_{\boldsymbol{X}_n}\mathrm{St}(p,N)$ to $\mathrm{St}(p,N)$ at the latest estimate $\boldsymbol{X}_n \in \mathrm{St}(p,N)$; (ii) assign $R_{\boldsymbol{X}_n}(\boldsymbol{D}_n) \in \mathrm{St}(p,N)$ to a new estimate $\boldsymbol{X}_{n+1}$ with a *retraction* $R_{\boldsymbol{X}_n}\colon T_{\boldsymbol{X}_n}\mathrm{St}(p,N) \to \mathrm{St}(p,N)$ as a certain approximation of *the exponential map*[b]. Many examples of retractions for $\mathrm{St}(p,N)$ are known, e.g., QR decomposition, polar decomposition and the Cayley transform [4]–[7]. Along this strategy, optimization algorithms designed originally on a vector space have been extended for (1) combined with a *vector transport* [4], which translates a tangent vector into one in the other tangent space. Such extensions include the steepest descent method, the conjugated gradient method, Newton's method, quasi-Newton's method, the Barzilai–Borwein method and the trust-region method [1], [5], [7]–[12].

Recently, to gain the speed of the convergence to a solution, the way for exploiting the ideas in *the Nesterov Accelerated Gradient (NAG)* [13] for its applications to (1) has been studied (see, e.g., [14]–[16]). The NAG [13] over a vector space, say $\mathcal{X}$, generates the sequence $(\boldsymbol{Y}_n)_{n=0}^{\infty} \subset \mathcal{X}$ as

$$\left.\begin{aligned} \boldsymbol{Y}_{n+1} &= \boldsymbol{Z}_n - \gamma_n \nabla f(\boldsymbol{Z}_n) \\ \boldsymbol{Z}_{n+1} &= \boldsymbol{Y}_{n+1} + \beta_n(\boldsymbol{Y}_{n+1} - \boldsymbol{Y}_n) \end{aligned}\right\} \tag{2}$$

with initial points $\boldsymbol{Y}_0 = \boldsymbol{Z}_0 \in \mathcal{X}$. This remarkably simple method accelerates the convergence of the gradient descent method although NAG does not guarantee a monotonically decrease of the value of f. In particular for convex problems, NAG is known to achieve the optimal convergence rate [17], in the first-order methods, with suitable stepsizes $\gamma_n > 0$ and momentum parameters $\beta_n \in \mathbb{R}$. In [18], an extension of NAG has been made for unconstrained smooth non-convex problems and its efficacy has also been examined.

Since the second step in (2) of NAG relies on the linear structure of $\mathcal{X}$, its extensions to be applicable to $\mathrm{St}(p,N) \subsetneq \mathcal{X} \coloneqq \mathbb{R}^{N\times p}$ has to face inherent difficulty caused by the loss of linearity. For example, [14], [15] proposed to overcome this issue by translating $\boldsymbol{Y}_n \in \mathrm{St}(p,N)$ into a point on the tangent space $T_{\boldsymbol{Y}_{n+1}}\mathrm{St}(p,N)$ with the inversion of $\mathrm{Exp}_{\boldsymbol{Y}_{n+1}}$, but the computation of the inversion map itself requires iterative algorithm [19] and therefore their methods are hard to be implemented for Problem I.1 as remarked in [16]. The author of [16] proposed to extend NAG for Problem I.1 by replacing the exponential map with the Cayley retraction in [7]. To ensure a monotonically decrease of the value of f, [16] employs an adaptive restart scheme, which starts the algorithm again with the latest estimate as an initial point, and numerical performance of [16] for Problem I.1 was verified experimentally. Although NAG in [16] guarantees the sequence of the gradients converges to zero for non-convex optimization on $\mathcal{X}$ [16, Theorem 4.1], the condition corresponding to [16, Eq. (4.6)] has not been justified for the algorithm [16, Alg. 4.1] because different tangent spaces

[a]The existence of a minimizer in (1) is automatically guaranteed by the compactness of $\mathrm{St}(p,N)$ and the continuity of f.

[b]The exponential map $\mathrm{Exp}_{\boldsymbol{X}}\colon T_{\boldsymbol{X}}\mathrm{St}(p,N) \to \mathrm{St}(p,N)$ at $\boldsymbol{X} \in \mathrm{St}(p,N)$ assigns a given direction $\boldsymbol{D} \in T_{\boldsymbol{X}}\mathrm{St}(p,N)$ to a point on the geodesic of $\mathrm{St}(p,N)$ with the initial velocity $\boldsymbol{D}$.

are employed at every iteration and therefore [16, Alg. 4.1], designed for Problem I.1, has no guarantee of convergence.

From a totally different view in utilizing the Cayley transform for Problem I.1 with $p = N$, we recently proposed an Adaptive Localized Cayley Parametrization technique (ALCP) in [20]. In a Cayley parametrization technique (CP) [3], [20], [21], we translate Problem I.1 with $p = N$ into the following problem[c)] with the inversion $\widetilde{\varphi}_{\boldsymbol{S}}^{-1}\colon Q_{N,N} \to \mathrm{O}(N)\setminus E_{N,N}(\boldsymbol{S})$ of a *localized Cayley transform* $\widetilde{\varphi}_{\boldsymbol{S}}$ for $\boldsymbol{S} \in \mathrm{O}(N)$ (see Def. II.1 for $\widetilde{\varphi}_{\boldsymbol{S}}, \widetilde{\varphi}_{\boldsymbol{S}}^{-1}, Q_{N,N}, E_{N,N}(\boldsymbol{S})$):

Problem I.2. *For an arbitrarily given $\epsilon > 0$, a continuously differentiable function $f\colon \mathbb{R}^{N\times N} \to \mathbb{R}$ and $\boldsymbol{S} \in \mathrm{SO}(N) \coloneqq \{\boldsymbol{X} \in \mathrm{O}(N) | \det(\boldsymbol{X}) = 1\}$,*

$$\text{Find } \boldsymbol{V}^\star \in Q_{N,N} \text{ s.t. } f\circ\widetilde{\varphi}_{\boldsymbol{S}}^{-1}(\boldsymbol{V}^\star) < \min f(\mathrm{SO}(N)) + \epsilon. \quad (3)$$

Problem I.2 is a sound relaxation of Problem I.1 because the existence of $\boldsymbol{V}^\star$ in (3) is guaranteed by the denseness of $\widetilde{\varphi}_{\boldsymbol{S}}^{-1}(Q_{N,N}) = \mathrm{O}(N) \setminus E_{N,N}(\boldsymbol{S})$ in $\mathrm{SO}(N)$ [21]. Since $Q_{N,N}$ is a vector space, we can enjoy for Problem I.2 various arts, even (2) of NAG, of optimization over a vector space. To overcome the raised issue (see [3], [21]) regarding the performance degradation around the singular point set $E_{N,N}(\boldsymbol{S})$, we proposed ALCP in [20] by changing a center point $\boldsymbol{S}$ to a certain $\boldsymbol{S}' \in \mathrm{O}(N)$ after detecting the performance degradation (see Sec. II). We can also exploit ALCP even for general case $p \le N$ by replacing $\widetilde{\varphi}_{\boldsymbol{S}}^{-1}$ and $\mathrm{SO}(N)$ with $\Xi \circ \widetilde{\varphi}_{\boldsymbol{S}}^{-1}$ and $\mathrm{St}(p, N)$ in (3), where $\Xi\colon \mathrm{O}(N) \to \mathrm{St}(p, N)$ is a canonical projection (see the beginning of Sec. III). However the computational complexity $\mathcal{O}(N^3)$ for $\Xi \circ \widetilde{\varphi}_{\boldsymbol{S}}^{-1}$ is expensive compared with that for retractions of $\mathrm{St}(p, N)$ even if $p \ll N$ as found in typical applications in signal processing and machine learning.

In this paper, we propose (i) a computationally efficient ALCP for $p \le N$ with a new *Generalized Left Localized Cayley Transform (G-L^{2}CT)* $\Phi_{\boldsymbol{S}}$; (ii) a NAG combined with the proposed ALCP.

More precisely, for parametrization of $\mathrm{St}(p, N)$, we propose G-L^{2}CT $\Phi_{\boldsymbol{S}} : \mathrm{St}(p, N) \setminus E_{N,p}(\boldsymbol{S}) \to Q_{N,p}$ centered at $\boldsymbol{S} \in \mathrm{O}(N)$ and its inversion $\Phi_{\boldsymbol{S}}^{-1} : Q_{N,p}(\boldsymbol{S}) \to \mathrm{St}(p, N)\setminus E_{N,p}(\boldsymbol{S})$ (see Def. III.1 and Prop. III.2 for $\Phi_{\boldsymbol{S}}, \Phi_{\boldsymbol{S}}^{-1}, Q_{N,p}, E_{N,p}(\boldsymbol{S})$), where $Q_{N,p}$ is a structured subspace of $Q_{N,N}$. Theorem III.3 ensures the image $\mathrm{St}(p, N) \setminus E_{N,p}(\boldsymbol{S})$ of $\Phi_{\boldsymbol{S}}^{-1}$ is dense in $\mathrm{St}(p, N)$, thus Problem I.1 with $p < N$ can be relaxed soundly into the following problem:

Problem I.3. *For an arbitrarily given $\epsilon > 0$, a continuously differentiable function $f\colon \mathbb{R}^{N\times p} \to \mathbb{R}$ and $\boldsymbol{S} \in \mathrm{O}(N)$,*

$$\text{Find } \boldsymbol{V}^\star \in Q_{N,p} \text{ s.t. } f \circ \Phi_{\boldsymbol{S}}^{-1}(\boldsymbol{V}^\star) < \min f(\mathrm{St}(p, N)) + \epsilon.$$

By imposing a special structure $\mathrm{O}_p(N) \subset \mathrm{O}(N)$ (see (4)) on center points $\boldsymbol{S}$, the computational complexity for $\Phi_{\boldsymbol{S}}^{-1}$ achieves $\mathcal{O}(Np^2+p^3)$, which is more efficient than that for $\Xi\circ\widetilde{\varphi}_{\boldsymbol{S}}^{-1}$ and is competitive to that for retractions [22] of $\mathrm{St}(p, N)$.

[c)] Strictly speaking, Problem I.2 is a relaxation of optimization over $\mathrm{SO}(N)$. For a relaxation of optimization over $\mathrm{O}(N)$, we need to solve additionally another Problem I.2 replaced $\mathrm{SO}(N)$ with $\mathrm{O}(N) \setminus \mathrm{SO}(N)$.

In Sec. IV, we propose NAG with ALCP for (1) by synchronizing the change of center points of ALCP and the adaptive restart scheme of NAG [16]. We also evaluate a Lipschitz constant of the gradient of f in the CP domain (see Lemma III.6) because the proposed NAG requires the Lipschitz constant. Unlike [16] for Problem I.1, the proposed NAG (i) has a guarantee to generate the gradient sequence of f, in the CP domain, which converges to zero (see Theorem. IV.1); (ii) does not require the inversion of neither the exponential map nor retractions.

Numerical experiments in Sec. V show that the proposed NAG outperforms the standard optimization methods provided in *Manopt* [23], e.g., the Barzilai-Borwein method and quasi-Newton's method (L-BFGS), for (1) in the scenarios of joint diagonalization and orthogonal Procrustes problem.

Notation The matrices $\boldsymbol{X}_{\mathrm{up}} \in \mathbb{R}^{p\times p}$ and $\boldsymbol{X}_{\mathrm{lo}} \in \mathbb{R}^{(N-p)\times p}$ denote the upper and the lower block matrices of $\boldsymbol{X} \in \mathbb{R}^{N\times p}$ respectively. The matrices $\boldsymbol{S}_{\mathrm{le}} \in \mathbb{R}^{N\times p}$ and $\boldsymbol{S}_{\mathrm{ri}} \in \mathbb{R}^{N\times(N-p)}$ denote the left and right block matrices of $\boldsymbol{S} \in \mathbb{R}^{N\times N}$ respectively. The matrix $\boldsymbol{I}_{N\times p} \in \mathrm{St}(p, N)$ denotes the first p columns of the identity matrix $\boldsymbol{I} \in \mathbb{R}^{N\times N}$. The norms $\|\cdot\|_2$ and $\|\cdot\|_F$ denote the spectral norm and the Frobenius norm respectively. The function $\sigma_{\min}(\cdot)$ denotes the smallest singular value of a given matrix. The function $o(\epsilon)$ denotes a matrix-valued function that satisfies $\lim_{\epsilon\to 0} \|o(\epsilon)\|_F/\epsilon = 0$.

II. Preliminary

Definition II.1 (Right localized Cayley transform for $\mathrm{O}(N)$ [20])**.** Let $\boldsymbol{S} \in \mathrm{O}(N)$, $Q_{N,N}(\boldsymbol{S}) \coloneqq \{\boldsymbol{V} \in \mathbb{R}^{N\times N} | \boldsymbol{V}^{\mathrm{T}} = -\boldsymbol{V}\}$, $E_{N,N}(\boldsymbol{S}) \coloneqq \{\boldsymbol{X} \in \mathrm{O}(N) | \det(\boldsymbol{I} + \boldsymbol{S}^{\mathrm{T}}\boldsymbol{X}) = 0\}$. The right localized Cayley transform $\widetilde{\varphi}_{\boldsymbol{S}} : \mathrm{O}(N) \setminus E_{N,N}(\boldsymbol{S}) \to Q_{N,N}(\boldsymbol{S})$ centered at $\boldsymbol{S}$ is defined by

$$(\boldsymbol{X} \in \mathrm{O}(N) \setminus E_{N,N}(\boldsymbol{S})) \quad \widetilde{\varphi}_{\boldsymbol{S}}(\boldsymbol{X}) = (\boldsymbol{S} - \boldsymbol{X})(\boldsymbol{S} + \boldsymbol{X})^{-1},$$

its inversion $\widetilde{\varphi}_{\boldsymbol{S}}^{-1}\colon Q_{N,N}(\boldsymbol{S}) \to \mathrm{O}(N) \setminus E_{N,N}(\boldsymbol{S})$ is given as

$$(\boldsymbol{V} \in Q_{N,N} \coloneqq Q_{N,N}(\boldsymbol{S})) \quad \widetilde{\varphi}_{\boldsymbol{S}}^{-1}(\boldsymbol{V}) = (\boldsymbol{I} - \boldsymbol{V})(\boldsymbol{I} + \boldsymbol{V})^{-1}\boldsymbol{S}.$$

Remark II.2 (On Def. II.1)**.** For $\boldsymbol{V} \in Q_{N,N}$, $\boldsymbol{I} + \boldsymbol{V}$ is invertible because every eigenvalue of $\boldsymbol{V}$ is pure imaginary [21]. A specialization $\widetilde{\varphi}_{\boldsymbol{I}}$ of $\widetilde{\varphi}_{\boldsymbol{S}}$ is the classical Cayley transform [20]. $Q_{N,N}$ is a vector space because it is closed under matrix addition and scalar multiplication. For every $\boldsymbol{S} \in \mathrm{O}(N)$, $Q_{N,N}(\boldsymbol{S})$ is the same set as $Q_{N,N}$, however we distinguish them later from the viewpoint of the parametrization with $\boldsymbol{S}$.

One of the first applications of the Cayley transform to Problem I.1 with $p = N$ is found in [21] where Problem I.1 is relaxed to Problem I.2. Since $Q_{N,N}(\boldsymbol{S})$ is a vector space, we can enjoy for Problem I.2 various optimization methods designed specially on a vector space. However, the convergence speed of optimization algorithms combined with CP tends to slow down severely in the case where the minimizer $\boldsymbol{X}^\star \in \mathrm{O}(N)$ is close to $E_{N,N}(\boldsymbol{S})$ [3], [21]. Since it is a priori unclear whether $\boldsymbol{X}^\star$ is close to $E_{N,N}(\boldsymbol{S})$ or not, the center point $\boldsymbol{S}$ was used as a hyper parameter in [3].

Through a quantitative analysis on the mobility of $\widetilde{\varphi}_{\boldsymbol{S}}^{-1}$ [20, Eq. (7)], we recently revealed that $\widetilde{\varphi}_{\boldsymbol{S}}^{-1}$ can be insensitive to the change on $Q_{N,N}(\boldsymbol{S})$ at points distant from zero, which is

certainly a major reason of of the performance degradation of optimization algorithms combined with CP. We also proposed, for $p = N$, the Adaptive Localized Cayley Parametrization Technique (ALCP), which changes a center point $\boldsymbol{S}$ to $\boldsymbol{S}' \coloneqq \widetilde{\varphi}_{\boldsymbol{S}}^{-1}(\boldsymbol{V}_n)$ adaptively and then restarts iterative suppression of $f \circ \widetilde{\varphi}_{\boldsymbol{S}'}^{-1}$ over $Q_{N,N}(\boldsymbol{S}')$ with the initial point $\boldsymbol{V}_n' \coloneqq \widetilde{\varphi}_{\boldsymbol{S}'} \circ \widetilde{\varphi}_{\boldsymbol{S}}^{-1}(\boldsymbol{V}_n) = \boldsymbol{0}$. More precisely, we change $\boldsymbol{S}$ in the case where the latest estimate $\boldsymbol{V}_n \in Q_{N,N}(\boldsymbol{S})$ is distant from zero. Thanks to this adaptive strategy, a center point $\boldsymbol{S}$ is no longer a hyper parameter because we can use the center point strategically not to loose a sufficient mobility in the update of the estimation.

III. Adaptive Localized Cayley Parametrization for Stiefel manifold

We can apply ALCP for $\mathrm{O}(N)$ to Problem I.1 with $p < N$ by replacing $\widetilde{\varphi}_{\boldsymbol{S}}^{-1}$ and $\mathrm{SO}(N)$ in (3) with $\Xi \circ \widetilde{\varphi}_{\boldsymbol{S}}^{-1}$ and $\mathrm{St}(p,N)$, where $\Xi\colon \mathrm{O}(N) \to \mathrm{St}(p,N)\colon \boldsymbol{X} \to \boldsymbol{X}\boldsymbol{I}_{N\times p}$. However, the computational complexity for $\Xi \circ \widetilde{\varphi}_{\boldsymbol{S}}^{-1}$ is $\mathcal{O}(N^3)$ that is expensive even in comparison with $\mathcal{O}(Np^2+p^3)$ required for retractions of $\mathrm{St}(p,N)$.

For a computationally efficient extension of ALCP to $\mathrm{St}(p,N)$, we newly propose G-L^{2}CT.

Definition III.1 (G-L^{2}CT)**.** Let $\boldsymbol{S} \in \mathrm{O}(N)$, $E_{N,p}(\boldsymbol{S}) \coloneqq \{\boldsymbol{X} \in \mathrm{St}(p,N) | \det(\boldsymbol{I} + \boldsymbol{S}_{\mathrm{le}}^{\mathrm{T}}\boldsymbol{X}) = 0\}$, and

$$Q_{N,p}(\boldsymbol{S}) \coloneqq Q_{N,p} \coloneqq \left\{ \begin{bmatrix} \boldsymbol{A} & -\boldsymbol{B}^{\mathrm{T}} \\ \boldsymbol{B} & \boldsymbol{0} \end{bmatrix} \middle| \begin{smallmatrix} \boldsymbol{A}^{\mathrm{T}} = -\boldsymbol{A} \in \mathbb{R}^{p\times p} \\ \boldsymbol{B} \in \mathbb{R}^{(N-p)\times p} \end{smallmatrix} \right\} \subset Q_{N,N}.$$

The G-L^{2}CT $\Phi_{\boldsymbol{S}}\colon \mathrm{St}(p,N) \setminus E_{N,p}(\boldsymbol{S}) \to Q_{N,p}(\boldsymbol{S})$ centered at $\boldsymbol{S}$ is defined by

$$(\boldsymbol{X} \in \mathrm{St}(p,N) \setminus E_{N,p}(\boldsymbol{S})) \quad \Phi_{\boldsymbol{S}}(\boldsymbol{X}) = \begin{bmatrix} \boldsymbol{A} & -\boldsymbol{B}^{\mathrm{T}} \\ \boldsymbol{B} & \boldsymbol{0} \end{bmatrix}$$

with $\boldsymbol{A} \coloneqq (\boldsymbol{I} + \boldsymbol{S}_{\mathrm{le}}^{\mathrm{T}}\boldsymbol{X})^{-\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{le}} - \boldsymbol{S}_{\mathrm{le}}^{\mathrm{T}}\boldsymbol{X})(\boldsymbol{I} + \boldsymbol{S}_{\mathrm{le}}^{\mathrm{T}}\boldsymbol{X})^{-1} \in \mathbb{R}^{p\times p}$ and $\boldsymbol{B} \coloneqq -\boldsymbol{S}_{\mathrm{ri}}^{\mathrm{T}}\boldsymbol{X}(\boldsymbol{I} + \boldsymbol{S}_{\mathrm{le}}^{\mathrm{T}}\boldsymbol{X})^{-1} \in \mathbb{R}^{(N-p)\times p}$.

Proposition III.2 (Inversion of G-L^{2}CT)**.** *The G-L^{2}CT $\Phi_{\boldsymbol{S}}$ centered at $\boldsymbol{S} \in \mathrm{O}(N)$ is a diffeomorphism between $\mathrm{St}(p,N)\setminus E_{N,p}(\boldsymbol{S})$ and $Q_{N,p}(\boldsymbol{S})$, and its inversion $\Phi_{\boldsymbol{S}}^{-1}\colon Q_{N,p}(\boldsymbol{S}) \to \mathrm{St}(p,N) \setminus E_{N,p}(\boldsymbol{S})$ is given by*

$$(\boldsymbol{V} \in Q_{N,p}(\boldsymbol{S})) \quad \Phi_{\boldsymbol{S}}^{-1}(\boldsymbol{V}) = \boldsymbol{S}(\boldsymbol{I} - \boldsymbol{V})(\boldsymbol{I} + \boldsymbol{V})^{-1}\boldsymbol{I}_{N\times p}.$$

We call $E_{N,p}(\boldsymbol{S})$ the singular point set of $\Phi_{\boldsymbol{S}}$. In the following, $\boldsymbol{A}(\boldsymbol{V}) \in \mathbb{R}^{p\times p}$ and $\boldsymbol{B}(\boldsymbol{V}) \in \mathbb{R}^{(N-p)\times p}$ denote the left upper and the left lower block matrices of $\boldsymbol{V} \in Q_{N,p}(\boldsymbol{S})$ respectively. When there is no possibility of confusion, we will use $\boldsymbol{A}$ and $\boldsymbol{B}$ instead of them respectively.

Theorem III.3 (Denseness of $\mathrm{St}(p,N)\setminus E_{N,p}(\boldsymbol{S})$)**.** *For $p < N$ and $\boldsymbol{S} \in \mathrm{O}(N)$, it holds $\mathrm{St}(p,N) \setminus E_{N,p}(\boldsymbol{S}) = \Xi(\mathrm{O}(N) \setminus E_{N,N}(\boldsymbol{S}))$ and $\mathrm{St}(p,N) \setminus E_{N,p}(\boldsymbol{S})$ is dense in $\mathrm{St}(p,N)$.*

Remark III.4 (On G-L^{2}CT)**.**

(a) (Specialization of $\Phi_{\boldsymbol{S}}$). A specialization of $\Phi_{\boldsymbol{S}}$ with $p = N$ and $\boldsymbol{S} = \boldsymbol{I}$ is the classical Cayley transform.

(b) (Sound relaxation of Problem I.1). By the denseness of $\mathrm{St}(p,N) \setminus E_{N,p}(\boldsymbol{S})$ in $\mathrm{St}(p,N)$, Problem I.3 is a sound relaxation of Problem I.1 with $p < N$.

(c) (Mobility of $\Phi_{\boldsymbol{S}}^{-1}$). As [20], for $\boldsymbol{V}, \boldsymbol{F} \in Q_{N,p}(\boldsymbol{S})$ and $\epsilon > 0$, we evaluate a mobility of $\Phi_{\boldsymbol{S}}^{-1}$ with $p < N$ at $\boldsymbol{V}$: $\|\Phi_{\boldsymbol{S}}^{-1}(\boldsymbol{V}+\epsilon\boldsymbol{F}) - \Phi_{\boldsymbol{S}}^{-1}(\boldsymbol{V})\|_F \le 2\epsilon r(\boldsymbol{B})\|\boldsymbol{F}\|_F + \|o(\epsilon)\|_F$, where $r(\boldsymbol{B}) \coloneqq (1 + \|\boldsymbol{B}\|_2^2)^{1/2}/(1 + \sigma_{\min}^2(\boldsymbol{B}))$. Since $r(\boldsymbol{B})$ is bounded as $(1+\|\boldsymbol{B}\|_2^2)^{-1/2} \le r(\boldsymbol{B})$ with holding equality when $\sigma_{\min}(\boldsymbol{B}) = \|\boldsymbol{B}\|_2$, the mobility shows $\Phi_{\boldsymbol{S}}^{-1}$ can be insensitive to change on $Q_{N,p}(\boldsymbol{S})$ in the case where $\|\boldsymbol{B}\|_2$ is large. This leads to slow convergence of naive optimization methods applied to Problem I.3.

(d) (Comparisons to other Cayley transforms on $\mathrm{St}(p,N)$). In the literature, we have found different ideas (e.g., [24], [25]) for parametrization of $\mathrm{St}(p,N)$ with the Cayley transform. [24] proposed to parameterize $\mathrm{St}(p,N)$ with a diffeomorphism $\widetilde{\Phi}_{\boldsymbol{S}}^{-1}\colon \widetilde{Q}_{N,p}(\boldsymbol{S}) \to \mathrm{St}(p,N) \setminus \widetilde{E}_{N,p}(\boldsymbol{S})\colon \boldsymbol{V} \mapsto \widetilde{\varphi}_{\boldsymbol{S}}^{-1}(\boldsymbol{V})\boldsymbol{I}_{N\times p}$, where $\boldsymbol{S} \in \mathrm{O}(N)$, $\widetilde{E}_{N,p}(\boldsymbol{S}) \coloneqq \{\boldsymbol{X} \in \mathrm{St}(p,N) | \det([\boldsymbol{S}_{\mathrm{le}}]_{\mathrm{up}} + \boldsymbol{X}_{\mathrm{up}}) = 0\}$ and $\widetilde{Q}_{N,p}(\boldsymbol{S}) \subset Q_{N,p}(\boldsymbol{S})$ is the domain of $\widetilde{\Phi}_{\boldsymbol{S}}^{-1}$. Although this map seems to be a natural extension of Def. II.1, the condition $\widetilde{Q}_{N,p}(\boldsymbol{S}) \ni \boldsymbol{0}$ for some $\boldsymbol{S}$ is not guaranteed[d], which implies $\widetilde{Q}_{N,p}(\boldsymbol{S})$ is not a vector subspace of $Q_{N,p}(\boldsymbol{S})$. To see this fact, choose $\boldsymbol{S} \in \mathrm{O}(N)$ such that $\det([\boldsymbol{S}_{\mathrm{le}}]_{\mathrm{up}}) = 0$, implying thus $\det([\boldsymbol{S}_{\mathrm{le}}]_{\mathrm{up}} + [\boldsymbol{S}_{\mathrm{le}}]_{\mathrm{up}}) = 0$ and $\boldsymbol{S}_{\mathrm{le}} \in \widetilde{E}_{N,p}(\boldsymbol{S})$. Assume also that $\boldsymbol{0} \in \widetilde{Q}_{N,p}(\boldsymbol{S})$. Then, we have $\boldsymbol{S}_{\mathrm{le}} = \widetilde{\Phi}_{\boldsymbol{S}}^{-1}(\boldsymbol{0}) \in \mathrm{St}(p,N) \setminus \widetilde{E}_{N,p}(\boldsymbol{S})$, which is absurd. On the other hand, $\Phi_{\boldsymbol{S}}^{-1}$ for every $\boldsymbol{S} \in \mathrm{O}(N)$ is a diffeomorphism between the vector space $Q_{N,p}(\boldsymbol{S})$ and $\mathrm{St}(p,N) \setminus E_{N,p}(\boldsymbol{S})$, thus we can enjoy various optimization methods designed over a vector space for minimization of $f_{\boldsymbol{S}} \coloneqq f \circ \Phi_{\boldsymbol{S}}^{-1}$ over $Q_{N,p}(\boldsymbol{S})$.

We consider the computational complexities for $\Phi_{\boldsymbol{S}}$ and $\Phi_{\boldsymbol{S}}^{-1}$. From the fact $(\boldsymbol{I} - \boldsymbol{V})(\boldsymbol{I} + \boldsymbol{V})^{-1} = 2(\boldsymbol{I} + \boldsymbol{V})^{-1} - \boldsymbol{I}$ for $\boldsymbol{V} \in Q_{N,p}(\boldsymbol{S})$ and the Schur complement formula [26, Sec. 0.8.5], the inversion $\Phi_{\boldsymbol{S}}^{-1}(\boldsymbol{V})$ can be expressed as $2(\boldsymbol{S}_{\mathrm{le}} - \boldsymbol{S}_{\mathrm{ri}}\boldsymbol{B})\boldsymbol{M}^{-1} - \boldsymbol{S}_{\mathrm{le}}$, where $\boldsymbol{M} \coloneqq \boldsymbol{I} + \boldsymbol{A} + \boldsymbol{B}^{\mathrm{T}}\boldsymbol{B} \in \mathbb{R}^{p\times p}$. The complexities for $\Phi_{\boldsymbol{S}}$ and $\Phi_{\boldsymbol{S}}^{-1}$ are $\mathcal{O}(N^2p + p^3)$, which are dominated by computations of $\boldsymbol{S}_{\mathrm{ri}}^{\mathrm{T}}\boldsymbol{X}$ and $\boldsymbol{S}_{\mathrm{ri}}\boldsymbol{B}$ respectively.

To reduce these complexities, we impose a structure

$$\mathrm{O}_p(N) \coloneqq \left\{ \Gamma_N(\boldsymbol{T}) \coloneqq \begin{bmatrix} \boldsymbol{T} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{N-p} \end{bmatrix} \middle| \boldsymbol{T} \in \mathrm{O}(p) \right\} \subset \mathrm{O}(N) \quad (4)$$

to center points. For $\boldsymbol{S} \in \mathrm{O}_p(N)$, we obtain $\boldsymbol{S}_{\mathrm{ri}}^{\mathrm{T}}\boldsymbol{X} = \boldsymbol{X}_{\mathrm{lo}}$ and $\boldsymbol{S}_{\mathrm{ri}}\boldsymbol{B} = \begin{bmatrix}\boldsymbol{0}_p & \boldsymbol{B}^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}$, thus the complexities for $\Phi_{\boldsymbol{S}}$ and $\Phi_{\boldsymbol{S}}^{-1}$ centered at $\boldsymbol{S} \in \mathrm{O}_p(N)$ can be reduced to $\mathcal{O}(Np^2 + p^3)$, which is competitive to that for retractions of $\mathrm{St}(p,N)$ [22].

We present in Lemma III.5 a computational choice of center points $\boldsymbol{S} \in \mathrm{O}_p(N)$ from $\boldsymbol{X} \in \mathrm{St}(p,N)$ such that the performance degradation hardly appears around $\Phi_{\boldsymbol{S}}(\boldsymbol{X})$ as in Remark III.4 (c). Since the complexity for the Singular Value Decomposition (SVD) of $\boldsymbol{X}_{\mathrm{up}} \in \mathbb{R}^{p\times p}$ is $\mathcal{O}(p^3)$, thus the choice of $\boldsymbol{S}$ in Lemma III.5 is computationally efficient.

Lemma III.5 (Choice of center points)**.** *Define a set $\tau(\boldsymbol{X}) \coloneqq \{\boldsymbol{Q}_1\boldsymbol{Q}_2^{\mathrm{T}} \in \mathrm{O}(p) | (\boldsymbol{Q}_1, \boldsymbol{\Sigma}, \boldsymbol{Q}_2) \in \mathrm{SVD}(\boldsymbol{X}_{up})\}$, where $\mathrm{SVD}(\boldsymbol{X}_{up})$ is the set of all tuples $(\boldsymbol{Q}_1, \boldsymbol{\Sigma}, \boldsymbol{Q}_2)$ such that*

[d] Another Cayley transform on $\mathrm{St}(p,N)$ is found in [25], however its image Γ^x in [25, Theorem A] is not guaranteed as well to be a vector space.

Algorithm 1 Nesterov accelerated gradient based on ALCP

Input: $\boldsymbol{X}_0 \in \mathrm{St}(p,N)$, $c \in (0, 2^{-1}]$, $T > 0$, L_{CP}
  $\boldsymbol{S}_0 \in \Gamma_N(\tau(\boldsymbol{X}_0))$, $\boldsymbol{Y}_0 = \boldsymbol{Z}_0 = \Phi_{\boldsymbol{S}_0}(\boldsymbol{X}_0)$, $k \leftarrow 0$
  for $n = 0, 1, \ldots, \widehat{n} - 1$ **do**
    $\boldsymbol{Y}_{n+1} = \boldsymbol{Z}_n - \gamma_n \nabla f_{\boldsymbol{S}_n}(\boldsymbol{Z}_n)$ (γ_n is the output of Alg. 2)
    $\boldsymbol{X}_{n+1} = \Phi_{\boldsymbol{S}_n}^{-1}(\boldsymbol{Y}_{n+1})$
    if (6) or (7) are satisfied **then**
      $\boldsymbol{X}_{n+1} = \boldsymbol{X}_n$, $\boldsymbol{S}_{n+1} \in \Gamma_N(\tau(\boldsymbol{X}_{n+1}))$
      $\boldsymbol{Y}_{n+1} = \boldsymbol{Z}_{n+1} = \Phi_{\boldsymbol{S}_{n+1}}(\boldsymbol{X}_{n+1})$, $k \leftarrow 0$ (reset β_n)
    else
      $\boldsymbol{Z}_{n+1} = \boldsymbol{Y}_{n+1} + \beta_n(\boldsymbol{Y}_{n+1} - \boldsymbol{Y}_n)$, $\boldsymbol{S}_{n+1} = \boldsymbol{S}_n$,
      $k \leftarrow k + 1$ (see Remark IV.2 for $\beta_n \in \mathbb{R}$)
    end if
  end for
Output: $\boldsymbol{X}_{\widehat{n}}$

$\boldsymbol{X}_{up} = \boldsymbol{Q}_1 \boldsymbol{\Sigma} \boldsymbol{Q}_2^{\mathrm{T}}$ is an SVD of $\boldsymbol{X}_{up}$ with $\boldsymbol{Q}_1, \boldsymbol{Q}_2 \in \mathrm{O}(p)$ and a non-negative diagonal matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$. For $\boldsymbol{X} \in \mathrm{St}(p,N)$, $\boldsymbol{T} \in \tau(\boldsymbol{X})$ and $\boldsymbol{S} \coloneqq \Gamma_N(\boldsymbol{T})$, we have $\boldsymbol{X} \in \mathrm{St}(p,N) \setminus E_{N,p}(\boldsymbol{S})$, $\boldsymbol{A}(\Phi_{\boldsymbol{S}}(\boldsymbol{X})) = \boldsymbol{0}$, $\|\boldsymbol{B}(\Phi_{\boldsymbol{S}}(\boldsymbol{X}))\|_2 \leq 1$, and $\|\Phi_{\boldsymbol{S}}(\boldsymbol{X})\|_F \leq \sqrt{2p}$.

From the discussion above, we propose a computationally efficient ALCP for Problem I.1 with $\Phi_{\boldsymbol{S}}^{-1}$. By letting $\mathcal{A}_n : Q_{N,p}(\boldsymbol{S}_n) \to Q_{N,p}(\boldsymbol{S}_n)$ be an update scheme at nth iteration and $T > 0$, ALCP generates estimates $(\boldsymbol{X}_n)_{n=0}^{\infty} \subset \mathrm{St}(p,N)$ of solutions to Problem I.1, center points $(\boldsymbol{S}_n)_{n=0}^{\infty} \subset \mathrm{O}_p(N)$ and $(\boldsymbol{V}_n)_{n=0}^{\infty} \subset Q_{N,p}$ iteratively as

$$\left.\begin{aligned} \boldsymbol{X}_{n+1} &\coloneqq \Phi_{\boldsymbol{S}_n}^{-1}(\mathcal{A}_n(\boldsymbol{V}_n)) \\ \boldsymbol{S}_{n+1} &\coloneqq \begin{cases} \boldsymbol{S}_n & (\|\mathcal{A}_n(\boldsymbol{V}_n)\|_F \leq T) \\ \boldsymbol{S} \in \Gamma_N(\tau(\boldsymbol{X}_{n+1})) & (\text{otherwise}) \end{cases} \\ \boldsymbol{V}_{n+1} &\coloneqq \Phi_{\boldsymbol{S}_{n+1}}(\boldsymbol{X}_{n+1}). \end{aligned}\right\} \quad (5)$$

We note that $\boldsymbol{V}_{n+1} = \mathcal{A}_n(\boldsymbol{V}_n)$ when $\boldsymbol{S}_{n+1} = \boldsymbol{S}_n$. A simple specialization of ALCP is a steepest descent method by $\mathcal{A}_n(\boldsymbol{V}) \coloneqq \boldsymbol{V} - \gamma_n \nabla f_{\boldsymbol{S}_n}(\boldsymbol{V})$ with a suitable stepsize $\gamma_n > 0$. In the same way as [3, Theorem 3.2], we obtain the gradient $\nabla f_{\boldsymbol{S}}(\boldsymbol{V}) = 2(\boldsymbol{W}_{\boldsymbol{S}}^f(\boldsymbol{V}) - \boldsymbol{W}_{\boldsymbol{S}}^f(\boldsymbol{V})^{\mathrm{T}})$, where

$$[\boldsymbol{W}_{\boldsymbol{S}}^f(\boldsymbol{V})]_{ij} \coloneqq \begin{cases} 0 & (p+1 \leq i, j \leq N) \\ [\overline{\boldsymbol{W}}_{\boldsymbol{S}}^f(\boldsymbol{V})]_{ij} & (\text{otherwise}), \end{cases}$$

$\overline{\boldsymbol{W}}_{\boldsymbol{S}}^f(\boldsymbol{V}) \coloneqq (\boldsymbol{I}+\boldsymbol{V})^{-1} \boldsymbol{I}_{N \times p} \nabla f(\boldsymbol{X})^{\mathrm{T}} \boldsymbol{S} (\boldsymbol{I}+\boldsymbol{V})^{-1}$ and $\boldsymbol{X} \coloneqq \Phi_{\boldsymbol{S}}^{-1}(\boldsymbol{V})$. For $\boldsymbol{T} \in \mathrm{O}(p)$, $\boldsymbol{W}_{\Gamma_N(\boldsymbol{T})}^f$ can be simplified from the Schur complement formula as

$$\boldsymbol{W}_{\Gamma_N(\boldsymbol{T})}^f(\boldsymbol{V}) \coloneqq \begin{bmatrix} \boldsymbol{P} & \boldsymbol{P}\boldsymbol{B}^{\mathrm{T}} + \boldsymbol{M}^{-1}[\nabla f(\boldsymbol{X})]_{\mathrm{lo}}^{\mathrm{T}} \\ -\boldsymbol{B}\boldsymbol{P} & \boldsymbol{0} \end{bmatrix} \in \mathbb{R}^{N \times N},$$

where $\boldsymbol{P} \coloneqq \boldsymbol{M}^{-1} \nabla f(\boldsymbol{X})^{\mathrm{T}} \begin{bmatrix} \boldsymbol{S}_p^{\mathrm{T}} & -\boldsymbol{B}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} \boldsymbol{M}^{-1} \in \mathbb{R}^{p \times p}$ and $\boldsymbol{M} \coloneqq \boldsymbol{I} + \boldsymbol{A} + \boldsymbol{B}^{\mathrm{T}}\boldsymbol{B} \in \mathbb{R}^{p \times p}$. Therefore, the complexity for $\nabla f_{\boldsymbol{S}}(\boldsymbol{V})$ with $\boldsymbol{S} \in \mathrm{O}_p(N)$ is $\mathcal{O}(Np^2 + p^3)$.

Moreover, we evaluate a Lipschitz constant $L_{\mathrm{CP}} > 0$ of $\nabla f_{\boldsymbol{S}}$ on $Q_{N,p}(\boldsymbol{S})$ under a mild assumption because the proposed NAG combined with ALCP in Section IV requires L_{CP}. Many applications of Problem I.1 satisfy our assumption, e.g., Sec. V, thus Lemma III.6 is useful for plugging optimization methods, exploiting the Lipschitz constant, into ALCP.

Lemma III.6 (Lipschitz constant L_{CP} of $\nabla f_{\boldsymbol{S}}$)**.** *We assume that ∇f is Lipschitz continuous with $L > 0$ on $\mathrm{St}(p,N)$, i.e., $(\boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathrm{St}(p,N))\ \|\nabla f(\boldsymbol{X}_1) - \nabla f(\boldsymbol{X}_2)\|_F \leq L\|\boldsymbol{X}_1 - \boldsymbol{X}_2\|_F$. Let $\mu \coloneqq \max \|\nabla f(\mathrm{St}(p,N))\|_2$[e]. Then, it holds with $L_{\mathrm{CP}} \coloneqq 8(L+\mu)$ for $\boldsymbol{S} \in \mathrm{O}(N)$ and $\boldsymbol{V}_1, \boldsymbol{V}_2 \in Q_{N,p}(\boldsymbol{S}))$ that*

$$\|\nabla f_{\boldsymbol{S}}(\boldsymbol{V}_1) - \nabla f_{\boldsymbol{S}}(\boldsymbol{V}_2)\|_F \leq L_{\mathrm{CP}} \|\boldsymbol{V}_1 - \boldsymbol{V}_2\|_F.$$

Algorithm 2 Backtracking algorithm

Input: $c \in (0, 2^{-1}]$, $\rho \in (0,1)$, $\gamma > 0$, L_{CP}
  while γ satisfies the condition (6) and $\gamma \geq L_{\mathrm{CP}}^{-1}$ **do**
    $\gamma \leftarrow \rho\gamma$
  end while
Output: γ

IV. Nesterov acceleration based on ALCP

To achieve a faster convergence speed for Problem I.1, by assuming a Lipschitz constant $L_{\mathrm{CP}} > 0$ of $\nabla f_{\boldsymbol{S}}$ is available, we present a NAG with a restart scheme based on ALCP. In the proposed NAG illustrated in Alg. 1, we synchronize the change of center points $\boldsymbol{S} \in \mathrm{O}_p(N)$ of ALCP in (5) and the adaptive restart scheme of NAG [16]. It is known that ripples in the sequence of the values of f in NAG lead to slow convergence of NAG, thus to avoid the ripples, adaptive restart schemes for NAG have been studied and their remarkable efficacies have been examined numerically [16], [27].

We change a center point for Alg. 1 when one of the following conditions hold true:

$$f_{\boldsymbol{S}_n}(\boldsymbol{Y}_{n+1}) > f_{\boldsymbol{S}_n}(\boldsymbol{Y}_n) - c\gamma_n \|\nabla f_{\boldsymbol{S}_n}(\boldsymbol{Z}_n)\|_F^2, \quad (6)$$

$$\|\boldsymbol{Y}_{n+1}\|_F > T \quad \text{and} \quad k > 0, \quad (7)$$

where $c \in (0, 2^{-1}]$, $T > 0$, $\gamma_n > 0$ is a stepsize and k is the number of iterations that the current center point has not been changed. The first criterion (6), which is proposed by [16], ensures for NAG to decrease monotonically the value of $f_{\boldsymbol{S}_n}$. The former of the second criterion (7) is derived from the mobility analysis of $\Phi_{\boldsymbol{S}}^{-1}$ in Remark III.4 (c). We need the latter condition $k > 0$ in (7) to work Alg. 1 well even if T is small. Indeed, if we do not use the condition $k > 0$, then the estimate $\boldsymbol{X}_n$ has not been changed in a case where the former condition in (7) is satisfied at every iteration due to small T.

To find γ_n that (6) does not hold with, we employ a backtracking algorithm in Alg. 2. Since the existence of such a stepsize is not always guaranteed, we use L_{CP}^{-1} as a lower bound of acceptable stepsizes. When Alg. 1 restarts at nth iteration, (6) at $n+1$st iteration does not hold with $\gamma_{n+1} = L_{\mathrm{CP}}^{-1}$ at least due to $\boldsymbol{Y}_{n+1} = \boldsymbol{Z}_{n+1}$ and the Lipschitz continuous of $\nabla f_{\boldsymbol{S}_{n+1}}$, thus Alg. 1 is well-defined.

Then, we also show that the gradient $\nabla f_{\boldsymbol{S}_n}(\boldsymbol{Y}_n)$ converges to zero from any initial point in Theorem IV.1.

Theorem IV.1 (Convergence analysis for Alg. 1)**.** *Suppose that $(\boldsymbol{S}_n)_{n=0}^{\infty}$ and $(\boldsymbol{Y}_n)_{n=0}^{\infty}$ are generated by Alg. 1. Then we have $\lim_{n\to\infty} \|\nabla f_{\boldsymbol{S}_n}(\boldsymbol{Y}_n)\| = 0$.*

Remark IV.2. Since we can enjoy the linear structure of $Q_{N,p}(\boldsymbol{S})$, Alg. 1 essentially requires only computation of (2)

[e] The existence of μ is guaranteed by the compactness of $\mathrm{St}(p,N)$ and the continuity of $\|\nabla f(\cdot)\|_2$.

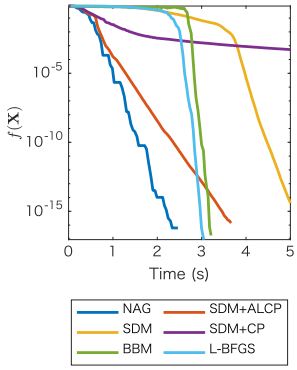

Fig. 1. $f(\boldsymbol{X})$ versus running time for the methods in the scenario of joint diagonalization.

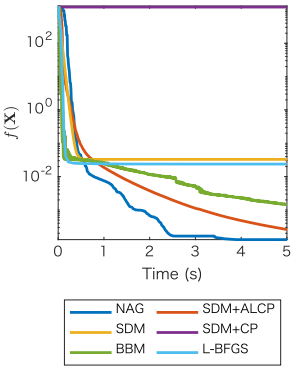

Fig. 2. $f(\boldsymbol{X})$ versus running time for the methods in the scenario of orthogonal Procrustes problem.

without requiring the inversion of neither the exponential map nor retractions unlike [14]–[16]. Moreover, Theorem IV.1 holds by employing any momentum parameter $\beta_n \in \mathbb{R}$.

V. Numerical experiments

To demonstrate the performance of Alg. 1 (NAG), we compare NAG to the steepest descent method with ALCP (SDM+ALCP) and the standard methods for (1) based on retractions, e.g., the steepest descent method (SDM), the Barzilai-Borwein method (BBM) and L-BFGS method, provided by *Manopt* [23], in two scenarios: A. joint diagonalization ($p = N$) and B. orthogonal Procrustes problem ($p \ll N$). Moreover, to examine the avoidance of the performance degradation of CP, we test the steepest descent method with CP (SDM+CP), i.e., SDM+ALCP with $T = \infty$ in (5). The parameters of NAG are given as $\beta_n = (\alpha_n - 1)/\alpha_{n-1}$ with $\alpha_n = (1+(1+4\alpha_{n-1}^2)^{1/2})/2$ for $n \geq 0$ and $\alpha_{-1} = 1$ as [13], and $T = 5$ for the problem A, $T = \sqrt{2p}$ for the problem B. For the backtracking algorithm, we employ $c = 2^{-13}$, $\rho = 2^{-1}$ and an initial stepsize $\gamma = 1$, which are the default of Manopt.

A. Joint diagonalization. We consider minimization of $f(\boldsymbol{X}) := \sum_{i=1}^{I} \|\mathrm{off}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}_i\boldsymbol{X})\|_F^2$ subject to $\boldsymbol{X} \in \mathrm{O}(N)$ for symmetric matrices $\boldsymbol{A}_i \in \mathbb{R}^{N\times N}$, where $\mathrm{off}(\boldsymbol{X}) := \boldsymbol{X} - \mathrm{diag}(\boldsymbol{X})$ and $\mathrm{diag}(\boldsymbol{X})$ denotes the diagonal matrix whose diagonal entries are those of $\boldsymbol{X}$. From $\nabla f(\boldsymbol{X}) = 4\sum_{i=1}^{I} \boldsymbol{A}_i\boldsymbol{X}\mathrm{off}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}_i\boldsymbol{X})$ [1], we obtain $L_{\mathrm{CP}} = 64(3+\sqrt{N})\sum_{i=1}^{I}\|\boldsymbol{A}_i\|_2^2$. Each $\boldsymbol{A}_i$ is generated as $\boldsymbol{X}^{\star\mathrm{T}}\boldsymbol{\Lambda}_i\boldsymbol{X}^{\star}$ with a random orthogonal matrix $\boldsymbol{X}^{\star} \in \mathrm{O}(N)$, where entries of a diagonal $\boldsymbol{\Lambda}_i \in \mathbb{R}^{N\times N}$ are uniformly chosen from $[0, 1]$.

B. Orthogonal Procrustes problem. We consider minimization of $f(\boldsymbol{X}) := \|\boldsymbol{C}\boldsymbol{X} - \boldsymbol{D}\|_F^2$ subject to $\boldsymbol{X} \in \mathrm{St}(p, N)$ for $\boldsymbol{C} \in \mathbb{R}^{q\times N}$ and $\boldsymbol{D} \in \mathbb{R}^{q\times p}$. A closed-form solution to the problem is still a challenge [2]. From $\nabla f(\boldsymbol{X}) = 2\boldsymbol{C}^{\mathrm{T}}(\boldsymbol{C}\boldsymbol{X} - \boldsymbol{D})$, we obtain $L_{\mathrm{CP}} = 16(2\|\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C}\|_2 + \|\boldsymbol{C}^{\mathrm{T}}\boldsymbol{D}\|_2)$. We generate an ill-conditioned matrix $\boldsymbol{C}$ as [2, Ex. 4 in Sec. 4.3] $(m_1 = N/4, m_2 = N/2, m_3 = N/2 - 10, m_4 = 10)$ and $\boldsymbol{D} := \boldsymbol{X}^{\star}\boldsymbol{C}$ with a random $\boldsymbol{X}^{\star} \in \mathrm{St}(p, N)$.

Fig. 1 and Fig. 2 show the evolution of $f(\boldsymbol{X})$ of the methods with respect to running time (seconds) for solving the problem A with $N = 100$ and $I = 100$, and the problem B with $N = 1000$ and $p = q = 10$ respectively. In both experiments, we can see that SDM+ALCP outperforms SDM+CP by changing center points adaptively and the convergence speed of NAG is faster than one of SDM+ALCP. Moreover, we can observe that NAG overwhelms the standard methods, e.g., BBM and L-BFGS, using a retraction in both experiments.

References

[1] M. Nikpour, J. H. Manton, and G. Hori, "Algorithms on the Stiefel manifold for joint diagonalisation," in *ICASSP*, 2002.

[2] J. B. Francisco and F. S. V. Bazán, "Nonmonotone algorithm for minimization on closed sets with applications to minimization on Stiefel manifolds," *J. Comput. Appl. Math.*, vol. 236, no. 10, 2012.

[3] K. Helfrich, D. Willmott, and Q. Ye, "Orthogonal recurrent neural networks with scaled Cayley transform," in *ICML*, vol. 80, 2018.

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

[5] T. E. Abrudan, J. E. Eriksson, and V. Koivunen, "Steepest descent algorithms for optimization under unitary matrix constraint," *IEEE Trans. Signal Proces.*, vol. 56, no. 3, 2008.

[6] S. Fiori, T. Kaneko, and T. Tanaka, "Learning on the compact Stiefel manifold by a Cayley-transform-based pseudo-retraction map," in *IJCNN*, 2012.

[7] Z. Wen and W. Yin, "A feasible method for optimization with orthogonality constraints," *Math. Program.*, vol. 142, no. 1, 2013.

[8] A. Edelman, T. A. Arias, and S. T. Smith, "The geometry of algorithms with orthogonality constraints," *SIAM J. Matrix Anal. Appl.*, vol. 20, no. 2, 1998.

[9] W. Ring and B. Wirth, "Optimization methods on Riemannian manifolds and their application to shape space," *SIAM J. Optim.*, vol. 22, no. 2, 2012.

[10] H. Sato and T. Iwai, "A new, globally convergent Riemannian conjugate gradient method," *Optimization*, vol. 64, no. 4, 2015.

[11] W. Huang, K. A. Gallivan, and P.-A. Absil, "A Broyden class of quasi-Newton methods for Riemannian optimization," *SIAM J. Optim.*, vol. 25, no. 3, 2015.

[12] H. Kasai and B. Mishra, "Inexact trust-region algorithms on Riemannian manifolds," in *NIPS*, 2018.

[13] Y. Nesterov, "A method for solving the convex programming problem with convergence rate $o(1/k^2)$," in *Dokl. Akad. nauk SSSR*, vol. 269, 1983.

[14] Y. Liu, F. Shang, J. Cheng, H. Cheng, and L. Jiao, "Accelerated first-order methods for geodesically convex optimization on Riemannian manifolds," in *NIPS*, 2017.

[15] H. Zhang and S. Sra, "An estimate sequence for geodesically convex optimization," in *ICML*, vol. 75, 2018.

[16] J. W. Siegel, "Accelerated optimization with orthogonality constraints," *arXiv preprint arXiv:1903.05204v03 (to appear in J. Comput. Math.)*, Oct 2019.

[17] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2013, vol. 87.

[18] S. Ghadimi and G. Lan, "Accelerated gradient methods for nonconvex nonlinear and stochastic programming," *Math. Program.*, vol. 156, 2016.

[19] R. Zimmermann, "A matrix-algebraic algorithm for the Riemannian logarithm on the Stiefel manifold under the canonical metric," *SIAM J. Matrix Anal. Appl.*, vol. 38, no. 2, 2017.

[20] K. Kume and I. Yamada, "Adaptive localized Cayley parametrization technique for smooth optimization over the Stiefel manifold," in *EUSIPCO*, 2019.

[21] I. Yamada and T. Ezaki, "An orthogonal matrix optimization by dual Cayley parametrization technique," in *ICA*, 2003.

[22] X. Zhu, "A Riemannian conjugate gradient method for optimization on the Stiefel manifold," *Comput. Optim. Appl.*, vol. 67, no. 1, 2017.

[23] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre, "Manopt, a Matlab toolbox for optimization on manifolds," *J. Mach. Learn. Res.*, vol. 15, 2014.

[24] C. Fraikin, K. Hüper, and P. V. Dooren, "Optimization over the Stiefel manifold," in *PAMM*, vol. 7, no. 1, 2007.

[25] E. Macías-Virgós, M. J. Pereira-Sáez, and D. Tanré, "Cayley transform on Stiefel manifolds," *J. Geom. Phys.*, vol. 123, 2018.

[26] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 2012.

[27] B. O'Donoghue and E. Candès, "Adaptive restart for accelerated gradient schemes," *Found. Comput. Math.*, vol. 15, no. 3, 2015.