Adaptive Localized Cayley Parametrization Technique for Smooth 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*—We propose a novel computational strategy, named the adaptive localized Cayley parametrization technique for acceleration of optimization over the Stiefel manifold. The proposed optimization algorithm is designed as a gradient descent type scheme for the composite of the original cost function and the inverse of the localized Cayley transform defined on the vector space of all skew-symmetric matrices. Thanks to the adaptive localized Cayley transform which is a computable diffeomorphism between the orthogonal group and the vector space of the skew-symmetric matrices, the proposed algorithm (i) is free from the singularity issue, which can cause performance degradation, observed in the *dual Cayley parametrization technique* [Yamada-Ezaki'03] as well as (ii) can enjoy powerful arts for acceleration on the vector space without suffering from the nonlinear nature of the Stiefel manifold. We also present a convergence analysis, for the prototype algorithm employing the Armijo's rule, that shows the gradient of the composite function at zero in the range space of the localized Cayley transform is guaranteed to converge to zero. Numerical experiments show excellent performance compared with major optimization algorithms designed essentially with retractions on the tangent space of the Stiefel manifold [Absil-Mahony-Sepulcher'08, Wen-Yin'13].**

***Index Terms*—Stiefel manifold optimization, orthogonal group optimization, Riemannian manifold optimization, Cayley transform, Anderson acceleration**

I. Introduction

In this paper, we consider an optimization problem over the Stiefel manifold $St(p,N) = \{\boldsymbol{W} \in \mathbb{R}^{N\times p} \mid \boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} = \boldsymbol{I}\}$:

$$\text{minimize } f(\boldsymbol{U}) \text{ subject to } \boldsymbol{U} \in St(p,N), \tag{1}$$

where $f : \mathbb{R}^{N\times p} \to \mathbb{R}$ is differential. This problem has rich applications in data science including signal processing and machine learning, such as nearest low-rank correlation matrix problem [1], nonlinear eigenvalue problem [2], sparse principal component analysis [3], 1-bit compressed sensing [4], dimension reduction for ICA [5], face descriptor for recognition [6] and improvement of generalization for deep neural network [7]. However, the problem (1) is not simple at all due to the non-convexity of $St(p,N)$, which bears many computational ideas to deal with $St(p,N)$ for (1).

A major strategies [8]–[15] deal with $St(p,N)$ as a Riemannian manifold endowed with a Riemannian metric to build optimization schemes in its tangent bundle $TSt(p,N) := \bigcup_{\boldsymbol{U}\in St(p,N)}(\{\boldsymbol{U}\} \times T_{\boldsymbol{U}}St(p,N))$, where $T_{\boldsymbol{U}}St(p,N)$ stands for $Np - p(p+1)/2$-dimensional tangent vector space whose zero corresponds to $\boldsymbol{U}$. Along these strategies, classical optimization algorithms, e.g., the steepest descent method, the Newton's method and the conjugate gradient method, designed originally on a vector space, have been extended to those on the tangent spaces of $St(p,N)$. At each step, the extended algorithms try to update the current estimate $\boldsymbol{U} \in St(p,N)$ of the optimal point ideally along the geodesic on $St(p,N)$ in the direction $\boldsymbol{D}$ determined on the tangent space. However the computation of the geodesic requires infinite calculations, its approximated map called *retraction* $R : TSt(p,N) \to St(p,N) : (\boldsymbol{U},\boldsymbol{D}) \mapsto R_{\boldsymbol{U}}(\boldsymbol{D})$, which is inherited two natures from the geodesic: (i) $R_{\boldsymbol{U}}(\boldsymbol{0}) = \boldsymbol{U}$ and (ii) $\frac{d}{dt}\big|_{t=0} R_{\boldsymbol{U}}(t\boldsymbol{D}) = \boldsymbol{D}$, is proposed, e.g., with QR decomposition/polar decomposition, the Euclidean projection [8], [13], [14], and the Cayley transform [9], [13], [16]. These strategies are realized by the following procedure: (i) Determine a search direction $\boldsymbol{D} \in T_{\boldsymbol{U}}St(p,N)$; (ii) Apply a retraction which assigns to the current estimate $\boldsymbol{U} \in St(p,N)$ a new estimate on $St(p,N)$ along the approximated geodesic in the search direction $\boldsymbol{D}$ determined in (i). Although these strategies can also be extended to optimization problems over general manifolds, not necessary over the Stiefel manifold, they do not seem to exploit inherently algebraic properties of the Stiefel manifold, e.g., $St(p,N)$ can be seen as a canonical projection of the orthogonal group $O(N) := St(N,N)$. This situation suggests a possibility toward alternative powerful strategy for optimization over the Stiefel manifold if we find some ideas to exploit algebraic property of $O(N)$.

A dense subset of $O(N)$ can be parameterized in terms of skew-symmetric matrices by using a diffeomorphism known as the *Cayley transform* [17], [18] $\Phi : O(N) \setminus E \to Q(N) : \boldsymbol{U} \mapsto (\boldsymbol{I} - \boldsymbol{U})(\boldsymbol{I} + \boldsymbol{U})^{-1}$, and its inverse $\Phi^{-1} : Q(N) \to O(N) \setminus E : \boldsymbol{V} \mapsto (\boldsymbol{I} - \boldsymbol{V})(\boldsymbol{I} + \boldsymbol{V})^{-1}$, where $Q(N) := \{\boldsymbol{V} \in \mathbb{R}^{N\times N} | \boldsymbol{V}^{\mathrm{T}} = -\boldsymbol{V}\}$ is the vector space of all skew-symmetric matrices and $E := \{\boldsymbol{U} \in O(N) | \det(\boldsymbol{I} + \boldsymbol{U}) = 0\}$ is the set of all singular points of Φ. Based on the fact that $Q(N)$ constitutes a real vector space, gradient descent type schemes including the steepest descent method and the Newton's method for the composite of f and Φ^{-1} defined on $Q(N)$, called the *dual Cayley parametrization technique*, were introduced in [17] as algorithms specialized for (1) in the case of $p = N$. The update of the schemes can be expressed by $\boldsymbol{U}_{n+1} = \Phi^{-1}(\mathcal{A}(\Phi(\boldsymbol{U}_n)))$, where $\mathcal{A}$ denotes the update rule from $\boldsymbol{V}_n := \Phi(\boldsymbol{U}_n) \in Q(N)$ to $\boldsymbol{V}_{n+1} := \mathcal{A}(\boldsymbol{V}_n) := \Phi(\boldsymbol{U}_{n+1}) \in Q(N)$. Any computational technique developed for minimization of $f \circ \Phi^{-1}$ over vector spaces, e.g., the gradient descent type method and the Newton's method, can be applied as building blocks of the update rule $\mathcal{A}$. However their

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

numerical performances tend to slow down severely around E due to the appearance of undesirable plateau [17].

In this paper, to overcome completely the weakness caused by the set E, while preserving the advantage in [17], we newly propose the *adaptive localized Cayley parametrization technique* as an adaptive extension and acceleration of the dual Cayley parametrization technique [17]. The proposed adaptive localized Cayley parametrization is a time varying modification of [17] for adjustment of the Cayley type parametrization itself in order for the point sequence not to approach the modified singular set. Similar modifications of the parametrization are found, e.g., in [19], where a measure of proximity to a singular point of the *standard Cayley transform* was introduced to judge whether a certain modification, called "pivoting", of Φ is to be applied or not, and in [9], where Φ^{-1} is used as a retraction but their approach can also be interpreted as a time varying modification, at every update of the orthogonal matrix, of the parametrization in [17]. However so far the existing strategies do not seem to have succeeded yet in exploiting maximally the potential advantage hidden in the translation of (1), achieved essentially by the dual Cayley parametrization, into the simplified task of optimization over the vector space. Motivated by these situations, in this paper, we propose to combine a time varying modification of the Cayley parametrization and powerful acceleration techniques [20] including *Anderson acceleration* [21] (See II.C) which was developed originally for fixed-point approximation and is applicable to optimization tasks but only over vector spaces. We also present a convergence analysis, for the adaptive localized Cayley parametrization for a gradient descent type algorithm with the *Armijo's rule* [22], that shows the gradient sequence of the composite function at zero in the range of space of the localized Cayley transform is guaranteed to converge to zero. Numerical experiments in the scenarios of a joint diagonalization for symmetric matrices and an eigenbasis extraction demonstrate that the proposed adaptive extension with acceleration of [17] successfully avoids undesirable plateau and enjoy dramatical improvement in the speed of convergence, outperforming major optimization algorithms designed with retractions on the tangent space of the Stiefel manifold.

II. Preliminaries

A. Notation

Let $\boldsymbol{I}_{N\times p} \in \mathbb{R}^{N\times p}$ stand for the first p-th columns of the identity matrix $\boldsymbol{I} \in \mathbb{R}^{N\times N}$. Let $SO(N) := \{\boldsymbol{U} \in O(N) \mid \det(\boldsymbol{U}) = 1\}$. $\sigma_{\min}(\boldsymbol{X})$ stands for the smallest singular value of $\boldsymbol{X} \in \mathbb{R}^{N\times N}$ and $o(\epsilon)$ is a matrix-valued function which satisfies $\lim_{\epsilon\to 0} \left\| \frac{o(\epsilon)}{\epsilon} \right\|_2 = 0$, where $\|\cdot\|_2$ stands for the spectral norm. We also use $\|\cdot\|$ to denote the standard Euclidean norm for vectors.

B. Dual Cayley parametrization techniques

The dual Cayley parametrization technique [17] was proposed to solve (1) in the case of $p = N$:

$$\text{minimize } f(\boldsymbol{U}) \text{ subject to } \boldsymbol{U} \in O(N). \tag{2}$$

The idea of the technique is to regard the optimization problem as optimization over the vector space $Q(N)$ by parametrization of the subset of $O(N)$ with $Q(N)$ using the Cayley transform Φ. Although Φ can parametrize $O(N) \setminus E \subsetneq SO(N)$, can not parametrize any subset of $O(N) \setminus SO(N) = \{\boldsymbol{U} \in O(N) | \det(\boldsymbol{U}) = -1\} = \{\boldsymbol{U}\boldsymbol{T} | \boldsymbol{U} \in SO(N)\}$, where $\boldsymbol{T} \in O(N) \setminus SO(N)$. To parametrize also a subset of $O(N) \setminus SO(N)$, the dual Cayley parametrization [17] was proposed.

Definition 1 (Dual Cayley parametrization)**.**

$$\begin{cases} \Phi^{-1} & : Q(N) \to O(N) \setminus E : \boldsymbol{V} \mapsto \Phi^{-1}(\boldsymbol{V}) \\ \boldsymbol{T}\Phi^{-1} & : Q(N) \to \boldsymbol{T}(O(N) \setminus E) : \boldsymbol{V} \mapsto \Phi^{-1}(\boldsymbol{V})\boldsymbol{T}, \end{cases} \tag{3}$$

where $\boldsymbol{T} \in O(N) \setminus SO(N)$ and $\boldsymbol{T}(O(N) \setminus E) := \{\boldsymbol{U}\boldsymbol{T} | \boldsymbol{U} \in O(N) \setminus E\} \subsetneq O(N) \setminus SO(N)$.

Since $(O(N)\setminus E)\cup \boldsymbol{T}(O(N)\setminus E)$ is dense in $O(N)$, [17] formulated an optimization problem over $(O(N)\setminus E)\cup \boldsymbol{T}(O(N)\setminus E)$ corresponding to (2) by translating the optimization further into

$$\begin{cases} \text{find} \quad \boldsymbol{U}^{*(1)} \in \Phi^{-1}\left(\underset{\boldsymbol{V}\in Q(N)}{\arg\min} \hat{f}_1(\boldsymbol{V}) \right) \\ \text{find} \quad \boldsymbol{U}^{*(2)} \in \boldsymbol{T}\left(\Phi^{-1}\left(\underset{\boldsymbol{V}\in Q(N)}{\arg\min} \hat{f}_2(\boldsymbol{V}) \right) \right), \end{cases} \tag{4}$$

where $\hat{f}_1(\boldsymbol{V}) := f(\Phi^{-1}(\boldsymbol{V}))$ and $\hat{f}_2(\boldsymbol{V}) := f(\Phi^{-1}(\boldsymbol{V})\boldsymbol{T})$ for $\boldsymbol{V} \in Q(N)$. Since $Q(N)$ is a vector space, standard optimization techniques, e.g., the steepest descent method, the Newton's method, the conjugate gradient method, can be applied to the optimization problem (4). Unfortunately, the numerical performance of the technique tends to slow down severely due to the appearance of undesirable plateau around the singular points E.

C. Anderson acceleration

The Anderson acceleration [20], [21], [23] is a well-established technique to accelerate Picard type iterative schemes: $\boldsymbol{x}_{n+1} = g(\boldsymbol{x}_n) \in \mathbb{R}^M$ for finding a fixed-point $\boldsymbol{x}^\star \in \mathbb{R}^M$ of wide range of nonlinear operators $g : \mathbb{R}^M \to \mathbb{R}^M$. This remarkably simple acceleration is known to be very effective [20], [23] for many algorithms using $g : \mathrm{Id} - \gamma\nabla J$ with $\gamma > 0$ to minimize of J. Anderson acceleration is a restarting technique of the Picard type update whose new initial point is computed as $\boldsymbol{x}_{\text{ext}} := \sum_{k=0}^{K-1} c_k^\star \boldsymbol{x}_k$, where $(\boldsymbol{x}_k)_{k=0}^{K-1}$ are the K latest estimates and $\boldsymbol{c}^\star$ is a minimizer of $\|\sum_{k=0}^{K-1} c_k(g(\boldsymbol{x}_k) - \boldsymbol{x}_k)\|$, as an approximation of $\|g(\sum_{k=0}^{K-1} c_k\boldsymbol{x}_k) - \sum_{k=0}^{K-1} c_k\boldsymbol{x}_k\|$, subject to $\sum_{k=0}^{K-1} c_k = 1$. Therefore we have to solve the approximated problem

$$\text{find } \boldsymbol{c}^\star \in \underset{\boldsymbol{c}^{\mathrm{T}}\boldsymbol{1}=1}{\arg\min} \|\boldsymbol{R}\boldsymbol{c}\|, \tag{5}$$

where $\boldsymbol{R} = [\nabla J(\boldsymbol{x}_0)\ \nabla J(\boldsymbol{x}_1)\ \ldots\ \nabla J(\boldsymbol{x}_{K-1})]$. However, since $\boldsymbol{R}$ tends to become singular or nearly-singular even if K is small, a Tikhonov type regularization technique has also been discussed as *regularized nonlinear acceleration* (RNA)

Algorithm 1 Regularized Nonlinear Acceleration [24] as an Anderson type acceleration [21]

Generate a pair of sequences $(\boldsymbol{x}_k)_{k=1}^K,\ (\boldsymbol{y}_k)_{k=0}^{K-1}$ from (6)
$\boldsymbol{R} = [\boldsymbol{x}_1 - \boldsymbol{y}_0\ \boldsymbol{x}_2 - \boldsymbol{y}_1\ \ldots\ \boldsymbol{x}_K - \boldsymbol{y}_{K-1}]$
$\boldsymbol{c}^\star = \frac{(\boldsymbol{R}^{\mathrm{T}}\boldsymbol{R}+\lambda\boldsymbol{I})^{-1}\mathbf{1}}{\mathbf{1}^{\mathrm{T}}(\boldsymbol{R}^{\mathrm{T}}\boldsymbol{R}+\lambda\boldsymbol{I})^{-1}\mathbf{1}}$.
$\boldsymbol{x}_{\mathrm{ext}} = \sum_{k=1}^K c_k^\star \boldsymbol{x}_k$

in [23], [24] of which a version is summarized in Algorithm 1, where a pair of sequences $(\boldsymbol{x}_n)_{n=1}^K,\ (\boldsymbol{y}_n)_{n=0}^{K-1}$ generated by

$$\begin{cases} \boldsymbol{x}_n &= g(\boldsymbol{y}_{n-1}) \\ \boldsymbol{y}_n &= \frac{\gamma_n}{\gamma}\boldsymbol{x}_n + \left(1 - \frac{\gamma_n}{\gamma}\right)\boldsymbol{y}_{n-1}. \end{cases} \tag{6}$$

III. Optimization algorithms via adaptive localized Cayley parametrization

A. The mobility bound of the Cayley parametrization

To see the influence of the singular points of Φ in the gradient type scheme algorithm in [17], we evaluate the mobility at $\boldsymbol{V}$ as

$$\|\Phi^{-1}(\boldsymbol{V}+\epsilon\boldsymbol{F}) - \Phi^{-1}(\boldsymbol{V})\|_2 \leq \frac{2\|\epsilon\boldsymbol{F}\|_2}{\sigma_{\min}^2(\boldsymbol{I}+\boldsymbol{V})} + \|o(\epsilon)\|_2, \tag{7}$$

where $\boldsymbol{V},\ \boldsymbol{F} \in Q(N),\ \epsilon \in \mathbb{R}$. From this inequality, the upper bound of the mobility is inversely proportional to the square of $\sigma_{\min}(\boldsymbol{I}+\boldsymbol{V})$ and its maximum is achieved by $\boldsymbol{V} = \mathbf{0}$, implying thus undesirable plateau hardly appears around $\mathbf{0}$. Unfortunately, the dual Cayley parametrization technique [17] can not obtain the estimate $\boldsymbol{V}_n \in Q(N)$ while keeping $\sigma_{\min}(\boldsymbol{I}+\boldsymbol{V}_n)$ small when $\boldsymbol{V}^\star$, which corresponds to the optimal point $\boldsymbol{U}^\star$, is located far away from $\mathbf{0}$.

B. Proposed algorithms

In the following, we present an extension of the Cayley transform in order to keep the mobility large in optimization algorithms for problem (2).

Definition 2 (*Localized Cayley transform*)**.** For a given $\boldsymbol{S} \in O(N)$, the localized Cayley transform $\Phi_{\boldsymbol{S}} : O(N) \setminus E(\boldsymbol{S}) \to Q_{\boldsymbol{S}}(N)$ centered at $\boldsymbol{S}$ is defined by

$$\Phi_{\boldsymbol{S}}(\boldsymbol{U}) := (\boldsymbol{S}-\boldsymbol{U})(\boldsymbol{S}+\boldsymbol{U})^{-1} \quad (\boldsymbol{U} \in O(N) \setminus E(\boldsymbol{S}))\,, \tag{8}$$

where $E(\boldsymbol{S}) := \{\boldsymbol{U} \in O(N) | \det(\boldsymbol{S}+\boldsymbol{U}) = 0\}$ is the set of all singular points of $\Phi_{\boldsymbol{S}}$, and $Q_{\boldsymbol{S}}(N) := \{\Phi_{\boldsymbol{S}}(\boldsymbol{U}) | \boldsymbol{U} \in O(N) \setminus E(\boldsymbol{S})\}$ is also the set of all skew-symmetric matrices but for parametrization of $O(N) \setminus E(\boldsymbol{S})$.

Moreover, we give the inverse of $\Phi_{\boldsymbol{S}}$ is given in the following lemma.

Lemma 1 (Inverse of the localized Cayley transform)**.** The localized Cayley transform $\Phi_{\boldsymbol{S}}$ centered at $\boldsymbol{S} \in O(N)$ is diffeomorphic and its inverse mapping $\Phi_{\boldsymbol{S}}^{-1} : Q_{\boldsymbol{S}}(N) \to O(N) \setminus E(\boldsymbol{S})$ is given by

$$\Phi_{\boldsymbol{S}}^{-1}(\boldsymbol{V}) := (\boldsymbol{I}-\boldsymbol{V})(\boldsymbol{I}+\boldsymbol{V})^{-1}\boldsymbol{S} \quad (\boldsymbol{V} \in Q_{\boldsymbol{S}}(N))\,. \tag{9}$$

Algorithm 2 Adaptive localized Cayley parametrization with a steepest descent method at every K-th iteration

Choose $\boldsymbol{U}_0^{(1)} \in SO(N)$, $\boldsymbol{U}_0^{(2)} \in \boldsymbol{T}(SO(N))$
$\boldsymbol{S}_0^{(t)} = \boldsymbol{U}_0^{(t)} \quad (t = 1,\ 2)$
for $n = 0, 1, 2, \ldots$ **do**
  $\boldsymbol{V}_{n+1}^{(t)} = \boldsymbol{V}_n^{(t)} - \gamma_n^{(t)} \nabla \left(f \circ \Phi_{\boldsymbol{S}_n^{(t)}}^{-1}\right)(\boldsymbol{V}_n^{(t)})$ $(\gamma_n^{(t)} \in [0, \infty))$
  $\boldsymbol{U}_{n+1}^{(t)} = \Phi_{\boldsymbol{S}_n^{(t)}}^{-1}(\boldsymbol{V}_{n+1}^{(t)})$
  if $n \bmod K = 0$ **then**
    $\boldsymbol{S}_{n+1}^{(t)} = \boldsymbol{U}_{n+1}^{(t)}$, $\boldsymbol{V}_{n+1}^{(t)} = \mathbf{0}$
  else
    $\boldsymbol{S}_{n+1}^{(t)} = \boldsymbol{S}_n^{(t)}$
  end if
end for $\quad (t = 1,\ 2)$

In a way similar to (7), we obtain

$$\|\Phi_{\boldsymbol{S}}^{-1}(\Phi_{\boldsymbol{S}}(\boldsymbol{U}) + \epsilon\boldsymbol{F}) - \Phi_{\boldsymbol{S}}^{-1}(\Phi_{\boldsymbol{S}}(\boldsymbol{U}))\|_2 \leq \frac{2\|\epsilon\boldsymbol{F}\|_2 + \|o(\epsilon)\|_2}{\sigma_{\min}^2(\boldsymbol{I} + \Phi_{\boldsymbol{S}}(\boldsymbol{U}))}$$

for any $\boldsymbol{U} \in O(N) \setminus E(\boldsymbol{S})$, which suggests that the upper bound of the mobility $\|\Phi_{\boldsymbol{S}}^{-1}(\Phi_{\boldsymbol{S}}(\boldsymbol{U})+\epsilon\boldsymbol{F}) - \Phi_{\boldsymbol{S}}^{-1}(\Phi(\boldsymbol{U}))\|_2$ at $\Phi_{\boldsymbol{S}}(\boldsymbol{U})$ is maximized by setting $\boldsymbol{S} := \boldsymbol{U}$ to ensure $\Phi_{\boldsymbol{S}}(\boldsymbol{U}) = \mathbf{0}$.

Based on this fact, we propose the *adaptive localized Cayley parametrization technique*: $\boldsymbol{U}_{n+1} := \Phi_{\boldsymbol{S}_n}^{-1}(\mathcal{A}(\Phi_{\boldsymbol{S}_n}(\boldsymbol{U}_n)))$ for (2), where $\mathcal{A}$ is the update rule from $\boldsymbol{V}_n := \Phi_{\boldsymbol{S}_n}(\boldsymbol{U}_n) \in Q_{\boldsymbol{S}_n}(N)$ to $\boldsymbol{V}_{n+1} := \mathcal{A}(\boldsymbol{V}_n) := \Phi_{\boldsymbol{S}_n}(\boldsymbol{U}_{n+1}) \in Q_{\boldsymbol{S}_n}(N)$ for finding $\boldsymbol{V}_{n+1} \in Q_{\boldsymbol{S}_n}(N)$ and $\boldsymbol{U}_{n+1} := \Phi_{\boldsymbol{S}_n}^{-1}(\boldsymbol{V}_{n+1})$ to suppress $f \circ \Phi_{\boldsymbol{S}_n}^{-1}$ over $Q_{\boldsymbol{S}_n}(N)$, equivalently f over $O(N) \setminus E(\boldsymbol{S}_n)$, and $\boldsymbol{S}_n$ is updated to $\boldsymbol{U}_n$ at every K-th iteration. Actually, it is ideal to keep the upper bound of the mobility the maximum, i.e., $\boldsymbol{S}_n$ is updated to $\boldsymbol{U}_n$ at every iteration, but it is expected that the upper bound remains near the maximum if K is not large. Thanks to the adaptive localized Cayley parametrization technique, any optimization technique over a vector space can be employed as the update rule $\mathcal{A}$, and as a simplest example, we present a steepest descent method for (2) in Algorithm 2, where (i) $\boldsymbol{U}_n^{(1)} \in SO(N)$ and $\boldsymbol{U}_n^{(2)} \in \boldsymbol{T}(SO(N))$ are generated simultaneously by using suitable stepsizes $\gamma_n^{(t)}$ $(t = 1, 2)$, e.g., the Armijo's rule (See Theorem 1) to achieve monotone decreasing of $f(\boldsymbol{U}_n^{(t)})$ $(n = 1, 2, \ldots)$ for $t = 1, 2$, and (ii) the gradient is given by $\nabla\left(f \circ \Phi_{\boldsymbol{S}}^{-1}\right)(\boldsymbol{V}) = 2(\boldsymbol{W} - \boldsymbol{W}^{\mathrm{T}})$ with $\boldsymbol{W} = (\boldsymbol{I}+\boldsymbol{V})^{-1}\boldsymbol{S}\nabla f\left(\Phi_{\boldsymbol{S}}^{-1}(\boldsymbol{V})\right)^{\mathrm{T}}(\boldsymbol{I}+\boldsymbol{V})^{-1}$.

Remark 1. A specialization of Algorithm 2 with $K = 1$ reproduces a Riemannian optimization technique for (1) with the Cayley transform as a retraction [9]. We emphasize that the strategies with $K > 1$ [17], [19] and with $K = 1$ [9] have distinct difference in the sense that the former can enjoy further special arts developed for optimization tasks over a common vector space during K iterations (See, e.g., Algorithm 3) but the latter can not.

Moreover, to make the best of the advantage of the proposed

Algorithm 3 Adaptive localized Cayley parametrization with an Anderson type acceleration at every K-th iteration

Choose $\boldsymbol{U}_0^{(1)} \in SO(N)$, $\boldsymbol{U}_0^{(2)} \in \boldsymbol{T}(SO(N))$, $\gamma \in [0, \infty)$
$\boldsymbol{S}_0^{(t)} = \boldsymbol{U}_0^{(t)}$, $\boldsymbol{Y}_0^{(t)} = \boldsymbol{0}$ $(t = 1,\ 2)$
for $n = 0, 1, 2, \ldots$ **do**
  $\boldsymbol{X}_{n+1}^{(t)} = \boldsymbol{Y}_n^{(t)} - \gamma \nabla \left(f \circ \Phi_{\boldsymbol{S}_n^{(t)}}^{-1} \right) (\boldsymbol{Y}_n^{(t)})$
  if n mod $K = 0$ **then**
    $\boldsymbol{X}_{\text{ext}}^{(t)} = \text{RNA}\left((\boldsymbol{X})_{k=n-K+1}^{K},\ (\boldsymbol{Y})_{k=n-K}^{K},\ \lambda \right)$
    $\boldsymbol{U}_{n+1}^{(t)} = \Phi_{\boldsymbol{S}_n^{(t)}}^{-1}(\boldsymbol{X}_{\text{ext}}^{(t)})$
    $\boldsymbol{S}_{n+1}^{(t)} = \boldsymbol{U}_{n+1}^{(t)}$, $\boldsymbol{Y}_{n+1}^{(t)} = \boldsymbol{0}$
  else
    $\boldsymbol{U}_{n+1}^{(t)} = \Phi_{\boldsymbol{S}_n^{(t)}}^{-1}(\boldsymbol{X}_{n+1}^{(t)})$
    $\boldsymbol{S}_{n+1}^{(t)} = \boldsymbol{S}_n^{(t)}$
    $\boldsymbol{Y}_{n+1}^{(t)} = \frac{\gamma_n^{(t)}}{\gamma} \boldsymbol{X}_{n+1}^{(t)} + \left(1 - \frac{\gamma_n^{(t)}}{\gamma} \right) \boldsymbol{Y}_n^{(t)}$ $(\gamma_n \in [0, \infty))$
  end if
end for $(t = 1,\ 2)$

parametrization with $K > 1$ in Remark 1, we also propose an enhancement as Algorithm 3 of the parametrization technique with an Anderson type acceleration, RNA [23], which can be applied to optimization tasks but only to over a vector spaces. Since both of these two computational ideas can be implemented at every K-th iteration, the combined algorithm in Algorithm 3 can enjoy excellent synergy effects (See sec. IV).

Remark 2. The singular point set $E(\boldsymbol{S})$ of $\Phi_{\boldsymbol{S}}$ is different from $E(\boldsymbol{S}')$ of $\Phi_{\boldsymbol{S}'}$ for all $\boldsymbol{S}' \in O(N) \setminus E(\boldsymbol{S})$. Moreover, since the maximum of the determinant $|\det(\boldsymbol{S} + \Phi_{\boldsymbol{S}}^{-1}(\boldsymbol{V}))| = \frac{2^N}{\det(\boldsymbol{I}+\boldsymbol{V})}$ is achieved at $\boldsymbol{V} = \boldsymbol{0}$, $\boldsymbol{S}$ is ensured to locate away from $E(\boldsymbol{S})$.

Next, we present an extension of the adaptive localized Cayley parametrization techniques for (1). To achieve this goal, we use the canonical projection: $\Xi : O(N) \to St(p, N) : \boldsymbol{S} \mapsto \boldsymbol{S}\boldsymbol{I}_{N\times p}$, for translating cost function f into $h := f \circ \Xi$. The problem (1) with $p < N$ can be translated into minimization of h over $SO(N)$ because $St(p, N)$ for $p < N$ is a connected manifold unlike $O(N)$. Thus, the proposed parametrization techniques can be applied to problem (1) for general case by passing through Ξ.

C. Convergence Analysis

To establish a convergence analysis for Algorithm 2, we employ the *backtracking algorithm* for stepsizes $\gamma_n^{(t)}$ to satisfy the *Armijo's rule* (See [22]).

Theorem 1 (Convergence Analysis of Algorithm 2)**.** Suppose that $f : \mathbb{R}^{N\times N} \to \mathbb{R}$ is differentiable and $(\boldsymbol{U}_n^{(t)})_{n=0}^{\infty}$ $(t = 1, 2)$ are sequences generated by Algorithm 2 using the backtracking algorithm [22] for stepsizes. Then we have

$$\lim_{n\to\infty} \left\| \nabla \left(f \circ \Phi_{\boldsymbol{U}_n^{(t)}}^{-1} \right) (\boldsymbol{0}) \right\| = 0. \tag{10}$$

IV. Numerical experiments

To demonstrate the performances of Algorithm 2 and Algorithm 3, we consider two optimization problems with many applications in data sciences. We compare Algorithm 2 and Algorithm 3 with (a) the original dual Cayley parametrization technique [17], which is equivalent to Algorithm 2 with $K = \infty$, (b) the steepest descent methods for optimization over a Riemannian manifold employing, as retractions, (b-1) the Cayley transform [9], which is equivalent to Algorithm 2 with $K = 1$ (See Remark 2), and (b-2) QR decomposition [8], called Algorithm 4 in this paper. Each stepsize of all algorithms are determined by the backtracking algorithm [22]. A parameter λ of Algorithm 3 for regularization is set as 10^{-17}.

A. Joint diagonalization. We consider

$$\text{minimize } f(\boldsymbol{U}) := \sum_{m=1}^{10} \text{off}(\boldsymbol{U}^{-1}\boldsymbol{A}_m\boldsymbol{U}) \text{ subject to } \boldsymbol{U} \in O(10),$$

where $\text{off}(\boldsymbol{X}) := \sum_{i\neq j} x_{ij}^2$, x_{ij} denotes the (i, j)-th entry of $\boldsymbol{X} \in \mathbb{R}^{10\times 10}$ and $\boldsymbol{A}_m \in \mathbb{R}^{10\times 10}$ $(m = 1, 2, \ldots, 10)$ are obtained by $\boldsymbol{U}^\star\boldsymbol{\Lambda}_m\boldsymbol{U}^{\star-1}$ with $\boldsymbol{U}^\star \in O(10)$ and a randomly chosen diagonal matrix $\boldsymbol{\Lambda}_m \in \mathbb{R}^{10\times 10}$. A fixed stepsize γ of Algorithm 3 is set as 0.0005. To check the avoidance of undesirable plateau and the performance for the proposed algorithms, we consider the following two settings: (i) the optimal point and the initial point are around singular set E of Φ:

$$\boldsymbol{U}^\star := \begin{bmatrix} -\boldsymbol{I}_2 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_8 \end{bmatrix},\ \boldsymbol{U}_0 := \begin{bmatrix} \boldsymbol{R}(15\pi/16) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_8 \end{bmatrix}$$

and (ii) the optimal point and the initial point are far from E:

$$\boldsymbol{U}^\star := \begin{bmatrix} \boldsymbol{R}(\pi/6) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_8 \end{bmatrix},\ \boldsymbol{U}_0 := \boldsymbol{I},$$

where $\boldsymbol{R}(\theta) \in \mathbb{R}^{2\times 2}$ stands for a rotation matrix of a given angle θ.

B. Eigenbasis extraction. We consider

$$\text{minimize } f(\boldsymbol{U}) := -\text{Tr}(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{U}) \text{ subject to } \boldsymbol{U} \in St(7, 500),$$

where $\boldsymbol{A} \in \mathbb{R}^{500\times 500}$ is randomly chosen such that $\boldsymbol{A}$ is symmetric (Note: Any solution of this problem is known to be an orthonormal eigenbasis associated with the 7 largest eigenvalues of $\boldsymbol{A}$ [8]). We also choose an initial point $\boldsymbol{U}_0 \in St(7, 500)$ randomly. A fixed stepsize γ of Algorithm 3 is set as 0.000001.

Results of our experiments of each algorithm for the *joint diagonalization* with the above settings (i), (ii) and for the *eigenbasis extraction* are illustrated respectively in Fig. 1, Fig. 2 and Fig. 3. Fig. 1 shows that the proposed algorithms (Algorithm 2 with $K = 5$ and Algorithm 3 with $K = 5$) succeed in avoiding the undesirable plateau which is observed in the original dual Cayley parametrization technique (Algorithm 2 with $K = \infty$). Compared with Algorithm 2 with $K = 1$ and Algorithm 4, Algorithm 2 with $K = 5$ in Fig. 1 and Fig. 2 shows better performance in the first scenario. Fig. 3 shows that Algorithm 2 ($K = 5$) has competitive

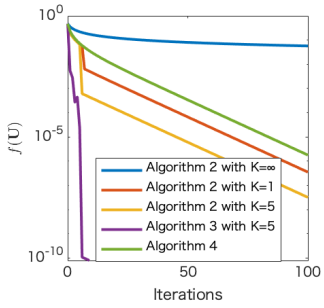

Fig. 1. The value of the cost function versus the number of iteration for different algorithms for the joint diagonalization with the setting (i) the initial point and the optimal point are around E.

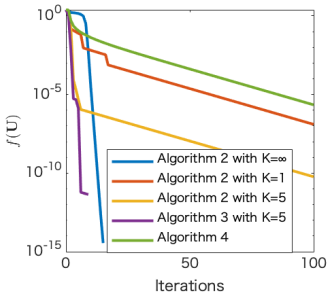

Fig. 2. The value of the cost function versus the number of iteration for different algorithms for the joint diagonalization with the setting (ii) the initial point and the optimal point are far from E.

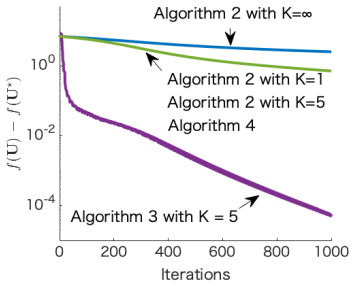

Fig. 3. The subtraction of the value of the cost function for the optimal point from one for each estimate versus the number of iteration for different algorithms for eigenbasis extraction.

performance with Algorithm 2 ($K = 1$) and Algorithm 4 in the second scenario. These results show that the prototype algorithm, Algorithm 2, has at least competitive performance with standard optimization techniques over Riemannian manifolds, which suggests great potential for improvement of the speed of convergence by introducing further advanced techniques applicable over vector spaces. Indeed, we can observe that Anderson type acceleration (Algorithm 3 with $K = 5$) succeeds in accelerating dramatically Algorithm 2 with $K = 5$, and Algorithm 3 with $K = 5$ overwhelms the others in all scenarios.

V. Conclusion

We presented a novel computational strategy, named the adaptive localized Cayley parametrization technique for acceleration of optimization over the Stiefel manifold. The proposed algorithm (i) is free from the singularity issue, which can cause performance degradation, observed in the dual Cayley parametrization technique [Yamada- Ezaki ' 03] as well as (ii) can enjoy powerful arts for acceleration on the vector space without suffering from the nonlinear nature of the Stiefel manifold. Numerical experiments show that the proposed algorithm outperforms major optimization algorithms designed with retractions on the tangent space of the Stiefel manifold.

References

[1] R. Pietersz and P. J. F. Groenen, "Rank reduction of correlation matrices by majorization," *Quantitative Finance*, vol. 4, no. 6, pp. 649–662, 2004.

[2] Z. Zhao, Z. Bai, and X. Jin, "A Riemannian Newton algorithm for nonlinear eigenvalue problems," *SIAM Journal on Matrix Analysis and Applications*, vol. 36, no. 2, pp. 752–774, 2015.

[3] Z. Lu and Y. Zhang, "An augmented Lagrangian approach for sparse principal component analysis," *Mathematical Programming*, vol. 135, no. 1, pp. 149–193, 2012.

[4] P. T. Boufounos and R. G. Baraniuk, "1-bit compressive sensing," in *CISS*, Princeton, N.J., 2008, pp. 16–21.

[5] F. J. Theis, T. P. Cason, and P. A. Absil, "Soft dimension reduction for ICA by joint diagonalization on the Stiefel manifold," in *ICA*. Springer Berlin Heidelberg, 2009, pp. 354–361.

[6] J. Lu, V. E. Liong, X. Zhou, and J. Zhou, "Learning compact binary face descriptor for face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 10, pp. 2041–2056, 2015.

[7] L. Huang, X. Liu, B. Lang, A. W. Yu, Y. Wang, and B. Li, "Orthogonal weight normalization: Solution to optimization over multiple dependent Stiefel manifolds in deep neural networks," in *AAAI*, 2018.

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

[9] Z. Wen and W. Yin, "A feasible method for optimization with orthogonality constraints," *Mathematical Programming*, vol. 142, no. 1, pp. 397–434, 2013.

[10] X. Zhu, "A Riemannian conjugate gradient method for optimization on the Stiefel manifold," *Computational Optimization and Applications*, vol. 67, no. 1, pp. 73–110, 2017.

[11] C. Fraikin, K. Hüper, and P. V. Dooren, "Optimization over the Stiefel manifold," *PAMM*, vol. 7, no. 1, pp. 1 062 205–1 062 206, 2007.

[12] B. Jiang and Y.-H. Dai, "A framework of constraint preserving update schemes for optimization on Stiefel manifold," *Mathematical Programming*, vol. 153, no. 2, pp. 535–575, 2015.

[13] T. E. Abrudan, J. Eriksson, and V. Koivunen, "Steepest descent algorithms for optimization under unitary matrix constraint," *IEEE Transactions on Signal Processing*, vol. 56, no. 3, pp. 1134–1147, 2008.

[14] P.-A. Absil and J. Malick, "Projection-like retractions on matrix manifolds," *SIAM Journal on Optimization*, vol. 22, no. 1, pp. 135–158, 2012.

[15] M. Nikpour, J. H. Manton, and G. Hori, "Algorithms on the Stiefel manifold for joint diagonalisation," in *ICASSP*, vol. 2. IEEE, 2002, pp. II–1481.

[16] S. Fiori, T. Kaneko, and T. Tanaka, "Learning on the compact Stiefel manifold by a cayley-transform-based pseudo-retraction map," in *IJCNN*, 2012, pp. 1–8.

[17] I. Yamada and T. Ezaki, "An orthogonal matrix optimization by dual Cayley parametrization technique," in *ICA*, Nara, Japan, 2003, pp. 35–40.

[18] I. Satake, *Linear algebra*. NY:Marcel Dekker, Inc., 1975.

[19] G. Hori and T. Tanaka, "Pivoting in Cayley tranform-based optimization on orthogonal groups," in *APSIPA*. Biopolis Singapore, 2010, pp. 181–184.

[20] V. Eyert, "A comparative study on methods for convergence acceleration of iterative vector sequences," *Journal of Computational Physics*, vol. 124, no. 2, pp. 271 – 285, 1996.

[21] D. G. Anderson, "Iterative procedures for nonlinear integral equations," *J. ACM*, vol. 12, no. 4, pp. 547–560, Oct. 1965.

[22] A. Wächter and L. T. Biegler, "On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming," *Mathematical Programming*, vol. 106, no. 1, pp. 25–57, 2006.

[23] D. Scieur, A. d'Aspremont, and F. Bach, "Regularized nonlinear acceleration," in *NIPS*, 2016, pp. 712–720.

[24] D. Scieur, "Acceleration in optimization," Ph.D. Thesis, PSL Research University, Sep. 2018.