A Global Cayley parametrization of Stiefel Manifold for Direct Utilization of Optimization Mechanisms over Vector Spaces

Keita Kume and Isao Yamada, Tokyo Institute of Technology

ICASSP 2021 TORONTO Canada, June 6-11, 2021, Metro Toronto Convention Centre

Background & Motivation

Summary

- We propose a novel strategy for solving optimization problem over the Stiefel manifold $St(p,N)$ via solving optimization problem over a vector space.
- To this end, we newly define a surjection $\mathfrak{C}: Q \to St(p,N)$ from a vector space Q onto $St(p,N)$, called a global Cayley parametrization.
- We present useful properties of $\mathfrak{C}$ for optimization, e.g., the Lipschitz continuity of the gradient of $f \circ \mathfrak{C}$.
- Our numerical experiment verifies efficacies of the proposed strategy.

Optimization over $St(p,N) := \{\boldsymbol{U} \in \mathbb{R}^{N\times p} \mid \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_p\}$

For a given continuous function $f: \mathbb{R}^{N\times p} \to \mathbb{R}$,

$$\text{find } \boldsymbol{U}^\star \in \underset{\boldsymbol{U}\in St(p,N)}{\operatorname{argmin}} f(\boldsymbol{U}) \cdots (1)$$

Applications

Sparse PCA, Joint diagonalization, Enhancement of generalization in Deep Neural Network …

Main difficulties

- $St(p,N)$ is not a vector space.
  - $\alpha\boldsymbol{U}_1 + \beta\boldsymbol{U}_2 \notin St(p,N)$ $(\boldsymbol{U}_1,\boldsymbol{U}_2 \in St(p,N), \alpha,\beta \in \mathbb{R})$
- Many optimization techniques heavily rely on the linearity.
  - Gradient descent method: $\boldsymbol{x}_{n+1} := \boldsymbol{x}_n + \gamma_n \boldsymbol{d}_n$ $(\boldsymbol{x}_{n+1}, \boldsymbol{x}_n, \boldsymbol{d}_n \in \mathbb{R}^l, \gamma_n \in \mathbb{R})$

Cayley-type transform $\Phi_{\boldsymbol{S}}$ for $St(p,N)$ [Kume-Yamada'20, EUSIPCO]

Definition

For $\boldsymbol{S} \in O(N) := St(N,N)$, let $E_{N,p}(\boldsymbol{S}) := \{\boldsymbol{U} \in St(p,N) \mid \det(\boldsymbol{I} + \boldsymbol{S}_{le}^T\boldsymbol{U}) = 0\}$.
We define a mapping $\Phi_{\boldsymbol{S}}$ as

$$\Phi_{\boldsymbol{S}}: St(p,N) \setminus E_{N,p}(\boldsymbol{S}) \to Q_{N,p} := \underbrace{\left\{ \begin{bmatrix} \boldsymbol{A} & -\boldsymbol{B}^T \\ \boldsymbol{B} & \boldsymbol{0} \end{bmatrix} \,\middle|\, \begin{matrix} \boldsymbol{A}^T = -\boldsymbol{A} \in \mathbb{R}^{p\times p} \\ \boldsymbol{B} \in \mathbb{R}^{(N-p)\times p} \end{matrix} \right\}}_{\text{vector space}}$$

$$\boldsymbol{X} \mapsto \begin{bmatrix} \boldsymbol{A}_{\boldsymbol{S}}(\boldsymbol{X}) & -\boldsymbol{B}_{\boldsymbol{S}}^T(\boldsymbol{X}) \\ \boldsymbol{B}_{\boldsymbol{S}}(\boldsymbol{X}) & \boldsymbol{0} \end{bmatrix} \qquad \boldsymbol{S} := [\underbrace{\boldsymbol{S}_{le}}_{p}\ \underbrace{\boldsymbol{S}_{ri}}_{N-p}]\}N$$

- $\boldsymbol{A}_{\boldsymbol{S}}(\boldsymbol{X}) := (\boldsymbol{I} + \boldsymbol{S}_{le}^T\boldsymbol{X})^{-T}(\boldsymbol{X}^T\boldsymbol{S}_{le} - \boldsymbol{S}_{le}^T\boldsymbol{X})(\boldsymbol{I} + \boldsymbol{S}_{le}^T\boldsymbol{X})^{-1} \in Q_{p,p}$
- $\boldsymbol{B}_{\boldsymbol{S}}(\boldsymbol{X}) := -\boldsymbol{S}_{ri}^T\boldsymbol{X}(\boldsymbol{I} + \boldsymbol{S}_{le}^T\boldsymbol{X})^{-1} \in \mathbb{R}^{(N-p)\times p}$.

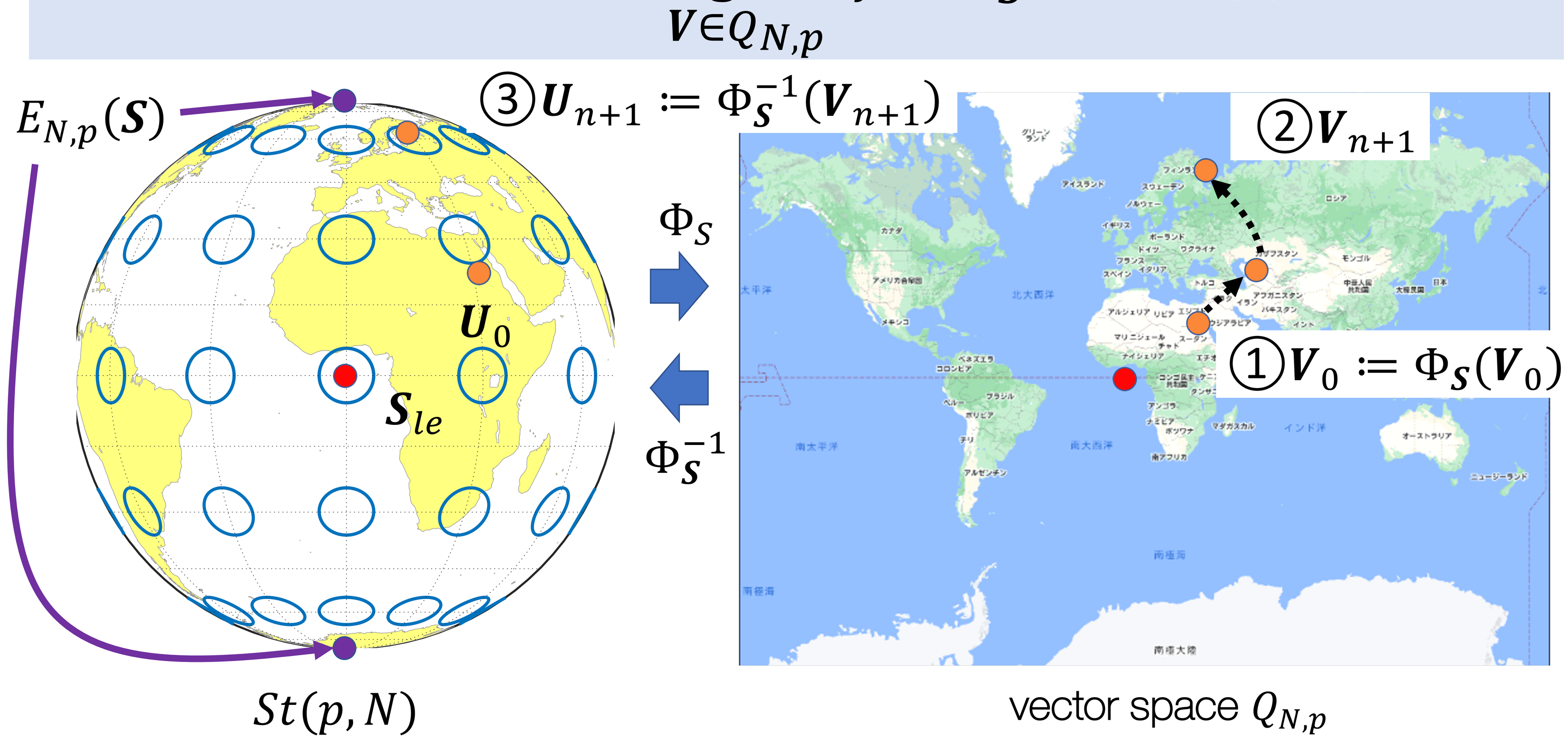
- $\Phi_{\boldsymbol{S}}$ is diffeomorphic between the vector space $Q_{N,p}$ and $St(p,N) \setminus E_{N,p}$ with $\Phi_{\boldsymbol{S}}^{-1}: Q_{N,p} \to St(p,N) \setminus E_{N,p}(\boldsymbol{S}): \boldsymbol{V} \mapsto \boldsymbol{S}(\boldsymbol{I} - \boldsymbol{V})(\boldsymbol{I} + \boldsymbol{V})^{-1}\boldsymbol{I}_{le}$.
- $\varphi := \Phi_{\boldsymbol{I}}$ with $p = N$ is the classical Cayley transform for $SO(N) := \{\boldsymbol{U} \in O(N) \mid \det(\boldsymbol{U}) = 1\}$.
- $\boldsymbol{S}$ of $\Phi_{\boldsymbol{S}}$ is called a center point because $\Phi_{\boldsymbol{S}}(\boldsymbol{S}_{le}) = \boldsymbol{0}$

Cayley parametrization (CP) strategy

Solve (1) via the following optimization problem over a vector space $Q_{N,p}$:

For a given continuous function $f: \mathbb{R}^{N\times p} \to \mathbb{R}$,

$$\text{find } \boldsymbol{V}^\star \in \underset{\boldsymbol{V}\in Q_{N,p}}{\operatorname{argmin}} f \circ \Phi_{\boldsymbol{S}}^{-1}(\boldsymbol{V}) \cdots (2)$$



1. Translate an initial point $\boldsymbol{U}_0 \in St(p,N)$ into $\boldsymbol{V}_0 := \Phi_{\boldsymbol{S}}(\boldsymbol{U}_0) \in Q(N,p)$.
2. Update candidate solutions $(\boldsymbol{V}_k)_{k=0}^{n+1} \subset Q_{N,p}$ in a vector space $Q_{N,p}$.
3. Translate the solution $\boldsymbol{V}_{n+1}$ into $\boldsymbol{U}_{n+1} := \Phi_{\boldsymbol{S}}^{-1}(\boldsymbol{V}_{n+1}) \in St(p,N)$.

- We can utilize for ② directly optimization techniques over a vector space.
- $\Phi_{\boldsymbol{S}}^{-1}(Q_{N,p}) \subsetneq St(p,N)$ for all $\boldsymbol{S} \in O(N)$ possibly induces $\min f(St(p,N)) \neq \min f \circ \Phi_{\boldsymbol{S}}^{-1}(Q_{N,p})$.

Natural question

Can we parameterize $St(p,N)$ by a single vector space entirely ?

This study

Global Cayley parametrization (G-CP) strategy

Solve (1) via the following optimization problem over a vector space Q:

For a given continuous function $f: \mathbb{R}^{N\times p} \to \mathbb{R}$,

$$\text{find } \boldsymbol{\mathcal{V}}^\star \in \min_{\boldsymbol{\mathcal{V}}\in Q} f \circ \mathfrak{C}(\boldsymbol{\mathcal{V}}) \cdots (3)$$

Definition

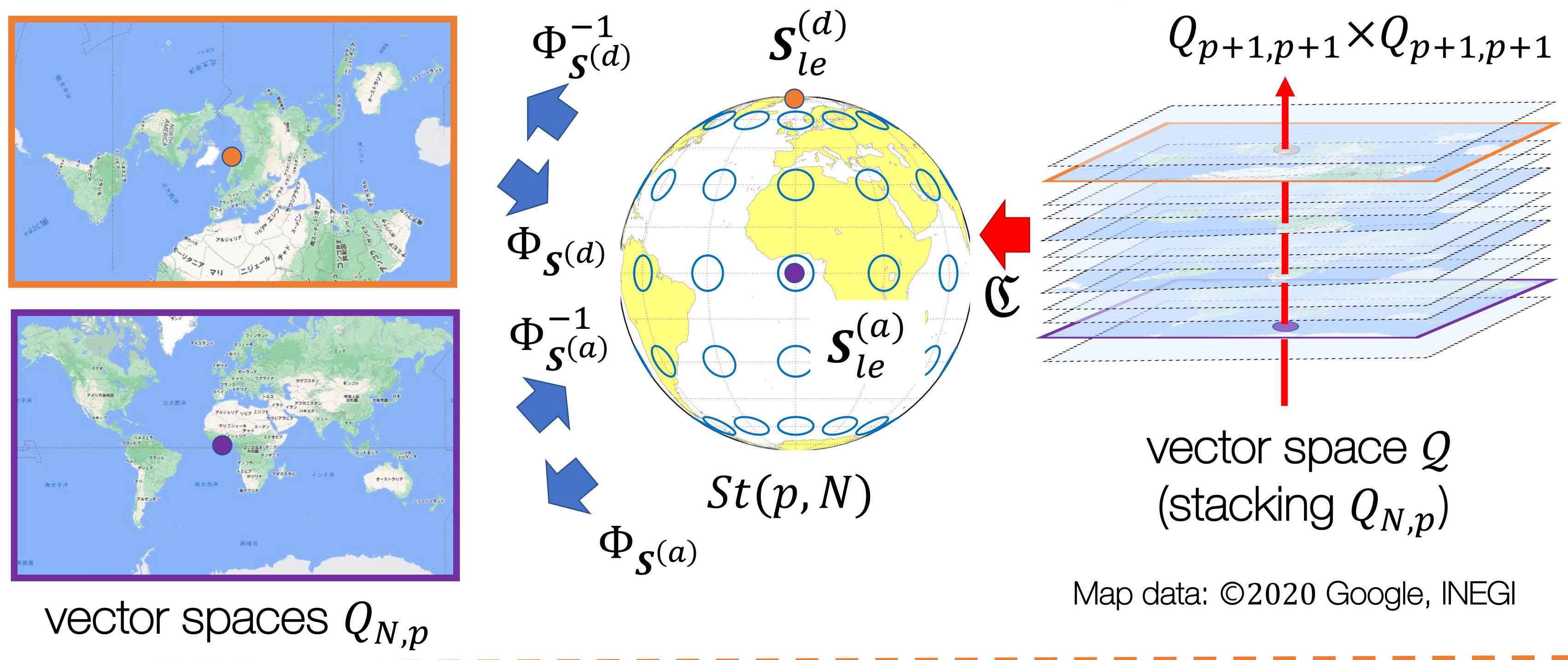
Let $Q := Q_{p+1,p+1} \times Q_{p+1,p+1} \times Q_{N,p}$.
Then, the global Cayley parametrization is defined by

$$\mathfrak{C}: Q \to St(p,N): \boldsymbol{\mathcal{V}} := (\boldsymbol{V}_1, \boldsymbol{V}_2, \boldsymbol{V}_3) \mapsto \Phi_{\boldsymbol{S}(\boldsymbol{V}_1,\boldsymbol{V}_2)}^{-1}(\boldsymbol{V}_3),$$

where $\boldsymbol{S}(\boldsymbol{V}_1,\boldsymbol{V}_2) := \begin{bmatrix} \varphi^{-1}(\boldsymbol{V}_1)\varphi^{-1}(\boldsymbol{V}_2) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{N-p-1} \end{bmatrix} \in SO(N)$ and

$\varphi^{-1}: Q_{p+1,p+1} \to SO(p+1) \setminus E_{p+1,p+1}(\boldsymbol{I}): \boldsymbol{V} \mapsto (\boldsymbol{I} - \boldsymbol{V})(\boldsymbol{I} + \boldsymbol{V})^{-1}$ is the inversion mapping of the classical Cayley transform φ.

Basic idea: parameterization of center points $\boldsymbol{S}$ by a single vector space



Let $\widehat{SO}_{p+1}(N) := \left\{ \begin{bmatrix} \widehat{\boldsymbol{T}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{N-p-1} \end{bmatrix} \in SO(N) \,\middle|\, \widehat{\boldsymbol{T}} \in SO(p+1) \right\}$.

It holds $\{\Phi_{\boldsymbol{S}}^{-1}(\boldsymbol{V}) \mid (\boldsymbol{S},\boldsymbol{V}) \in \widehat{SO}_p(N) \times Q_{N,p} \cong SO(p+1) \times Q_{N,p}\} = St(p,N)$.

Combined with the following lemma, we can derive **the surjection of $\mathfrak{C}$.**

Lemma [Weyl'46, The Classical Groups]

$\{\varphi^{-1}(\boldsymbol{V}_1)\varphi^{-1}(\boldsymbol{V}_2) \mid (\boldsymbol{V}_1,\boldsymbol{V}_2) \in Q_{p+1,p+1} \times Q_{p+1,p+1}\} = SO(p+1)$

Characterization of local minimizer and stationary point by $f \circ \mathfrak{C}$

Key properties for $\mathfrak{C}$ (Theorem 3.2)

1. For $\boldsymbol{U}^\star \in St(p,N)$, $\boldsymbol{U}^\star$ is a local minimizer of (1) $\Leftrightarrow$ $\boldsymbol{\mathcal{V}}^\star \in Q$ s.t. $\boldsymbol{U}^\star = \mathfrak{C}(\boldsymbol{\mathcal{V}}^\star)$ are local minimizers of (3).
2. Let f be differentiable. If $\nabla(f \circ \mathfrak{C})(\boldsymbol{\mathcal{V}}) = \boldsymbol{0}$ for $\boldsymbol{\mathcal{V}} \in Q$, then $\mathfrak{C}(\boldsymbol{\mathcal{V}})$ is a stationary point of (1).

We can find a local minimizer (stationary point) of the problem (1) via the problem (3) defined over the vector space Q.

Useful properties of $f \circ \mathfrak{C}$ for optimization (Proposition 3.5)

Ex: Lipschitz continuity of $\nabla(f \circ \mathfrak{C})$

Let $f: \mathbb{R}^{N\times p} \to \mathbb{R}$ be continuously differentiable and $\max \|\nabla f(St(p,N))\|_F \leq \mu$.
Suppose ∇f is Lipschitz continuous with $L > 0$ over $St(p,N)$.
Then, $\nabla(f \circ \mathfrak{C})$ is Lipschitz continuous with $24(L+\mu)$ over Q.

Numerical experiments

- Optimization technique: gradient descent method (GDM)
- Comparison with
  1. CP strategy [Kume-Yamada'20]
  2. Retraction-based strategy with QR decomposition (QR) [Boumal et al.'14, J. Mach. Res.]

1. Joint diagonalizaton problem

$$f(\boldsymbol{U}) := \sum_{i\in\mathcal{I}} \|\boldsymbol{U}^T\boldsymbol{A}_i\boldsymbol{U} - \mathrm{Diag}(\boldsymbol{U}^T\boldsymbol{A}_i\boldsymbol{U})\|_F^2$$

- $N = 1000, p = 10, |\mathcal{I}| = 10$

2. Eigenvalue problem

$$f(\boldsymbol{U}) := -\mathrm{Tr}(\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{U})$$

- $N = 2000, p = 10$

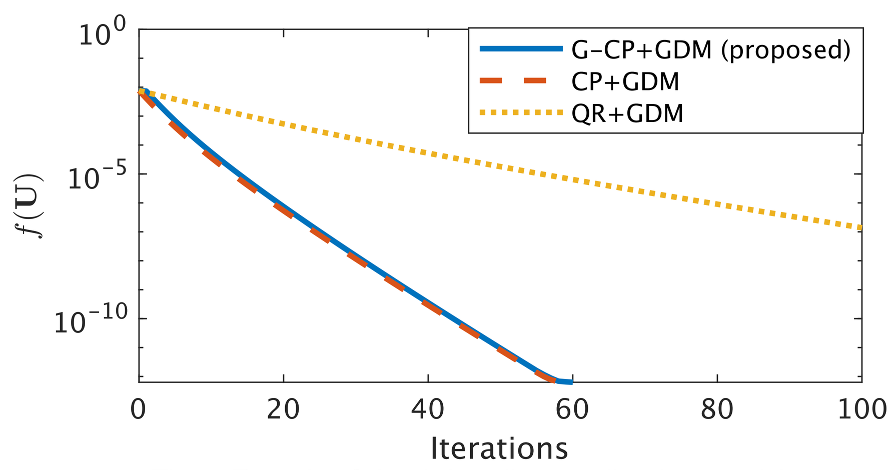


Fig.1 Result of the joint diagonalization problem

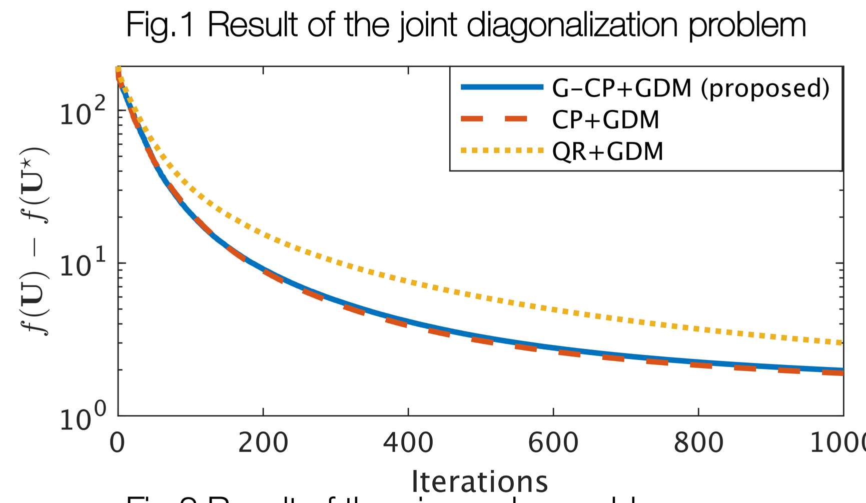


Fig.2 Result of the eigenvalue problem

G-CP strategy has potential to bring numerous optimization mechanisms over a vector space to (1) without losing the performance compared with CP strategy and the retraction-based strategy.

[Kume-Yamada'20] K. Kume and I. Yamada, "A Nesterov-type acceleration with adaptive localized Cayley parametrization for optimization over the Stiefel manifold," *in EUSIPCO*, 2020.
[Gallier'13] J. Gallier, "Remarks on the Cayley representation of orthogonal matrices and on perturbing the diagonal of a matrix to make it invertible," *arXiv preprint arXiv:math/0606320v2*, 2013.
[Boumal et al.'14] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre, "Manopt, a matlab toolbox for optimization on manifolds," *J. Mach. Res.*, 2014.