ISS-SP-005

Adaptive Localized Cayley Parametrization を利用した Stiefel 多様体制約付き確率的勾配分散縮小法

久米啓太 †　　山田功 †

† 東京工業大学工学院情報通信系

1　はじめに

Stiefel 多様体 $\mathrm{St}(p,N) := \{\boldsymbol{X} \in \mathbb{R}^{N\times p} | \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \boldsymbol{I}\}$ 上の最適化は主成分分析 [1] やディープニューラルネットワークの学習 [2] 等，信号処理や機械学習の領域で重要性が高まっている．これらの応用では各サンプルに対応づけられた $f_i : \mathbb{R}^{N\times p} \to \mathbb{R}$ の和 $f(\boldsymbol{X}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{X})$ の最小化が制約条件 $\boldsymbol{X} \in \mathrm{St}(p,N)$ 下で求められる．ビッグデータ解析応用 (n が大) では，計算量削減のため，一部の f_i のみを推定値更新に利用する確率的最適化戦略 (例: 確率的分散勾配縮小法 (SVRG) [3,4]) が望まれるが，既存の「$\mathrm{St}(p,N)$ 制約付き SVRG」[1] には無制約の SVRG の特長が十分に継承されていない．小文では，「Adaptive Localized Cayley Parametrization(ALCP) [5] を利用した $\mathrm{St}(p,N)$ 制約付き SVRG」を提案する．

2　確率的勾配分散縮小法 (SVRG)

2.1　$\mathbb{R}^{N\times p}$ 上の非凸最適化のための SVRG

$\mathbb{R}^{N\times p}$ 上の SVRG [4] は，確率的勾配降下法で問題となる確率勾配の分散抑圧のため，基準点 $\tilde{\boldsymbol{X}}$ を設け，一様ランダムに選ばれた $i \in \{1,2,\ldots,n\}$ を用いた更新:

$$\boldsymbol{X}_{k+1} = \boldsymbol{X}_k - \gamma_k(\nabla f_i(\boldsymbol{X}_k) - \nabla f_i(\tilde{\boldsymbol{X}}) + \nabla f(\tilde{\boldsymbol{X}})) \quad (1)$$

を行う．基準点は，一定回数の反復の後，最新の推定値に更新される．基準点と推定値での勾配の相乗効果により，確率勾配 $\nabla f_i(\boldsymbol{X}_k)$ のみを用いた素朴な確率的勾配降下法の収束性能が大きく改善されるため，SVRG は**近年の確率的最適化法の代表的な戦略基盤**となっている．

2.2　接空間を利用した「$\mathrm{St}(p,N)$ 制約付き SVRG」

「標準的な $\mathrm{St}(p,N)$ 制約付き最適化戦略」の各反復では，推定値 $\boldsymbol{X}_k \in \mathrm{St}(p,N)$ は，その近傍で $\mathrm{St}(p,N)$ を局所近似する接空間 $T_{\boldsymbol{X}_k}\mathrm{St}(p,N)$ 上に制限された負勾配 $-\nabla_{T_{\boldsymbol{X}_k}} f(\boldsymbol{X}_k)$ を初速度とする測地線上の 1 点に更新される [6]．既存の「$\mathrm{St}(p,N)$ 制約付き SVRG」[1] も，この戦略に倣って実現されてきたが，**基準点の勾配と推定値の勾配が異なる接空間上に制限されるため，両者の相乗効果を引き出すことが困難**であった．実際に [1] は，基準点の勾配を最新の接空間に vector transport [6] する方針を提案しているが，収束性能劣化を招く「0 に収束するステップ幅 γ_k」を用いる必要がある．

3　ALCP を利用した「$\mathrm{St}(p,N)$ 制約付き SVRG」

最近，筆者らは「$\mathrm{O}(N) := \mathrm{St}(N,N)$ 制約付き最適化戦略」として ALCP [5] を提案している．[2,7,8] には $\mathrm{O}(N)$ 制約付き最適化問題を，逆 Cayley 変換 $\Phi_{\boldsymbol{I}}^{-1}[p=N, \boldsymbol{S}=\boldsymbol{I}$ の場合の式 (2)] を用いた**ベクトル空間** $Q_{N,N}(\boldsymbol{I})$ **上の** $f\circ\Phi_{\boldsymbol{I}}^{-1}$ **の最適化問題に帰着**する最適化戦略が提案されているが，特異点集合 $E_{N,N}(\boldsymbol{I})$ 周辺での縮尺拡大による収束性能劣化が問題となっていた (図 1) [7,8]．[5] は中心点 $\boldsymbol{S} \in \mathrm{O}(N)$ ごとに定義される逆局所 Cayley 変換 $\Phi_{\boldsymbol{S}}^{-1}$ [$p=N$ の場合の式 (2)] を導入し，最新の推定値 $\boldsymbol{X}_k \in \mathrm{O}(N)$ が $E_{N,N}(\boldsymbol{S})$ に接近しないよう $\boldsymbol{S}$ を適応的に更新することにより収束性能の改善を図るアイディアを示している (図 2)．**ALCP は「ベクトル空間上で開発された最先端の最適化戦略」を $\mathrm{O}(N)$ 上の最適化問題に直接導入可能にする**大きな特長を備えている．

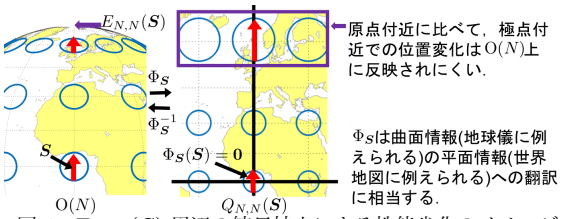

図 1: $E_{N,N}(\boldsymbol{S})$ 周辺の縮尺拡大による性能劣化のイメージ

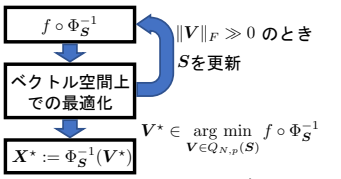

図 2: ALCP のアイディア

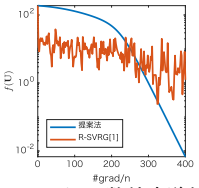

図 3: 数値実験例

小文では，$\mathrm{St}(p,N)$ へ逆局所 Cayley 変換の一般化[1]:

$$\begin{aligned}\Phi_{\boldsymbol{S}}^{-1} &: Q_{N,p}(\boldsymbol{S}) \to \mathrm{St}(p,N) \setminus E_{N,p}(\boldsymbol{S}) \\ &: \boldsymbol{V} \mapsto \boldsymbol{S}(\boldsymbol{I}-\boldsymbol{V})(\boldsymbol{I}+\boldsymbol{V})^{-1}\boldsymbol{I}_{N\times p} \end{aligned} \quad (2)$$

を与えると共に，ALCP を $\mathrm{St}(p,N)$ に適用することを提案している．**中心点 $\boldsymbol{S}$ と基準点を同じタイミングで変更する**ことにより，ALCP を利用した「$\mathrm{St}(p,N)$ 制約付き SVRG」が自然に実現できる．提案法は基準点更新までの間，同一のベクトル空間上で点列更新 [式 (1)] されるため，**[4] で示された収束解析法も自然に継承できる**．なお、最近に至るまで SVRG にはいくつかの改良工夫が見られるが，いずれの工夫も小文のアイディアで $\mathrm{St}(p,N)$ 上に直接導入可能である．また，提案法のステップ幅 γ_k の決定に必要な $\nabla(f\circ\Phi_{\boldsymbol{S}}^{-1})$ の **Lipschitz 定数評価法**も得ており，この情報を活用する SVRG 以外の多くの最適化戦略も $\mathrm{St}(p,N)$ 上に翻訳できる．$\mathrm{O}(10)$ 上の同時対角化問題 [5]($n=10^4$) を例とする数値実験を新たに実施し，提案法 (ALCP を利用した「$\mathrm{St}(p,N)$ 制約付き SVRG」) が極めて優れた収束性能を達成することを確認している (図 3:∇f_i の評価回数に対する性能)．

参考文献

[1] H. Sato, H. Kasai, and B. Mishra, "Riemannian stochastic variance reduced gradient algorithm with retraction and vector transport," *SIAM J. Optim.*, vol. 29, no. 2, pp. 1444–1472, 2019.

[2] K. Helfrich, D. Willmott, and Q. Ye, "Orthogonal recurrent neural networks with scaled Cayley transform," in *ICML2018*.

[3] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *NIPS2013*.

[4] S. e. a. Reddi, "Stochastic variance reduction for nonconvex optimization," in *ICML2016*.

[5] K. Kume and I. Yamada, "Adaptive localized Cayley parametrization technique for smooth optimization over the Stiefel manifold," in *EUSIPCO2019*.

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

[7] I. Yamada and T. Ezaki, "An orthogonal matrix optimization by dual Cayley parametrization technique," in *ICA2003*.

[8] G. Hori and T. Tanaka, "Pivoting in Cayley tranform-based optimization on orthogonal groups," in *APSIPA2010*.

[1] ベクトル空間 $Q_{N,p}(\boldsymbol{S}) := \left\{ \begin{bmatrix} \boldsymbol{A} & -\boldsymbol{B}^{\mathrm{T}} \\ \boldsymbol{B} & \boldsymbol{0} \end{bmatrix} \middle| \begin{matrix} \boldsymbol{A}^{\mathrm{T}}=-\boldsymbol{A}\in\mathbb{R}^{p\times p} \\ \boldsymbol{B}\in\mathbb{R}^{N-p\times p} \end{matrix} \right\}$, 特異点集合 $E_{N,p}(\boldsymbol{S}) := \{\boldsymbol{U} \in \mathrm{St}(p,N) | \det(\boldsymbol{I}+(\boldsymbol{S}\boldsymbol{I}_{N\times p})^{\mathrm{T}}\boldsymbol{U}) = 0\}$, $\boldsymbol{I}_{N\times p} \in \mathrm{St}(p,N)$ は単位行列 $\boldsymbol{I} \in \mathrm{O}(N)$ の左から p 列を並べた行列．