Vol.2019-AVM-106 No.20
2019/9/20

非線形構造制約付き大規模機械学習問題への取り組み
～ 非厳密リーマン多様体最適化の理論と応用 ～

笠井 裕之 †1,a)

概要：本講演では，非線形構造制約付き大規模機械学習問題への取り組みについて紹介する．具体的には，大規模データを対象とした非厳密リーマン多様体最適化理論の基本と研究成果について紹介する．講演の前半部では，機械学習の研究について階層モデルにより整理し，近年の大規模機械学習の課題について最適化問題の視点から整理する．その後，構造制約問題に着目し，リーマン多様体最適化が活躍する問題設定と実際の適用事例を紹介する．そして，あらためてリーマン多様体最適化問題を定義し，その最適化手法について紹介する．後半は，大規模機械学習問題へのアプローチとして，講演者らが取り組む非厳密リーマン多様体最適化理論の基本と，近年の研究成果について紹介する．最後に，最適化処理を考慮した多様体の幾何空間についても触れる．

1. Introduction

Nonlinear and nonconvex constraints have attracted much attention recently in machine learning applications. They include, for example, orthogonality, fixed rankness, and symmetric positive definiteness. One versatile framework to tackle the problems under such constraints is *Riemannian optimization* or *manifold optimization*. This presentation provides a basic concept, fundamentals and applications of Riemannian optimization in machine learning field. In addition, some recent progresses on *inexact* approaches for large-scale learning problems are mainly addressed.

2. Riemannian optimization

We consider the optimization problem

$$\min_{x\in\mathcal{M}} f(x), \tag{1}$$

where $f : \mathcal{M} \to \mathbb{R}$ is a smooth real-valued function on a *Riemannian manifold* $\mathcal{M}$ [1], [2]. The particular focus is when f has a *finite-sum* structure, which frequently arises as big-data problems in machine learning applications. Specifically, we consider the form $f(x) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_i(x)$, where n is the total number of samples and $f_i(x)$ is the cost function for the i-th ($i \in [n]$) sample.

"Riemannian optimization translates the constrained optimization problem (1) into an unconstrained optimization problem over the manifold $\mathcal{M}$". This viewpoint has shown benefits in many applications. The principal component analysis and subspace tracking problems are defined on the *Grassmann* manifold [3]. The low-rank matrix & tensor completion problems are examples on the manifold of *fixed-rank* matrices and tensors [4], [5]. The linear regression problem is defined on the manifold of the fixed-rank matrices. The independent component analysis problem requires a whitening step that is posed as a joint diagonalization problem on the *Stiefel* manifold.

†1 現在，電気通信大学
Presently with The University of Electro-Communications
a) kasai@is.uec.ac.jp

3. Inexact Riemannian optimization

3.1 First-order stochastic optimization

A popular approach to solve (1) is the *Riemannian steepest descent* (RSD) algorithm [1], which is traced back to Luenberger's work in 1972. It calculates the *Riemannian full gradient* $\mathrm{grad} f(x)$ every iteration, which can be computationally heavy when the data size n is extremely large. To address this issue, the *Riemannian stochastic gradient descent* (RSGD) algorithm, which is a counterpart of the *stochastic gradient descent* (SGD) in the Euclidean space [6], becomes a computationally efficient approach [7]. The advantage of RSGD is that it calculates only *Riemannian stochastic gradient* $\mathrm{grad} f_i(x)$ for an i-th sample every iteration, which results in that the complexity per iteration is *independent* of n.

However, similarly to SGD, RSGD suffers from slow convergence due to a *decaying stepsize* . For this issue, *variance reduction* (VR) methods on Riemannian manifolds, including **RSVRG** [8], [9], [10] and **RSRG** [11], have achieved a faster convergence rate, which are generalization of the algorithms in the Euclidean space [12]. The core idea is to reduce the variance of *noisy* stochastic gradients by periodical full gradient estimations, resulting in a *linear* convergent rate. It should, however, be pointed out that such Riemannian VR methods require *retraction* and *vector transport* operations at *every iteration*.

Besides, a class of algorithms including, for example, Adam, AdaGrad, and RMProp, that has become increasingly common lately, especially in *deep learning*, adapts the learning rate of each coordinate of the past gradients. However, such explorations on Riemannian manifolds are relatively new and challenging. This is because of the intrinsic nonlinear structure of the underlying manifold and the absence of a canonical coordinate system. In machine learning applications, however, most manifolds of interest form as *matrix* with notions of row and column subspaces. To this end, such a rich structure should not be lost by transforming matrices to just a stack of vectors. For this particular purpose, **RASA** has been very recently proposed for problems on Riemannian matrix manifolds by adapting the row and column subspaces of gradients [13].

3.2 Second-order stochastic optimization

All the above algorithms are *first-order* algorithms, which guarantee convergence to the *first-order optimality condition*, i.e., $||\mathrm{grad} f(x)||_x = 0$, using only the gradient information. Thus, their performance in ill-conditioned problems suffers due to poor curvature approximation. *Second-order* algorithms, on the other hand, alleviate such effects by exploiting curvature information effectively. Therefore, they are expected to converge to a solution that satisfies the *second-order optimality conditions*, i.e., $||\mathrm{grad} f(x)||_x = 0$ and $\mathrm{Hess} f(x) \succeq 0$, where $\mathrm{Hess} f(x)$ is the *Riemannian Hessian* of f at x. The *Riemannian Newton* method is a second-order algorithm, which has a *superlinear local* convergence rate [1]. It, however, lacks *global convergence* and its practical variants are computationally expensive to implement. A popular alternative is the *Riemannian limited memory BFGS* algorithm (RLBFGS) that requires lower memory. It, however, exhibits only a linear convergence rate and requires many vector transports of curvature information pairs. Finally, the *Riemannian trust-region* algorithm (RTR) comes with a global convergence property [1] and a superlinear local convergence rate [1].

A common issue among second-order algorithms is higher computational costs for dealing with exact or approximate Hessian matrices, which is computationally prohibitive in a large-scale setting. To address this issue, **inexact RTR** has been proposed [14], where **Sub-RTR** adopts *sub-sampling* techniques that have recently been proposed in the Euclidean space. On the stochastic front, the VR methods have been recently extended to take curvature information into account as **R-SQN-VR** [15].

4. Geometries for optimization

Recent many applications in data science and engineering increasingly have *multidimensional* or *multi-array* data structure called as *tensors*. The *tensor-based learning problem* or simply the *tensor learning problem* has gained much attentions in machine learning fields. Examples of such tensor learning problems include, to name a few, tensor completion and decomposition, low-rank regression, multilinear multitask learning, and spatiotemporal regression. **Preconditioned Tucker manifold** with a *quotient manifold* structure that represents the *Tucker decomposition* of tensor has been proposed [5]. It especially uses *manifold preconditioning* with a *tailored metric* (inner product) in the *Riemannian optimization* framework on quotient manifolds. More concretely, a novel *Riemannian metric* or inner product is proposed that exploits the *second-order information* as well as the *structured symmetry* in the Tucker decomposition.

Lastly, a novel Riemannian geometry of a generalization of the *simplex* constraint to constraints with matrices, i.e., the matrix simplex constraint $\mathbf{X}_1 + \mathbf{X}_2 + \ldots + \mathbf{X}_K = \mathbf{I}$, has been proposed [16], where $\mathbf{X}_i \succeq \mathbf{0}$ is a symmetric positive semidefinite. It is called as the **simplex of positive definite matrices**. Although the constraint is a natural generalization of the simplex constraint, its study is rather limited. It focuses on developing optimization-related ingredients that allow to propose optimization algorithms on this constraint set. The expressions of the ingredients extend to the case of Hermitian positive definite matrices.

5. Furthermore

There exist active efforts for numerical tools of Riemannian optimization. Among them, a *de-facto* standard tool is Manopt [17] [*1], where some works above are integrated. In addition, McTorch [*2] has been recently released [18] for a manifold optimization library for deep learning. They support various *ready-to-use* manifold factories and Riemannian optimization solvers.

References

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

[2] 佐藤 and 笠井, "リーマン多様体上の最適化の基本と最新動向（解説論文）," **システム制御情報学会誌**, vol. 62, no. 1, 2018.

[3] B. Mishra, H. Kasai, P. Jawanpuria, and A. Saroop, "A Riemannian gossip approach to subspace learning on Grassmann manifold," *Mach. Learn.*, pp. 1–21, 2019.

[4] B. Mishra and R. Sepulchre, "R3MC: A Riemannian three-factor algorithm for low-rank matrix completion," in *IEEE CDC*, 2014, pp. 1137–1142.

[5] H. Kasai and B. Mishra, "Low-rank tensor completion: a Riemannian manifold preconditioning approach," in *ICML*, 2016.

[6] H. Kasai, "SGDLibrary: A MATLAB library for stochastic optimization algorithms," *JMLR*, vol. 18, no. 215, pp. 1–5, 2018.

[7] S. Bonnabel, "Stochastic gradient descent on Riemannian manifolds," *IEEE Trans. on Automatic Control*, vol. 58, no. 9, pp. 2217–2229, 2013.

[8] H. Sato, H. Kasai, and B. Mishra, "stochastic variance reduced gradient algorithm with retraction and vector transport," *SIAM Journal on Optimization*, vol. 29, no. 2, pp. 1444–1472, 2019.

[9] H. Kasai, H. Sato, and B. Mishra, "Riemannian stochastic variance reduced gradient on Grassmann manifold," *NIPS workshop OPT*, 2016.

[10] H. Zhang, S. J. Reddi, and S. Sra, "Riemannian SVRG: fast stochastic optimization on Riemannian manifolds," in *NIPS*, 2016.

[11] H. Kasai, H. Sato, and B. Mishra, "Riemannian stochastic recursive gradient algorithm," in *ICML*, 2018.

[12] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *NIPS*, 2013.

[13] H. Kasai, P. Jawanpuria, and B. Mishra, "Riemannian adaptive stochastic gradient algorithms on matrix manifolds," in *ICML*, 2019.

[14] H. Kasai and B. Mishra, "Inexact trust-region algorithm on Riemannian manifolds," in *NeurIPS*, 2018.

[15] H. Kasai, H. Sato, and B. Mishra, "Riemannian stochastic quasi-Newton algorithm with variance reduction and its convergence analysis," in *AISTATS*, 2018.

[16] B. Mishra, H. Kasai, and J. P., "Riemannian optimization on the simplex of positive definite matrices," *arXiv preprint arXiv:1906.10436*, 2019.

[17] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre, "Manopt, a Matlab toolbox for optimization on manifolds," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1455–1459, 2014.

[18] M. Meghwanshi, P. Jawanpuria, A. Kunchukuttan, H. Kasai, and B. Mishra, "McTorch, a manifold optimization library for deep learning," in *arXiv preprint arXiv:1810.01811*, 2018.

[*1] https://www.manopt.org/
[*2] https://github.com/mctorch/mctorch