日本オペレーションズ・リサーチ学会
2021年 春季研究発表会

適応的学習率最適化アルゴリズムのRiemann多様体への拡張と自然言語処理への応用

05001362　明治大学　*酒井 裕行　SAKAI Hiroyuki
01016200　明治大学　飯塚 秀明　IIDUKA Hideaki

1. はじめに

制約付き最適化問題の探索領域がRiemann多様体と呼ばれる幾何学的構造を持つとき、その問題はRiemann多様体上の制約なし最適化問題と捉えることができる [1]。既存の研究 [3] では、Riemann多様体が積多様体の構造を持つ場合にAMSGrad[2]の拡張であるRAMSGradが提案され、その収束解析が行われている。しかし、そこでは減少学習率の場合のregretの上限が与えられているだけであり、理論と実用の両方の観点から改善の余地があるといえる。

本発表ではRAMSGradの修正版であるアルゴリズムを紹介し、その収束率に関して議論する。

2. Riemann多様体上の適応的学習率最適化アルゴリズム

$M_1, \cdots, M_N$ を完備かつ単連結で各点での断面曲率が $\kappa_i \leq 0$ 以下であるようなRiemann多様体とし、$M := M_1 \times \cdots \times M_N$ とする。このとき、M 上の点 $x \in M$ を各成分を用いて $x = (x^1, \cdots, x^N)$ と表記する。Riemann多様体 M_i の点 x^i における接空間のRiemann計量による内積を $\langle \cdot, \cdot \rangle_{x^i}$ と表し、接ベクトル $\xi^i \in T_{x^i} M_i$ のノルムを $\left\| \xi^i \right\|_{x^i} := \sqrt{\langle \xi^i, \xi^i \rangle_{x^i}}$ と定める。M_i の点 x^i における指数写像を $\exp_{x^i} : T_{x^i} M_i \to M_i$ と表す。空でない測地的凸集合 $X_i \subset M_i$ に対して、$\Pi_{X_i} : M_i \to X_i$ を X_i への距離射影とする[1]。また、$X := X_1 \times \cdots \times X_N$ とする。さらに、$x^i, y^i \in M_i$ に対して $\varphi_{x^i \to y^i} : T_{x^i} M_i \to T_{y^i} M_i$ を接ベクトルのノルムを保つ写像[2]とする。

このとき、次の最適化問題を解くことを考える。

問題 1　$f_t : M \to \mathbb{R}$ $(t = 1, \cdots, T)$ を測地的凸関数とし、$f(x) := (1/T) \sum_{t=1}^{T} f_t(x)$ としたとき、

$$x_* \in X_* := \left\{ x_* \in X : f(x_*) = \inf_{x \in X} f(x) \right\}$$

なる点 x_* を求める。

問題 1 を解くためのアルゴリズムとして、RAMSGrad[3] が提案されている。これは、ニューラルネットワークの学習等で幅広く使われている適応的学習率最適化アルゴリズムのひとつであるAMSGrad[2] のRiemann多様体への拡張である。ただし、これは減少学習率にのみ対応しているうえ、収束解析もregretの上限が与えられただけである [3, Theorem 1]。そのため、[4] では、定数学習率にも対応するようなRAMSGradの修正版であるアルゴリズム 1 が提案され、問題 1 を解くことを保証する収束解析が与えられている。ここで、$M_i = \mathbb{R}$ とすると通常のAMSGradと一致する。

3. 収束解析

アルゴリズム 1 の問題 1 に関する収束率を命題 1 と 2 に示す。

命題 1 (定数学習率に対する収束率)　$\alpha_n := \alpha > 0$ および $\beta_{1n} := \beta \in [0, 1)$ とする。このとき、

$$\mathbb{E}\left[\frac{1}{n} \sum_{k=1}^{n} f(x_k) - f(x_*) \right] \leq \mathcal{O}\left(\frac{1}{n} \right) + C_1 \alpha + C_2 \beta$$

が成り立つ。ただし、C_1, C_2 は正の定数である。

命題 1 は、アルゴリズム 1 は定数学習率を用いる場合、$\alpha, \beta > 0$ を小さく選べば問題 1 の解を近似することを意味している。

[1] Hadamard 多様体 (単連結で完備かつ非正曲率な Riemann 多様体) では、測地的凸集合への距離射影が存在することが知られている [7]。

[2] 平行移動を用いるのが最も自然である。

アルゴリズム 1 Modified RAMSGrad [4]

Require: $(\alpha_n)_{n\in\mathbb{N}} \subset [0,1)$, $(\beta_{1n})_{n\in\mathbb{N}} \subset [0,1)$, $\beta_2 \in [0,1)$.

1: $n \leftarrow 1, x_1 \in X, \tau_0 = m_0 = 0 \in T_{x_0}M, v_0^i, \hat{v}_0^i = 0 \in \mathbb{R}$
2: **loop**
3: $t_n \in \{1, \cdots, T\}$ を一様にランダムに選ぶ。
4: $g_{t_n} = (g_{t_n}^i) = \operatorname{grad} f_{t_n}(x_n)$
5: **for** $i = 1, 2, \cdots, N$ **do**
6: $m_n^i = \beta_{1n}\tau_{n-1}^i + (1-\beta_{1n})g_{t_n}^i$
7: $v_n^i = \beta_2 v_{n-1}^i + (1-\beta_2)\left\|g_{t_n}^i\right\|_{x_n^i}^2$
8: $\hat{v}_n^i = \max\{\hat{v}_{n-1}^i, v_n^i\}$
9: $x_{n+1}^i = \Pi_{X_i}\left[\exp_{x_n^i}\left(-\alpha_n \dfrac{m_n^i}{\sqrt{\hat{v}_n^i}}\right)\right]$
10: $\tau_n^i = \varphi_{x_n^i \to x_{n+1}^i}^i(m_n^i)$
11: **end for**
12: $n \leftarrow n+1$
13: **end loop**

命題 2 (減少学習率に対する収束率) $\alpha_n := 1/n^\eta$ $(\eta \in [1/2, 1))$ および $\sum_{k=1}^{\infty} \beta_{1k} < \infty$ とする。このとき、

$$\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n} f(x_k) - f(x_*)\right] = \mathcal{O}\left(\frac{1}{n^{1-\eta}}\right).$$

が成り立つ。

命題 2 は、アルゴリズム 1 は減少学習率を用いる場合、問題 1 の解へ収束することを意味している。

4. Poincaré 埋め込み

アルゴリズム 1 の応用先として、Poincaré 埋め込み [5] がある。Poincaré 埋め込みとは、自然言語のような階層構造をもつ木構造データを双曲空間のモデルのひとつである Poincaré 円盤などに埋め込む手法である。木構造データを双曲空間に埋め込むのは、双曲空間が Euclid 空間と微分同相であり対応付けがしやすいことや、そもそも双曲空間が木構造データを埋め込むのに適しているという理由からである。

ここでは、WordNet の単語データの 5 次元 Poincaré 円盤 $\mathcal{B}^5$ への埋め込み $\Theta = \{u_i\}_{i=1}^N$ を見つける。用いる目的関数は、

$$\mathcal{L}(\Theta) = \sum_{(u,v)\in\mathcal{D}} \log \frac{e^{-d(u,v)}}{\sum_{v'\in\mathcal{N}(u)} e^{-d(u,v')}}$$

である。ここで、$\mathcal{D} = \{(u,v)\}$ は階層構造のある単語 u と v の全ての対であり、$\mathcal{N}(u) := \{v' : (u,v') \notin \mathcal{D}\} \cup \{v\}$ である。また、$d : \mathcal{B}^5 \times \mathcal{B}^5 \to \mathbb{R}$ は距離関数である。

本発表においては、アルゴリズム 1 は、[6] などの既存手法と比較して少ない反復回数と短い実行時間で、高い精度の埋め込みを達成していることを示す。

参考文献

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

[2] S. J. Reddi, S. Kale, and S. Kumar, On the convergence of Adam and beyond, Proceedings of The International Conference on Learning Representations, pp. 1-23, 2018.

[3] G. Bécigneul and O.-E. Ganea, Riemannian adaptive optimization methods, Proceedings of The International Conference on Learning Representations, 2019.

[4] H. Sakai, H. Iiduka, Riemannian Adaptive Optimization Algorithm and Its Application to Natural Language Processing, arXiv preprint arXiv:2004.00897, 2020.

[5] M. Nickel and D. Kiela, Poincaré embeddings for learning hierarchical representations, in Advances in neural information processing systems, pp. 6338-6347, 2017.

[6] S. Bonnabel, Stochastic gradient descent on Riemannian manifolds, IEEE Transactions on Automatic Control, vol. 58, no. 9, pp. 2217-2229, 2013.

[7] C. Li, G. López, and V. Martquez, Iterative algorithms for nonexpansive mappings on Hadamard manifolds, Taiwanese Journal of Mathematics, vol. 14, no. 2, pp. 541-559, 2010.