日本オペレーションズ・リサーチ学会
2022年 秋季研究発表会

全勾配を用いない勾配推定量: 1 次及び 2 次の最適性と連合学習

東京大学 / 理研 AIP　*大古 一聡　OKO Kazusato
東京大学　秋山 俊太　AKIYAMA Shunta
東京大学 / 理研 AIP　鈴木 大慈　SUZUKI Taiji

本発表では，分散縮小型確率的最適化法において全勾配を用いない新たな勾配推定量を提案し，種々の問題設定でその有用性に理論保証を与える．

1. 導入

機械学習における n データの経験損失最小化問題などは，

$$\min_{x\in\mathbb{R}^d}\left\{f(x):=\frac{1}{n}\sum_{i=1}^{n}f_i(x)\right\} \tag{1}$$

という形の最適化問題に定式化される．解 x が $\|\nabla f(x)\| \le \varepsilon$ を満たすとき ε-1 次最適であると言い，加えて $\lambda_{\min}(\nabla^2 f(x)) \ge -\delta$ を満たすとき (ε,δ)-2 次最適であると言う．

近年は大量の学習データを用いる場面が頻繁に現れるが，n が大きい場合に毎回全ての ∇f_i を計算するのは非効率である．各ステップで一部データをサンプルして勾配を計算する手法の中でも，過去の情報を利用して全勾配 $\nabla f(x)$ の推定を行う分散縮小型確率的最適化手法が，勾配計算回数の削減に成功を収めてきた [6]．

代表的な手法のうち，特に非凸最適化では，SAGA [5] は全勾配計算を必要としない一方，ε-1 次最適解の計算に $O\left(\frac{n^{2/3}}{\varepsilon^2}\right)$ 回の勾配計算を要する．一方，SARAH [4] は定期的な全勾配計算によって推定量を初期化することにより，$O\left(\frac{n^{1/2}}{\varepsilon^2}\right)$ 回の勾配計算で済み，これは下界と一致する．このトレードオフは，例えば分散学習への適用には全勾配計算はボトルネックとなるという理由から，いくつかの研究により解消が試みられてきた．その一つ，Li et al. [2] は途中の全勾配計算が不要かつ $O\left(\frac{n^{1/2}}{\varepsilon^2}\right)$ のレートを達成する ZeroSARAH を提案した．しかし，推定量の構成は複雑で，2 次収束性保証や分散学習への適用といった典型的な拡張が難しかった．[1]

[1] Li らは同時に分散学習の手法を提案しているが，彼らの証明は途中で f_i に強い均一性の仮定が必要である．

本発表では，SAGA [5] 及び SARAH [4] の中間点として，問題 (1) のための新しい勾配推定量を提案する．提案手法は ZeroSARAH と同様に途中の全勾配計算が不要かつほぼ最適な勾配計算量を達成する，1 重ループのシンプルなアルゴリズムである．それに留まらず，2 次最適性保証が可能で，さらに分散学習の一種である連合学習の設定では，通信計算量が既存手法 [3] よりも $\sqrt{n}$ 倍（n はクライアントの数）改善する応用を与える．

2. 新たな勾配推定量: SL-SARAH

提案アルゴリズムを以下に紹介する．

Algorithm 1 SL-SARAH(x^0,η,b,T,r)

Randomly sample b data I^0
$y_i^0 \leftarrow \frac{1}{b}\sum_{i\in I^0}\nabla f_i(x^0)$ $(i=1,\cdots,n)$
for $t=1$ to T **do**
　Randomly sample b data I^t
　$x^t \leftarrow x^{t-1} - \dfrac{\eta}{n}\displaystyle\sum_{i=1}^{n} y_i^{t-1}$
　$y_i^t \leftarrow \begin{cases} \nabla f_i(x^t) & \text{for } j\in I^t \\ \dfrac{1}{b}\displaystyle\sum_{j\in I^t}(\nabla f_j(x^t)-\nabla f_j(x^{t-1})) + y_i^{t-1} & \\ & \text{for } i\notin I^t \end{cases}$

提案手法の特徴は，SAGA に倣い過去の勾配を保存することに加え，SARAH に着想を得て，毎ステップミニバッチで取られたサンプルの勾配を用いて，選ばれていない他の全ての i について y_i^t を更新する点にある．この更新は y_i^t が $\nabla f_i(x^t)$ の近似と考えると一見不合理に見えるものの，$\frac{1}{n}\sum_{i=1}^{n} y_i^t$ で $\nabla f(x^t)$ を近似するには合理的であることが解析により分かる．

3. 理論保証: 1 次及び 2 次最適性

各 f_i の L-平滑性，初期点における勾配の差の有界性 ($\|\nabla f_i(x^0)-\nabla f(x^0)\| \le \sigma_0$)，及び f の有界

性 ($\Delta = f(x^0) - \inf_{x\in\mathbb{D}^d} f(x) < \infty$) を仮定した下で，以下を得る．

定理 1. $b = \sqrt{n}$ かつ $\eta = \tilde{\Theta}\left(\frac{1}{L}\right)$ の時，Algorithm 1 は確率 $1-\nu$ ($\nu \in (0,1)$) で，ε-1 次最適解に

$$\tilde{O}\left(\frac{L\Delta\sqrt{n} + \sigma_0^2\sqrt{n}}{\varepsilon^2}\right)$$

回の勾配計算量で到達する．

これは [2] と同等の結果である．

加えて各 f_i の Hessian の ρ-Lipschitz 連続性の成立を仮定し，Algorithm 1 の 5 行目の更新を原点を中心とし半径 r の d 次元球の内部からの一様分布 $B(r)$ に従う ξ^t を用いて $x^t \leftarrow x^{t-1} - \frac{\eta}{n}\sum_{i=1}^n y_i^{t-1} + \xi^t$ に変更すると，以下の定理を得る．

定理 2. $b \geq \sqrt{n} + \frac{L^2}{\delta^2}, \eta = \tilde{\Theta}\left(\frac{1}{L}\right), r \ll 1$ の時，Algorithm 1 は確率 $1-\nu$ で，(ε,δ)-2 次最適解に

$$\tilde{O}\left(\left(L\Delta + \sigma_0^2\right)\left(\frac{\sqrt{n}}{\varepsilon^2} + \frac{\rho^2\sqrt{n}}{\delta^4} + \frac{L^2}{\varepsilon^2\delta^2} + \frac{L^2\rho^2}{\delta^6}\right)\right)$$

回の勾配計算量で到達する．

提案手法は鞍点を抜け出すためのサブルーチンを必要とせず，1 重ループで 2 次最適性を保証する初の手法であり，特定の設定では既存手法 [1] よりも効率的である．更にアルゴリズムを工夫すると，$\frac{L^2}{\varepsilon^2\delta^2} + \frac{L^2\rho^2}{\delta^6}$ の部分を $\frac{L}{\varepsilon^2\delta} + \frac{L\rho^2}{\delta^5}$ に改善できる．

これらに加え，PL 条件下での解析も可能である．

4. 連合学習への応用

Algorithm 1 の拡張性は高く，分散学習の一種である連合学習の手法を作ることができる．ここでは，n 個のクライアントがそれぞれ m 個のデータを持つ状況を考え，問題 (1) で特に f_i が

$$f_i(x) := \frac{1}{m}\sum_{j=1}^m f_{i,j}(x)$$

で定まるとする．スペースの都合上結果のみ示す．

定理 1 の仮定に加え，$\{f_i\}_{i=1}^n$ は 2 次の不均一性を満たす，即ち任意の i, x に対して $\|\nabla^2 f_i(x) - \nabla^2 f(x)\| \leq \zeta$ が成り立つとする．各クライアントで $\|\nabla f_{i,j}(x) - \nabla f_i(x)\| \leq \sigma$ ($\forall j, x$) であるとする．さらに簡単のため，$\rho, \Delta, L = \Theta(1)$ とする．

定理 3. 提案アルゴリズムは，途中 1 回あたり $\sqrt{n}$ の通信量を用い，確率 $1-\nu$ で，通信回数

$$\tilde{O}\left(1 + \frac{\zeta}{\varepsilon^2} + \frac{\sigma^2}{n\mathcal{B}\varepsilon^4} + \frac{\sigma^2}{\mathcal{B}^{\frac{1}{2}}\varepsilon^2}\right)$$

で ε-1 次最適解に到達する．ただし $\mathcal{B}$ は 1 ラウンドに各クライアント i が計算できる勾配の数．

これは毎回の通信を全てのクライアントで行っていた [3] の結果を改善する．スペースの都合上詳細は省くが，定理 2 と 3 の解析を組み合わせ，連合学習の状況で 2 次最適性を付与することもできる．

謝辞

本研究の過程では，村田智也氏に貴重な助言を頂いた．また本研究は，JST-CREST (JPMJCR2015, JPMJCR2115) の支援を受けたものである．

参考文献

[1] Zhize Li. SSRGD: Simple stochastic recursive gradient descent for escaping saddle points. *Advances in Neural Information Processing Systems*, 32, 2019.

[2] Zhize Li, Slavomír Hanzely, and Peter Richtárik. ZeroSARAH: Efficient nonconvex finite-sum optimization with zero full gradient computation. *arXiv preprint arXiv:2103.01447*, 2021.

[3] Tomoya Murata and Taiji Suzuki. Bias-variance reduced local SGD for less heterogeneous federated learning. In *International Conference on Machine Learning*, pages 7872–7881. PMLR, 2021.

[4] Lam M Nguyen, Marten van Dijk, Dzung T Phan, Phuong Ha Nguyen, Tsui-Wei Weng, and Jayant R Kalagnanam. Finite-sum smooth optimization with SARAH. *Computational Optimization and Applications*, pages 1–33, 2022.

[5] Sashank J Reddi, Suvrit Sra, Barnabás Póczos, and Alex Smola. Fast incremental method for smooth nonconvex optimization. In *2016 IEEE 55th conference on decision and control (CDC)*, pages 1971–1977. IEEE, 2016.

[6] Nicolas Roux, Mark Schmidt, and Francis Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. *Advances in neural information processing systems*, 25:2663–2671, 2012.