PEKING UNIVERSITY 北京大学

UNIVERSITY of ROCHESTER

Efficient Smooth Non-Convex Stochastic Compositional Optimization via Stochastic Recursive Gradient Descent

Huizhuo Yuan, Xiangru Lian, Chris Junchi Li, and Ji Liu

快手 KUAISHOU TECHNOLOGY

Tencent 腾讯

Generalized Eigenvector Estimation

Composition of two expectations of stochastic functions:

$$\min_{x\in\mathbb{R}^d} \quad \{\Phi(x) \equiv (f\circ g)(x)\} \tag{1}$$

- Outer function $f:\mathbb{R}^l \to \mathbb{R}$ is defined as $f(y) := \mathbb{E}_v[f_v(y)]$
- Inner function $g:\mathbb{R}^d \to \mathbb{R}^l$ is $g(x) := \mathbb{E}_w[g_w(y)]$
- Each stochastic component f_v, g_w are smooth but *not* necessarily convex.
- Compositional optimization can be used to formulate many important machine learning problems, e.g. reinforcement learning, risk management, multi-stage stochastic programming and deep neural net, etc.

In this paper, simplified to

$$\Phi(x) = \frac{1}{n}\sum_{i=1}^{n} f_i\left(\frac{1}{m}\sum_{j=1}^{m} g_j(x)\right) \tag{2}$$

Derivation of Online GEV

$$\Phi(x) = \frac{1}{n}\sum_{i=1}^{n} f_i\left(\frac{1}{m}\sum_{j=1}^{m} g_j(x)\right) \tag{3}$$

We can conduct (via the chain rule) the gradient descent iteration

$$x_{t+1} = x_t - \eta[\partial g(x_t)]^\top \nabla f(g(x_t)) \tag{4}$$

where $\partial g(x)$ is the Jacobian matrix of $g(x)$ and $\nabla f(y)$ is the gradient of $f(y)$

- Involves computing $g(x_t) = \frac{1}{m}\sum_{j=1}^{m} g_j(x_t)$ at each interation, which is often time-consuming in big data applications
- SCGD [6] introduce a two-time-scale algorithm called Stochastic Compositional Gradient Descent (SCGD) along with its accelerated (in Nesterov's sense) variant Acc-SCGD
- Many other follow-up works [7, 4, 3, 2]

We design a novel algorithm called `SARAH-Compositional` based on Stochastic Compositional Variance Reduced Gradient method (see [3]), hybriding with the stochastic recursive gradient method [5]

Stochastic Scaled Gradient Descent

Informal `SARAH-Compositional` algorithm:

$$\boldsymbol{g}_t = g_{j_{2,t}}(x_t) - g_{j_{2,t}}(x_{t-1}) + \boldsymbol{g}_{t-1}$$
$$\boldsymbol{G}_t = \partial g_{j_{2,t}}(x_t) - \partial g_{j_{2,t}}(x_{t-1}) + \boldsymbol{G}_{t-1}$$
$$\boldsymbol{F}_t = (\boldsymbol{G}_t)^\top \nabla f_{i_{2,t}}(\boldsymbol{g}_t)$$

once every q steps update using a large minibatch

- For appropriately chosen constant stepsize $\eta > 0$, update the iteration via $x_{t+1} = x_t - \eta \boldsymbol{F}_t$
- Output $\tilde{x}$ chosen uniformly at random from $\{x_t\}_{t=0}^{T-1}$

Strict-Saddle Property

Theorem. Let some smoothness and boundedness assumptions hold, as well as some finite variance assumptions (online case).

(1) **Finite-sum case:** Let $q = (2m+n)/3$ and set the stepsize $\eta \asymp 1/\sqrt{2m+n}$. The IFO complexity for `SARAH-Compositional` to achieve an ε-accurate solution is bounded by

$$\lesssim 2m + n + (2m+n)^{1/2}\varepsilon^{-2} \tag{5}$$

(2) **Online case:** Once every q iterates we sample a large mini-batches $\mathcal{A}_1, \mathcal{B}_1, \mathcal{C}_1$ of size $\asymp \sigma^2/\varepsilon^2$.[a] Let $q \asymp \sigma^2/\varepsilon^2$ (depending on variance of noise) and set the stepsize $\eta \asymp \varepsilon/\sigma$. The IFO complexity for `SARAH-Compositional` to achieve an ε-accurate

$$\lesssim \sigma^2\varepsilon^{-2} + \sigma\cdot\varepsilon^{-3}. \tag{6}$$

[a] To estimate the (products of) derivatives of the ground truth

Convergence Rate Results for SSGD

Remark. (1) `SARAH-Compositional` algorithm achieve a reduced IFO complexities of $\mathcal{O}\left((m+n)^{1/2}\varepsilon^{-2} \wedge \varepsilon^{-3}\right)$ for both finite-sum and online cases. [a]

(2) Experimentally, we compare our new compositional optimization method with a few rival algorithms, and show `SARAH-Compositional` can be a useful algorithm for tasks including portfolio management & reinforcement learning

Future directions include: (1) non-smooth case (2) theory of lower bounds for stochastic compositional optimization

[a] Similar form shared by the complexity of `SPIDER-SFO` (SARAH variant) [1, 8]) and is *optimal* since it matches the theoretical lower bound. In need of new lower-bound results to justify the optimality of `SARAH-Compositional` due to different assumptions

Thanks For Your Attention

References

[1] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path integrated differential estimator. *arXiv preprint arXiv:1807.01695*, 2018.

[2] Zhouyuan Huo, Bin Gu, Ji Liu, and Heng Huang. Accelerated method for stochastic composition optimization with nonsmooth regularization. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[3] Tianyi Lin, Chenyou Fan, Mengdi Wang, and Michael I Jordan. Improved oracle complexity for stochastic compositional variance reduced gradient. *arXiv preprint arXiv:1806.00458*, 2018.

[4] Liu Liu, Ji Liu, and Dacheng Tao. Variance reduced methods for non-convex composition optimization. *arXiv preprint arXiv:1711.04416*, 2017.

[5] Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pages 2613–2621, 2017.

[6] Mengdi Wang, Ethan X Fang, and Han Liu. Stochastic compositional gradient descent: Algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161(1-2):419–449, 2017.

[7] Mengdi Wang, Ji Liu, and Ethan X Fang. Accelerating stochastic composition optimization. *Journal of Machine Learning Research*, 18:1–23, 2017.

[8] Zhe Wang, Kaiyi Ji, Yi Zhou, Yingbin Liang, and Vahid Tarokh. Spiderboost: A class of faster variance-reduced algorithms for nonconvex optimization. *arXiv preprint arXiv:1810.10690*, 2018.