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

Peking University 北京大学 · University of Rochester · 快手 Kuaishou Technology · Tencent 腾讯

Wenqing Hu* Chris Junchi Li* Xiangru Lian* Ji Liu* Huizhuo Yuan*

Stochastic Compositional Optimization

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 $g(x) := \mathbb{E}_w[g_w(y)]$
- f_v, g_w are smooth but *not* necessarily convex.
- Important machine learning problems, e.g. reinforcement learning, risk management, multi-stage stochastic programming and deep neural net, etc.

Real World Applications

For notation simplicity, write

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

Risk management problem

$$\min_{x\in\mathbb{R}^N} \quad -\frac{1}{T}\sum_{t=1}^{T}\langle r_t, x\rangle + \frac{1}{T}\sum_{t=1}^{T}\left(\langle r_t, x\rangle - \frac{1}{T}\sum_{s=1}^{T}\langle r_s, x\rangle\right)^2 \tag{3}$$

Value function evaluation in reinforcement learning

$$\mathbb{E}\left(V^\pi(s_1) - \mathbb{E}[r_{s_1,s_2} + \gamma V^\pi(s_2)|s_1]\right)^2 \tag{4}$$

Algorithm

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

$$x_{t+1} = x_t - \eta[\partial g(x_t)]^\top \nabla f(g(x_t)) \tag{5}$$

- 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 Wang et al. (2017a) 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 Wang et al. (2017b); Liu et al. (2017); Lin et al. (2018); Huo et al. (2018)

We design a novel algorithm called `SARAH-Compositional` based on Stochastic Compositional Variance Reduced Gradient method (see Lin et al. (2018)), hybriding with the stochastic recursive gradient method Nguyen et al. (2017)

`SARAH-Compositional` Algorithm

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 $\{x_t\}_{t=0}^{T-1}$

Convergence Theorem

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

$$\lesssim 2m + n + (2m+n)^{1/2}\varepsilon^{-2} \tag{6}$$

(2) **Online case:** Once every q iterates we sample a large minibatches $\mathcal{A}_1, \mathcal{B}_1, \mathcal{C}_1$ 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

$$\lesssim \sigma^2\varepsilon^{-2} + \sigma \cdot \varepsilon^{-3}. \tag{7}$$

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

Experimental Results

- Risk management problem
- Finite-sum case
- Search optimal stepsize in each model.

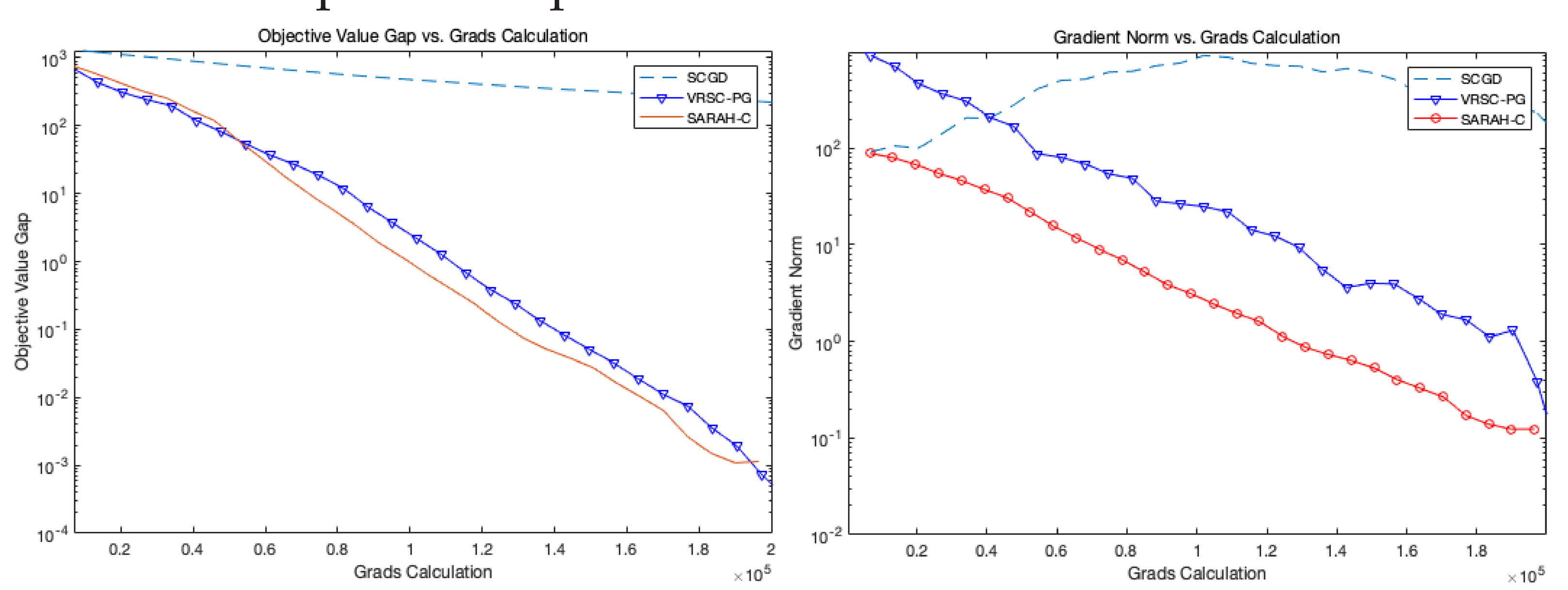

Figure 1: Experiment on the portfolio management. The x-axis is the number of gradients calculations, the y-axis is the function value gap and the norm of gradient respectively. The risk matrix are generated by a Gaussian distribution with covariance matrix Σ, $\kappa(\Sigma) = 20$

Convergence Rate of `SARAH-Compositional`

Algorithm	Finite-sum	Online
SCGD (Wang et al., 2017a)	unknown	ε^{-8}
Acc-SCGD (Wang et al., 2017a)	unknown	ε^{-7}
SCGD (Wang et al., 2017b)	unknown	$\varepsilon^{-4.5}$
SCVR / SC-SCSG (Liu et al., 2017)	$(n+m)^{4/5}\varepsilon^{-2}$	$\varepsilon^{-3.6}$
VRSC-PG (Huo et al., 2018)	$(n+m)^{2/3}\varepsilon^{-2}$	unknown
`SARAH-Compositional`[a]	$(n+m)^{1/2}\varepsilon^{-2}$	ε^{-3}

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) Fang et al. (2018); Wang et al. (2018)) 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

Fang, C., Junchi Li, C., Lin, Z., and Zhang, T. (2018). Spider: Near-optimal non-convex optimization via stochastic path integrated differential estimator. *arXiv preprint arXiv:1807.01695*.

Huo, Z., Gu, B., Liu, J., and Huang, H. (2018). Accelerated method for stochastic composition optimization with nonsmooth regularization. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Lin, T., Fan, C., Wang, M., and Jordan, M. I. (2018). Improved oracle complexity for stochastic compositional variance reduced gradient. *arXiv preprint arXiv:1806.00458*.

Liu, L., Liu, J., and Tao, D. (2017). Variance reduced methods for non-convex composition optimization. *arXiv preprint arXiv:1711.04416*.

Nguyen, L. M., Liu, J., Scheinberg, K., and Takáč, M. (2017). Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pages 2613–2621.

Wang, M., Fang, E. X., and Liu, H. (2017a). Stochastic compositional gradient descent: Algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161(1-2):419–449.

Wang, M., Liu, J., and Fang, E. X. (2017b). Accelerating stochastic composition optimization. *Journal of Machine Learning Research*, 18:1–23.

Wang, Z., Ji, K., Zhou, Y., Liang, Y., and Tarokh, V. (2018). Spiderboost: A class of faster variance-reduced algorithms for nonconvex optimization. *arXiv preprint arXiv:1810.10690*.