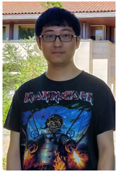
Dimension-free Complexity Bounds for High-order Nonconvex Finite-sum Optimization

Dongruo Zhou [1] Quanquan Gu [1]



[1]Department of Computer Science, UCLA

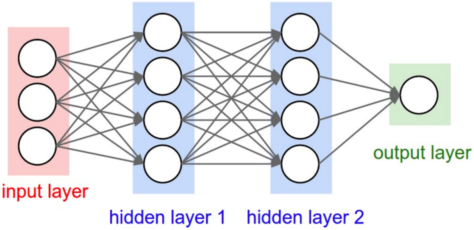
July 18, 2022

Nonconvex Finite-sum Optimization

Optimization problem:

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}), \ \ f_i \text{ can be nonconvex.}$$

Very common in machine learning!



NP-hard to solve in general (Hillar and Lim, 2013)

First-Order Stationary Points

We aim to find a first-order stationary point $\mathbf{x}$, where

$$\nabla F(\mathbf{x}) = \mathbf{0}.$$

Why stationary points? In some cases, stationary points are *global minima*!

For instance, gradient dominant functions!

First-Order Stationary Points

We aim to find a first-order stationary point $\mathbf{x}$, where

$$\nabla F(\mathbf{x}) = \mathbf{0}.$$

Why stationary points? In some cases, stationary points are *global minima*!

For instance, gradient dominant functions!

Low-rank matrix factorization (Ge, Lee, and Ma, 2016)

$$f_{i,j}(\mathbf{X}) = (\mathbf{M}_{i,j} - [\mathbf{X}\mathbf{X}^\top]_{i,j})^2$$

Training deep linear networks (Hardt and Ma, 2016; Zou, Long, and Gu, 2020)

$$f_i(\mathbf{A}_1, \ldots, \mathbf{A}_L) = \|\mathbf{y}_i - (\mathbf{I} + \mathbf{A}_1) \cdots (\mathbf{I} + \mathbf{A}_L)\mathbf{x}_i\|_2^2$$

Goal: To find an ϵ-stationary point $\mathbf{x}$, where $\|\nabla F(\mathbf{x})\| \leq \epsilon$

Complexity measure: Number of calls to each f_i, obtain $(\nabla f_i, \nabla^2 f_i, \ldots)$

High-Order Regularization Method (Birgin et al., 2017)

Starting from $\mathbf{x}_0$, at round t, given current iterate $\mathbf{x}_t$,

- Construct the p-th order Taylor approximation at $\mathbf{x}_t$, that is

$$F(\mathbf{x}_t + \mathbf{h}) \approx F(\mathbf{x}_t) + m_t^p(\mathbf{h}), \ m_t^p(\mathbf{h}) = \sum_{i=1}^{p} \langle \nabla^i F(\mathbf{x}_t), \mathbf{h}^{\otimes i} \rangle + \frac{M_t}{(p+1)!} \|\mathbf{h}\|^{p+1}.$$

- Compute $\mathbf{h}_t$ as the approximate minimizer of $m_t^p(\mathbf{h})$, where

$$\mathbf{h}_t \approx \operatorname*{argmin}_{\mathbf{h} \in \mathbb{R}^d} m_t^p(\mathbf{h}).$$

- Update iterate $\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{h}_t$
- Special cases: gradient descent ($p = 1$), cubic regularization of Newton method (Nesterov and Polyak, 2006) ($p = 2$)

Convergence of p-th order regularization method

Theorem (Birgin et al. 2017)

p-th order regularization method converges to a ϵ-stationary point within

$$O(n\epsilon^{-(p+1)/p})$$

number of oracle calls.

Convergence of p-th order regularization method

Theorem (Birgin et al. 2017)

p-th order regularization method converges to a ϵ-stationary point within

$$O(n\epsilon^{-(p+1)/p})$$

number of oracle calls.

Can we design an algorithm whose p-th order oracle complexity has a sublinear dependence on n, and the best dependence on ϵ?

Stochastic p-th Order Method: Derivative Estimators

We construct semi-stochastic estimations $\mathbf{J}_t^{(i)} \approx \nabla^i F(\mathbf{x}_t), i = 1, \cdots, p$, then let

$$\mathbf{h}_t = \operatorname*{argmin}_{\mathbf{h} \in \mathbb{R}^d} \widehat{m}_t^p(\mathbf{h}) = \sum_{i=1}^{p} \langle \mathbf{J}_t^{(i)}, \mathbf{h}^{\otimes i} \rangle + \frac{M_t}{(p+1)!} \|\mathbf{h}\|^{p+1}$$

Then update

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{h}_t$$

Stochastic p-th Order Method: Derivative Estimators

We construct semi-stochastic estimations $\mathbf{J}_t^{(i)} \approx \nabla^i F(\mathbf{x}_t), i = 1, \cdots, p$, then let

$$\mathbf{h}_t = \operatorname*{argmin}_{\mathbf{h} \in \mathbb{R}^d} \widehat{m}_t^p(\mathbf{h}) = \sum_{i=1}^{p} \langle \mathbf{J}_t^{(i)}, \mathbf{h}^{\otimes i} \rangle + \frac{M_t}{(p+1)!} \|\mathbf{h}\|^{p+1}$$

Then update

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{h}_t$$

How to construct estimated tensor $\mathbf{J}_t^{(i)}$?

- One-point Taylor expansion estimator (OP-TE)
  - Inspired by variance-reduced gradient/Hessian (Johnson and Zhang, 2013; Zhou, Xu, and Gu, 2018)
- Two-point Taylor expansion estimator (TP-TE)
  - Inspired by recursive variance-reduced gradient/Hessian (Nguyen et al., 2017; Fang et al., 2018; Shen et al., 2019)

Theoretical Results

Theorem (p-th order oracle complexity for OP-TE and TP-TE)

With specific parameter choices, OP-TE will find an ϵ-stationary point within

$$\widetilde{O}(n^{(3p-1)/(3p)}\epsilon^{-(p+1)/p})$$

number of stochastic p-th order oracle calls, TP-TE will find an ϵ-stationary point within

$$\widetilde{O}(n^{(2p-1)/(2p)}\epsilon^{-(p+1)/p})$$

number of stochastic p-th order oracle calls. $\widetilde{O}(\cdot)$ hides logarithmic terms and polynomial term of p.

- Dimension-free bounds!

Complexity Comparison

Algorithm	p-th order oracle complexity
HR (Birgin et al., 2017)	$O\left(\frac{n}{\epsilon^{(p+1)/p}}\right)$
OP-TE (This work)	$O\left(\frac{n^{(3p-1)/(3p)}}{\epsilon^{(p+1)/p}}\right)$
TP-TE (This work)	$O\left(\frac{n^{(2p-1)/(2p)}}{\epsilon^{(p+1)/p}}\right)$
Lower bound (Emmenegger, Kyng, and Zehmakan, 2021)	$\Omega\left(\frac{n^{(p-1)/(2p)}}{\epsilon^{(p+1)/p}}\right)$

- Improves HR by a $n^{1/(2p)}$ factor!
- Still $\sqrt{n}$ away from lower bound ...

Reference I

Birgin, Ernesto G et al. (2017). "Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models". In: *Mathematical Programming* 163.1-2, pp. 359–368.

Emmenegger, Nicolas, Rasmus Kyng, and Ahad N Zehmakan (2021). "On the Oracle Complexity of Higher-Order Smooth Non-Convex Finite-Sum Optimization". In: *arXiv preprint arXiv:2103.05138*.

Fang, Cong et al. (2018). "SPIDER: Near-Optimal Non-Convex Optimization via Stochastic Path-Integrated Differential Estimator". In: *Advances in Neural Information Processing Systems*, pp. 686–696.

Ge, Rong, Jason D Lee, and Tengyu Ma (2016). "Matrix completion has no spurious local minimum". In: *Advances in Neural Information Processing Systems*, pp. 2973–2981.

Hardt, Moritz and Tengyu Ma (2016). "Identity matters in deep learning". In: *arXiv preprint arXiv:1611.04231*.

Hillar, Christopher J and Lek-Heng Lim (2013). "Most tensor problems are NP-hard". In: *Journal of the ACM (JACM)* 60.6, p. 45.

Johnson, Rie and Tong Zhang (2013). "Accelerating stochastic gradient descent using predictive variance reduction". In: *Advances in neural information processing systems*, pp. 315–323.

Nesterov, Yurii and B. T. Polyak (2006). "Cubic regularization of Newton method and its global performance". In: *Mathematical Programming* 108.1, pp. 177–205.

Nguyen, Lam M et al. (2017). "SARAH: A Novel Method for Machine Learning Problems Using Stochastic Recursive Gradient". In: *International Conference on Machine Learning*, pp. 2613–2621.

Shen, Zebang et al. (2019). "A Stochastic Trust Region Method for Non-convex Minimization". In: *arXiv preprint arXiv:1903.01540*.

Reference II

Zhou, Dongruo, Pan Xu, and Quanquan Gu (2018). "Stochastic Variance-Reduced Cubic Regularized Newton Methods". In: *Proceedings of the 35th International Conference on Machine Learning*. PMLR, pp. 5990–5999.

Zou, Difan, Philip M. Long, and Quanquan Gu (2020). "On the Global Convergence of Training Deep Linear ResNets". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=HJxEhREKDH.