SSRGD: Simple Stochastic Recursive Gradient Descent for Escaping Saddle Points

Tsinghua University

King Abdullah University of Science and Technology

Zhize Li

Tsinghua University, and KAUST

Problem

We consider two types of **nonconvex** problems.
1) The **finite-sum** problem:

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

where $f(x)$ and all individual $f_i(x)$ are possibly nonconvex.
2) The **online (expectation)** problem:

$$\min_{x\in\mathbb{R}^d} f(x) := \mathbb{E}_{\zeta\sim\mathcal{D}}[F(x,\zeta)], \qquad (2)$$

where $f(x)$ and $F(x,\zeta)$ are possibly nonconvex.

Definition

Assumption 1 (**Gradient Lipschitz**)

$$\|\nabla f_i(x_1) - \nabla f_i(x_2)\| \le L\|x_1 - x_2\|, \quad \forall x_1, x_2.$$

Assumption 2 (**Hessian Lipschitz**)

$$\|\nabla^2 f_i(x_1) - \nabla^2 f_i(x_2)\| \le \rho\|x_1 - x_2\|, \quad \forall x_1, x_2.$$

Convergence guarantee:
- **ϵ-first-order** stationary point: $\|\nabla f(x)\| \le \epsilon$.
- **(ϵ,δ)-second-order** stationary point: $\|\nabla f(x)\| \le \epsilon$ and $\lambda_{\min}(\nabla^2 f(x)) \ge -\delta$.

Note that **$\nabla f(x) = 0$ and $\nabla^2 f(x) \succ 0$** $\Rightarrow x$ is a **local minimum**.

Convergence Result

Table 1: Stochastic gradient complexity of optimization algorithms for nonconvex **finite-sum** problem (1)

Algorithm	Stochastic gradient complexity	Convergence guarantee	Negative-curvature search subroutine
GD [Nesterov, 2004]	$O(\frac{n}{\epsilon^2})$	ϵ-first-order	No
SVRG [Reddi et al., 2016], [Allen-Zhu and Hazan, 2016]; SCSG [Lei et al., 2017]; SVRG+ [Li and Li, 2018]	$O(n+\frac{n^{2/3}}{\epsilon^2})$	ϵ-first-order	No
SNVRG [Zhou et al., 2018b]; SPIDER [Fang et al., 2018]; SpiderBoost [Wang et al., 2018]; SARAH [Pham et al., 2019]	$O(n+\frac{n^{1/2}}{\epsilon^2})$	ϵ-first-order	No
SSRGD (this paper)	$O(n+\frac{n^{1/2}}{\epsilon^2})$	ϵ-first-order	No
PGD [Jin et al., 2017]	$\widetilde{O}(\frac{n}{\epsilon^2}+\frac{n}{\delta^4})$	(ϵ,δ)-second-order	No
Neon2+FastCubic/CDHS [Agarwal et al., 2016, Carmon et al., 2016]	$\widetilde{O}(\frac{n}{\epsilon^{1.5}}+\frac{n}{\delta^3}+\frac{n^{3/4}}{\epsilon^{1.75}}+\frac{n^{3/4}}{\delta^{3.5}})$	(ϵ,δ)-second-order	Needed
Neon2+SVRG [Allen-Zhu and Li, 2018]	$\widetilde{O}(\frac{n^{2/3}}{\epsilon^2}+\frac{n}{\delta^3}+\frac{n^{3/4}}{\delta^{3.5}})$	(ϵ,δ)-second-order	Needed
Stabilized SVRG [Ge et al., 2019]	$\widetilde{O}(\frac{n^{2/3}}{\epsilon^2}+\frac{n}{\delta^3}+\frac{n^{2/3}}{\delta^4})$	(ϵ,δ)-second-order	No
SNVRG$^+$+Neon2 [Zhou et al., 2018a]	$\widetilde{O}(\frac{n^{1/2}}{\epsilon^2}+\frac{n}{\delta^3}+\frac{n^{3/4}}{\delta^{3.5}})$	(ϵ,δ)-second-order	Needed
SPIDER-SFO$^+$(+Neon2) [Fang et al., 2018]	$\widetilde{O}(\frac{n^{1/2}}{\epsilon^2}+\frac{n^{1/2}}{\epsilon\delta^2}+\frac{1}{\epsilon\delta^3}+\frac{1}{\delta^5})$	(ϵ,δ)-second-order	Needed
SSRGD (this paper)	$\widetilde{O}(\frac{n^{1/2}}{\epsilon^2}+\frac{n^{1/2}}{\delta^4}+\frac{n}{\delta^3})$	(ϵ,δ)-second-order	No

- We improve the result of Stabilized SVRG [Ge et al., 2019] to almost optimal, i.e., from $n^{2/3}/\epsilon^2$ to $n^{1/2}/\epsilon^2$ since [Fang et al., 2018] gave a lower bound $\Omega(n^{1/2}/\epsilon^2)$ for finding even just an ϵ-first-order stationary point. Also, our SSRGD is better than SPIDER-SFO$^+$ if δ is very small (e.g., $\delta \le 1/\sqrt{n}$).
- Note that the other two $n^{1/2}$ algorithms (SNVRG$^+$ and SPIDER-SFO$^+$) need the negative curvature search subroutine (e.g., Neon/Neon2) for escaping the saddle points while our SSRGD only needs to add random perturbations.
- Besides, we also prove the convergence results for nonconvex **online (expectation)** problem (2).

References

Rong Ge, Zhize Li, Weiyao Wang, and Xiang Wang. Stabilized SVRG: Simple Variance Reduction for Nonconvex Optimization. In *COLT*, 2019.

Lam M. Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: A Novel Method for Machine Learning Problems Using Stochastic Recursive Gradient. In *ICML*, 2017.

Algorithm

Algorithm 1: Simple Stochastic Recursive Gradient Descent (SSRGD)

input: initial point x_0, epoch length m, minibatch size b, step size η, perturbation radius r, threshold gradient g_{thres}, threshold function value f_{thres}, super epoch length t_{thres}.
$\mathit{super\_epoch} \leftarrow 0$
for $s = 0, 1, 2, \ldots$ **do**
  if $\mathit{super\_epoch} = 0$ and $\|\nabla f(x_{sm})\| \le g_{\text{thres}}$ **then**
    $\widetilde{x} \leftarrow x_{sm},\ t_{\text{init}} \leftarrow sm,\ \mathit{super\_epoch} \leftarrow 1$
    $x_{sm} \leftarrow \widetilde{x} + \xi$, where ξ uniformly $\sim \mathbb{B}_0(r)$ // we use super epoch since we do not want to add the perturbation too often near a saddle point
  $v_{sm} \leftarrow \nabla f(x_{sm})$ // compute full gradient every m steps
  for $k = 1, 2, \ldots, m$ **do**
    $t \leftarrow sm + k$
    $x_t \leftarrow x_{t-1} - \eta v_{t-1}$
    $v_t \leftarrow \frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x_t) - \nabla f_i(x_{t-1})\big) + v_{t-1}$ // I_b are i.i.d. samples with $|I_b| = b$
    if $\mathit{super\_epoch} = 1$ and $(f(\widetilde{x}) - f(x_t) \ge f_{\text{thres}}$ or $t - t_{\text{init}} \ge t_{\text{thres}})$ **then**
      $\mathit{super\_epoch} \leftarrow 0$
  $x_{(s+1)m} \leftarrow x_t$

Parameters: $m = \sqrt{n},\ b = \sqrt{n},\ \eta = \widetilde{O}(\frac{1}{L}),\ r = \widetilde{O}\big(\min(\frac{\delta^3}{\rho^2\epsilon}, \frac{\delta^{3/2}}{\rho\sqrt{L}})\big),\ g_{\text{thres}} = \epsilon,\ f_{\text{thres}} = \widetilde{O}(\frac{\delta^3}{\rho^2}),\ t_{\text{thres}} = \widetilde{O}(\frac{1}{\eta\delta})$

Proof Overview

- **1. Large gradients: $\|\nabla f(x)\|^2 > g_{\text{thres}} = \epsilon$**

Key relation between $f(x_t)$ and $f(x_{t-1})$, where $x_t = x_{t-1} - \eta v_{t-1}$.

$$f(x_t) \le f(x_{t-1}) - \frac{\eta}{2}\|\nabla f(x_{t-1})\|^2 - \Big(\frac{1}{2\eta} - \frac{L}{2}\Big)\|x_t - x_{t-1}\|^2 + \frac{\eta}{2}\|\nabla f(x_{t-1}) - v_{t-1}\|^2. \qquad (3)$$

Observ.: cancel the last two terms $\Rightarrow$ get an ϵ-first-order stationary point ($\|\nabla f(x)\| \le \epsilon$) in $\frac{2(f(x_0)-f^*)}{\eta\epsilon^2}$ steps.

▷ First consider the gradient estimator v_t in SVRG papers (convergence result $O(n^{2/3}/\epsilon^2)$):

$v_t \leftarrow \frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x_t) - \nabla f_i(\widetilde{x})\big) + \nabla f(\widetilde{x})$ (reuse the fixed snapshot full gradient $\nabla f(\widetilde{x})$).

Bound the third term (variance): $\mathbb{E}\big[\|\nabla f(x_{t-1}) - v_{t-1}\|^2\big] \le \frac{L^2}{b}\mathbb{E}[\|x_{t-1} - \widetilde{x}\|^2]$.

Connect with the second term by using Young's inequality: $-\|x_t - x_{t-1}\|^2 \le \frac{1}{\alpha}\|x_{t-1} - \widetilde{x}\|^2 - \frac{1}{1+\alpha}\|x_t - \widetilde{x}\|^2$.

Sum up (3) for each epoch s (m steps) to **cancel the last two terms**:

$\mathbb{E}[f(x_{(s+1)m})] \le \mathbb{E}[f(x_{sm})] - \frac{\eta}{2}\sum_{j=sm+1}^{sm+m}\mathbb{E}[\|\nabla f(x_{j-1})\|^2]$ (need $b \ge m^2$ due to Young's inequality).

Thus the **convergence result** is $T(b + \frac{n}{m}) = \frac{1}{\epsilon^2}(b + \frac{n}{m}) = \frac{n^{2/3}}{\epsilon^2}$ by choosing $b = m^2 = n^{2/3}$ due to $b \ge m^2$.

▷ Now consider the recursive gradient estimator (originally introduced by [Nguyen et al. 2017]) in Algo 1:

Only need to bound the third term: $\mathbb{E}\big[\|\nabla f(x_{t-1}) - v_{t-1}\|^2\big] \le \frac{L^2}{b}\sum_{j=sm+1}^{t-1}\mathbb{E}[\|x_j - x_{j-1}\|^2]$.

Already connect with the second term, and **sum up** (3) for each epoch to cancel the last two terms.

Thus the **convergence result** is $T(b + \frac{n}{m}) = \frac{1}{\epsilon^2}(b + \frac{n}{m}) = \frac{n^{1/2}}{\epsilon^2}$ by choosing $b = m = n^{1/2}$.

- **2. Around saddle points: $\|\nabla f(\widetilde{x})\|^2 \le \epsilon$ and $\lambda_{\min}(\nabla^2 f(\widetilde{x})) \le -\delta$**

i) Localization: $\forall t,\ \|x_t - x_0\| \le \sqrt{t(f(x_0) - f(x_t))}$. If function value does not decrease so much, then all iteration points are not far from the start point.

ii) Small stuck region in the random perturbation ball: $\exists t \le t_{\text{thres}},\ \|x_t - x_0\| \ge \Omega(\delta)$. After the perturbation $x_0 = \widetilde{x} + \xi$, x_0 will escape this saddle point in a super epoch, i.e., within t_{thres} steps.