Inexact SARAH for Solving Stochastic Optimization Problems

Lam M. Nguyen[1,2] · Katya Scheinberg[1] · Martin Takáč[1]

[1]Lehigh University · [2]IBM Thomas J. Watson Research Center

SARAH

The Problem and Assumptions

The Problem:

$$\min_{w \in \mathbb{R}^d} \left\{ F(w) = \mathbb{E}[f(w;\xi)] \right\}$$

– ξ is a random variable obeying some distribution

Assumptions:

- $f(w;\xi)$ is L-smooth for every realization of ξ $\exists L > 0$ such that: $\|\nabla f(w;\xi) - \nabla f(w';\xi)\| \le L\|w - w'\|$, $\forall w, w' \in \mathbb{R}^d$
- $F : \mathbb{R}^d \to \mathbb{R}$ is a μ-strongly convex, i.e., $\exists \mu > 0$ such that: $F(w) \ge F(w') + \langle \nabla F(w'), (w - w') \rangle + \frac{\mu}{2}\|w - w'\|^2$, $\forall w, w' \in \mathbb{R}^d$
- $f(w;\xi)$ is convex for every realization of ξ
- We can compute unbiased gradient $\mathbb{E}[\nabla f(w_t;\xi_t)] = \nabla F(w_t)$

Finite-sum Problem:

$$\min_{w \in \mathbb{R}^d} \left\{ F(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w) \right\}$$

Existing Complexity Results for Finite-sum

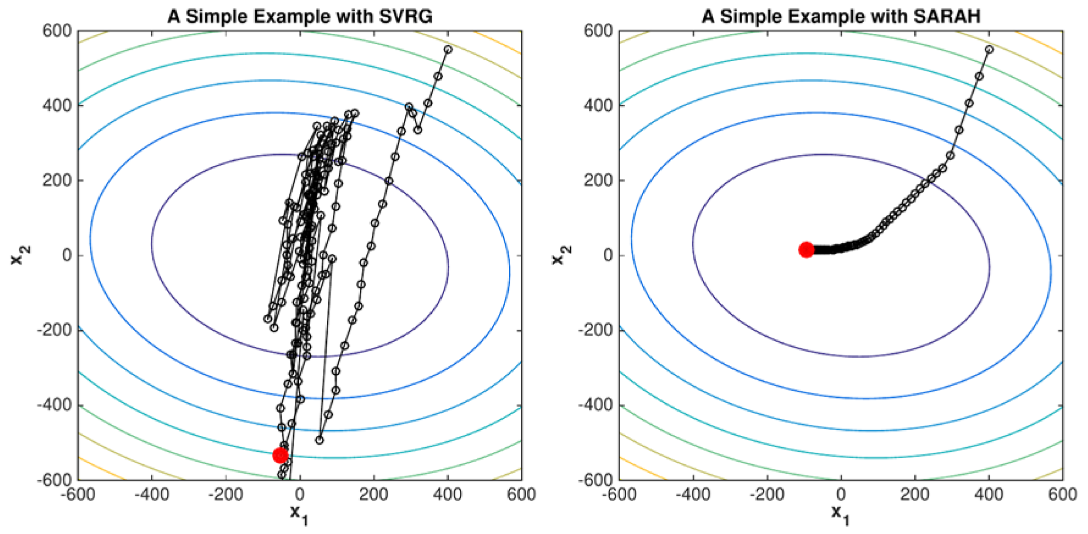
Complexity (Strongly convex) for finite-sum ($\kappa = L/\mu$)

Method	Complexity	Fixed LR	Low Storage
GD	$\mathcal{O}(n\kappa \log(1/\epsilon))$	✓	✓
SGD [6, 1]	$\mathcal{O}(\kappa/\epsilon)$	✗	✓
SVRG [3]	$\mathcal{O}((n+\kappa)\log(1/\epsilon))$	✓	✓
SAG/SAGA [8, 2]	$\mathcal{O}((n+\kappa)\log(1/\epsilon))$	✓	✗
SARAH [7]	$\mathcal{O}((n+\kappa)\log(1/\epsilon))$	✓	✓

SARAH vs. SVRG

- Both methods require **restarting**. Computing a full gradient for every outer loop $v_0 = \nabla F(w_0)$
- The difference is the **stochastic gradient update**
  - **SVRG**: $v_t = \nabla f_{i_t}(w_t) - \nabla f_{i_t}(w_0) + v_0$
  - **SARAH**: $v_t = \nabla f_{i_t}(w_t) - \nabla f_{i_t}(w_{t-1}) + v_{t-1}$

One outer loop behavior of SVRG and SARAH:



Inexact SARAH Algorithm (iSARAH)

Inexact SARAH (iSARAH):

Parameters: the learning rate $\eta > 0$ and the inner loop size m, the sample set size b.
Initialize: $\tilde{w}_0$.
Iterate:
for $s = 1, 2, \dots, \mathcal{T}$, **do**
$\tilde{w}_s = \text{iSARAH-IN}(\tilde{w}_{s-1}, \eta, m, b)$.
end for
Output: $\tilde{w}_{\mathcal{T}}$.

---

iSARAH-IN(w_0, η, m, b):

Input: $w_0 (= \tilde{w}_{s-1})$ the learning rate $\eta > 0$, the inner loop size m, the sample set size b.
Generate random variables $\{\zeta_i\}_{i=1}^{b}$ i.i.d.
Compute $v_0 = \frac{1}{b}\sum_{i=1}^{b} \nabla f(w_0;\zeta_i)$.
$w_1 = w_0 - \eta v_0$.
Iterate:
for $t = 1, \dots, m-1$, **do**
Generate a random variable ξ_t
$v_t = \nabla f(w_t;\xi_t) - \nabla f(w_{t-1};\xi_t) + v_{t-1}$.
$w_{t+1} = w_t - \eta v_t$.
end for
Set $\tilde{w} = w_t$ with t chosen uniformly at random from $\{0, 1, \dots, m\}$
Output: $\tilde{w}$

Theoretical Results (Strongly Convex)

Theorem: Suppose that F is μ-strongly convex and $f(w;\xi)$ is L-smooth and convex for every realization of ξ. Consider iSARAH with the choice of η, m, and b such that

$$\alpha = \frac{1}{\mu\eta(m+1)} + \frac{\eta L}{2 - \eta L} + \frac{4\kappa - 2}{b(2 - \eta L)} < 1.$$

(Note that $\kappa = L/\mu$.) Then, we have

$$\mathbb{E}[\|\nabla F(\tilde{w}_s)\|^2] - \Delta \le \alpha^s(\|\nabla F(\tilde{w}_0)\|^2 - \Delta), \qquad (1)$$

where

$$\Delta = \frac{\delta}{1 - \alpha} \text{ and } \delta = \frac{4}{b(2 - \eta L)}\mathbb{E}\left[\|\nabla f(w_*;\xi)\|^2\right].$$

Corollary: Let $\eta = \mathcal{O}\left(\frac{1}{L}\right)$, $m = \mathcal{O}(\kappa)$, $b = \mathcal{O}\left(\max\left\{\frac{1}{\epsilon}, \kappa\right\}\right)$ and $s = \mathcal{O}\left(\log\left(\frac{1}{\epsilon}\right)\right)$ in Theorem above. Then, the total work complexity to achieve $\mathbb{E}[\|\nabla F(\tilde{w}_s)\|^2] \le \epsilon$ is $\mathcal{O}\left(\left(\max\left\{\frac{1}{\epsilon}, \kappa\right\} + \kappa\right)\log\left(\frac{1}{\epsilon}\right)\right)$.

Complexity Comparisons

Strongly convex: ($\kappa = L/\mu$)

Method	Bound	Problem type
SARAH	$\mathcal{O}\left((n+\kappa)\log\left(\frac{1}{\epsilon}\right)\right)$	Finite-sum
SVRG	$\mathcal{O}\left((n+\kappa)\log\left(\frac{1}{\epsilon}\right)\right)$	Finite-sum
SCSG	$\mathcal{O}\left(\left(\min\left\{\frac{\kappa}{\epsilon}, n\right\} + \kappa\right)\log\left(\frac{1}{\epsilon}\right)\right)$	Finite-sum
SCSG	$\mathcal{O}\left(\left(\frac{\kappa}{\epsilon} + \kappa\right)\log\left(\frac{1}{\epsilon}\right)\right)$	Expectation
SGD	$\mathcal{O}\left(\frac{\kappa}{\epsilon}\right)$	Expectation
iSARAH	$\mathcal{O}\left(\left(\max\left\{\frac{1}{\epsilon}, \kappa\right\} + \kappa\right)\log\left(\frac{1}{\epsilon}\right)\right)$	Expectation

General convex:

Method	Bound	Problem type
SCSG	$\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$	Expectation
SGD	$\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$	Expectation
iSARAH (one loop)	$\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$	Expectation
iSARAH (multiple loop)	$\mathcal{O}\left(\frac{1}{\epsilon}\log\left(\frac{1}{\epsilon}\right)\right)$	Expectation

Nonconvex:

Method	Bound	Problem type
SCSG	$\mathcal{O}\left(\frac{1}{\epsilon^{5/3}}\right)$	Expectation
SGD	$\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$	Expectation
iSARAH (one loop)	$\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$	Expectation

References

[1] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization Methods for Large-scale Machine Learning. SIAM Review, 2018
[2] A. Defazio, F. Bach, S. Lacoste-Julien. SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives. NIPS 2014
[3] R. Johnson and T. Zhang. Accelerating Stochastic Gradient Descent using Predictive Variance Reduction. NIPS 2013.
[4] L. Lei and M. Jordan. Less than a Single Pass: Stochastically Controlled Stochastic Gradient. AISTATS 2017
[5] L. Lei, C. Ju, J. Chen, and M. I. Jordan. Non-convex finite-sum optimization via scsg methods. NIPS 2017
[6] H. Robbins and S. Monro. A Stochastic Approximation Method. 1951
[7] L. Nguyen, J. Liu, K. Scheinberg, and M. Takac. SARAH: A Novel Method for Machine Learning Problems Using Stochastic Recursive Gradient. ICML 2017
[8] M. Schmidt, N. Le Roux, and F. Bach. Minimizing Finite Sums with the Stochastic Average Gradient. Mathematical Programming 2017