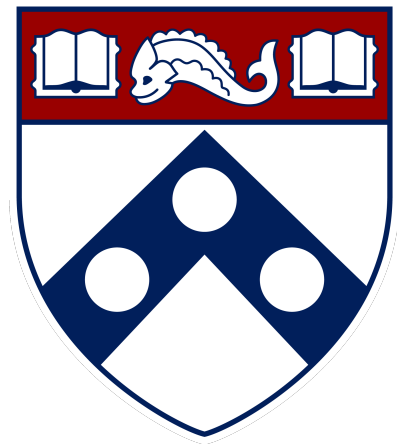
Variance reduction for stochastic gradient methods



Yuxin Chen

Wharton Statistics & Data Science, Fall 2023

Outline

- Stochastic variance reduced gradient (SVRG)
  - Convergence analysis for strongly convex problems
- Stochastic recursive gradient algorithm (SARAH)
  - Convergence analysis for nonconvex problems
- Other variance reduced stochastic methods
  - Stochastic dual coordinate ascent (SDCA)

Finite-sum optimization

Stochastic gradient descent (SGD)

Algorithm 12.1 Stochastic gradient descent (SGD)

1: **for** $t = 1, 2, \ldots$ **do**
2: pick $i_t \sim \mathsf{Unif}(1, \ldots, n)$
3: $\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \nabla f_{i_t}(\boldsymbol{x}^t)$

Recall: SGD theory with fixed stepsizes

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \, \boldsymbol{g}^t$$

Recall: SGD theory with fixed stepsizes

$$\mathbb{E}[F(\boldsymbol{x}^t) - F(\boldsymbol{x}^*)] \leq \frac{\eta L \sigma_{\mathrm{g}}^2}{2\mu} + (1 - \eta\mu)^t \big(F(\boldsymbol{x}^0) - F(\boldsymbol{x}^*)\big)$$

A simple idea

Reducing variance via gradient aggregation

Stochastic variance reduced gradient (SVRG)

Strongly convex and smooth problems (no regularization)

Stochastic variance reduced gradient (SVRG)

— Johnson, Zhang '13

Stochastic variance reduced gradient (SVRG)

SVRG algorithm (Johnson, Zhang '13)

Algorithm 12.2 SVRG for finite-sum optimization

Remark

Convergence analysis of SVRG

Theorem 12.1

Convergence analysis of SVRG

Theorem 12.1

Proof of Theorem 12.1

Proof of Theorem 12.1

Proof of Theorem 12.1

Proof of Theorem 12.1

Proof of Theorem 12.1 (cont.)

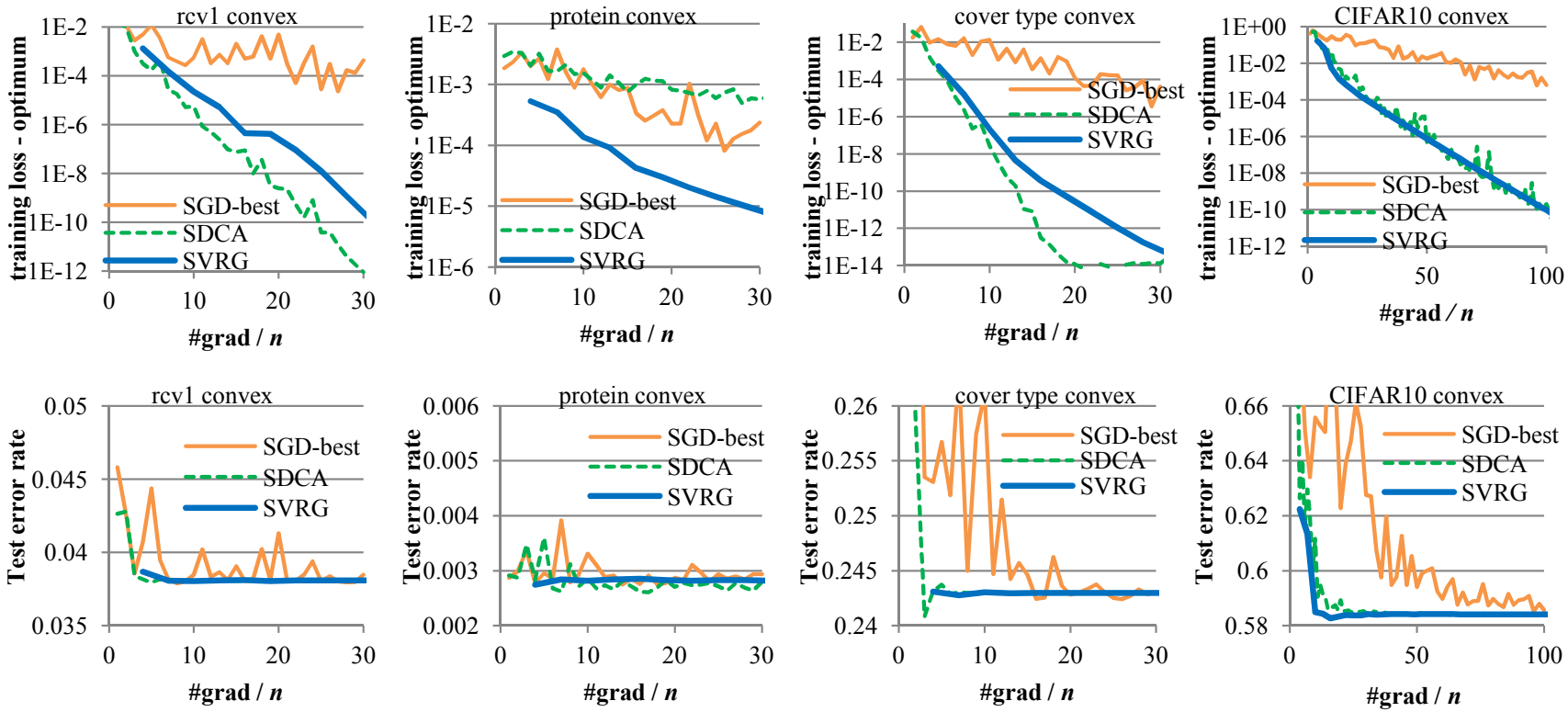
Proof of Theorem 12.1 (cont.)

Proof of Lemma 12.2

Proof of Lemma 12.2 (cont.)

Numerical example: logistic regression

— *Johnson, Zhang '13*



ℓ_2-regularized logistic regression on CIFAR-10

Comparisons with GD and SGD

	SVRG	GD	SGD
comp. cost	$(n+\kappa)\log\frac{1}{\varepsilon}$	$n\kappa\log\frac{1}{\varepsilon}$	$\frac{\kappa^2}{\varepsilon}$ (practically often $\frac{\kappa}{\varepsilon}$)

Proximal extension

Proximal extension (Xiao, Zhang '14)

Algorithm 12.3 Prox-SVRG for finite-sum optimization

Stochastic recursive gradient algorithm (SARAH)

Nonconvex and smooth problems

Recursive stochastic gradient estimates

— *Nguyen, Liu, Scheinberg, Takac '17*

Restarting gradient estimate every epoch

Bias of gradient estimates

StochAstic Recursive grAdient algoritHm

Algorithm 12.4 SARAH (Nguyen et al. '17)

Convergence analysis of SARAH (nonconvex)

Convergence analysis of SARAH (nonconvex)

Proof of Theorem 12.3

Proof of Theorem 12.3 (cont.)

Proof of Theorem 12.3 (cont.)

Proof of Lemma 12.4

Proof of Lemma 12.4 (cont.)

Proof of Lemma 12.5

Stochastic dual coordinate ascent (SDCA)

— a dual perspective

A class of finite-sum optimization

Dual formulation

Derivation of the dual formulation

Randomized coordinate ascent on dual problem

— Shalev-Shwartz, Zhang '13

Stochastic dual coordinate ascent (SDCA)

Algorithm 12.5 SDCA for finite-sum optimization

A variant of SDCA without duality

A variant of SDCA without duality

A variant of SDCA without duality

SDCA as SGD

SDCA as variance-reducedSGD

Convergence guarantees of SDCA

Theorem 12.6 (informal, Shalev-Shwartz '16)

Reference

1. "*Recent advances in stochastic convex and non-convex optimization*," Z. Allen-Zhu, *ICML Tutorial*, 2017.
2. "*Accelerating stochastic gradient descent using predictive variance reduction*," R. Johnson, T. Zhang, *NIPS*, 2013.
3. "*Barzilai-Borwein step size for stochastic gradient descent*," C. Tan, S. Ma, Y.H. Dai, Y. Qian, *NIPS*, 2016.
4. "*A proximal stochastic gradient method with progressive variance reduction*," L. Xiao, T. Zhang, *SIAM Journal on Optimization*, 2014.
5. "*Minimizing finite sums with the stochastic average gradient*," M. Schmidt, N. Le Roux, F. Bach, *Mathematical Programming*, 2013.
6. "*SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives*," A. Defazio, F. Bach, and S. Lacoste-Julien, *NIPS*, 2014.

Reference

[7] "*Variance reduction for faster non-convex optimization*," Z. Allen-Zhu, E. Hazan, *ICML*, 2016.

[8] "*Katyusha: The first direct acceleration of stochastic gradient methods*," Z. Allen-Zhu, *STOC*, 2017.

[9] "*SARAH: A novel method for machine learning problems using stochastic recursive gradient*," L. Nguyen, J. Liu, K. Scheinberg, M. Takac, *ICML*, 2017.

[10] "*Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator*," C. Fang, C. Li, Z. Lin, T. Zhang, *NIPS*, 2018.

[11] "*SpiderBoost and momentum: Faster variance reduction algorithms*," Z. Wang, K. Ji, Y. Zhou, Y. Liang, V. Tarokh, *NIPS*, 2019.

Reference

[12] "*Optimal finite-Sum smooth non-convex optimization with SARAH*," L. Nguyen, M. vanDijk, D. Phan, P. Nguyen, T. Weng, J. Kalagnanam, arXiv:1901.07648, 2019.

[13] "*Stochastic dual coordinate ascent methods for regularized loss minimization*," S. Shalev-Shwartz, T. Zhang, *Journal of Machine Learning Research*, 2013.

[14] "*SDCA without duality, regularization, and individual convexity*," S. Shalev-Shwartz, ICML, 2016.

[15] "*Optimization methods for large-scale machine learning*," L. Bottou, F. Curtis, J. Nocedal, 2016.