Variance Reduction on Adaptive Stochastic Mirror Descent

Wenjie Li[1], Zhanyu Wang[1], Yichen Zhang[2], and Guang Cheng[1]

[1]Department of Statistics, Purdue University

[2]Krannert School of Management, Purdue University

Introduction

Background

- Variance reduction can improve the convergence of SGD-like algorithms in non-convex optimization problems
- Mirror Descent algorithms are useful in non-smooth optimization problems, especially general adaptive mirror descent algorithms.

Contributions

- In this paper, we prove that variance reduction can reduce the gradient complexity of the general adaptive SMD algorithms, which makes them converge faster. So it means any existing mirror descent algorithm can work well with variance reduction.

Algorithm

- We study the following general variance reduced adaptive stochastic mirror descent algorithm, where in line 7, a large batch gradient is used to reduce the variance of a small batch gradient.

Algorithm 1 General Adaptive Stochastic Mirror Descent with Variance Reduction Algorithm

1: **Input:** Number of stages T, initial x_1, step sizes $\{\alpha_t\}_{t=1}^T$, batch, mini-batch sizes $\{B_t, b_t\}_{t=1}^T$
2: **for** $t = 1$ **to** T **do**
3: Randomly sample a batch $\mathcal{I}_t$ with size B_t
4: $g_t = \nabla f_{\mathcal{I}_t}(x_t)$; $y_1^t = x_t$
5: **for** $k = 1$ **to** K **do**
6: Randomly pick sample $\tilde{\mathcal{I}}_t$ of size b_t
7: $v_k^t = \nabla f_{\tilde{\mathcal{I}}_t}(y_k^t) - \nabla f_{\tilde{\mathcal{I}}_t}(y_1^t) + g_t$
8: $y_{k+1}^t = \text{argmin}_y\{\alpha_t\langle v_k^t, y\rangle + \alpha_t h(x) + B_{\psi_{tk}}(y, y_k^t)\}$
9: **end for**
10: $x_{t+1} = y_{K+1}^t$
11: **end for**
12: **Return** (Smooth case) Uniformly sample t^* from $\{t\}_{t=1}^T$ and output x_{t^*}; (P-L case) $x_{t^*} = x_{T+1}$

- We assume the proximal functions $\psi_t(x)$ are all m-strongly convex with respect to $\|\cdot\|_2$, i.e.,

$$\psi_t(y) \geq \psi_t(x) + \langle \nabla\psi_t(x), y - x\rangle + \frac{m}{2}\|y - x\|_2^2, \forall t > 0$$

- The standard Lipschitzness, unbiasedness, and bounded variance assumptions on the gradients are also assumed.

Results

Theorem 1: Convergence of General Adaptive SMD with VR

Suppose that $\psi_{tk}(x)$ satisfy the m-strong convexity assumption and f satisfies the Lipschitz gradients and bounded variance assumptions. Further assume that the learning rate, the batch sizes, the mini-batch sizes, the number of outer and inner loop iterations are set to be $\alpha_t = m/L, B_t = n \wedge (20\sigma^2/m^2\epsilon^2), b_t = b, T = 1 \vee 16\Delta_F L/(m^2\epsilon^2 K), K = \left\lfloor\sqrt{b/20}\right\rfloor \vee 1$, where Δ_F is a constant. Then the output of Algorithm 1 converges with gradient computations

$$O(\frac{n}{\epsilon^2\sqrt{b}} \wedge \frac{\sigma^2}{\epsilon^4\sqrt{b}} + \frac{b}{\epsilon^2})$$

- We list the SFO complexity of a few relevant algorithms. "VR" stands for variance reduction. As can be observed in the table, when a correct mini-batch size b is chosen, variance redcution helps the convergence of any Stochastic Mirror Descent algorithm.

Algorithms	SFO computations
SVRG [5]	$O(n^{2/3}/\epsilon^2)$
SCSG [2]	$O(n/\epsilon^2 \wedge 1/\epsilon^{10/3})$
ProxGD [1]	$O(n/\epsilon^2)$
ProxSVRG/SAGA [4]	$O(n/(\epsilon^2\sqrt{b}) + n)$
ProxSVRG+ [3]	$O(n/(\epsilon^2\sqrt{b}) \wedge (1/(\epsilon^4\sqrt{b})) + b/\epsilon^2)$
Adaptive SMD	$O(n/\epsilon^2 \wedge 1/\epsilon^4)$
Adaptive SMD + VR	$O(n/(\epsilon^2\sqrt{b}) \wedge 1/(\epsilon^4\sqrt{b}) + b/\epsilon^2)$

Corollary 1: Convergence of General Adaptive SMD with VR

With all the assumptions and parameter settings in Theorem 1, further assume that $b = \epsilon^{-4/3}$, where $\epsilon^{-4/3} \leq n$. Then the output of algorithm 1 converges with gradient computations

$$O(\frac{n}{\epsilon^{4/3}} \wedge \frac{1}{\epsilon^{10/3}} + \frac{1}{\epsilon^{10/3}}) \quad (1)$$

- A similar argument can be made in the PL-condition case where a slightly different choice of b is chosen. Variance reduction reduces the gradient complexity of the general adaptive stochastic mirror descent algorithm in both cases.

Experiments

- We choose AdaGrad and RMSProp as two special examples of our general algorithm to examine the effectiveness of variance reduction.

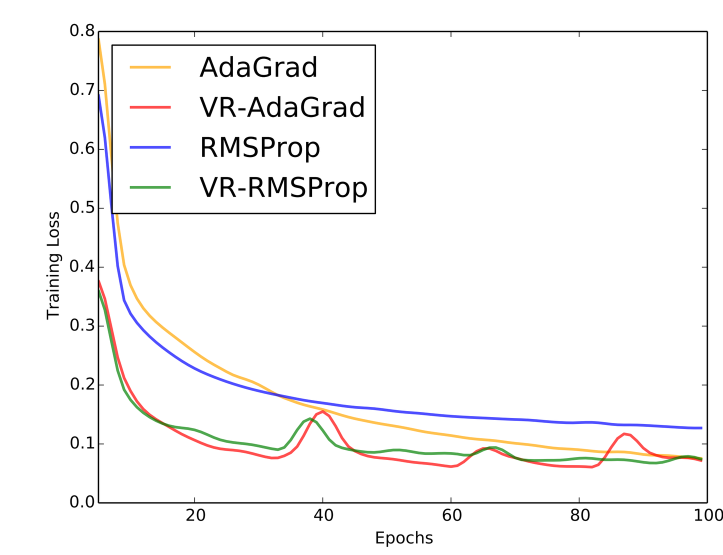

Figure 1: Fully Connected/MNIST/Train loss

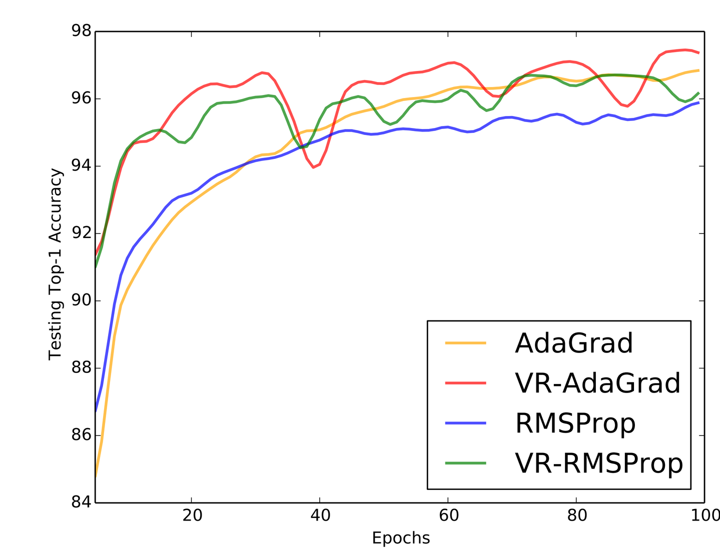

Figure 2: Fully Connected/MNIST/Test Acc

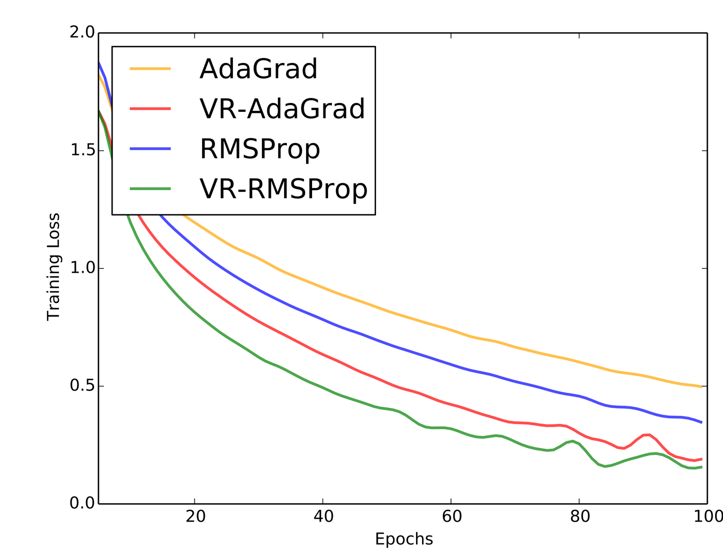

Figure 3: LeNet/CIFAR-10/Train loss

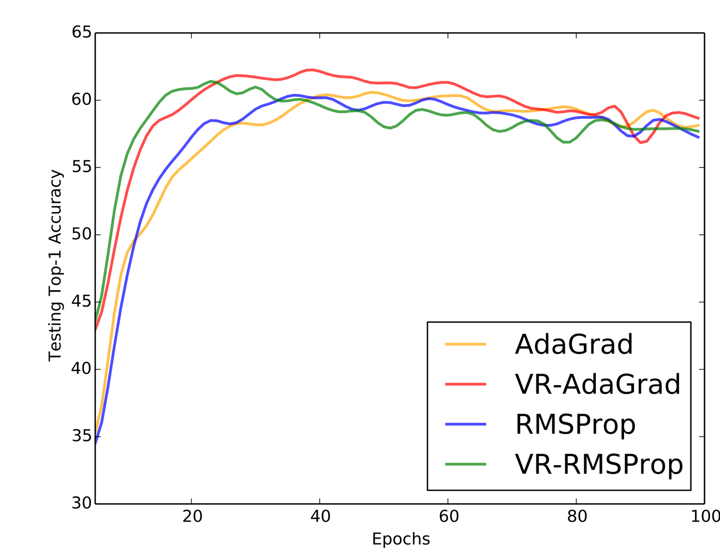

Figure 4: LeNet/CIFAR-10/Test Acc

- The upper row shows the training loss and the testing accuracy of the original algorithms and the variance reduced ones on the MNIST dataset.
- The lower row shows the training loss and the testing accuracy of the original algorithms and the variance reduced ones on the CIFAR-10 dataset.
- In both cases variance reduction is effective in boosting the convergence.

References

[1] Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. "Mini-batch Stochastic Approximation Methods for Nonconvex Stochastic Composite Optimization". In: *arXiv preprint arXiv:1308.6594* (2016).

[2] Lihua Lei et al. "Non-convex Finite-Sum Optimization Via SCSG Methods". In: *Advances in Neural Information Processing Systems 30* (2017), pp. 2348–2358.

[3] Zhize Li and Jian Li. "A Simple Proximal Stochastic Gradient Method for Nonsmooth Nonconvex Optimization". In: *Advances in Neural Information Processing Systems 31* (2018), pp. 5564–5574.

[4] Sashank Reddi et al. "Proximal Stochastic Methods for Nonsmooth Nonconvex Finite-Sum Optimization". In: *Advances in Neural Information Processing Systems 29* (2016b).

[5] Sashank J. Reddi et al. *Stochastic Variance Reduction for Nonconvex Optimization.* 2016a. arXiv: 1603.06160 [math.OC].