Stochastic Reweighted Gradient Descent Presentation

Hsiu-I Liao[1]*, Jimmy Wu[1]*

[1] Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan

* Equal Contribution

Introduction

- We propose SRG, a stochastic gradient method based solely on importance sampling that can reduce the variance of the gradient estimator and improve on the asymptotic error of SGD in the strongly convex and smooth case.

SGD vs. SRG

- Stochastic Gradient (SGD)

$$x_{k+1} = x_k - \alpha_k \nabla f_{i_k}(x_k)$$

- Stochastic Reweighted Gradient Descent (SRG)

$$x_{k+1} = x_k - \alpha_k \frac{1}{n p_k^{i_k}} \nabla f_{i_k}(x_k)$$

Importance Sampling Property

$$\mathbb{E}_p[f(x)] = \mathbb{E}_q\left[f(x)\frac{p(x)}{q(x)}\right]$$

for any distributions $p(x)$ and $q(x)$ holds. We take $p(x) = \frac{1}{n}$ and $q(x) = p_k^{i_k}$ in SRG.

The sampling q(x) has the same expectation but different variance compared to p(x). We aim to find a lower variance distribution in SRG

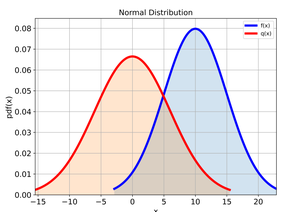


Figure: implying distribution of q(x) based on p(x)

Choose Proper Distributions

- SRG Distribution

$$p_k = \underbrace{(1-\theta_k)q_k}_{\text{greedy strategy}} + \underbrace{\frac{\theta_k}{n}}_{\text{uniformly sampling}}$$

- SRG+ Distribution

$$p'_k = \underbrace{(1-\eta_k-\theta_k)q_k}_{\text{greedy strategy}} + \underbrace{\eta_k v}_{\text{relative smoothness}} + \underbrace{\frac{\theta_k}{n}}_{\text{uniformly sampling}}$$

Algorithm of SRG

In each iteration of SRG, we first mix distribution at step 4, doing SRG update with importance sampling at step 7, and finally updating the gradient norm table at step 8

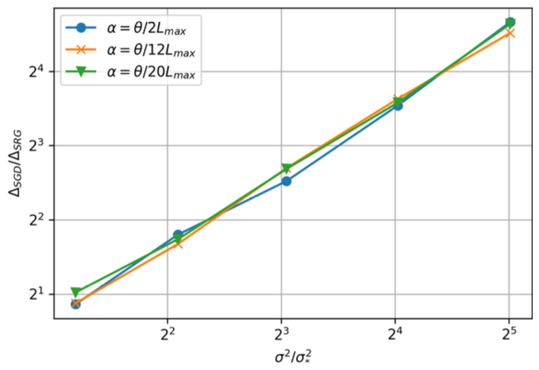
Algorithm 1 SRG

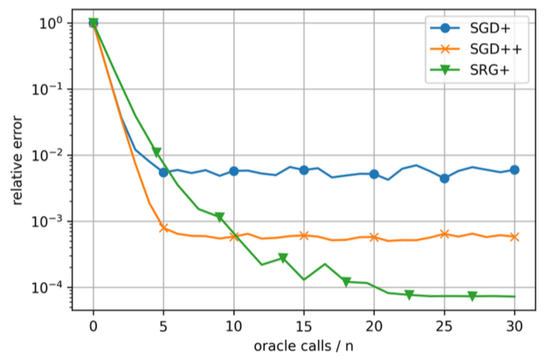
1: **Parameters:** step sizes $(\alpha_k)_{k=0}^{\infty} > 0$, mixture coefficients $(\theta_k)_{k=0}^{\infty} \in (0,1]$
2: **Initialization:** $x_0 \in \mathbb{R}^d, (\|g_0^i\|_2)_{i=1}^n \in \mathbb{R}^n$
3: **for** $k = 0, 1, 2, \ldots$ **do**
4: $\quad p_k = (1-\theta_k)q_k + \theta_k/n \qquad \{q_k \text{ is defined in (6)}\}$
5: $\quad b_k \sim \text{Bernoulli}(\theta_k)$
6: $\quad$ **if** $b_k = 1$ **then** $i_k \sim 1/n$ **else** $i_k \sim q_k$
7: $\quad x_{k+1} = x_k - \alpha_k \frac{1}{n p_k^{i_k}} \nabla f_{i_k}(x_k)$
8: $\quad \|g_{k+1}^i\|_2 = \begin{cases} \|\nabla f_i(x_k)\|_2 & \text{if } b_k = 1 \text{ and } i_k = i \\ \|g_k^i\|_2 & \text{otherwise} \end{cases}$
9: **end for**

Figure: SRG algorithm

Experimental Results - Finite-Sum Problem

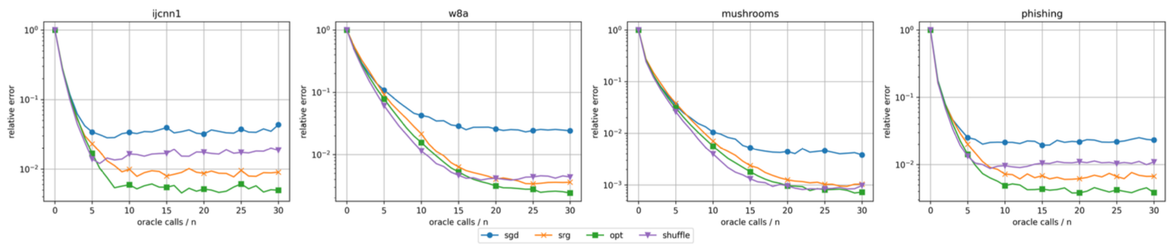
The following left figure show relationship between the two ratios is linear, and very close to identity. The following right figure show all three algorithms are able to converge even when using the large $O(1/\bar{L})$ step sizes.





Experimental Results - Comparison with Benchmarks

The following figures show the performance of SRG and other benchmark methods algorithms in LIBSVM. We can observe that SRG has the best performance in *ijcnn1, w8a, mushrooms, phishing*. We also see that SRG is more robust across tasks than SGD.



Variance Reduction

To do variance reduction, our algorithm has to minimize σ term

$$\sigma^2(x_k, p) \le \frac{2}{n^2}\sum_{i=1}^n \frac{\|\nabla f_i(x_k) - g_k^i\|_2^2}{p^i} + 2\tilde{\sigma}^2(x_k, p)$$

(1) For right-hand side

$$\tilde{\sigma}^2(x_k, p) := \frac{1}{n^2}\sum_{i=1}^n \frac{1}{p^i}\left\|g_k^i\right\|_2^2$$

this approximation is minimized at (Zhao Zhang, 2015):

$$q_k = \arg\min_p \tilde{\sigma}^2(x_k, p) = \left(\frac{\|g_k^i\|_2}{\sum_{j=1}^n \left\|g_k^j\right\|_2}\right)_{i=1}^n$$

(2) For the left-hand side, we minimize the discrepancy $||\nabla f_i(x_k) - g_k^i||^2$ by updating g_k^i frequently

Complexity of SGD and SGD Under Strongly-convex and Smooth

Corollary 4.4 : The convergence rate of SRG

$$\mathbb{E}\left[T^k\right] \le (1-\rho)^k T^0 + (1+2\theta)\frac{6\alpha^2\sigma_*^2}{\rho}$$

Algorithm	Complexity
SGD	$O\left(\kappa_{\max} + \frac{\sigma^2}{\mu^2\varepsilon}\right)\log\left(\frac{1}{\varepsilon}\right)$
SRG	$O\left(n + \sqrt{\frac{n\sigma_*^2}{\mu^2\varepsilon}} + \kappa_{\max} + \frac{\sigma_*^2}{\mu^2\varepsilon}\right)\log\left(\frac{1}{\varepsilon}\right)$
SRG+	$O\left(n + \sqrt{\frac{n\sigma_*^2}{\mu\varepsilon}} + \bar{\kappa} + \frac{\sigma_*^2}{\mu^2\varepsilon}\right)\log\left(\frac{1}{\varepsilon}\right)$

Conclusion

- We introduced SRG, a new importance-sampling based stochastic optimization algorithm for finite-sum problems that reduces the variance of the gradient estimator.
- We analyzed its convergence rate in the strongly convex and smooth case, and showed that it can improve on the asymp- totic error of SGD.