Stochastic Reweighted Gradient Descent

Ayoub El Hanchi [1]   David A. Stephens [2]   Chris J. Maddison [1]

July 16, 2022

[1]University of Toronto and Vector Institute

[2]McGill University

Finite-sum Optimization

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

Stochastic Gradient Descent

Pros: Low per-iteration cost leading to fast initial convergence.

Cons: High variance of gradient estimator leading to high asymptotic error.

Variance Reduction

Pros: Vanishing variance of gradient estimator leading to zero asymptotic error.

Cons: High(er) per-iteration cost leading to slow initial convergence.

Question

Can we get the best of both worlds ?

- Fast initial convergence. (through low per-iteration cost)
- Small asymptotic error (through some form of variance reduction).

Stochastic Reweighted Gradient Descent

Main idea: use importance sampling instead of control variates to greedily reduce the variance of the gradient estimator.

$$p_k = (1-\theta_k)q_k + \frac{\theta_k}{n}$$

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

$$\|g_{k+1}^i\|_2 = \begin{cases} \|\nabla f_i(x_k)\|_2 & \text{if } i = i_k \\ \|g_k^i\|_2 & \text{otherwise} \end{cases}$$

where $q_k \propto \|g_k^i\|$ and $(\theta_k)_{k=0}^{\infty} \in (0,1)$

Open Problem

Pros: Almost no overhead compared to SGD, preserving fast initial convergence.

Cons: Non-zero asymptotic variance leading non-zero asymptotic error.

Variants of this idea already explored in many (15+) previous works. What's missing ? A clean and direct convergence rate analysis under standard assumptions.

Main result

Slight modification:

$$b_k \sim \text{Bernoulli}(\theta_k)$$

$$\|g_{k+1}^i\|_2 = \begin{cases} \|\nabla f_i(x_k)\|_2 & \text{if } i = i_k \text{ and } b_k = 1 \\ \|g_k^i\|_2 & \text{otherwise} \end{cases}$$

Main result

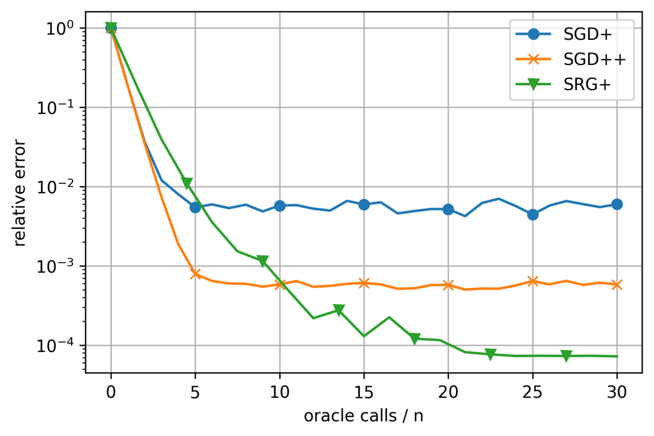
Under strong convexity of the function, and smoothness and convexity of the component functions, SRG has similar non-asymptotic convergence rate as SGD, but asymptotic error is better:

$$\lim_{k\to\infty} \mathbb{E}\left[\|x_k^{SGD} - x^*\|_2^2\right] = O(\sigma^2)$$
$$\lim_{k\to\infty} \mathbb{E}\left[\|x_k^{SRG} - x^*\|_2^2\right] = O(\sigma_*^2)$$

where:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x^*)\|_2^2$$
$$\sigma_*^2 = \frac{1}{n^2}\left(\sum_{i=1}^{n}\|\nabla f_i(x^*)\|_2\right)^2$$

Illustration



Conclusion

We propose and analyze an SGD-like algorithm that enjoys both:

- Negligible per-iteration overhead over SGD leading to fast initial convergence.
- Smaller asymptotic error through importance-sampling-based variance reduction.

Possible future direction: Efficient implementation in deep learning frameworks ? What about other forms of variance reduction ?