Estimate Sequences for Variance-Reduced Stochastic Composite Optimization

Andrei Kulunchakov Julien Mairal

andrei.kulunchakov@inria.fr julien.mairal.@inria.fr

International Conference on Machine Learning, 2019

Poster event-4062, (Jun 12th, Pacific Ballroom 204)

Problem statement

Assumptions

We solve a stochastic composite optimization problem

$$F(x) = f(x) + \psi(x) \quad \text{where} \quad f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) \quad \text{with} \quad f_i(x) = \mathbb{E}_{\xi}\left[\tilde{f}_i(x,\xi)\right],$$

where $\psi(x)$ is a convex penalty, each f_i is L-smooth and μ-strongly convex.

Variance in gradient estimates

Stochastic realizations of gradients are available for each i

$$\tilde{\nabla} f_i(x) = \nabla f_i(x) + \xi_i \quad \text{with } \mathbb{E}[\xi_i] = 0 \quad \text{and } \mathrm{Var}\left[\xi_i\right] \leq \sigma^2.$$

Main contribution (I)

Optimal incremental algorithm robust to noise

Optimal incremental algorithm with a complexity

$$O\left(\left(n+\sqrt{\frac{nL}{\mu}}\right)\log\left(\frac{F(x_0)-F^\star}{\varepsilon}\right)\right)+O\left(\frac{\sigma^2}{\mu\varepsilon}\right),$$

based on the SVRG gradient estimator with random sampling.

Algorithm

Briefly, the algorithm is an incremental hybrid of the heavy-ball method with randomly updated SVRG anchor point and two auxiliary sequences, controlling the extrapolation.

Main contribution (II)

Novelty

- When $\sigma^2 = 0$, we recover the same complexity as Katyusha [Allen-Zhu, 2017].
- Novelty: accelerated incremental algorithm robust to $\sigma^2 > 0$ with the optimal term $\sigma^2/\mu\varepsilon$.

Another contributions

- Generic proofs for incremental methods (SVRG, SAGA, MISO, SDCA) to show their robustness to noise

$$O\left(\left(n + \frac{L}{\mu}\right) \log\left(\frac{F(x_0) - F^\star}{\varepsilon}\right)\right) + O\left(\frac{\sigma^2}{\mu\varepsilon}\right).$$

- When $\mu = 0$, we recover optimal rates in fixed horizon and known σ^2.
- Provide a support for non-uniform sampling.

Side contributions

Adaptivity to strong convexity parameter μ

When $\sigma = 0$, we show adaptivity to μ for all above-mentioned non-accelerated methods. This property is new for SVRG.

Accelerated SGD

A version of robust accelerated SGD with complexity similar to [Ghadimi and Lan, 2012, 2013]

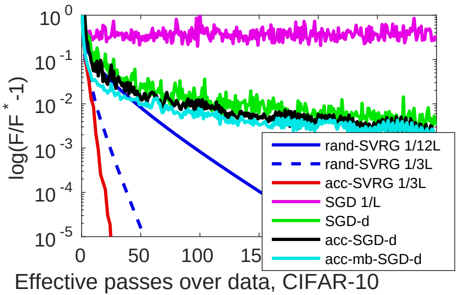
$$O\left(\sqrt{\frac{L}{\mu}} \log\left(\frac{F(x_0) - F^\star}{\varepsilon}\right)\right) + O\left(\frac{\sigma^2 + \sigma_n^2}{\mu\varepsilon}\right),$$

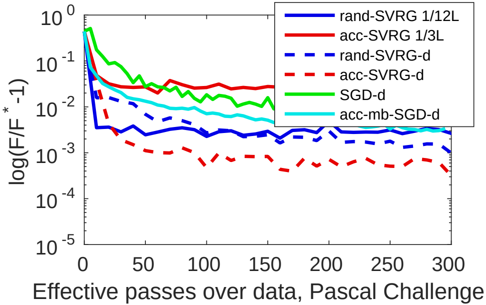
where σ_n^2 is due to sampling the data points.

Experiments with three datasets in the experiments

— Pascal Large Scale Learning Challenge ($n = 25 \cdot 10^4$)
— Light gene expression data for breast cancer ($n = 295$)
— CIFAR-10 (images represented by features from a network) with $n = 5 \cdot 10^4$

Examples with zero noise ($\sigma = 0$) and stochastic case ($\sigma > 0$)





Poster event-4062, (Jun 12th, Pacific Ballroom 204)