Estimate Sequences for Variance-Reduced Stochastic Composite Optimization

Andrei Kulunchakov and Julien Mairal - Inria Grenoble

A long overview

- **Generic complexity analysis** for proximal SVRG, SAGA, MISO.
- Robustness to noise of these algorithms, e.g., to solve

$$\min_{x \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^{n} f_i(x) + \psi(x) \right\} \quad \text{with} \quad f_i(x) = \mathbb{E}_{\rho_i}\left[\tilde{f}_i(x, \rho_i)\right],$$

which is a **stochastic finite-sum problem**,
- or when assuming one has access only to the stochastic oracle

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

- The f_i's are μ-strongly convex and L-smooth and ψ is convex.
- A simple strategy **with averaging** gives the iteration complexity

$$O\left(\left(n + \frac{L}{\mu}\right) \log\left(\frac{F(x_0) - F^\star}{\varepsilon}\right)\right) + O\left(\frac{\sigma^2}{\mu\varepsilon}\right), \qquad (1)$$

for the criterion $\mathbb{E}[F(x_k) - F^\star] \le \varepsilon$.
- We also obtain **new algorithms** with the same complexity.
- We also obtain an **accelerated proximal SGD** with complexity

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

for $n = 1$ (simple stochastic composite problem).
- We also obtain an **accelerated SVRG** algorithm, with complexity

$$\underbrace{O\left(\left(n + \sqrt{\frac{nL}{\mu}}\right) \log\left(\frac{F(x_0) - F^\star}{\varepsilon}\right)\right)}_{\text{optimal for finite sums}} + \underbrace{O\left(\frac{\sigma^2}{\mu\varepsilon}\right)}_{\text{optimal for the noise}}, \qquad (3)$$

- we also treat the **convex but not strongly convex case** ($\mu = 0$),
- ... and study **non-uniform sampling strategies** for different L_i.

Two generic schemes

A first classical scheme:

$$x_k \leftarrow \text{Prox}_{\eta_k\psi}[x_{k-1} - \eta_k g_k] \quad \text{with} \quad \mathbb{E}[g_k | \mathcal{F}_{k-1}] = \nabla f(x_{k-1}), \qquad (A)$$

and another less-classical one:

$$\bar{x}_k \leftarrow (1 - \mu\eta_k)\bar{x}_{k-1} + \mu\eta_k x_{k-1} - \eta_k g_k \quad \text{and} \quad x_k = \text{Prox}_{\frac{\psi}{\gamma_k}}[\bar{x}_k]. \qquad (B)$$

Both approaches can be interpreted with **estimate sequences**:

$$d_k(x) = (1 - \delta_k) d_{k-1}(x) + \delta_k l_k(x) \text{ (always minimized by } x_k\text{)},$$

with, for (A),

$$l_k(x) = f(x_{k-1}) + g_k^\top (x - x_{k-1}) + \frac{\mu}{2}\|x - x_{k-1}\|^2 + \psi(x_k) + \psi'(x_k)^\top (x - x_k).$$

or, for (B),

$$l_k(x) = f(x_{k-1}) + g_k^\top (x - x_{k-1}) + \frac{\mu}{2}\|x - x_{k-1}\|^2 + \psi(x).$$

Gradient estimators and algorithms

- **exact gradient**, with $g_k = \nabla f(x_{k-1})$ (when $\sigma = 0$).
- **SGD**, when we assume that g_k has bounded variance.
- **random-SVRG**: draw randomly one index i_k and

$$g_k = \tilde{\nabla} f_{i_k}(x_{k-1}) - \tilde{\nabla} f_{i_k}(\tilde{x}_{k-1}) + \tilde{\nabla} f(\tilde{x}_{k-1}), \qquad (4)$$

where $\tilde{x}_{k-1}$ is an anchor point updated with probability $1/n$ at iteration k (as in [4]) and $\tilde{\nabla}$ denotes noisy gradients.
- **SAGA/MISO/SDCA**:

$$g_k = \tilde{\nabla} f_{i_k}(x_{k-1}) - z_{k-1}^{i_k} + \bar{z}_{k-1}, \qquad (5)$$

with $\bar{z}_{k-1} = \frac{1}{n}\sum_{i=1}^{n} z_{k-1}^i$. Then, choose β in $[0, \mu]$ and update

$$z_k^{i_k} = \tilde{\nabla} f_{i_k}(x_{k-1}) - \beta x_k \quad \text{and} \quad z_k^i = z_{k-1}^i \quad \text{for all} \quad i \ne i_k.$$

- **Links with existing approaches when** $\sigma = 0$: (A)+(4) $\approx$ SVRG; (A) + (5) with $\beta = 0 \Rightarrow$ SAGA; (B) + (5) with $\beta = \mu \approx$ SDCA/MISO.
- **Other combinations are new algorithms**.

An accelerated proximal SGD

Consider the parameter sequence

$$\delta_k = \sqrt{\eta_k \gamma_k} \quad \text{and} \quad \gamma_k = (1 - \delta_k)\gamma_{k-1} + \delta_k \mu.$$

Then, perform the iteration

$$\begin{aligned} x_k &= \text{Prox}_{\eta_k\psi}[y_{k-1} - \eta_k g_k] \quad \text{with} \quad \mathbb{E}[g_k | \mathcal{F}_{k-1}] = \nabla f(y_{k-1}) \\ y_k &= x_k + \beta_k (x_k - x_{k-1}) \quad \text{with} \quad \beta_k = \frac{\delta_k (1 - \delta_k)\eta_{k+1}}{\eta_k \delta_{k+1} + \eta_{k+1}\delta_k^2}, \end{aligned} \qquad (C)$$

An accelerated random-SVRG algorithm

- After appropriate initializations for $v_0, \tilde{x}_0, \gamma_0$.
- Find (δ_k, γ_k) such that $\gamma_k = (1 - \delta_k)\gamma_{k-1} + \delta_k \mu$ and $\delta_k = \sqrt{\frac{5\eta_k\gamma_k}{3n}}$.
- Choose the extrapolation point

$$y_{k-1} = \theta_k v_{k-1} + (1 - \theta_k)\tilde{x}_{k-1} \quad \text{with} \quad \theta_k = \frac{3n\delta_k - 5\mu\eta_k}{3 - 5\mu\eta_k};$$

- Compute the noisy gradient estimator

$$g_k = \tilde{\nabla} f_{i_k}(y_{k-1}) - \tilde{\nabla} f_{i_k}(\tilde{x}_{k-1}) + \bar{z}_{k-1};$$

- Obtain the new iterate

$$x_k \leftarrow \text{Prox}_{\eta_k\psi}[y_{k-1} - \eta_k g_k];$$

- Find the minimizer v_k of the estimate sequence d_k:

$$v_k = \left(1 - \frac{\mu\delta_k}{\gamma_k}\right) v_{k-1} + \frac{\mu\delta_k}{\gamma_k} y_{k-1} + \frac{\delta_k}{\gamma_k \eta_k}(x_k - y_{k-1});$$

- Update the anchor point $\tilde{x}_k$ and gradient $\bar{z}_k$ with prob $1/n$.
- Output x_k (**no averaging needed**).

$\mu > 0$, constant step sizes

Consider the online averaging strategy $\hat{x}_k = (1 - \delta_k)\hat{x}_{k-1} + \delta_k x_k$. Then, with a step size of order $1/L$ (up to a constant factor),

Method	Complexity $O(.)$	Bias
SGD	$(L/\mu)\log(1/\varepsilon)$	$(\sigma^2 + \sigma_n^2)/L$
acc-SGD	$\sqrt{L/\mu}\log(1/\varepsilon)$	$(\sigma^2 + \sigma_n^2)/\sqrt{\mu L}$
SVRG/SAGA/MISO	$(n + L/\mu)\log(1/\varepsilon)$	σ^2/L
acc-SVRG	$(n + \sqrt{nL/\mu})\log(1/\varepsilon)$	$\sigma^2/(\sqrt{\mu n L} + \mu n)$

- Note that the step size for acc-SVRG is of order $\min(1/L, 1/\mu n)$, the rest are adaptive to μ.
- σ_n^2 is due to sampling the data points.
- The bias of acc-SGD is **potentially huge**.

$\mu > 0$, decreasing step sizes

The complexities (1), (2), and (3) are obtained by
- first running algorithms with constant step sizes, as above,
- restart using decreasing step sizes:
  - SVRG/SAGA/MISO: $\eta_k = \min\left(\frac{1}{12L}, \frac{1}{5\mu n}, \frac{2}{\mu(k+1)}\right)$;
  - acc-SVRG: $\eta_k = \min\left(\frac{1}{3L}, \frac{1}{15\mu n}, \frac{4}{\mu(k+1)^2}\right)$;

Experiments on logistic regression ($\sigma^2 = 0$ on top)

- (left) Pascal Large Scale Learning Challenge ($n = 250000$);
- (right) CIFAR-10 represented by using a two-layer unsupervised convolutional neural network ($n = 50000$).