Stochastic Optimization with Variance Reduction for Infinite Datasets with Finite-Sum Structure

Alberto Bietti Julien Mairal - Inria Grenoble

Context and Problem Setting

Typical optimization settings for machine learning:

- **Stochastic approximation** (SGD) for infinite datasets: $\min_x \mathbb{E}_{\zeta\sim\mathcal{D}}[f(x,\zeta)]$ ($\mathcal{D}$: data distribution).
- **Variance reduction** (SAG/SVRG/SDCA/...) for finite datasets: $\min_x \frac{1}{n}\sum_{i=1}^n f_i(x)$ (where $f_i(x) = f(x,\zeta_i)$ with $\zeta_i \sim \mathcal{D}$).

Useful **hybrid setting** in between: include **random perturbations** ($\rho \sim \Gamma$) of each example (e.g. for regularization, stable feature selection, privacy). We consider the objective:

$$\min_{x\in\mathbb{R}^p}\left\{ f(x) := \frac{1}{n}\sum_{i=1}^n f_i(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\rho\sim\Gamma}[\tilde{f}_i(x,\rho)]\right\}.$$

Main examples considered:

- **Image "data augmentation"**: add random transformations of each image in the training set (crop, scale, brightness, contrast, etc)
- **Dropout**: set coordinates of feature vectors to 0 independently with some probability δ.

Key observation: variance from perturbations only is small compared to variance across all examples. Informally,

$$\mathrm{Var}_\rho \nabla\tilde{f}_i(x,\rho) \ll \mathrm{Var}_{i,\rho}\nabla\tilde{f}_i(x,\rho).$$

Contribution: improve convergence of SGD by exploiting the finite-sum structure using **variance reduction**. We obtain a $O(1/t)$ convergence with much smaller constant term depending on **variance from perturbations only**.

Gradient Variance Decomposition

- Total gradient variance σ_{tot}^2:

$$\sigma_{\mathrm{tot}}^2 := \mathrm{Var}_{i,\rho}\nabla\tilde{f}_i(x^*,\rho) = \mathbb{E}_{i,\rho}[\|\nabla\tilde{f}_i(x^*,\rho)\|^2]$$

- Variance from perturbations σ_p^2:

$$\sigma_p^2 := \mathbb{E}_i \mathrm{Var}_\rho \nabla\tilde{f}_i(x^*,\rho) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_\rho\left[\|\nabla\tilde{f}_i(x^*,\rho) - \nabla f_i(x^*)\|^2\right].$$

- Variance decomposition:

$$\sigma_{\mathrm{tot}}^2 = \sigma_p^2 + \mathbb{E}_i[\|\nabla f_i(x^*)\|^2] \gg \sigma_p^2.$$

- **Variance reduction** (goal): $O(\sigma_{\mathrm{tot}}^2/\mu\epsilon)$ for SGD $\to O(\sigma_p^2/\mu\epsilon)$

Examples of practical gains:

Application case	Estimated ratio $\sigma_{\mathrm{tot}}^2/\sigma_p^2$
Additive Gaussian noise $\mathcal{N}(0,\alpha^2 I)$	$\approx 1 + 1/\alpha^2$
Dropout with probability δ	$\approx 1 + 1/\delta$
Feature rescaling by s in $\mathcal{U}(1-w, 1+w)$	$\approx 1 + 3/w^2$
ResNet-50, color perturbation	21.9
ResNet-50, rescaling + crop	13.6
Unsupervised CNN, rescaling + crop	9.6
Scattering, gamma correction	9.8

The Stochastic MISO Algorithm

Assumptions:

- **global strong convexity**: f is μ-strongly convex;
- **smoothness**: $\tilde{f}_i(\cdot,\rho)$ is L-smooth for all i and ρ (i.e., differentiable with L-Lipschitz gradients);

Composite case (for non-smooth regularizers h): $F(x) := f(x) + h(x)$.

Algorithm 1 S-MISO

Input: step-size sequence $(\alpha_t)_{t\geq 1}$;
Initialize $x_0 = \frac{1}{n}\sum_i z_i^0$ for some $(z_i^0)_{i=1,\ldots,n}$;
for $t = 1,\ldots$ **do**
Sample index $i_t \sim \{1..n\}$, perturbation $\rho_t \sim \Gamma$, and update:

$$z_i^t = \begin{cases}(1-\alpha_t)z_i^{t-1} + \alpha_t(x_{t-1} - \frac{1}{\mu}\nabla\tilde{f}_i(x_{t-1},\rho_t)), & \text{if } i = i_t\\ z_i^{t-1}, & \text{otherwise.}\end{cases}$$

$$x_t = \frac{1}{n}\sum_{i=1}^n z_i^t \quad \text{or} \quad x_t = \mathrm{prox}_{h/\mu}\left(\frac{1}{n}\sum_{i=1}^n z_i^t\right) \quad \text{(composite).}$$

end for

- Reduces to **MISO** when $\sigma^2 = 0$ (no perturbations) and $\alpha_t = const$.
- Reduces to **SGD** or a variant of **RDA** when $n = 1$.

Link with MISO/Finito. [Defazio et al., 2014; Mairal, 2015; Lin et al., 2015]

- Incrementally updates *approximate quadratic lower bounds* to each f_i of the form $d_i^t(x) = c_i^t + \frac{\mu}{2}\|x - z_i^t\|^2$ as follows:

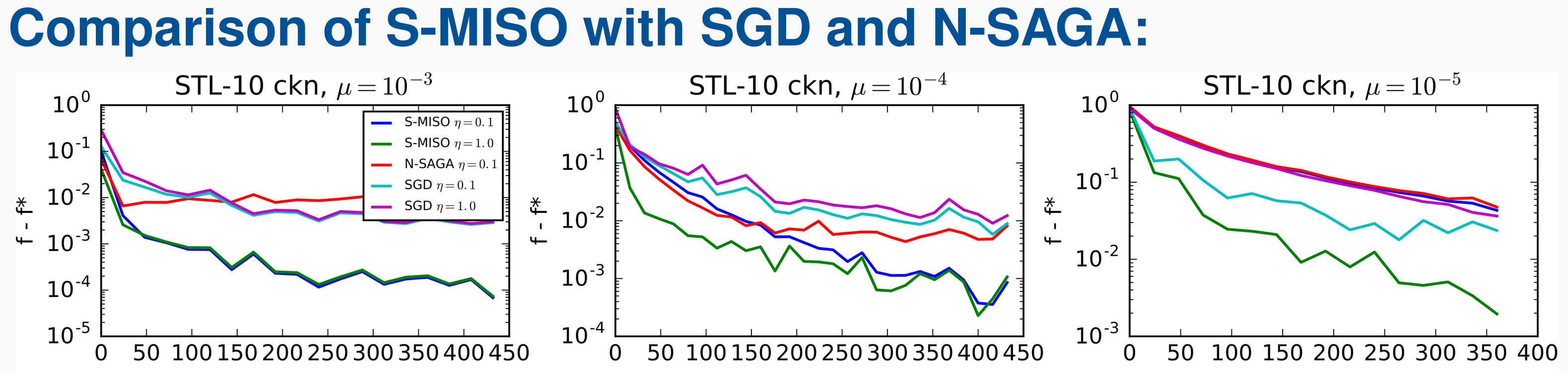
$$d_i^t(x) = \begin{cases}(1-\alpha_t)d_i^{t-1}(x) + \alpha_t\tilde{d}_i^t(x), & \text{if } i = i_t\\ d_i^{t-1}(x), & \text{otherwise,}\end{cases}$$

with $\tilde{d}_i^t(x) = \tilde{f}_i(x_{t-1},\rho_t) + \langle\nabla\tilde{f}_i(x_{t-1},\rho_t), x - x_{t-1}\rangle + \frac{\mu}{2}\|x - x_{t-1}\|^2$.

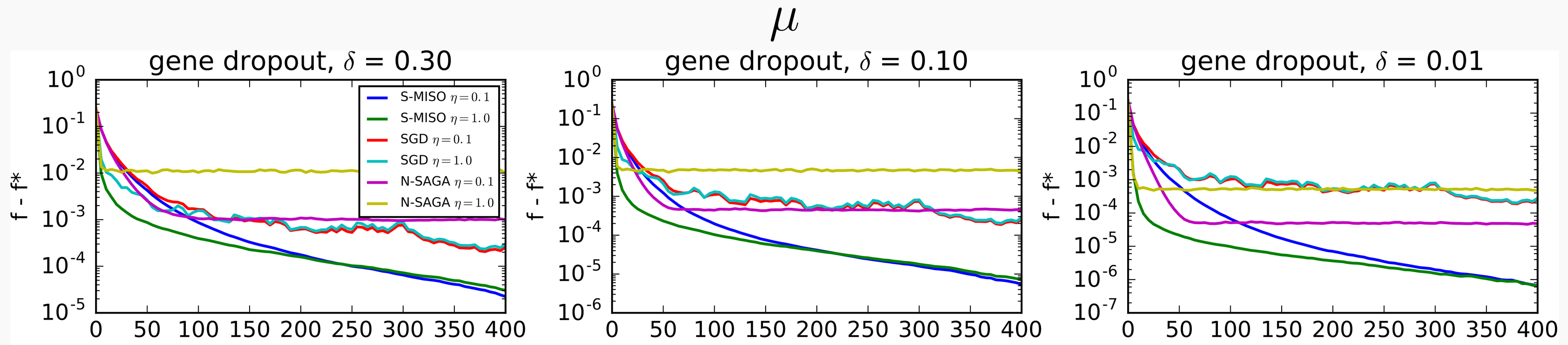
- Without perturbations, the updates are also similar to SDCA without duality [Shalev-Shwartz, 2016].

Experiments

Comparison of S-MISO with SGD and N-SAGA:



data augmentation on STL-10 image dataset with different values of μ



Dropout with different probabilities δ on gene expression data.

Convergence Results

Define the **Lyapunov function**

$$C_t = \frac{1}{2}\|x_t - x^*\|^2 + \frac{\alpha_t}{n^2}\sum_{i=1}^n\|z_i^t - z_i^*\|^2,$$

with $z_i^* := x^* - \frac{1}{\mu}\nabla f_i(x^*)$.

Theorem (Recursion on C_t, smooth case)

If $(\alpha_t)_{t\geq 1}$ are positive, non-increasing step-sizes with

$$\alpha_1 \leq \min\left\{\frac{1}{2}, \frac{n}{2(2\kappa - 1)}\right\},$$

with $\kappa = L/\mu$, then C_t obeys the recursion

$$\mathbb{E}[C_t] \leq \left(1 - \frac{\alpha_t}{n}\right)\mathbb{E}[C_{t-1}] + 2\left(\frac{\alpha_t}{n}\right)^2\frac{\sigma_p^2}{\mu^2}.$$

In the **composite case** with **non-uniform sampling** using distribution $(q_i)_i$, we obtain a similar recursion on

$$C_t^q = F(x^*) - D_t(x_t) + \frac{\mu\alpha_t}{n^2}\sum_{i=1}^n\frac{1}{q_i n}\|z_i^t - z_i^*\|^2,$$

where $D_t(x) = \frac{1}{n}\sum_{i=1}^n d_i^t(x) + h(x)$.

Theorem (Convergence of C_t, decreasing step-sizes)

Let the sequence of step-sizes $(\alpha_t)_{t\geq 1}$ be defined by

$$\alpha_t = \frac{2n}{\gamma + t} \quad \text{for } \gamma \geq 0 \text{ s.t. (1) holds.}$$

For all $t \geq 0$, it holds that

$$\mathbb{E}[C_t] \leq \frac{\nu}{\gamma + t + 1},$$

where

$$\nu := \max\left\{\frac{8\sigma_p^2}{\mu^2}, (\gamma+1)C_0\right\}.$$

Step-size strategy. (See [Bottou, Curtis and Nocedal, 2016] for SGD)

- Keep constant $= \alpha$ for a few epochs to "forget" dependence on C_0
- Decay with $\alpha_t = 2n/(\gamma + t)$, with γ such that $\alpha_1 \approx \alpha$.

Iterate averaging. From $O(L\sigma_p^2/\mu^2\epsilon)$ to $O(\sigma_p^2/\mu\epsilon)$ complexity.

Iteration **complexity** comparison:

Method	Asymptotic error	Iteration complexity
SGD	0	$O\left(\frac{L}{\mu}\log\frac{1}{\bar\epsilon} + \frac{\sigma_{\mathrm{tot}}^2}{\mu\epsilon}\right)$ with $\bar\epsilon = O\left(\frac{\sigma_{\mathrm{tot}}^2}{\mu}\right)$
N-SAGA	$\epsilon_0 = O\left(\frac{\sigma_p^2}{\mu}\right)$	$O\left(\left(n + \frac{L}{\mu}\right)\log\frac{1}{\epsilon}\right)$ with $\epsilon > \epsilon_0$
S-MISO	0	$O\left(\left(n + \frac{L}{\mu}\right)\log\frac{1}{\bar\epsilon} + \frac{\sigma_p^2}{\mu\epsilon}\right)$ with $\bar\epsilon = O\left(\frac{\sigma_p^2}{\mu}\right)$

Code. `https://github.com/albietz/stochs` (C++/Eigen/Cython).