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

Alberto Bietti Julien Mairal

Inria Grenoble (Thoth)

March 21, 2017

Stochastic optimization in machine learning

- **Stochastic approximation**: $\min_x \mathbb{E}_{\zeta \sim \mathcal{D}}[f(x, \zeta)]$
  - Infinite datasets (expected risk, $\mathcal{D}$: data distribution), or "single pass"
  - SGD, stochastic mirror descent, FOBOS, RDA
  - $O(1/\epsilon)$ complexity
- **Incremental methods with variance reduction**: $\min_x \frac{1}{n} \sum_{i=1}^n f_i(x)$
  - Finite datasets (empirical risk): $f_i(x) = \ell(y_i, x^\top \xi_i) + (\mu/2)\|x\|^2$
  - SAG, SDCA, SVRG, SAGA, MISO, etc.
  - $O(\log 1/\epsilon)$ complexity

Data perturbations in machine learning

- Perturbations of data useful for regularization, stable feature selection, privacy aware learning
- We focus on *data augmentation* of a finite training set, for regularization purposes (better performance on test data), e.g.:
  - **Image data augmentation**: add random transformations of each image in the training set (crop, scale, rotate, brightness, contrast, etc.)
  - **Dropout**: set coordinates of feature vectors to 0 with probability δ.

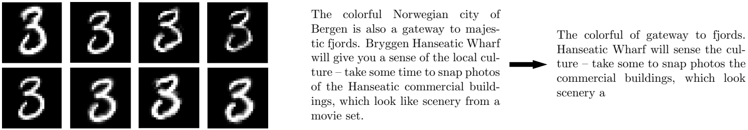


Figure: Data augmentation on MNIST digit (left), Dropout on text (right).

Optimization objective with perturbations

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

- $f_i(x) = \mathbb{E}_{\rho \sim \Gamma}[\tilde{f}_i(x, \rho)]$
- ρ: perturbation
- $\tilde{f}_i(\cdot, \rho)$ is convex with L-Lipschitz gradients
- F is μ-strongly convex
- h: convex, possibly non-smooth, penalty, e.g. ℓ_1 norm

Can we do better than SGD?

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

- SGD is a natural choice
  - Sample index i_t, perturbation $\rho_t \sim \Gamma$
  - Update: $x_t = x_{t-1} - \eta_t \nabla \tilde{f}_{i_t}(x_{t-1}, \rho_t)$
- $O(\sigma_{tot}^2/\mu t)$ convergence, with $\sigma_{tot}^2 := \mathbb{E}_{i,\rho}[\|\nabla \tilde{f}_i(x^*, \rho)\|^2]$
- **Key observation**: variance from perturbations only is small compared to variance across all examples
- **Contribution**: improve convergence of SGD by exploiting the finite-sum structure using variance reduction. Yields $O(\sigma^2/\mu t)$ convergence with

$$\mathbb{E}_{\rho}\left[\|\nabla \tilde{f}_i(x^*, \rho) - \nabla f_i(x^*)\|^2\right] \leq \sigma^2 \ll \sigma_{tot}^2$$

Background: MISO algorithm (Mairal, 2015)

- Finite sum problem: $\min_x f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x)$
- Maintains a **quadratic lower bound model** $d_i^t(x) = \frac{\mu}{2}\|x - z_i^t\|^2 + c_i^t$ on each f_i
- d_i^t is updated using a strong convexity lower bound on f_i:

$$f_i(x) \geq f_i(x_{t-1}) + \langle \nabla f_i(x_{t-1}), x - x_{t-1} \rangle + \frac{\mu}{2}\|x - x_{t-1}\|^2 =: l_i^t(x)$$

- Two steps:
  - Select i_t, update: $d_i^t(x) = \begin{cases} (1-\alpha) d_i^{t-1}(x) + \alpha l_i^t(x), & \text{if } i = i_t \\ d_i^{t-1}(x), & \text{otherwise} \end{cases}$
  - Minimize the model: $x_t = \arg\min_x \{D_t(x) = \frac{1}{n} \sum_{i=1}^n d_i^t(x)\}$

MISO algorithm (Mairal, 2015)

- Final algorithm: at iteration t, choose index i_t at random and update:

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

$$x_t = \frac{1}{n}\sum_{i=1}^{n} z_i^t$$

- Complexity $O((n + L/\mu)\log 1/\epsilon)$, typical of variance reduction
- Similar to SDCA without duality (Shalev-Shwartz, 2016)

Stochastic MISO

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

- With perturbations, we cannot compute exact strong convexity lower bounds on $f_i = \mathbb{E}_\rho[\tilde{f}_i(\cdot, \rho)]$
- Instead, use *approximate* lower bounds using stochastic gradient estimates $\nabla \tilde{f}_{i_t}(x_{t-1}, \rho_t)$
- Allow decreasing step-sizes α_t in order to guarantee convergence as in stochastic approximation

Stochastic MISO: algorithm

Input: step-size sequence $(\alpha_t)_{t\geq 1}$;
for $t = 1, \dots$ **do**
Sample i_t uniformly at random, $\rho_t \sim \Gamma$, and update:

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

$$x_t = \frac{1}{n} \sum_{i=1}^n z_i^t = x_{t-1} + \frac{1}{n}(z_{i_t}^t - z_{i_t}^{t-1}).$$

end for

Note: reduces to MISO for $\sigma^2 = 0, \alpha_t = \alpha$, and to SGD for $n = 1$.

Stochastic MISO: convergence analysis

Define the Lyapunov function (with $z_i^* := x^* - \frac{1}{\mu}\nabla f_i(x^*)$)

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

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

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

Note: Similar recursion for SGD with σ_{tot}^2 instead of σ^2.

Stochastic MISO: convergence with decreasing step-sizes

Similar to SGD (Bottou et al., 2016).

Theorem (Convergence of Lyapunov function)

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

$$\alpha_t = \frac{2n}{\gamma + t} \quad \text{for } \gamma \geq 0 \text{ s.t. } \alpha_1 \leq \min\left\{\frac{1}{2}, \frac{n}{2(2\kappa - 1)}\right\}.$$

For $t \geq 0$,

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

where

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

Q: How can we get rid of the dependence on C_0?

Practical step-size strategy

- Following Bottou et al. (2016), we keep the step-size constant for a few epochs in order to quickly "forget" the initial condition C_0
- Using a **constant step-size** $\bar{\alpha}$, we can converge linearly near a constant error $\bar{C} = \frac{2\bar{\alpha}\sigma^2}{n\mu^2}$ (in practice: a few epochs)
- We then **start decreasing step-sizes** with γ large enough s.t. $\alpha_1 = 2n/(\gamma + 1) \approx \bar{\alpha}$, no more C_0 in the convergence rate!
- Overall, complexity for reaching $\mathbb{E}[\|x_t - x^*\|^2] \leq \epsilon$:

$$O\left((n + L/\mu) \log \frac{C_0}{\bar{\epsilon}}\right) + O\left(\frac{\sigma^2}{\mu^2 \epsilon}\right).$$

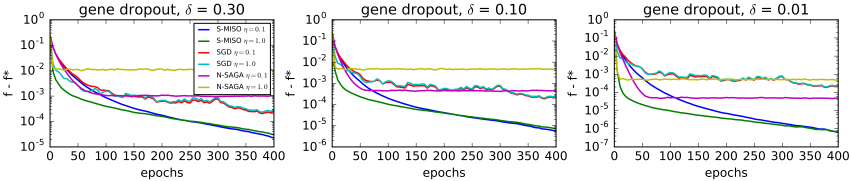
- For $\mathbb{E}[f(x_t) - f(x^*)] \leq \epsilon$, the second term becomes $O(L\sigma^2/\mu^2\epsilon)$ via smoothness. Iterate averaging brings this down to $O(\sigma^2/\mu\epsilon)$.

Extensions

- Composite objectives ($h \neq 0$, e.g., ℓ_1 penalty)
  - MISO extends to this case by adding h to lower bound model (Lin et al., 2015)
  - Different Lyapunov function ($\|x_t - x^*\|^2$ replaced by an upper bound)
  - Similar to Regularized Dual Averaging when $n = 1$
- Non-uniform sampling
  - Smoothness constants L_i of each $\tilde{f}_i$ can vary a lot in heterogeneous datasets
  - Sampling "difficult" examples more often can improve dependence in L from L_{max} to $L_{average}$
- Same convergence results apply (same Lyapunov recursion, decreasing step-sizes, iterate averaging)

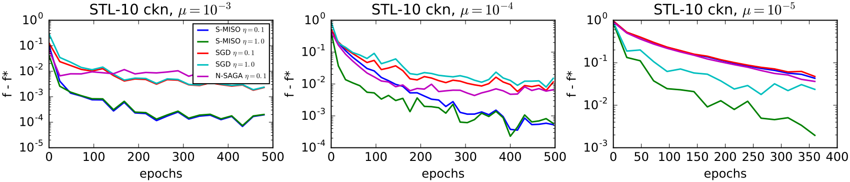
Experiments: dropout

Dropout rate δ controls the variance of the perturbations.



Experiments: image data augmentation

Random image crops and scalings, encoding with an unsupervised deep convolutional network. Different conditioning, controlled by μ.



Conclusion

- Exploit underlying finite-sum structures in stochastic optimization problems using variance reduction
- Bring SGD variance term down to the variance induced by *perturbations only*
- Useful for data augmentation (e.g. random image transformations, Dropout)
- Future work: application to stable feature selection?
- C++/Eigen library with Cython extension available: `http://github.com/albietz/stochs`

References

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization Methods for Large-Scale Machine Learning. *arXiv:1606.04838*, 2016.

S. Lacoste-Julien, M. Schmidt, and F. Bach. A simpler approach to obtaining an $O(1/t)$ convergence rate for the projected stochastic subgradient method. *arXiv:1212.2002*, 2012.

H. Lin, J. Mairal, and Z. Harchaoui. A Universal Catalyst for First-Order Optimization. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

J. Mairal. Incremental Majorization-Minimization Optimization with Application to Large-Scale Machine Learning. *SIAM Journal on Optimization*, 25(2):829–855, 2015.

S. Shalev-Shwartz. SDCA without Duality, Regularization, and Individual Convexity. In *International Conference on Machine Learning (ICML)*, 2016.

Acceleration by iterate averaging

- For function values, averaging helps bring the complexity term $O(L\sigma^2/\mu^2\epsilon)$ down to $O(\sigma^2/\mu\epsilon)$
- Similar technique to Lacoste-Julien et al. (2012), but allows small initial step-sizes

Theorem (Convergence under iterate averaging)

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

$$\alpha_t = \frac{2n}{\gamma + t} \quad \textit{for } \gamma \geq 1 \textit{ s.t. } \alpha_1 \leq \min\left\{\frac{1}{2}, \frac{n}{4(2\kappa - 1)}\right\}.$$

We have

$$\mathbb{E}[f(\bar{x}_T) - f(x^*)] \leq \frac{2\mu\gamma(\gamma-1)C_0}{T(2\gamma + T - 1)} + \frac{16\sigma^2}{\mu(2\gamma + T - 1)},$$

where $\bar{x}_T := \frac{2}{T(2\gamma+T-1)} \sum_{t=0}^{T-1} (\gamma + t) x_t$.

Stochastic MISO (composite, non-uniform sampling)

Input: step-sizes $(\alpha_t)_{t \geq 1}$, sampling distribution q;
for $t = 1, \ldots$ **do**

Sample an index $i_t \sim q$, a perturbation $\rho_t \sim \Gamma$, and update:

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

$$\bar{z}_t = \frac{1}{n} \sum_{i=1}^{n} z_i^t = \bar{z}_{t-1} + \frac{1}{n}(z_{i_t}^t - z_{i_t}^{t-1})$$

$$x_t = \mathrm{prox}_{h/\mu}(\bar{z}_t).$$

end for

Note: Similar to RDA for $n = 1$ when $\alpha_t = 1/t$.

General S-MISO: analysis

- Lyapunov function

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

- Bound on the iterates

$$\frac{\mu}{2}\,\mathbb{E}[\|x_t - x^*\|^2] \leq \mathbb{E}[F(x^*) - D_t(x_t)].$$

- Recursion

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

with $\sigma_q^2 = \frac{1}{n}\sum_i \frac{\sigma_i^2}{q_i n}$.