A Family of Stochastic Surrogate Optimization Algorithms

Julien Mairal **Alberto Bietti**

Inria Grenoble

May 24, 2017

Motivation: large-scale machine learning

Minimizing large finite sums of functions

Given data points $\mathbf{x}_i$, $i = 1, \ldots, n$, learn some **model parameters** θ in $\mathbb{R}^p$ by minimizing

$$\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{x}_i, \theta) + \psi(\theta),$$

where ℓ measures the **data fit**, and ψ is a **regularization function**.

Minimizing expectations

If the amount of data is infinite, we may want to directly minimize the **expected cost**

$$\min_{\theta \in \mathbb{R}^p} \mathbb{E}_{\mathbf{x}}[\ell(\mathbf{x}, \theta)] + \psi(\theta),$$

leading to a **stochastic optimization problem**.

Methodology

We will consider optimization methods that iteratively build a **model** of the objective before updating the variable:

$$\theta_t \in \arg\min_{\theta \in \mathbb{R}^p} g_t(\theta),$$

where g_t is **easy to minimize** and exploits the objective structure: **large finite sum, expectation, (strong) convexity, composite**?

Methodology

We will consider optimization methods that iteratively build a **model** of the objective before updating the variable:

$$\theta_t \in \arg\min_{\theta \in \mathbb{R}^p} g_t(\theta),$$

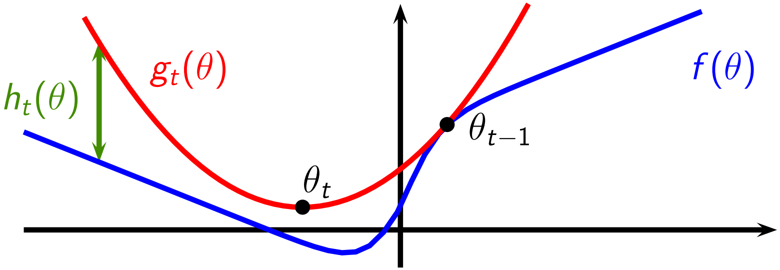
where g_t is **easy to minimize** and exploits the objective structure: **large finite sum, expectation, (strong) convexity, composite**?

There is a large body of related work

- Kelley's and bundle methods;
- incremental and online EM algorithms;
- incremental and stochastic proximal gradient methods;
- variance-reduction techniques for minimizing finite sums.

[Neal and Hinton, 1998; Duchi and Singer, 2009; Bertsekas, 2011; Schmidt et al., 2017; Defazio et al., 2014a; Shalev-Shwartz and Zhang, 2013; Lan and Zhou, 2015]...

Setting: MM with first-order surrogate functions



- $g_t(\theta_t) \geq f(\theta_t)$ for θ_t in $\arg\min_{\theta \in \Theta} g_t(\theta)$;
- the **approximation error** $h_t := g_t - f$ is differentiable, and ∇h_t is L-Lipschitz. Moreover, $h_t(\theta_{t-1}) = 0$ and $\nabla h_t(\theta_{t-1}) = 0$;
- we may also need g_t to be strongly convex;
- example: quadratic upper bound from smoothness.

Theoretical guarantees of the basic MM algorithm

When using first-order surrogates,

- for **convex** problems: $O(L/\epsilon)$ iterations for $f(\theta_t) - f^\star \leq \epsilon$.
- for μ-**strongly convex** ones: $O((L/\mu)\log(1/\epsilon))$.
- for **non-convex** problems: $f(\theta_t)$ monotonically decreases and

$$\liminf_{t \to +\infty} \inf_{\theta \in \Theta} \frac{\nabla f(\theta_t, \theta - \theta_t)}{\|\theta - \theta_t\|_2} \geq 0, \tag{1}$$

  which we call asymptotic stationary point condition.

Directional derivative:

$$\nabla f(\theta, \kappa) = \lim_{\varepsilon \to 0^+} \frac{f(\theta + \varepsilon\kappa) - f(\theta)}{\varepsilon}.$$

- when $\Theta = \mathbb{R}^p$ and f is smooth, (1) is equivalent to $\nabla f(\theta_t) \to 0$.

Outline

1. Stochastic MM algorithm
2. Incremental MM algorithm
3. Faster algorithm for smooth and strongly convex functions
4. Hybrid incremental/stochastic algorithm

Stochastic majorization minimization [Mairal, 2013]

Assume that f is an expectation:

$$f(\theta) = \mathbb{E}_{\mathbf{x}}[\ell(\theta, \mathbf{x})].$$

Recipe

- Draw **a single function** $f_t : \theta \mapsto \ell(\theta, \mathbf{x}_t)$ at iteration t;
- Choose a **first-order surrogate function** $\tilde{g}_t$ for f_t at θ_{t-1};
- **Update the model** $g_t = (1 - w_t)g_{t-1} + w_t\tilde{g}_t$ with appropriate w_t;
- Update θ_t by minimizing g_t.

Related work:

- online EM
- online matrix factorization

[Neal and Hinton, 1998; Cappé and Moulines, 2009; Mairal et al., 2010; Razaviyayn et al., 2016]...

Stochastic majorization minimization [Mairal, 2013]

Theoretical Guarantees - Non-Convex Problems

under a set of reasonable assumptions,

- $f(\theta_t)$ **almost surely converges**;
- the function g_t asymptotically behaves as a first-order surrogate;
- **asymptotic stationary point conditions** hold almost surely.

Theoretical Guarantees - Convex Problems

under a few assumptions, for proximal gradient surrogates, we obtain similar expected rates as SGD with averaging: $O(1/t)$ for **strongly convex problems**, $O(\log(t)/\sqrt{t})$ for **convex ones**.

The most interesting feature of this principle is probably the ability to deal with some non-smooth non-convex problems.

Outline

1. Stochastic MM algorithm
2. Incremental MM algorithm
3. Faster algorithm for smooth and strongly convex functions
4. Hybrid incremental/stochastic algorithm

MISO (MM) for non-convex optimization [Mairal, 2015]

Assume that f is a finite sum:

$$f(\theta) = \frac{1}{n} \sum_{i=1}^{n} f^i(\theta).$$

Recipe

- Draw at random **a single index** i_t at iteration t;
- Compute a **first-order surrogate** $g_t^{i_t}$ of f^{i_t} at θ_{t-1};
- **Incrementally update** the approximate surrogate

$$g_t := \frac{1}{n} \sum_{i=1}^{n} g_t^i = g_{t-1} + \frac{1}{n}(g_t^{i_t} - g_{t-1}^{i_t}).$$

- Update θ_t by minimizing g_t.

MISO (MM) for non-convex optimization [Mairal, 2015]

Theoretical Guarantees - Non-Convex Problems
same as the basic MM algorithm with probability one.

Theoretical Guarantees - Convex Problems
when using proximal gradient surrogates,

- for **convex problems**, $O(nL/\epsilon)$.
- for μ-**strongly convex problems**, $O((nL/\mu)\log(1/\epsilon))$.

The computational complexity is the same as ISTA.
Related work for non-convex problems:

- incremental EM
- more specific incremental MM algorithms.

[Neal and Hinton, 1998; Ahn et al., 2006].

Outline

MISO-μ [Mairal, 2015; Lin et al., 2015]

μ-strongly convex, L-smooth functions f^i, objective:

$$f(\theta) := \frac{1}{n} \sum_{i=1}^{n} f^i(\theta) + \psi(\theta),$$

Strong convexity provides simple **quadratic surrogate lower bounds**:

$$g_t^i : \theta \mapsto f^i(\theta_{t-1}) + \nabla f^i(\theta_{t-1})^\top (\theta - \theta_{t-1}) + \frac{\mu}{2} \|\theta - \theta_{t-1}\|_2^2 + \psi(\theta). \quad (\star)$$

MISO-μ [Mairal, 2015; Lin et al., 2015]

μ-strongly convex, L-smooth functions f^i, objective:

$$f(\theta) := \frac{1}{n}\sum_{i=1}^{n} f^i(\theta) + \psi(\theta),$$

Strong convexity provides simple **quadratic surrogate lower bounds**:

$$g_t^i : \theta \mapsto f^i(\theta_{t-1}) + \nabla f^i(\theta_{t-1})^\top(\theta - \theta_{t-1}) + \frac{\mu}{2}\|\theta - \theta_{t-1}\|_2^2 + \psi(\theta). \quad (\star)$$

Recipe

- Draw at random **a single index** i_t at iteration t;
- Update $g_t^{i_t} = (1-\alpha)g_{t-1}^{i_t} + \alpha(\star)$, with $\alpha = \min\left(1, \frac{\mu n}{2(L-\mu)}\right)$
- **Incrementally update** the full surrogate

$$g_t := \frac{1}{n}\sum_{i=1}^{n} g_t^i = g_{t-1} + \frac{1}{n}(g_t^{i_t} - g_{t-1}^{i_t}).$$

- Update θ_t by minimizing g_t.

MISO-μ [Mairal, 2015; Lin et al., 2015]

Convergence of MISO-μ

When the functions f_i are μ-strongly convex, L-smooth:

$$\mathbb{E}[f(\theta_t)] - f^* \leq \frac{1}{\tau}(1-\tau)^{t+1}\left(f(\theta_0) - g_0(\theta_0)\right) \quad \text{with} \quad \tau \geq \min\left\{\frac{\mu}{4L}, \frac{1}{2n}\right\}.$$

Furthermore, we also have fast convergence of the **certificate**

$$\mathbb{E}[f(\theta_t) - g_t(\theta_t)] \leq \frac{1}{\tau}(1-\tau)^{t}\left(f^* - g_0(\theta_0)\right).$$

MISO-μ [Mairal, 2015; Lin et al., 2015]

Convergence of MISO-μ

When the functions f_i are μ-strongly convex, L-smooth:

$$\mathbb{E}[f(\theta_t)] - f^* \leq \frac{1}{\tau}(1-\tau)^{t+1}\left(f(\theta_0) - g_0(\theta_0)\right) \quad \text{with} \quad \tau \geq \min\left\{\frac{\mu}{4L}, \frac{1}{2n}\right\}.$$

Furthermore, we also have fast convergence of the **certificate**

$$\mathbb{E}[f(\theta_t) - g_t(\theta_t)] \leq \frac{1}{\tau}(1-\tau)^{t}\left(f^* - g_0(\theta_0)\right).$$

Complexity: $O((n + L/\mu)\log(1/\epsilon))$. (Like SAG/SAGA/SVRG/...)

Note: similar to variants of SDCA.

[Shalev-Shwartz and Zhang, 2013; Shalev-Shwartz, 2016; Defazio et al., 2014b]

Outline

1. Stochastic MM algorithm
2. Incremental MM algorithm
3. Faster algorithm for smooth and strongly convex functions
4. Hybrid incremental/stochastic algorithm

Hybrid stochastic/incremental optimization: motivation

Hybrid setting: finite sum + random perturbations ρ

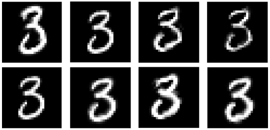
$$f(\theta) := \frac{1}{n}\sum_{i=1}^{n} f^i(\theta) + \psi(\theta) \quad \text{with } f^i(\theta) := \mathbb{E}_\rho[\tilde{f}^i(\theta, \rho)]$$

Applications in machine learning

- improve generalization
- increase robustness
- augment datasets using prior knowledge
- stable feature selection
- privacy ?

Hybrid stochastic/incremental optimization: examples

- **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.



The colorful Norwegian city of Bergen is also a gateway to majestic fjords. Bryggen Hanseatic Wharf will give you a sense of the local culture – take some time to snap photos of the Hanseatic commercial buildings, which look like scenery from a movie set.

$\longrightarrow$

The colorful of gateway to fjords. Hanseatic Wharf will sense the culture – take some to snap photos the commercial buildings, which look scenery a

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

Can we do better than SGD?

$$f(\theta) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\rho[\tilde{f}^i(\theta, \rho)] + \psi(\theta)$$

- Proximal SGD: $O(\sigma_{tot}^2 / \mu\epsilon)$ complexity with

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

- **Can we do better?** if perturbation variance is “small”

Can we do better than SGD?

$$f(\theta) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_\rho[\tilde{f}^i(\theta, \rho)] + \psi(\theta)$$

- Proximal SGD: $O(\sigma_{tot}^2/\mu\epsilon)$ complexity with

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

- **Can we do better?** if perturbation variance is "small"
- Variance decomposition: $\sigma_{tot}^2 = {\sigma_p}^2 + \mathbb{E}_i[\|\nabla f^i(\theta^*) - \nabla f(\theta^*)\|^2]$,

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

Can we do better than SGD?

$$f(\theta) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_\rho[\tilde{f}^i(\theta, \rho)] + \psi(\theta)$$

- Proximal SGD: $O(\sigma_{tot}^2/\mu\epsilon)$ complexity with

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

- **Can we do better?** if perturbation variance is "small"
- Variance decomposition: $\sigma_{tot}^2 = {\sigma_p}^2 + \mathbb{E}_i[\|\nabla f^i(\theta^*) - \nabla f(\theta^*)\|^2]$,

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

- **Stochastic MISO** [Bietti and Mairal, 2017]: remove dependency on variance over i with variance reduction. Complexity $O({\sigma_p}^2/\mu\epsilon)$.

Examples of perturbation variance σ_ρ^2

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

Stochastic MISO [Bietti and Mairal, 2017]

- $\tilde{f}^i(\cdot, \rho)$ are μ-strongly convex, L-smooth
- Similar lower bound surrogates to MISO, but *approximate*

$$\tilde{f}^i(\theta_{t-1}, \rho_t) + \nabla \tilde{f}^i(\theta_{t-1}, \rho_t)^\top (\theta - \theta_{t-1}) + \frac{\mu}{2}\|\theta - \theta_{t-1}\|_2^2 + \psi(\theta). \quad (\star)$$

Recipe

- Draw at random **a single index** i_t at iteration t;
- Update $g_t^{i_t} = (1 - \alpha_t) g_{t-1}^{i_t} + \alpha_t (\star)$;
- **Incrementally update** the full surrogate

$$g_t := \frac{1}{n} \sum_{i=1}^{n} g_t^i = g_{t-1} + \frac{1}{n}(g_t^{i_t} - g_{t-1}^{i_t}).$$

- Update θ_t by minimizing g_t.

Stochastic MISO: convergence analysis ($\psi = 0$)

- Quadratic lower bounds $g_i^i(\theta) = c_t^i + \frac{\mu}{2}\|\theta - z_t^i\|^2$
- Define the **Lyapunov function** (with $z_*^i := \theta^* - \frac{1}{\mu}\nabla f^i(\theta^*)$)

$$C_t = \frac{1}{2}\|\theta_t - \theta^*\|^2 + \frac{\alpha_t}{n^2}\sum_{i=1}^n \|z_t^i - z_*^i\|^2.$$

Stochastic MISO: convergence analysis ($\psi = 0$)

- Quadratic lower bounds $g_i^i(\theta) = c_t^i + \frac{\mu}{2}\|\theta - z_t^i\|^2$
- Define the **Lyapunov function** (with $z_*^i := \theta^* - \frac{1}{\mu}\nabla f^i(\theta^*)$)

$$C_t = \frac{1}{2}\|\theta_t - \theta^*\|^2 + \frac{\alpha_t}{n^2}\sum_{i=1}^{n}\|z_t^i - z_*^i\|^2.$$

- If $(\alpha_t)_t$ decreasing with $\alpha_1 \leq \min\left\{\frac{1}{2}, \frac{n\mu}{4(L-\mu)}\right\}$, then

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

Stochastic MISO: convergence analysis ($\psi = 0$)

- Quadratic lower bounds $g_i^i(\theta) = c_t^i + \frac{\mu}{2}\|\theta - z_t^i\|^2$
- Define the **Lyapunov function** (with $z_*^i := \theta^* - \frac{1}{\mu}\nabla f^i(\theta^*)$)

$$C_t = \frac{1}{2}\|\theta_t - \theta^*\|^2 + \frac{\alpha_t}{n^2}\sum_{i=1}^n \|z_t^i - z_*^i\|^2.$$

- If $(\alpha_t)_t$ decreasing with $\alpha_1 \leq \min\left\{\frac{1}{2}, \frac{n\mu}{4(L-\mu)}\right\}$, then

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

Note:

- Similar recursion for SGD with σ_{tot}^2 instead of σ_p^2;
- Same recursion for composite case, with different C_t.
- See also [Shalev-Shwartz, 2016]

Stochastic MISO: complexity

Two phases

- Constant step-size $\bar{\alpha}$ down to noise level $\bar{\epsilon}$
- Then decay as $\alpha_t = 2n/(\gamma + t)$ with $\alpha_1 \approx \bar{\alpha}$
- [Bottou et al., 2016] for SGD
- Iterate averaging: from $O(L\sigma_p^2/\mu^2\epsilon)$ to $O(\sigma_p^2/\mu\epsilon)$

Stochastic MISO: complexity

Two phases

- Constant step-size $\bar{\alpha}$ down to noise level $\bar{\epsilon}$
- Then decay as $\alpha_t = 2n/(\gamma + t)$ with $\alpha_1 \approx \bar{\alpha}$
- [Bottou et al., 2016] for SGD
- Iterate averaging: from $O(L\sigma_p^2/\mu^2\epsilon)$ to $O(\sigma_p^2/\mu\epsilon)$

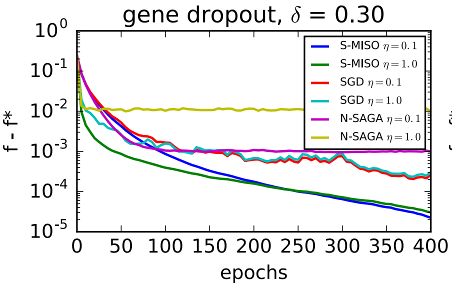
Complexity results

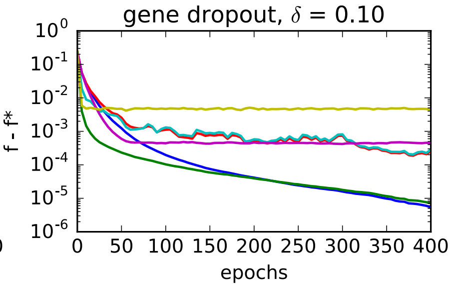
Method	Asymptotic error	Iteration complexity
SGD	0	$O\left(\frac{L}{\mu}\log\frac{1}{\bar{\epsilon}} + \frac{\sigma_{\text{tot}}^2}{\mu\epsilon}\right)$ with $\bar{\epsilon} = O\left(\frac{\sigma_{\text{tot}}^2}{\mu}\right)$
N-SAGA	$\epsilon_0 = O\left(\frac{\sigma_p^2}{\mu}\right)$	$O\left(\left(n +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

[Bottou et al., 2016; Hofmann et al., 2015]

S-MISO experiments: dropout

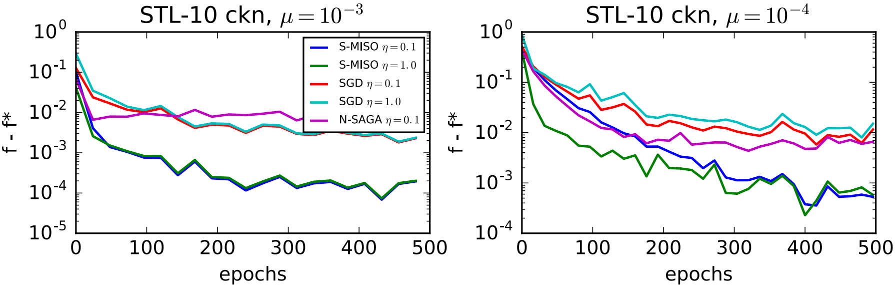
Dropout rate δ controls the variance of the perturbations.





S-MISO experiments: image data augmentation

Random image crops and rescalings, CNN features. Different conditioning, controlled by μ.



Conclusion

- a large class of **majorization-minimization** algorithms for **non-convex, possibly non-smooth, optimization**;
- fast algorithms for minimizing **large sums of convex functions** (using lower bounds).
- a **hybrid algorithm** that interpolates between stochastic and incremental settings and accelerates the hybrid setting

Conclusion

- a large class of **majorization-minimization** algorithms for **non-convex, possibly non-smooth, optimization**;
- fast algorithms for minimizing **large sums of convex functions** (using lower bounds).
- a **hybrid algorithm** that interpolates between stochastic and incremental settings and accelerates the hybrid setting

Related publications

- J. Mairal. Optimization with First-Order Surrogate Functions. *ICML*, 2013.
- J. Mairal. Stochastic Majorization-Minimization Algorithms for Large-Scale Optimization. *NIPS*, 2013.
- J. Mairal. Incremental Majorization-Minimization Optimization with Application to Large-Scale Machine Learning. *SIAM Journal on Optimization*, 2015;
- H. Lin, J. Mairal, and Z. Harchaoui. A Universal Catalyst for First-Order Optimization. *NIPS*, 2015;
- A. Bietti, J. Mairal. Stochastic Optimization with Variance Reduction for Infinite Datasets with Finite-Sum Structure. *arXiv 1610.00970*, 2017.

Stochastic MISO: convergence analysis

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

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

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

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

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

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

For $t \geq 0$,

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

where

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

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*

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

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

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!

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 $O(\sigma^2/\mu\epsilon)$.

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 $\rho_t \sim$ J, and update:

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

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

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

end for

Stochastic MISO (composite, non-uniform sampling)

Input: step-sizes $(\alpha_t)_{t \geq 1}$, sampling distribution q;
for $t = 1, \dots$ **do**
Sample an index $i_t \sim q$, a perturbation $\rho_t \sim$ J, and update:

$$
\begin{aligned}
z_i^t &= \begin{cases} (1 - \frac{\alpha_t}{q_i n}) z_i^{t-1} + \frac{\alpha_t}{q_i n}(x_{t-1} - \frac{1}{\mu} \nabla \tilde{f}_{i_t}(x_{t-1}, \rho_t)), & \text{if } i = i_t \\ z_i^{t-1}, & \text{otherwise} \end{cases} \\
\bar{z}_t &= \frac{1}{n} \sum_{i=1}^n z_i^t = \bar{z}_{t-1} + \frac{1}{n}(z_{i_t}^t - z_{i_t}^{t-1}) \\
x_t &= \mathrm{prox}_{h/\mu}(\bar{z}_t).
\end{aligned}
$$

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

References I

S. Ahn, J. A. Fessler, D. Blatt, and A. O. Hero. Convergent incremental optimization transfer algorithms: Application to tomography. *IEEE Transactions on Medical Imaging*, 25(3):283–296, 2006.

D. P. Bertsekas. Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. *Optimization for Machine Learning*, 2010 (1-38):3, 2011.

A. Bietti and J. Mairal. Stochastic optimization with variance reduction for infinite datasets with finite sum structure. *arXiv:1610.00970*, 2017.

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

O. Cappé and E. Moulines. Online expectation–maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):593–613, June 2009.

A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems (NIPS)*, 2014a.

References II

A. Defazio, J. Domke, and T. S. Caetano. Finito: A faster, permutable incremental gradient method for big data problems. In *International Conference on Machine Learning (ICML)*, 2014b.

J. C. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research (JMLR)*, 10: 2899–2934, 2009.

T. Hofmann, A. Lucchi, S. Lacoste-Julien, and B. McWilliams. Variance Reduced Stochastic Gradient Descent with Neighbors. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

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.

G. Lan and Y. Zhou. An optimal randomized incremental gradient method. *arXiv:1507.02000*, 2015.

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

References III

J. Mairal. Stochastic majorization-minimization algorithms for large-scale optimization. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

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

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research (JMLR)*, 11(Jan):19–60, 2010.

R. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, pages 355–368. Kluwer Academic Publishers, 1998.

M. Razaviyayn, M. Sanjabi, and Z.-Q. Luo. A stochastic successive minimization method for nonsmooth nonconvex optimization with applications to transceiver design in wireless communication networks. *Mathematical Programming*, 157 (2):515–545, 2016.

References IV

M. Schmidt, N. Le Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1):83–112, 2017.

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

S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research (JMLR)*, 14:567–599, 2013.