Stochastic Composite Optimization:
Variance Reduction, Acceleration, and Robustness to Noise

Julien Mairal

Inria Grenoble

Autrans, l'anniversaire d'Anatoli

Collaborator and Former Student

Andrei Kulunchakov

- A. Kulunchakov and J. Mairal. Estimate Sequences for Stochastic Composite Optimization: Variance Reduction, Acceleration, and Robustness to Noise. *Journal of Machine Learning Research (JMLR)*. 2020.
- A. Kulunchakov and J. Mairal. A Generic Acceleration Framework for Stochastic Composite Optimization. *Adv. Neural Information Processing Systems (NeurIPS)*. 2019.

Context of this presentation

We consider **composite** optimization problem

$$\min_{x \in \mathbb{R}^p} \{F(x) := f(x) + \psi(x)\},$$

where f is L-smooth and convex, ψ is convex.

Context of this presentation

We consider **composite** optimization problem

$$\min_{x \in \mathbb{R}^p} \{F(x) := f(x) + \psi(x)\},$$

where f is L-smooth and convex, ψ is convex.

Two settings of interest

Particularly interesting structures in machine learning are

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) \quad \text{or} \quad f(x) = \mathbb{E}[\tilde{f}(x,\xi)].$$

Those can typically be addressed with

- variants of SGD for the general stochastic case.
- variance-reduced algorithms such as SVRG, SAGA, MISO, SARAH, SDCA, Katyusha...

Part I: A few tricks I learned from Anatoli

Part I: A few tricks I learned from Anatoli

and also from reading G. Lan's papers

Trick 1: From sub-linear to linear rates with restarts

Consider a μ-strongly convex function F. Assume that an algorithm $\mathcal{M}$ produces a sequence of iterates $(x_k)_{k \geq 0}$ such that

$$F(x_k) - F^\star \leq \frac{L\|x_0 - x^\star\|^2}{2k^\alpha}.$$

Trick 1: From sub-linear to linear rates with restarts

Consider a μ-strongly convex function F. Assume that an algorithm $\mathcal{M}$ produces a sequence of iterates $(x_k)_{k \geq 0}$ such that

$$\frac{\mu}{2}\|x_k - x^\star\|^2 \leq \; F(x_k) - F^\star \leq \frac{L\|x_0 - x^\star\|^2}{2k^\alpha}.$$

With $t_0 = (2L/\mu)^{1/\alpha}$ iterations, we reduce the error such that $\|x_{t_0} - x^\star\|^2 \leq \frac{1}{2}\|x_0 - x^\star\|^2$.

Trick 1: From sub-linear to linear rates with restarts

Consider a μ-strongly convex function F. Assume that an algorithm $\mathcal{M}$ produces a sequence of iterates $(x_k)_{k\geq 0}$ such that

$$\frac{\mu}{2}\|x_k - x^\star\|^2 \leq \; F(x_k) - F^\star \leq \frac{L\|x_0 - x^\star\|^2}{2k^\alpha}.$$

With $t_0 = (2L/\mu)^{1/\alpha}$ iterations, we reduce the error such that $\|x_{t_0} - x^\star\|^2 \leq \frac{1}{2}\|x_0 - x^\star\|^2$.

Basic multi-stage scheme

This suggests a simple restart strategy with frequency t_0. Up to a few details, for $k = st_0$,

$$F(x_k) - F^\star \leq \frac{F(x_0) - F^\star}{2^s} \leq \left(1 - \frac{1}{2t_0}\right)^k (F(x_0) - F^\star).$$

Note: with $\alpha = 2$, we obtain the complexity of accelerated gradient descent methods.

Trick 1 bis: same idea in a stochastic environment

Consider a μ-strongly convex function F. Assume that an algorithm $\mathcal{M}$ produces a sequence of iterates $(x_k)_{k\geq 0}$ such that

$$\mathbb{E}[F(x_k) - F^\star] \leq \frac{L\|x_0 - x^\star\|^2}{2t^\alpha} + \frac{B\sigma^2}{2}.$$

Trick 1 bis: same idea in a stochastic environment

Consider a μ-strongly convex function F. Assume that an algorithm $\mathcal{M}$ produces a sequence of iterates $(x_k)_{k\geq 0}$ such that

$$\frac{\mu}{2}\mathbb{E}\|x_k - x^\star\|^2 \leq \mathbb{E}[F(x_k) - F^\star] \leq \frac{L\|x_0 - x^\star\|^2}{2t^\alpha} + \frac{B\sigma^2}{2}.$$

Basic multi-stage scheme

Same story: With $t_0 = (2L/\mu)^{1/\alpha}$ and a restarting strategy with frequency t_0, with $k = st_0$,

$$\mathbb{E}[F(x_k) - F^\star] \leq \frac{F(x_0) - F^\star}{2^s} \leq \left(1 - \frac{1}{2t_0}\right)^k (F(x_0) - F^\star) + B\sigma^2.$$

Trick 2: from non-converging to converging algorithms with restarts

Consider a μ-strongly convex function F. Assume that an algorithm $\mathcal{M}$ produces a sequence of iterates $(x_k)_{k\geq 0}$ such that

$$\mathbb{E}[F(x_k) - F^\star] \leq (1 - \tau\eta)^k \left(F(x_0) - F^\star\right) + \eta\sigma^2,$$

where η is a parameter (*e.g.*, a step size) with $0 < \tau\eta < 1$. For instance, a proximal stochastic gradient descent method, with stepsize $\eta \leq 1/L$ and averaging, $\approx$ yields $\tau = \mu$.

Trick 2: from non-converging to converging algorithms with restarts

Consider a μ-strongly convex function F. Assume that an algorithm $\mathcal{M}$ produces a sequence of iterates $(x_k)_{k\geq 0}$ such that

$$\mathbb{E}[F(x_k) - F^\star] \leq (1-\tau\eta)^k \left(F(x_0) - F^\star\right) + \eta\sigma^2,$$

where η is a parameter (*e.g.*, a step size) with $0 < \tau\eta < 1$. For instance, a proximal stochastic gradient descent method, with stepsize $\eta \leq 1/L$ and averaging, $\approx$ yields $\tau = \mu$.

Multi-stage scheme

Choose a sequence $\eta_t = \eta_0/2^t$, restart, while solving each sub-problem with accuracy $2\eta_t\sigma^2$.

Trick 2: from non-converging to converging algorithms with restarts

Consider a μ-strongly convex function F. Assume that an algorithm $\mathcal{M}$ produces a sequence of iterates $(x_k)_{k \geq 0}$ such that

$$\mathbb{E}[F(x_k) - F^\star] \leq (1 - \tau\eta)^k \left(F(x_0) - F^\star\right) + \eta\sigma^2,$$

where η is a parameter (*e.g.*, a step size) with $0 < \tau\eta < 1$. For instance, a proximal stochastic gradient descent method, with stepsize $\eta \leq 1/L$ and averaging, $\approx$ yields $\tau = \mu$.

Multi-stage scheme

Choose a sequence $\eta_t = \eta_0/2^t$, restart, while solving each sub-problem with accuracy $2\eta_t\sigma^2$. Then, let us compute the complexity to achieve $\mathbb{E}[F(x_k) - F^\star] \leq \varepsilon$ (with $\varepsilon \leq 2\eta_0\sigma^2$).

- **first stage:** to obtain $\mathbb{E}[F(x_k) - F^\star] \leq 2\eta_0\sigma^2$, the complexity is

$$O\left(\frac{1}{\tau\eta_0} \log\left(\frac{F(x_0) - F^\star}{\varepsilon}\right)\right)$$

Trick 2: from non-converging to converging algorithms with restarts

Consider a μ-strongly convex function F. Assume that an algorithm $\mathcal{M}$ produces a sequence of iterates $(x_k)_{k \geq 0}$ such that

$$\mathbb{E}[F(x_k) - F^\star] \leq (1 - \tau\eta)^k \left(F(x_0) - F^\star\right) + \eta\sigma^2,$$

where η is a parameter (*e.g.*, a step size) with $0 < \tau\eta < 1$. For instance, a proximal stochastic gradient descent method, with stepsize $\eta \leq 1/L$ and averaging, $\approx$ yields $\tau = \mu$.

Multi-stage scheme

Choose a sequence $\eta_t = \eta_0 / 2^t$, restart, while solving each sub-problem with accuracy $2\eta_t\sigma^2$. Then, let us compute the complexity to achieve $\mathbb{E}[F(x_k) - F^\star] \leq \varepsilon$ (with $\varepsilon \leq 2\eta_0\sigma^2$).

- **next stages:** each stage reduces the error by a factor 2 and the total complexity becomes

$$O\left(\frac{1}{\tau\eta_0}\log\left(\frac{F(x_0) - F^\star}{\varepsilon}\right)\right) + \sum_{t=1}^{T} O\left(\frac{1}{\tau\eta_t}\right).$$

Trick 2: from non-converging to converging algorithms with restarts

Consider a μ-strongly convex function F. Assume that an algorithm $\mathcal{M}$ produces a sequence of iterates $(x_k)_{k\geq 0}$ such that

$$\mathbb{E}[F(x_k) - F^\star] \leq (1-\tau\eta)^k \left(F(x_0) - F^\star\right) + \eta\sigma^2,$$

where η is a parameter (*e.g.*, a step size) with $0 < \tau\eta < 1$. For instance, a proximal stochastic gradient descent method, with stepsize $\eta \leq 1/L$ and averaging, $\approx$ yields $\tau = \mu$.

Multi-stage scheme

Choose a sequence $\eta_t = \eta_0/2^t$, restart, while solving each sub-problem with accuracy $2\eta_t\sigma^2$. Then, let us compute the complexity to achieve $\mathbb{E}[F(x_k) - F^\star] \leq \varepsilon$ (with $\varepsilon \leq 2\eta_0\sigma^2$).

- **next stages:** each stage reduces the error by a factor 2 and the total complexity becomes

$$O\left(\frac{1}{\tau\eta_0}\log\left(\frac{F(x_0)-F^\star}{\varepsilon}\right)\right) + O\left(\frac{\sigma^2}{\tau\varepsilon}\right).$$

Trick 3: importance of iterate averaging

Consider for instance proximal SGD with fixed step-size $1/L$ **without averaging**

$$\mathbb{E}\left[F(x_k) - F^\star + \frac{L}{2}\|x_k - x^\star\|^2\right] \leq \left(1 - \frac{\mu}{L}\right)^k \frac{L\|x_0 - x^\star\|^2}{2} + \frac{\sigma^2}{\mu},$$

Trick 3: importance of iterate averaging

Consider for instance proximal SGD with fixed step-size $1/L$ **with averaging**

$$\mathbb{E}\left[F(\hat{x}_k) - F^\star + \frac{\mu}{2}\|x_k - x^\star\|^2\right] \leq \left(1 - \frac{\mu}{L}\right)^k \left(F(x_0) - F^\star + \frac{\mu}{2}\|x_0 - x^\star\|^2\right) + \frac{\sigma^2}{L},$$

Trick 3: importance of iterate averaging

Consider for instance proximal SGD with fixed step-size $1/L$ **with averaging**

$$\mathbb{E}\left[F(\hat{x}_k) - F^\star + \frac{\mu}{2}\|x_k - x^\star\|^2\right] \leq \left(1 - \frac{\mu}{L}\right)^k \left(F(x_0) - F^\star + \frac{\mu}{2}\|x_0 - x^\star\|^2\right) + \frac{\sigma^2}{L},$$

With restart, we achieve the complexity

$$O\left(\frac{L}{\mu}\log\left(\frac{C_0}{\varepsilon}\right)\right) + O\left(\frac{\sigma^2}{\mu\varepsilon}\right).$$

Here, iterate averaging improves the dependence on σ^2.

Trick 3: importance of averaging

Consider another algorithm that achieves

$$\mathbb{E}[F(x_k) - F^\star] \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^k \left(F(x_0) - F^\star + \frac{\mu}{2}\|x_0 - x^\star\|^2\right) + \frac{\sigma^2}{\sqrt{\mu L}},$$

Trick 3: importance of averaging

Consider another algorithm that achieves

$$\mathbb{E}[F(x_k) - F^\star] \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^k \left(F(x_0) - F^\star + \frac{\mu}{2}\|x_0 - x^\star\|^2\right) + \frac{\sigma^2}{\sqrt{\mu L}},$$

With restart, we achieve the complexity

$$O\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{C_0}{\varepsilon}\right)\right) + O\left(\frac{\sigma^2}{\mu\varepsilon}\right).$$

This is the optimal complexity for stochastic first-order optimization (see Ghadimi and Lan, 2013). Note that we did not mention averaging here....

Part II: Stochastic Composite Optimization with Estimate Sequences

Reminder: Context of this presentation

We consider **composite** optimization problem

$$\min_{x \in \mathbb{R}^p} \{F(x) := f(x) + \psi(x)\},$$

where f is L-smooth and convex, ψ is convex.

Reminder: Context of this presentation

We consider **composite** optimization problem

$$\min_{x \in \mathbb{R}^p} \{F(x) := f(x) + \psi(x)\},$$

where f is L-smooth and convex, ψ is convex.

Two settings of interest

Particularly interesting structures in machine learning are

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) \qquad \text{or} \qquad f(x) = \mathbb{E}[\tilde{f}(x, \xi)].$$

Those can typically be addressed with

- variants of SGD for the general stochastic case.
- variance-reduced algorithms such as SVRG, SAGA, MISO, SARAH, SDCA, Katyusha...

Complexity of SGD variants for composite functions

We consider the worst-case complexity for finding a point $\bar{x}$ such that $\mathbb{E}[F(\bar{x}) - F^\star] \leq \varepsilon$ for

$$\min_{x \in \mathbb{R}^p} \{F(x) := \mathbb{E}[\tilde{f}(x, \xi)] + \psi(x)\},$$

In this talk, we consider the μ-strongly convex case only.

Complexity of SGD with iterate averaging

$$O\left(\frac{L}{\mu} \log\left(\frac{C_0}{\varepsilon}\right)\right) + O\left(\frac{\sigma^2}{\mu\varepsilon}\right),$$

under the (strong) assumption that the gradient estimates have **bounded variance** σ^2.

Complexity of SGD variants for composite functions

We consider the worst-case complexity for finding a point $\bar{x}$ such that $\mathbb{E}[F(\bar{x}) - F^\star] \leq \varepsilon$ for

$$\min_{x \in \mathbb{R}^p} \{F(x) := \mathbb{E}[\tilde{f}(x, \xi)] + \psi(x)\},$$

In this talk, we consider the μ-strongly convex case only.

Complexity of SGD with iterate averaging

$$O\left(\frac{L}{\mu}\log\left(\frac{C_0}{\varepsilon}\right)\right) + O\left(\frac{\sigma^2}{\mu\varepsilon}\right),$$

under the (strong) assumption that the gradient estimates have **bounded variance** σ^2.

Complexity of accelerated SGD [Ghadimi and Lan, 2013]

$$O\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{C_0}{\varepsilon}\right)\right) + O\left(\frac{\sigma^2}{\mu\varepsilon}\right).$$

Complexity for (deterministic) finite sums

We consider the worst-case complexity for finding a point $\bar{x}$ such that $\mathbb{E}[F(\bar{x}) - F^\star] \leq \varepsilon$ for

$$\min_{x \in \mathbb{R}^p} \left\{ F(x) := \frac{1}{n} \sum_{i=1}^n f_i(x) + \psi(x) \right\},$$

Complexity of SAGA/SVRG/SDCA/MISO/S2GD

$$O\left(\left(n + \frac{\bar{L}}{\mu}\right) \log\left(\frac{C_0}{\varepsilon}\right)\right) \quad \text{with} \quad \bar{L} = \frac{1}{n}\sum_{i=1}^n L_i.$$

Complexity of GD and acc-GD

$$O\left(\left(n\frac{L}{\mu}\right) \log\left(\frac{C_0}{\varepsilon}\right)\right) \quad \text{vs.} \quad O\left(\left(n\sqrt{\frac{L}{\mu}}\right) \log\left(\frac{C_0}{\varepsilon}\right)\right).$$

see also SDCA [Shalev-Shwartz and Zhang, 2014] and Catalyst [Lin, Mairal, and Harchaoui, 2018].

Complexity for (deterministic) finite sums

We consider the worst-case complexity for finding a point $\bar{x}$ such that $\mathbb{E}[F(\bar{x}) - F^\star] \leq \varepsilon$ for

$$\min_{x \in \mathbb{R}^p} \left\{ F(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x) + \psi(x) \right\},$$

Complexity of SAGA/SVRG/SDCA/MISO/S2GD

$$O\left(\left(n + \frac{\bar{L}}{\mu} \right) \log \left(\frac{C_0}{\varepsilon} \right) \right) \quad \text{with} \quad \bar{L} = \frac{1}{n} \sum_{i=1}^{n} L_i.$$

Complexity of Katyusha [Allen-Zhu, 2017]

$$O\left(\left(n + \sqrt{\frac{n\bar{L}}{\mu}} \right) \log \left(\frac{C_0}{\varepsilon} \right) \right).$$

see also SDCA [Shalev-Shwartz and Zhang, 2014] and Catalyst [Lin, Mairal, and Harchaoui, 2018].

Variance reduction

Variance reduction

Consider two random variables X, Y and define

$$Z = X - Y + \mathbb{E}[Y].$$

Then,

- $\mathbb{E}[Z] = \mathbb{E}[X]$
- $\mathrm{Var}(Z) = \mathrm{Var}(X) + \mathrm{Var}(Y) - 2\mathrm{cov}(X, Y)$.

The variance of Z may be smaller if X and Y are positively correlated.

Variance reduction

Variance reduction

Consider two random variables X, Y and define

$$Z = X - Y + \mathbb{E}[Y].$$

Then,

- $\mathbb{E}[Z] = \mathbb{E}[X]$
- $\text{Var}(Z) = \text{Var}(X) + \text{Var}(Y) - 2\text{cov}(X, Y)$.

The variance of Z may be smaller if X and Y are positively correlated.

Why is it useful for stochastic optimization?

- step-sizes for SGD have to decrease to ensure convergence.
- with variance reduction, one may use **larger constant** step-sizes.

Contributions of our work without acceleration

We extend and generalize the concept of **estimate sequences** introduced by Y. Nesterov to

- provide a **unified proof of convergence** for SAGA/random-SVRG/MISO.
- provide them **adaptivity for unknown μ** (known before for SAGA only).
- make them **robust to stochastic noise**, *e.g.*, for solving

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) \quad \text{with} \quad f_i(x) = \mathbb{E}[\tilde{f}_i(x,\xi)].$$

  with complexity

$$O\left(\left(n + \frac{\bar{L}}{\mu}\right)\log\left(\frac{C_0}{\varepsilon}\right)\right) + O\left(\frac{\tilde{\sigma}^2}{\mu\varepsilon}\right) \quad \text{with} \quad \tilde{\sigma}^2 \ll \sigma^2,$$

  where $\tilde{\sigma}^2$ is the variance due to small perturbations.
- obtain **new variants** of the above algorithms with the same guarantees.

Contributions of our work with acceleration

- we propose a **simple accelerated SGD algorithm** for composite optimization with optimal complexity
$$O\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{C_0}{\varepsilon}\right)\right) + O\left(\frac{\sigma^2}{\mu\varepsilon}\right),$$
- we propose an **accelerated variant** of SVRG for the stochastic finite-sum problem with complexity
$$O\left(\left(n+\sqrt{\frac{n\bar{L}}{\mu}}\right)\log\left(\frac{C_0}{\varepsilon}\right)\right) + O\left(\frac{\tilde{\sigma}^2}{\mu\varepsilon}\right) \quad \text{with} \quad \tilde{\sigma}^2 \ll \sigma^2.$$

When $\tilde{\sigma}=0$, the complexity matches that of Katyusha.

Estimate sequences

Definition [Nesterov].

A pair of sequences $(\varphi_k)_{t\geq 0}$ and $(\lambda_k)_{t\geq 0}$, with $\lambda_k \geq 0$ and $\varphi_k : \mathbb{R}^p \to \mathbb{R}$, is called an **estimate sequence** of function f if $\lambda_k \to 0$ and

$$\text{for any } x \in \mathbb{R}^p \text{ and all } k \geq 0, \qquad \varphi_k(x) \leq (1-\lambda_k) f(x) + \lambda_k \varphi_0(x).$$

In addition, if for some sequence $(x_k)_{k\geq 0}$ we have

$$f(x_k) \leq \varphi_k^\star \triangleq \min_{x\in\mathbb{R}^p} \varphi_k(x),$$

then

$$f(x_k) - f^\star \leq \lambda_k(\varphi_0(x^\star) - f^\star),$$

where $x^\star$ is a minimizer of f.

Estimate sequences

In summary, we need two properties

1. $\varphi_k(x) \leq (1 - \lambda_k) f(x) + \lambda_k \varphi_0(x)$;
2. $f(x_k) \leq \varphi_k^\star \stackrel{\triangle}{=} \min_{x \in \mathbb{R}^p} \varphi_k(x)$.

Remarks

- φ_k is neither an upper-bound, nor a lower-bound;
- Finding the right estimate sequence is often nontrivial.

Estimate sequences

In summary, we need two properties

1. $\varphi_k(x) \leq (1 - \lambda_k) f(x) + \lambda_k \varphi_0(x)$;
2. $f(x_k) \leq \varphi_k^\star \triangleq \min_{x \in \mathbb{R}^p} \varphi_k(x)$.

How to build an estimate sequence?

Define φ_k recursively

$$\varphi_k(x) \triangleq (1 - \alpha_k)\varphi_{k-1}(x) + \alpha_k d_k(x),$$

where d_k is a **lower-bound**, e.g., if f is smooth,

$$d_k(x) \triangleq f(y_k) + \nabla f(y_k)^\top (x - y_k) + \frac{\mu}{2}\|x - y_k\|_2^2,$$

Then, work hard to choose α_k as large as possible, and y_k and x_k such that property 2 holds. Subsequently, $\lambda_k = \prod_{t=1}^t (1 - \alpha_k)$.

Estimate sequences

In summary, we need two properties

1. $\varphi_k(x) \leq (1-\lambda_k) f(x) + \lambda_k \varphi_0(x)$;
2. $f(x_k) \leq \varphi_k^\star \stackrel{\triangle}{=} \min_{x \in \mathbb{R}^p} \varphi_k(x)$.

Example: if $\alpha_k = \frac{2}{k+2}$, then $\lambda_k = \prod_{t=1}^{t}(1-\alpha_t) = \frac{2}{(k+1)(k+2)} = O(1/k^2)$.

- Proofs based on estimates sequences are typically **constructive** and build the algorithm at the same time as they prove convergence, while **describing** the underlying model φ_k.
- But they lead to tedious calculations (about 2 pages).
- What we will need to do is to handle stochastic estimates of the gradients.

Estimate sequences

In summary, we need two properties

1. $\varphi_k(x) \leq (1-\lambda_k) f(x) + \lambda_k \varphi_0(x)$;
2. $f(x_k) \leq \varphi_k^\star \stackrel{\triangle}{=} \min_{x \in \mathbb{R}^p} \varphi_k(x)$.

Example: if $\alpha_k = \frac{2}{k+2}$, then $\lambda_k = \prod_{t=1}^{t}(1-\alpha_t) = \frac{2}{(k+1)(k+2)} = O(1/k^2)$.

- Proofs based on estimates sequences are typically **constructive** and build the algorithm at the same time as they prove convergence, while **describing** the underlying model φ_k.
- But they lead to tedious calculations (about 2 pages).
- What we will need to do is to handle stochastic estimates of the gradients.

A classical iteration

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

A classical iteration

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

covers SGD, SAGA, SVRG, and composite variants.

A classical iteration

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

covers SGD, SAGA, SVRG, and composite variants.

Interpretation

x_k minimizes the quadratic function φ_k, defined as

$$\varphi_k(x) = (1-\delta_k)\varphi_{k-1}(x) + \delta_k\Big(f(x_{k-1}) + g_k^\top(x - x_{k-1}) + \frac{\mu}{2}\|x - x_{k-1}\|^2 \\ \ldots + \psi(x_k) + \psi'(x_k)^\top(x - x_k)\Big),$$

where $\delta_k = \mu\eta_k$, $\psi'(x_k)$ is a subgradient in $\partial\psi(x_k)$, and $\varphi_0(x) = \varphi_0^\star + \frac{\mu}{2}\|x - x_0\|^2$.

A classical iteration

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

covers SGD, SAGA, SVRG, and composite variants.

Interpretation

x_k minimizes the quadratic function φ_k, defined as

$$\varphi_k(x) = (1-\delta_k)\varphi_{k-1}(x) + \delta_k\Big(f(x_{k-1}) + g_k^\top(x - x_{k-1}) + \frac{\mu}{2}\|x - x_{k-1}\|^2 \\ \ldots + \psi(x_k) + \psi'(x_k)^\top(x - x_k)\Big),$$

where $\delta_k = \mu\eta_k$, $\psi'(x_k)$ is a subgradient in $\partial\psi(x_k)$, and $\varphi_0(x) = \varphi_0^\star + \frac{\mu}{2}\|x - x_0\|^2$.

This is similar to the construction of **estimate sequences** by Y. Nesterov.
see also [Devolder, 2011, Lin et al., 2014] for stochastic problems.

A less classical iteration

$$\boxed{x_k = \text{Prox}_{\psi/\mu}\left[\bar{x}_k\right] \quad \text{with} \quad \bar{x}_k \leftarrow (1-\delta_k)\bar{x}_{k-1} + \delta_k x_k - \eta_k g_k \quad \text{and} \quad \mathbb{E}[g_k|\mathcal{F}_k] = \nabla f(x_{k-1}),}$$

covers MISO/Finito/primal SDCA with $\delta_k = \mu\eta_k$.

Interpretation

x_k minimizes the function φ_k, defined as

$$\varphi_k(x) \;=\; (1-\delta_k)\varphi_{k-1}(x) + \delta_k\Big(f(x_{k-1}) + g_k^\top(x - x_{k-1}) + \frac{\mu}{2}\|x - x_{k-1}\|^2 + \psi(x)\Big).$$

Estimate sequences will provide identical convergence proofs for both types of iterations.

Our convergence result with stochastic estimate sequences

General convergence result (no acceleration yet)

if $\eta_t \leq 1/L$ for all $t \geq 0$, then for all $k \geq 1$,

$$\mathbb{E}\left[F(\hat{x}_k) - F^\star + \frac{\mu}{2}\|x_k - x^\star\|^2\right] \leq \Gamma_k \left(F(x_0) - F^\star + \frac{\mu}{2}\|x_0 - x^\star\|^2 + \sum_{t=1}^{k} \frac{\delta_t \eta_t \sigma_t^2}{\Gamma_t}\right).$$

where $\Gamma_k = \prod_{t=1}^{k}(1-\delta_t)$, $\hat{x}_k = (1-\delta_k)\hat{x}_{k-1} + \delta_k x_k$, and $\sigma_t^2 = \mathbb{E}[\|g_t - \nabla f(x_{t-1})\|^2]$.

Our convergence result with stochastic estimate sequences

General convergence result (no acceleration yet)

if $\eta_t \leq 1/L$ for all $t \geq 0$, then for all $k \geq 1$,

$$\mathbb{E}\left[F(\hat{x}_k) - F^\star + \frac{\mu}{2}\|x_k - x^\star\|^2\right] \leq \Gamma_k \left(F(x_0) - F^\star + \frac{\mu}{2}\|x_0 - x^\star\|^2 + \sum_{t=1}^{k} \frac{\delta_t \eta_t \sigma_t^2}{\Gamma_t} \right).$$

where $\Gamma_k = \prod_{t=1}^{k}(1-\delta_t)$, $\hat{x}_k = (1-\delta_k)\hat{x}_{k-1} + \delta_k x_k$, and $\sigma_t^2 = \mathbb{E}[\|g_t - \nabla f(x_{t-1})\|^2]$.

Corollary: SGD with constant step size $\eta_k = 1/L$, with averaging

$$\mathbb{E}\left[F(\hat{x}_k) - F^\star + \frac{\mu}{2}\|x_k - x^\star\|^2\right] \leq 2\left(1 - \frac{\mu}{L}\right)^k (F(x_0) - F^\star) + \frac{\sigma^2}{L}.$$

Our convergence result with stochastic estimate sequences

General convergence result (no acceleration yet)

if $\eta_t \leq 1/L$ for all $t \geq 0$, then for all $k \geq 1$,

$$\mathbb{E}\left[F(\hat{x}_k) - F^\star + \frac{\mu}{2}\|x_k - x^\star\|^2\right] \leq \Gamma_k \left(F(x_0) - F^\star + \frac{\mu}{2}\|x_0 - x^\star\|^2 + \sum_{t=1}^{k} \frac{\delta_t \eta_t \sigma_t^2}{\Gamma_t}\right).$$

where $\Gamma_k = \prod_{t=1}^{k}(1-\delta_t)$, $\hat{x}_k = (1-\delta_k)\hat{x}_{k-1} + \delta_k x_k$, and $\sigma_t^2 = \mathbb{E}[\|g_t - \nabla f(x_{t-1})\|^2]$.

Corollary: SGD with constant step size $\eta_k = 1/L$, with averaging

$$\#\text{Comp} = O\left(\frac{L}{\mu}\log\left(\frac{C_0}{\varepsilon}\right)\right) \quad \text{with} \quad \text{Bias} = \frac{\sigma^2}{L}.$$

Our convergence result with stochastic estimate sequences

General convergence result (no acceleration yet)

if $\eta_t \leq 1/L$ for all $t \geq 0$, then for all $k \geq 1$,

$$\mathbb{E}\left[F(\hat{x}_k) - F^\star + \frac{\mu}{2}\|x_k - x^\star\|^2\right] \leq \Gamma_k \left(F(x_0) - F^\star + \frac{\mu}{2}\|x_0 - x^\star\|^2 + \sum_{t=1}^{k} \frac{\delta_t \eta_t \sigma_t^2}{\Gamma_t}\right).$$

where $\Gamma_k = \prod_{t=1}^{k}(1-\delta_t)$, $\hat{x}_k = (1-\delta_k)\hat{x}_{k-1} + \delta_k x_k$, and $\sigma_t^2 = \mathbb{E}[\|g_t - \nabla f(x_{t-1})\|^2]$.

Corollary: two-stage SGD with (i) constant step size; then (ii) decreasing step sizes

$$\#\text{Comp} = O\left(\frac{L}{\mu}\log\left(\frac{C_0}{\varepsilon}\right)\right) + O\left(\frac{\sigma^2}{\mu\varepsilon}\right).$$

An accelerated SGD algorithm

An algorithm derived from the estimate sequence method.

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

Interpretation

x_k minimizes the quadratic function φ_k, defined as

$$\varphi_k(x) = (1-\delta_k)\varphi_{k-1}(x) + \delta_k\Big(f(y_{k-1}) + g_k^\top(x - y_{k-1}) + \frac{\mu}{2}\|x - y_{k-1}\|^2$$
$$\ldots + \psi(x_k) + \psi'(x_k)^\top(x - x_k)\Big),$$

where $\delta_k = \mu\eta_k$, $\psi'(x_k)$ is a subgradient in $\partial\psi(x_k)$, and $\varphi_0(x) = \varphi_0^\star + \frac{\mu}{2}\|x - x_0\|^2$.

An accelerated SGD algorithm

An algorithm derived from the estimate sequence method.

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

Complexity: acc-SGD with constant step size $\eta_k = 1/L$

$$\mathbb{E}\left[F(x_k) - F^\star\right] \leq 2\left(1 - \sqrt{\frac{\mu}{L}}\right)^k (F(x_0) - F^\star) + \frac{\sigma^2}{\sqrt{\mu L}}.$$

Note that the bias is larger than regular SGD by $\sqrt{L/\mu}$.

An accelerated SGD algorithm

An algorithm derived from the estimate sequence method.

$$x_k = \text{Prox}_{\eta_k \psi}\left[y_{k-1} - \eta_k g_k\right] \quad \text{with} \quad \mathbb{E}[g_k|\mathcal{F}_{k-1}] = \nabla f(y_{k-1})$$

$$y_k = x_k + \beta_k(x_k - x_{k-1}) \quad \text{with} \quad \beta_k = \frac{\delta_k(1-\delta_k)\eta_{k+1}}{\eta_k \delta_{k+1} + \eta_{k+1}\delta_k^2},$$

Corollary: acc-SGD with constant step size $\eta_k = 1/L$, without averaging

$$\#\text{Comp} = O\left(\sqrt{\frac{L}{\mu}} \log\left(\frac{C_0}{\varepsilon}\right)\right) \quad \text{with} \quad \text{Bias} = \frac{\sigma^2}{\sqrt{\mu L}}.$$

An accelerated SGD algorithm

An algorithm derived from the estimate sequence method.

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

Corollary: two-stage acc-SGD with (i) constant step size; then (ii) decreasing step sizes

$$\#\text{Comp} = O\left(\sqrt{\frac{L}{\mu}} \log\left(\frac{C_0}{\varepsilon}\right)\right) + O\left(\frac{\sigma^2}{\mu\varepsilon}\right).$$

An accelerated SVRG algorithm for stochastic finite-sum problems

- Choose the extrapolation point

$$y_{k-1} = \theta_k v_{k-1} + (1-\theta_k)\tilde{x}_{k-1};$$

- Compute the noisy gradient estimator

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

- Obtain the new iterate

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

- Find the minimizer v_k of the estimate sequence:

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

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

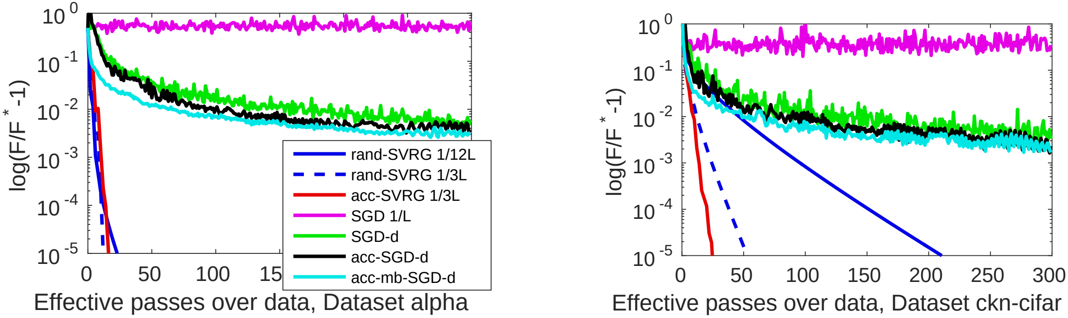
An accelerated SVRG algorithm for stochastic finite-sum problems

Remarks

- design of the algorithm and convergence proofs are based on estimate sequences.
- with two stages, the algorithm achieves the optimal complexity

$$O\left(\left(n + \sqrt{\frac{n\bar{L}}{\mu}}\right)\log\left(\frac{C_0}{\varepsilon}\right)\right) + O\left(\frac{\tilde{\sigma}^2}{\mu\varepsilon}\right) \qquad \text{with} \qquad \tilde{\sigma}^2 \ll \sigma^2.$$

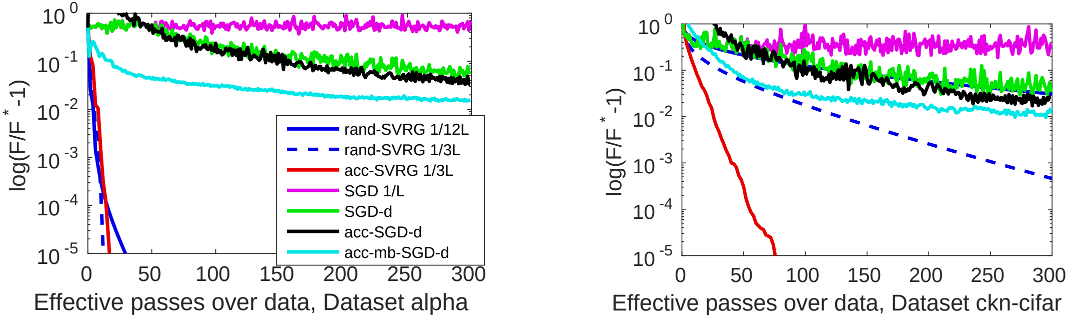
A few experiments



ℓ_2-logistic regression on two datasets, with $\mu = 1/10n$.

- no big difference between the variants of SGD with decreasing step sizes;
- variance reduction makes a huge difference.
- acceleration helps on ckn-cifar.

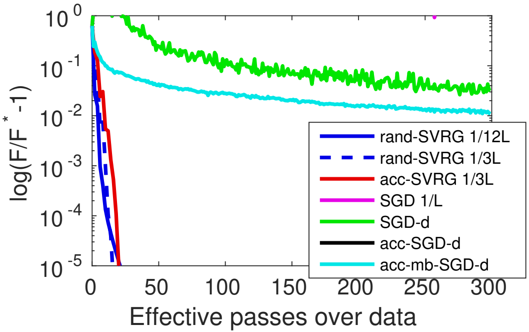
A few experiments

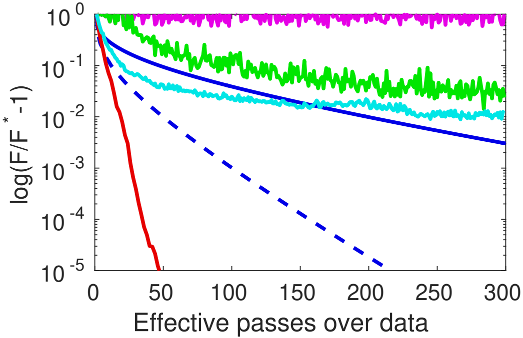


ℓ_2-logistic regression on two datasets, with $\mu = 1/100n$.

- as conditioning worsens, the benefits of acceleration are larger.
- accelerated SGD with mini-batches take the lead among SGD methods.

A few experiments





SVM with squared hinge loss on two datasets, with $\mu = 1/10n$.

- here, gradients are potentially unbounded and accelerated SGD diverges!
- accelerated SGD with mini-batches is stable and faster than SGD.

Remark about accelerated SGD

It does not always work. Why?

- the bounded noise variance assumption is not safe.
- the accelerated algorithm with constant step size (which is used to forget the initial condition) has much worth dependency in σ^2 (see next slide).

Remark about accelerated SGD

It does not always work. Why?

- the bounded noise variance assumption is not safe.
- the accelerated algorithm with constant step size (which is used to forget the initial condition) has much worth dependency in σ^2 (see next slide).

Convergence of SGD with $\eta_t = 1/L$

$$\mathbb{E}[f(\hat{x}_t) - f^\star] \leq 2\left(1 - \frac{\mu}{L}\right)^t (f(x_0) - f^\star) + \frac{\sigma^2}{L}.$$

Convergence of accelerated SGD with $\eta_t = 1/L$

$$\mathbb{E}[f(\hat{x}_t) - f^\star] \leq 2\left(1 - \sqrt{\frac{\mu}{L}}\right)^t (f(x_0) - f^\star) + \frac{\sigma^2}{\sqrt{\mu L}}.$$

Remark about accelerated SGD

It does not always work. Why?

- the bounded noise variance assumption is not safe.
- the accelerated algorithm with constant step size (which is used to forget the initial condition) has much worth dependency in σ^2 (see next slide).

Is it worthless?

- **removing the need for averaging** is great for sparse problems.
- with a **mini-batch** of size $\sqrt{L/\mu}$, we obtain the same complexity as the unaccelerated algorithm and the same stability w.r.t. σ^2, and we can parallelize for free!

References from this talk

The botany of incremental methods

- SAG [Schmidt et al., 2017].
- SAGA [Defazio et al., 2014a].
- SVRG [Xiao and Zhang, 2014].
- SDCA [Shalev-Shwartz and Zhang, 2014].
- Finito [Defazio et al., 2014b].
- MISO [Mairal, 2015].
- S2GD [Konečný and Richtárik, 2017].
- SARAH [Nguyen et al., 2017].
- MiG [Zhou et al., 2018].
- Katyusha [Allen-Zhu, 2017].
- Catalyst [Lin et al., 2018].
- . . .

Conclusion

- The estimate sequence method is a **generic tool**, which can be applied to stochastic optimization problems, including finite-sums.
- We use it to develop and analyze algorithms **without and with** acceleration.
- We discuss empirical findings regarding the **stability** of accelerated stochastic algorithms.
- . . . but stability issues can be fixed with mini-batching.

References I

Z. Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. In *Proceedings of Symposium on Theory of Computing (STOC)*, 2017.

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.

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

Olivier Devolder. Stochastic first order methods in smooth convex optimization. CORE Discussion Papers 2011070, UniversitÃ© catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2011.

Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, ii: shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 23(4):2061–2089, 2013.

References II

Jakub Konečný and Peter Richtárik. Semi-stochastic gradient descent methods. *Frontiers in Applied Mathematics and Statistics*, 3:9, 2017.

H. Lin, J. Mairal, and Z. Harchaoui. Catalyst acceleration for first-order convex optimization: from theory to practice. *Journal of Machine Learning Research (JMLR)*, 18(212):1–54, 2018.

Qihang Lin, Xi Chen, and Javier Peña. A sparsity preserving stochastic gradient methods for sparse regression. *Computational Optimization and Applications*, 58(2):455–482, 2014.

J. Mairal. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 25(2):829–855, 2015.

Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

S. Shalev-Shwartz and T. Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Mathematical Programming*, pages 1–41, 2014.

References III

L. Xiao and T. Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

Kaiwen Zhou, Fanhua Shang, and James Cheng. A simple stochastic variance reduced algorithm with fast convergence rates. *arXiv preprint arXiv:1806.11027*, 2018.

Variance reduction for finite sums (2/2)

SVRG (non-composite variante)

$$x_t = x_{t-1} - \gamma \left(\nabla f_{i_t}(x_{t-1}) - \nabla f_{i_t}(y) + \nabla f(y) \right),$$

where y is updated every epoch and $\mathbb{E}[\nabla f_{i_t}(y)|\mathcal{F}_{t-1}] = \nabla f(y)$.

SAGA

$$x_t = x_{t-1} - \gamma \left(\nabla f_{i_t}(x_{t-1}) - y_{i_t}^{t-1} + \frac{1}{n}\sum_{i=1}^n y_i^{t-1} \right),$$

where $\mathbb{E}[y_{i_t}^{t-1}|\mathcal{F}_{t-1}] = \frac{1}{n}\sum_{i=1}^n y_i^{t-1}$ and $y_i^t = \begin{cases} \nabla f_i(x_{t-1}) & \text{if } i = i_t \\ y_i^{t-1} & \text{otherwise.} \end{cases}$

MISO/Finito: for $n \geq L/\mu$, same form as SAGA but

$$\frac{1}{n}\sum_{i=1}^n y_i^{t-1} = -\mu x_{t-1} \quad \text{and} \quad y_i^t = \begin{cases} \nabla f_i(x_{t-1}) - \mu x_{t-1} & \text{if } i = i_t \\ y_i^{t-1} & \text{otherwise.} \end{cases}$$

The stochastic finite sum problem

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

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.

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

Data augmentation on digits (left); Dropout on text (right).