

Estimate Sequences for Variance-Reduced Stochastic Composite Optimization

Andrei Kulunchakov, Julien Mairal

Inria Grenoble

ICCOPT, Berlin





Andrei Kulunchakov

- A. Kulunchakov and J. Mairal. Estimate Sequences for Variance-Reduced Stochastic Composite Optimization. International Conference on Machine Learning (ICML). 2019.
- A. Kulunchakov and J. Mairal. Estimate Sequences for Stochastic Composite Optimization: Variance Reduction, Acceleration, and Robustness to Noise. preprint arXiv:1901.08788. 2019.

Context

Many subspace identification approaches require solving a **composite** optimization problem

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

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

Context

Many subspace identification approaches require solving a **composite** optimization problem

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

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

Two settings of interest

As seen in the previous talks, 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. . .

Complexity of SGD variants

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

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

We consider the worst-case complexity for finding a point \bar{x} such that $\mathbb{E}[F(\bar{x}) - F^*] \leq \varepsilon$ 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 et al., 2018].

Complexity for finite sums

We consider the worst-case complexity for finding a point \bar{x} such that $\mathbb{E}[F(\bar{x}) - F^*] \leq \varepsilon$ 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 et al., 2018].

Contributions without acceleration

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

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).

Contributions with acceleration

- we propose a **new 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.

A classical iteration

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

A classical iteration

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

covers SGD, SAGA, SVRG, and composite variants.

A classical iteration

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

covers SGD, SAGA, SVRG, and composite variants.

Interpretation

x_k minimizes the quadratic function d_k , defined as

$$d_k(x) = (1 - \delta_k)d_{k-1}(x) + \delta_k \left(f(x_{k-1}) + g_k^\top (x - x_{k-1}) + \frac{\mu}{2} \|x - x_{k-1}\|^2 \right. \\ \left. \dots + \psi(x_k) + \psi'(x_k)^\top (x - x_k) \right),$$

where $\delta_k = \mu\eta_k$, $\psi'(x_k)$ is a subgradient in $\partial\psi(x_k)$, and $d_0(x) = d_0^* + \frac{\mu}{2} \|x - x_0\|^2$.

A classical iteration

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

covers SGD, SAGA, SVRG, and composite variants.

Interpretation

x_k minimizes the quadratic function d_k , defined as

$$d_k(x) = (1 - \delta_k)d_{k-1}(x) + \delta_k \left(f(x_{k-1}) + g_k^\top (x - x_{k-1}) + \frac{\mu}{2} \|x - x_{k-1}\|^2 \right. \\ \left. \dots + \psi(x_k) + \psi'(x_k)^\top (x - x_k) \right),$$

where $\delta_k = \mu\eta_k$, $\psi'(x_k)$ is a subgradient in $\partial\psi(x_k)$, and $d_0(x) = d_0^* + \frac{\mu}{2} \|x - x_0\|^2$.

This is similar to the construction of **estimate sequences** by Nesterov.

see also [Devolder, 2011, Lin et al., 2014] for stochastic problems.

A less classical iteration

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

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

Interpretation

x_k minimizes the function d_k , defined as

$$d_k(x) = (1 - \delta_k)d_{k-1}(x) + \delta_k \left(f(x_{k-1}) + g_k^\top (x - x_{k-1}) + \frac{\mu}{2} \|x - x_{k-1}\|^2 + \psi(x) \right).$$

With estimate sequences, convergence proofs for both types of iterations are identical.

Convergence results

General convergence result

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

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

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 $\sigma_t^2 = \mathbb{E}[\|g_t - \nabla f(x_{t-1})\|^2]$.

Convergence results

General convergence result

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

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

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 $\eta_k = 1/L$

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

Convergence results

General convergence result

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

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

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 $\eta_k = 1/L$

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

Convergence results

General convergence result

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

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

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.

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

Interpretation

x_k minimizes the quadratic function d_k , defined as

$$\begin{aligned}d_k(x) &= (1 - \delta_k) d_{k-1}(x) + \delta_k \left(f(y_{k-1}) + g_k^\top (x - y_{k-1}) + \frac{\mu}{2} \|x - y_{k-1}\|^2 \right. \\&\quad \left. \dots + \psi(x_k) + \psi'(x_k)^\top (x - x_k) \right),\end{aligned}$$

where $\delta_k = \mu \eta_k$, $\psi'(x_k)$ is a subgradient in $\partial \psi(x_k)$, and $d_0(x) = d_0^* + \frac{\mu}{2} \|x - x_0\|^2$.

An accelerated SGD algorithm

An algorithm derived from the estimate sequence method.

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

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

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

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.

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

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

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

An accelerated SGD algorithm

An algorithm derived from the estimate sequence method.

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

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

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

- 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**).

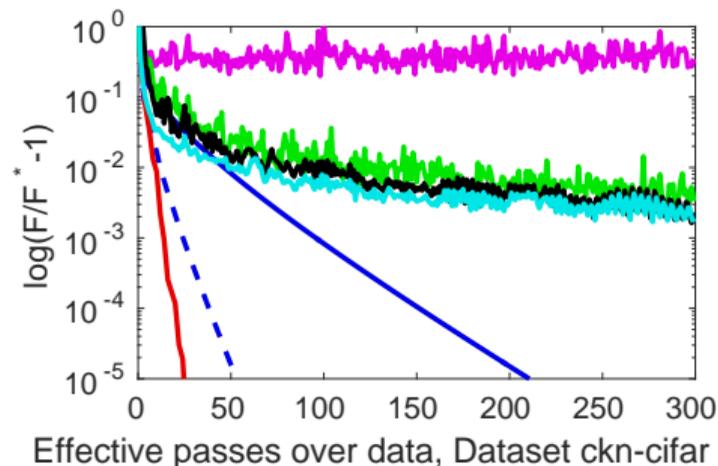
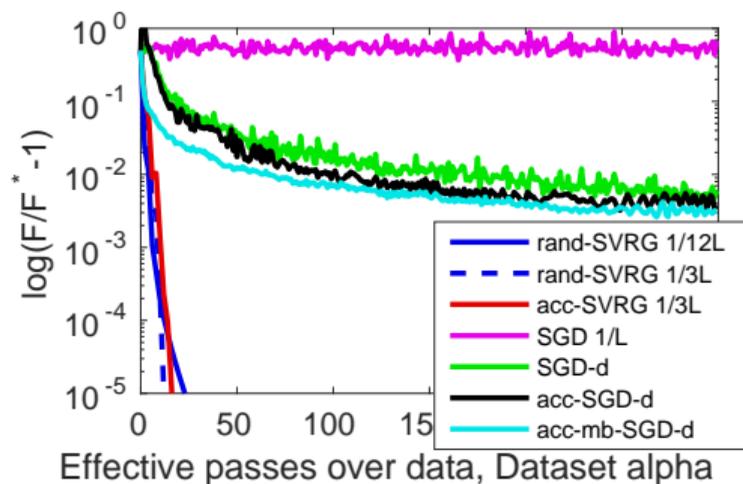
An accelerated SVRG algorithm for stochastic finite-sum problems

Remarks

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

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

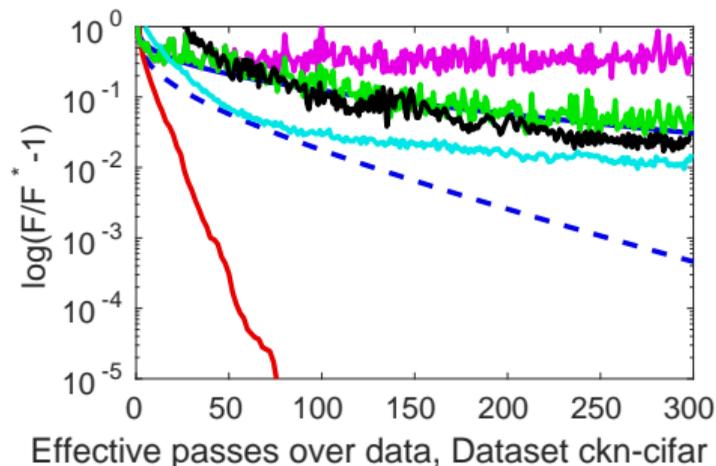
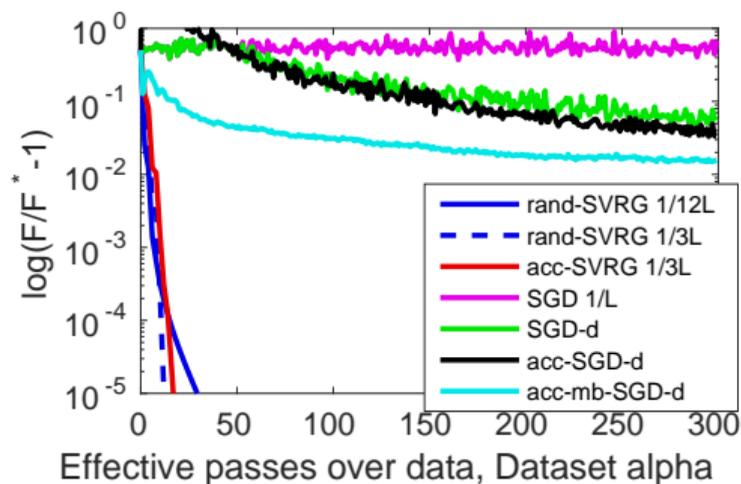
A few experiments



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.

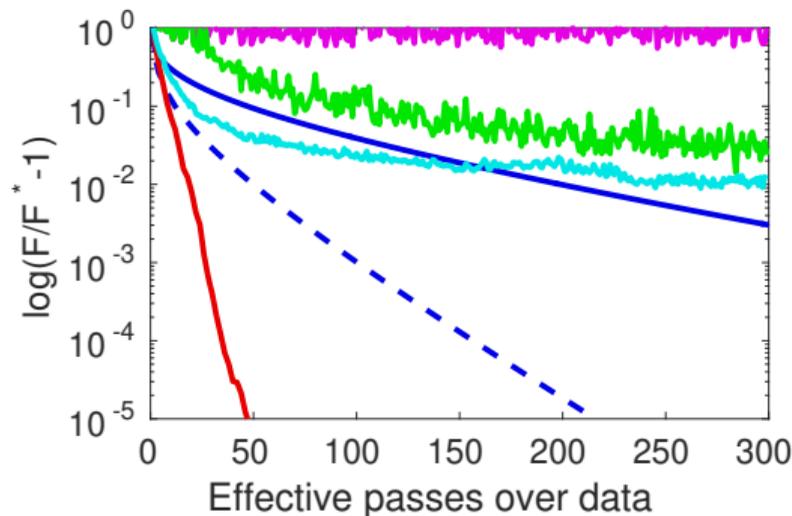
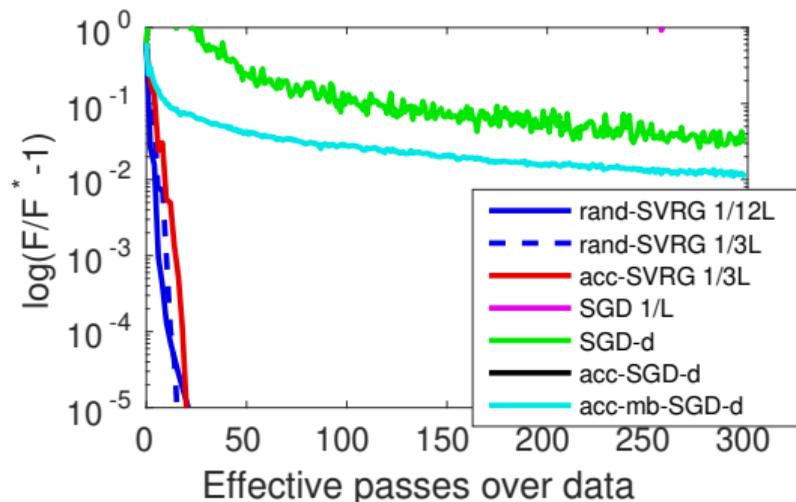
A few experiments



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 worse 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 worse dependency in σ^2 (see next slide).

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

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

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

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

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 worse 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, Universit  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.