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

Andrei Kulunchakov, Julien Mairal
Inria Grenoble

ML in the real world, Criteo
Publications

Andrei Kulunchakov

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 \( \psi \) 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 \( \psi \) 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...
Basics of gradient-based optimization

Smooth vs non-smooth

An important quantity to quantify smoothness is the Lipschitz constant of the gradient:

\[ \| \nabla f(x) - \nabla f(y) \| \leq L \| x - y \|. \]
Basics of gradient-based optimization

Smooth vs non-smooth

An important quantity to quantify smoothness is the \textbf{Lipschitz constant} of the gradient:

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|.$$

If \(f\) is twice differentiable, \(L\) may be chosen as the \textbf{largest eigenvalue} of the Hessian \(\nabla^2 f\). This is an upper-bound on the function curvature.
Basics of gradient-based optimization

Convex vs non-convex

(a) non-convex

(b) convex

(c) strongly-convex

An important quantity to quantify convexity is the **strong-convexity** constant

\[
f(x) \geq f(y) + \nabla f(y)^T (x - y) + \frac{\mu}{2} \|x - y\|^2,
\]
An important quantity to quantify convexity is the **strong-convexity** constant

\[
f(x) \geq f(y) + \nabla f(y)^\top (x - y) + \frac{\mu}{2} \|x - y\|^2,
\]

If \( f \) is twice differentiable, \( \mu \) may be chosen as the **smallest eigenvalue** of the Hessian \( \nabla^2 f \). This is a lower-bound on the function curvature.
Basics of gradient-based optimization

Picture from F. Bach

Why is the condition number $L/\mu$ important?

(small $\kappa = L/\mu$) (large $\kappa = L/\mu$)
Basics of gradient-based optimization

Picture from F. Bach

Trajectory of gradient descent with optimal step size.

(small $\kappa = L/\mu$) (large $\kappa = L/\mu$)
Variance reduction (1/2)

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.
Variance reduction (1/2)

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.
Variance reduction for smooth functions (2/2)

SVRG

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

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

SAGA

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

where \( \mathbb{E}[y_{i_t}^{t-1}|\mathcal{F}_{t-1}] = \frac{1}{n} \sum_{i=1}^{n} y_{i_t}^{t-1} \) 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

\[ \frac{1}{n} \sum_{i=1}^{n} y_{i_t}^{t-1} = -\mu x_{t-1} \quad \text{and} \quad y_i^t = \begin{cases} \nabla f_i(x_{t-1}) - \mu x_{t-1} & \text{if } i = i_t \\ y_i^{t-1} & \text{otherwise} \end{cases} \]
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 $\mu$-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 $\sigma^2$. 

Julien Mairal

Stochastic Composite Optimization
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 $\mu$-strongly convex case only.

Complexity of SGD with iterate averaging

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

under the (strong) assumption that the gradient estimates have bounded variance $\sigma^2$.

Complexity of accelerated SGD [Ghadimi and Lan, 2013]

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

$$O \left( \left( nL \frac{1}{\mu} \right) \log \left( \frac{C_0}{\varepsilon} \right) \right) \quad \text{vs.} \quad O \left( \left( n \sqrt{\frac{L}{\mu}} \right) \log \left( \frac{C_0}{\varepsilon} \right) \right).$$

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 $\mu$ (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

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

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

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

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

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

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.

Interpretation
\[ x_k \text{ minimizes the quadratic function } d_k, \text{ defined as} \]
\[ d_k(x) = (1 - \delta_k)d_{k-1}(x) + \delta_k (f(x_k - 1) + g_k^\top (x - x_k - 1) + \mu^2 \|x - x_k - 1\|^2 + \psi(x_k) + \psi'(x_k) \top (x - x_k)), \]

where \[ \delta_k = \mu \eta_k, \]
\[ \psi'(x_k) \text{ is a subgradient in } \partial \psi(x_k), \]
and \[ d_0(x) = d^\star_0 + \mu^2 \|x - x_0\|^2. \]

This is similar to the construction of estimate sequences by Nesterov. See also [Devolder, 2011, Lin et al., 2014] for stochastic problems.
A classical iteration

\[
x_k \leftarrow \text{Prox}_{\eta_k \psi} [x_{k-1} - \eta_k g_k]
\]
with

\[
\mathbb{E}[g_k | \mathcal{F}_k] = \nabla f(x_{k-1}),
\]

covers SGD, SAGA, SVRG, and composite variants.

**Interpretation**

\(x_k\) minimizes the quadratic function \(d_k\), defined as

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

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

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

covers SGD, SAGA, SVRG, and composite variants.

Interpretation

\( x_k \) minimizes the quadratic function \( d_k \), defined as

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

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 $k \geq 1$,

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

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 $k \geq 1$,

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

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

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.

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

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

Interpretation

\(x_k\) minimizes the quadratic function \(d_k\), defined as

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

\[
\left. \ldots + \psi(x_k) + \psi'(x_k)^\top (x - x_k) \right),
\]

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

An algorithm derived from the estimate sequence method.

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

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

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.

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

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

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.

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

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

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

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

\[
O \left( \left( n + \sqrt{\frac{nL}{\mu}} \right) \log \left( \frac{C_0}{\varepsilon} \right) \right) + O \left( \frac{\tilde{\sigma}^2}{\mu \varepsilon} \right) \quad \text{with} \quad \tilde{\sigma}^2 \ll \sigma^2.
\]
A few experiments

\[ \ell_2 \text{-logistic regression on two datasets, with } \mu = 1/10n. \]

- 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

\begin{align*}
\ell_2\text{-logistic regression on two datasets, with } \mu = 1/100n.
\end{align*}

\begin{itemize}
\item as conditioning worsens, the benefits of acceleration are larger.
\item accelerated SGD with mini-batches take the lead among SGD methods.
\end{itemize}
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 $\sigma^2$ (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 $\sigma^2$ (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 worth dependency in $\sigma^2$ (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. $\sigma^2$, 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 [Konečnỳ and Richtárik, 2017].
- SARAH [Nguyen et al., 2017].
- MiG [Zhou et al., 2018].
- 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


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

References II


