Variance reduction for stochastic gradient methods

Yuxin Chen
Princeton University, Fall 2019
Outline

- Stochastic variance reduced gradient (SVRG)
  - Convergence analysis for strongly convex problems
- Stochastic recursive gradient algorithm (SARAH)
  - Convergence analysis for nonconvex problems
- Other variance reduced stochastic methods
  - Stochastic dual coordinate ascent (SDCA)
Finite-sum optimization

\[
\text{minimize}_{x \in \mathbb{R}^d} \quad F(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x) + \psi(x)
\]

common task in machine learning

- linear regression: \( f_i(x) = \frac{1}{2} (a_i^T x - y_i)^2, \psi(x) = 0 \)
- logistic regression: \( f_i(x) = \log(1 + e^{-y_i a_i^T x}), \psi(x) = 0 \)
- Lasso: \( f_i(x) = \frac{1}{2} (a_i^T x - y_i)^2, \psi(x) = \lambda \|x\|_1 \)
- SVM: \( f_i(x) = \max\{0, 1 - y_i a_i^T x\}, \psi(x) = \frac{\lambda}{2} \|x\|_2^2 \)
- ...
Algorithm 12.1 Stochastic gradient descent (SGD)

1: for $t = 1, 2, \ldots$ do
2: pick $i_t \sim \text{Unif}(1, \ldots, n)$
3: $x^{t+1} = x^t - \eta_t \nabla f_{i_t}(x^t)$

As we have shown in the last lecture

- large stepsizes poorly suppress variability of stochastic gradients
  $\implies$ SGD with $\eta_t \asymp 1$ tends to oscillate around global mins
- choosing $\eta_t \asymp 1/t$ mitigates oscillation, but is too conservative
Recall: SGD theory with fixed stepsizes

\[ x^{t+1} = x^t - \eta_t g^t \]

- \( g^t \): an unbiased estimate of \( F(x^t) \)
- \( \mathbb{E}[\|g^t\|_2^2] \leq \sigma_g^2 + c_g \|\nabla F(x^t)\|_2^2 \)
- \( F(\cdot) \): \( \mu \)-strongly convex; \( L \)-smooth

From the last lecture, we know

\[
\mathbb{E}[F(x^t) - F(x^*)] \leq \frac{\eta L \sigma_g^2}{2\mu} + (1 - \eta\mu)^t (F(x^0) - F(x^*))
\]
Recall: SGD theory with fixed stepsizes

\[
\mathbb{E}[F(x^t) - F(x^*)] \leq \frac{\eta L \sigma_g^2}{2\mu} + (1 - \eta \mu)^t (F(x^0) - F(x^*))
\]

- vanilla SGD: \( g^t = \nabla f_{i_t}(x^t) \)
  - issue: \( \sigma_g^2 \) is non-negligible even when \( x^t = x^* \)
- question: it is possible to design \( g^t \) with reduced variability \( \sigma_g^2 \)?
A simple idea

Imagine we take some $\mathbf{v}^t$ with $\mathbb{E}[\mathbf{v}^t] = \mathbf{0}$ and set

$$
\mathbf{g}^t = \nabla f_{i_t}(\mathbf{x}^t) - \mathbf{v}^t
$$

— so $\mathbf{g}^t$ is still an unbiased estimate of $\nabla F(\mathbf{x}^t)$

**question:** how to reduce variability (i.e. $\mathbb{E}[\|\mathbf{g}^t\|_2^2] < \mathbb{E}[\|\nabla f_{i_t}(\mathbf{x}^t)\|_2^2]$)?

**answer:** find some zero-mean $\mathbf{v}^t$ that is positively correlated with $\nabla f_{i_t}(\mathbf{x}^t)$ (i.e. $\langle \mathbf{v}^t, \nabla f_{i_t}(\mathbf{x}^t) \rangle > 0$) (why?)
Reducing variance via gradient aggregation

If the current iterate is not too far away from previous iterates, then historical gradient info might be useful in producing such a $\nu^t$ to reduce variance.

**main idea of this lecture:** aggregate previous gradient info to help improve the convergence rate.
Stochastic variance reduced gradient (SVRG)
Strongly convex and smooth problems (no regularization)

\[ \text{minimize}_{x \in \mathbb{R}^d} \quad F(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) \]

- \( f_i \): convex and \( L \)-smooth
- \( F \): \( \mu \)-strongly convex
- \( \kappa := \frac{L}{\mu} \): condition number
Stochastic variance reduced gradient (SVRG)

— Johnson, Zhang ’13

**key idea:** if we have access to a history point $x^{\text{old}}$ and $\nabla F(x^{\text{old}})$, then

$$\nabla f_{i_t}(x^t) - \nabla f_{i_t}(x^{\text{old}}) + \nabla F(x^{\text{old}}) \quad \text{with } i_t \sim \text{Unif}(1, \cdots, n)$$

$\rightarrow 0$ if $x^t \approx x^{\text{old}}$  
$\rightarrow 0$ if $x^{\text{old}} \approx x^*$

- is an unbiased estimate of $\nabla F(x^t)$
- converges to 0 if $x^t \approx x^{\text{old}} \approx x^*$

variability is reduced!
Stochastic variance reduced gradient (SVRG)

- operate in epochs
- in the $s^{th}$ epoch
  - **very beginning**: take a snapshot $x_{s}^{\text{old}}$ of the current iterate, and compute the **batch gradient** $\nabla F(x_{s}^{\text{old}})$
  - **inner loop**: use the snapshot point to help reduce variance
    \[
    x_{s}^{t+1} = x_{s}^{t} - \eta \left\{ \nabla f_{i_{t}}(x_{s}^{t}) - \nabla f_{i_{t}}(x_{s}^{\text{old}}) + \nabla F(x_{s}^{\text{old}}) \right\}
    \]

**a hybrid approach**: the batch gradient is computed only once per epoch
SVRG algorithm (Johnson, Zhang ’13)

Algorithm 12.2 SVRG for finite-sum optimization

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:</td>
<td>for $s = 1, 2, \ldots$ do</td>
</tr>
<tr>
<td>2:</td>
<td>$\mathbf{x}<em>s^{\text{old}} \leftarrow \mathbf{x}</em>{s-1}^m$, and compute $\nabla F(\mathbf{x}_s^{\text{old}})$ // update snapshot</td>
</tr>
<tr>
<td>3:</td>
<td>initialize $\mathbf{x}_0^0 \leftarrow \mathbf{x}_s^{\text{old}}$</td>
</tr>
<tr>
<td>4:</td>
<td>for $t = 0, \ldots, m - 1$ do</td>
</tr>
<tr>
<td>5:</td>
<td>each epoch contains $m$ iterations</td>
</tr>
<tr>
<td></td>
<td>choose $i_t$ uniformly from ${1, \ldots, n}$, and</td>
</tr>
<tr>
<td></td>
<td>$\mathbf{x}_s^{t+1} = \mathbf{x}<em>s^t - \eta { \nabla f</em>{i_t}(\mathbf{x}<em>s^t) - \nabla f</em>{i_t}(\mathbf{x}_s^{\text{old}}) + \nabla F(\mathbf{x}_s^{\text{old}}) }$</td>
</tr>
</tbody>
</table>

Variance reduction
Remark

- constant stepsize $\eta$
- each epoch contains $2m + n$ gradient computations
  - the batch gradient is computed only once every $m$ iterations
  - the average per-iteration cost of SVRG is comparable to that of SGD if $m \gtrsim n$
Convergence analysis of SVRG

Theorem 12.1

Assume each $f_i$ is convex and $L$-smooth, and $F$ is $\mu$-strongly convex. Choose $m$ large enough s.t. $\rho = \frac{1}{\mu \eta (1 - 2L\eta) m} + \frac{2L\eta}{1 - 2L\eta} < 1$, then

$$\mathbb{E}[F(x_s^{\text{old}}) - F(x^*)] \leq \rho^s [F(x_0^{\text{old}}) - F(x^*)]$$

- **Linear convergence:** choosing $m \gtrsim L/\mu = \kappa$ and constant stepsizes $\eta \asymp 1/L$ yields $0 < \rho < 1/2$

  $$\Rightarrow O(\log \frac{1}{\varepsilon})$$ epochs to attain $\varepsilon$ accuracy
Convergence analysis of SVRG

Theorem 12.1

Assume each \( f_i \) is convex and \( L \)-smooth, and \( F \) is \( \mu \)-strongly convex. Choose \( m \) large enough s.t. \( \rho = \frac{1}{\mu \eta (1 - 2L \eta) m} + \frac{2L \eta}{1 - 2L \eta} < 1 \), then

\[
\mathbb{E}[F(x_{s \text{old}}) - F(x^*)] \leq \rho^s [F(x_0^{\text{old}}) - F(x^*)]
\]

- total computational cost:

\[
(m + n) \log \frac{1}{\varepsilon} \quad \asymp \quad (n + \kappa) \log \frac{1}{\varepsilon}
\]

# grad computation per epoch \quad if \( m \asymp \max\{n, \kappa\} \)
Proof of Theorem 12.1

Here, we provide the proof for an alternative version, where in each epoch,

\[ x_{s+1}^{\text{old}} = x_s^j \quad \text{with } j \sim \text{Unif}(0, \cdots, m - 1) \]

rather than \( j = m \) (12.1)

The interested reader is referred to Tan et al. ’16 for the proof of the original version.
Proof of Theorem 12.1

Let $g_t^s := \nabla f_{it}(x_t^s) - \nabla f_{it}(x_s^{\text{old}}) + \nabla F(x_s^{\text{old}})$ for simplicity. As usual, conditional on everything prior to $x_s^{t+1}$, one has

$$
\mathbb{E}[\|x_s^{t+1} - x^*\|^2_2] = \mathbb{E}[\|x_s^t - \eta g_s^t - x^*\|^2_2]
= \|x_s^t - x^*\|^2_2 - 2\eta(x_s^t - x^*)^\top \mathbb{E}[g_s^t] + \eta^2 \mathbb{E}[\|g_s^t\|^2_2]
\leq \|x_s^t - x^*\|^2_2 - 2\eta(x_s^t - x^*)^\top \nabla F(x_s^t) + \eta^2 \mathbb{E}[\|g_s^t\|^2_2]
$$

since $g_s^t$ is an unbiased estimate of $\nabla F(x_s^t)$

$$
\leq \|x_s^t - x^*\|^2_2 - 2\eta(F(x_s^t) - F(x^*)) + \eta^2 \mathbb{E}[\|g_s^t\|^2_2]
\quad \text{by convexity}
$$

- **key step:** control $\mathbb{E}[\|g_s^t\|^2_2]$
  — we’d like to upper bound it via the (relative) objective value

Variance reduction
Proof of Theorem 12.1

**main pillar:** control \( \mathbb{E}[\|g_s^t\|_2^2] \) via . . .

**Lemma 12.2**

\[
\mathbb{E}[\|g_s^t\|_2^2] \leq 4L \left[ F(x_s^t) - F(x^*) + F(x_{s}^{old}) - F(x^*) \right]
\]

This means if \( x_s^t \approx x_s^{old} \approx x^* \), then \( \mathbb{E}[\|g_s^t\|_2^2] \approx 0 \) (reduced variance)
Proof of Theorem 12.1

**main pillar:** control $\mathbb{E}[\|g_s^t\|^2_2]$ via . . .

**Lemma 12.2**

$$\mathbb{E}[\|g_s^t\|^2_2] \leq 4L [F(x_s^t) - F(x^*) + F(x_{s_{old}}) - F(x^*)]$$

this allows one to obtain: conditional on everything prior to $x_{s+1}^t$,

$$\mathbb{E}[\|x_{s+1}^t - x^*\|^2_2] \leq (12.2)$$

$$\leq \|x_s^t - x^*\|^2_2 - 2\eta [F(x_s^t) - F(x^*)]$$

$$+ 4L\eta^2 [F(x_s^t) - F(x^*) + F(x_{s_{old}}) - F(x^*)]$$

$$= \|x_s^t - x^*\|^2_2 - 2\eta (1 - 2L\eta) [F(x_s^t) - F(x^*)]$$

$$+ 4L\eta^2 [F(x_{s_{old}}) - F(x^*)] \quad (12.3)$$
Proof of Theorem 12.1 (cont.)

Taking expectation w.r.t. all history, we have

\[ 2\eta(1 - 2L\eta)m \mathbb{E}[F(x_{s+1}^{old}) - F(x^*)] \]

\[ = 2\eta(1 - 2L\eta) \sum_{t=0}^{m-1} \mathbb{E}[F(x_s^t) - F(x^*)] \]

\[ \leq \mathbb{E}[\|x_{s+1}^m - x^*\|^2_2] + 2\eta(1 - 2L\eta) \sum_{t=0}^{m-1} \mathbb{E}[F(x_s^t) - F(x^*)] \]

\[ \geq 0 \]

\[ \leq \mathbb{E}[\|x_{s+1}^0 - x^*\|^2_2] + 4Lm\eta^2[F(x_{s}^{old}) - F(x^*)] \text{ (apply (12.3) recursively)} \]

\[ = \mathbb{E}[\|x_{s}^{old} - x^*\|^2_2] + 4Lm\eta^2\mathbb{E}[F(x_{s}^{old}) - F(x^*)] \]

\[ \leq \frac{2}{\mu} \mathbb{E}[F(x_{s}^{old}) - F(x^*)] + 4Lm\eta^2\mathbb{E}[F(x_{s}^{old}) - F(x^*)] \text{ (strong convexity)} \]

\[ = \left(\frac{2}{\mu} + 4Lm\eta^2\right) \mathbb{E}[F(x_{s}^{old}) - F(x^*)] \]
Proof of Theorem 12.1 (cont.)

Consequently,

\[
\mathbb{E}[F(x_{s+1}^{\text{old}}) - F(x^*)] \\
\leq \frac{2}{\mu} + 4Lm\eta^2 \mathbb{E}[F(x_s^{\text{old}}) - F(x^*)] \\
= \left( \frac{1}{\mu \eta (1 - 2L\eta)m} + \frac{2L\eta}{1 - 2L\eta} \right) \mathbb{E}[F(x_s^{\text{old}}) - F(x^*)]
\]

\[
= \rho
\]

Applying this bound recursively establishes the theorem.
Proof of Lemma 12.2

\[
\mathbb{E}[\|\nabla f_{i_t}(x^t_s) - \nabla f_{i_t}(x_{s}^{\text{old}}) + \nabla F(x_{s}^{\text{old}})\|^2_2] \\
= \mathbb{E}[\|\nabla f_{i_t}(x^t_s) - \nabla f_{i_t}(x^*) - (\nabla f_{i_t}(x_{s}^{\text{old}}) - \nabla f_{i_t}(x^*) - \nabla F(x_{s}^{\text{old}}))\|^2_2] \\
\leq 2\mathbb{E}[\|\nabla f_{i_t}(x^t_s) - \nabla f_{i_t}(x^*)\|^2_2] + 2\mathbb{E}[\|\nabla f_{i_t}(x_{s}^{\text{old}}) - \nabla f_{i_t}(x^*) - \nabla F(x_{s}^{\text{old}})\|^2_2] \\
= 2\mathbb{E}[\|\nabla f_{i_t}(x^t_s) - \nabla f_{i_t}(x^*)\|^2_2] \\
\quad + 2\mathbb{E}[\|\nabla f_{i_t}(x_{s}^{\text{old}}) - \nabla f_{i_t}(x^*) - \mathbb{E}[\nabla f_{i_t}(x_{s}^{\text{old}}) - \nabla f_{i_t}(x^*)]\|^2_2] \\
\quad \text{since } \mathbb{E}[\nabla f_{i_t}(x^*)] = \nabla F(x^*) = 0 \\
\leq 2\mathbb{E}[\|\nabla f_{i_t}(x^t_s) - \nabla f_{i_t}(x^*)\|^2_2] + 2\mathbb{E}[\|\nabla f_{i_t}(x_{s}^{\text{old}}) - \nabla f_{i_t}(x^*)\|^2_2] \\
\leq 4L[F(x^t_s) - F(x^*) + F(x_{s}^{\text{old}}) - F(x^*)] \\
\text{where the last inequality would hold if we could justify} \\
\frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(x) - \nabla f_i(x^*)\|^2_2 \leq 2L[F(x) - F(x^*)] \quad (12.4) \\
\text{relies on both smoothness and convexity of } f_i
Proof of Lemma 12.2 (cont.)

To establish (12.4), observe from smoothness and convexity of $f_i$ that

$$\frac{1}{2L} \left\| \nabla f_i(x) - \nabla f_i(x^*) \right\|_2^2 \leq f_i(x) - f_i(x^*) - \nabla f_i(x^*)^\top (x - x^*)$$

an equivalent characterization of $L$-smoothness

Summing over all $i$ and recognizing that $\nabla F(x^*) = 0$ yield

$$\frac{1}{2L} \sum_{i=1}^n \left\| \nabla f_i(x) - \nabla f_i(x^*) \right\|_2^2 \leq nF(x) - nF(x^*) - n(\nabla F(x^*))^\top (x - x^*)$$

$$= nF(x) - nF(x^*)$$

as claimed
Numerical example: logistic regression

— Johnson, Zhang ’13

$\ell_2$-regularized logistic regression on CIFAR-10
## Comparisons with GD and SGD

<table>
<thead>
<tr>
<th>comp. cost</th>
<th>SVRG</th>
<th>GD</th>
<th>SGD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$(n + \kappa) \log \frac{1}{\varepsilon}$</td>
<td>$n\kappa \log \frac{1}{\varepsilon}$</td>
<td>$\frac{\kappa^2}{\varepsilon}$ (practically often $\frac{\kappa}{\varepsilon}$)</td>
</tr>
</tbody>
</table>
Proximal extension

\[ \text{minimize}_{\mathbf{x} \in \mathbb{R}^d} \quad F(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}) + \psi(\mathbf{x}) \]

- \( f_i \): convex and \( L \)-smooth
- \( F \): \( \mu \)-strongly convex
- \( \kappa := L/\mu \): condition number
- \( \psi \): potentially non-smooth
Algorithm 12.3 Prox-SVRG for finite-sum optimization

1: for $s = 1, 2, \ldots$ do
2: $x_{s}^{old} \leftarrow x_{s-1}^{m}$, and compute $\nabla F(x_{s}^{old})$ // update snapshot
3: initialize $x_{s}^{0} \leftarrow x_{s}^{old}$
4: for $t = 0, \ldots, m - 1$ do
   each epoch contains $m$ iterations
5: choose $i_t$ uniformly from $\{1, \ldots, n\}$, and

\[
x_{s}^{t+1} = \text{prox}_{\eta \psi}(x_{s}^{t} - \eta \{ \nabla f_{i_t}(x_{s}^{t}) - \nabla f_{i_t}(x_{s}^{old}) + \nabla F(x_{s}^{old}) \})
\]

stochastic gradient
Stochastic recursive gradient algorithm (SARAH)
Nonconvex and smooth problems

\[
\text{minimize}_{x \in \mathbb{R}^d} \quad F(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x)
\]

- \( f_i \): \( L \)-smooth, potentially nonconvex
Recursive stochastic gradient estimates

— Nguyen, Liu, Scheinberg, Takac ’17

**key idea:** recursive / adaptive updates of gradient estimates

\[ g^t = \nabla f_{i,t}(x^t) - \nabla f_{i,t}(x^{t-1}) + g^{t-1} \]  
\[ x^{t+1} = x^t - \eta g^t \] (12.5)

**comparison to SVRG** (use a fixed snapshot point for the entire epoch)

(SVRG) \[ g^t = \nabla f_{i,t}(x^t) - \nabla f_{i,t}(x^{\text{old}}) + \nabla F(x^{\text{old}}) \]
Restarting gradient estimate every epoch

For many (e.g. strongly convex) problems, recursive gradient estimate $g^t$ may decay fast (variance ↓; bias (relative to $\nabla F(x^t)$) ↑)

- $g^t$ may quickly deviate from the target gradient $\nabla F(x^t)$
- progress stalls as $g^t$ cannot guarantee sufficient descent

**solution:** reset $g^t$ every few iterations to calibrate with the true batch gradient
Bias of gradient estimates

Unlike SVRG, $g^t$ is NOT an unbiased estimate of $\nabla F(x^t)$

$$\mathbb{E}[g^t \mid \text{everything prior to } x^s] = \nabla F(x^t) - \nabla F(x^{t-1}) + g^{t-1} \neq 0$$

But if we average out all randomness, we have (exercise!)

$$\mathbb{E}[g^t] = \mathbb{E}[\nabla F(x^t)]$$
Algorithm 12.4 SARAH (Nguyen et al. ’17)

1: for $s = 1, 2, \ldots, S$ do
2: $x_0^s \leftarrow x_{s-1}^{m+1}$, and compute $g_0^s = \nabla F(x_0^s)$ // restart $g$ anew

3: $x_1^s = x_0^s - \eta g_0^s$
4: for $t = 1, \ldots, m$ do
5: choose $i_t$ uniformly from $\{1, \ldots, n\}$
6: $g_t^s = \nabla f_{i_t}(x_t^s) - \nabla f_{i_t}(x_{t-1}^s) + g_{t-1}^s$

Variance reduction
Convergence analysis of SARAH (nonconvex)

Theorem 12.3 (Nguyen et al. ’19)

Suppose each $f_i$ is $L$-smooth. Then SARAH with $\eta \lesssim \frac{1}{L \sqrt{m}}$ obeys

$$
\frac{1}{(m+1)S} \sum_{s=1}^{S} \sum_{t=0}^{m} \mathbb{E} \left[ \| \nabla F(x^t_s) \|_2^2 \right] \leq \frac{2}{\eta(m+1)S} \left[ F(x^0_0) - F(x^*) \right]
$$

• iteration complexity for finding $\varepsilon$-approximate stationary point (i.e. $\| \nabla F(x) \|_2 \leq \varepsilon$):

$$
O \left( n + \frac{L \sqrt{n}}{\varepsilon^2} \right) \quad (\text{setting } m \asymp n, \eta \asymp \frac{1}{L \sqrt{m}})
$$
Convergence analysis of SARAH (nonconvex)

**Theorem 12.3 (Nguyen et al. ’19)**

Suppose each $f_i$ is $L$-smooth. Then SARAH with $\eta \lesssim \frac{1}{L \sqrt{m}}$ obeys

$$
\frac{1}{(m + 1)S} \sum_{s=1}^{S} \sum_{t=0}^{m} \mathbb{E} \left[ \left\| \nabla F(x_t^s) \right\|_2^2 \right] \leq \frac{2}{\eta(m + 1)S} \left[ F(x_0^0) - F(x^*) \right]
$$

- also derived by Fang et al. ’18 (for a SARAH-like algorithm “Spider”) and improved by Wang et al. ’19 (for “SpiderBoost”)
Proof of Theorem 12.3

Theorem 12.3 follows immediately from the following claim on the total objective improvement in one epoch (why?)

$$\mathbb{E}[F(x_{m+1})] \leq \mathbb{E}[F(x_0)] - \frac{\eta}{2} \sum_{t=0}^{m} \mathbb{E}[\|\nabla F(x_t)\|_2^2]$$ \hspace{1cm} (12.6)

We will then focus on establishing (12.6)
Proof of Theorem 12.3 (cont.)

To establish (12.6), recall that the smoothness assumption gives

$$\mathbb{E}[F(x_{s+1}^t)] \leq \mathbb{E}[F(x_s^t)] - \eta \mathbb{E}[\nabla F(x_{s}^t)^\top g_s] + \frac{L\eta^2}{2} \mathbb{E}[\|g_s\|^2]$$  \hspace{1cm} (12.7)

Since $g_s^t$ is not an unbiased estimate of $\nabla F(x_s^t)$, we first decouple

$$2\mathbb{E}[\nabla F(x_s^t)^\top g_s] = \mathbb{E}[\|\nabla F(x_s^t)\|^2] + \mathbb{E}[\|g_s\|^2] - \mathbb{E}[\|\nabla F(x_s^t) - g_s\|^2]$$

desired gradient estimate variance squared bias of gradient estimate

Substitution into (12.7) with straightforward algebra gives

$$\mathbb{E}[F(x_{s+1}^t)] \leq \mathbb{E}[F(x_s^t)] - \frac{\eta}{2} \mathbb{E}[\|\nabla F(x_s^t)\|^2] + \frac{\eta}{2} \mathbb{E}[\|\nabla F(x_s^t) - g_s\|^2]$$

$$- \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right) \mathbb{E}[\|g_s\|^2]$$
Proof of Theorem 12.3 (cont.)

Sum over $t = 0, \ldots, m$ to arrive at

$$\mathbb{E}[F(x_s^{m+1})] \leq \mathbb{E}[F(x_s^0)] - \frac{\eta}{2} \sum_{t=0}^{m} \mathbb{E}[\|\nabla F(x_t^s)\|_2^2]$$

$$+ \frac{\eta}{2} \left\{ \sum_{t=0}^{m} \mathbb{E}[\|\nabla F(x_t^s) - g_t^s\|_2^2] - (1 - L\eta) \sum_{t=0}^{m} \mathbb{E}[\|g_t^s\|_2^2] \right\} \geq 1/2$$

The proof of (12.6) is thus complete if we can justify

**Lemma 12.4**

If $\eta \leq \frac{1}{L\sqrt{m}}$, then (for fixed $\eta$, the epoch length $m$ cannot be too large)

$$\sum_{t=0}^{m} \mathbb{E}[\|\nabla F(x_t^s) - g_t^s\|_2^2] \leq \frac{1}{2} \sum_{t=0}^{m} \mathbb{E}[\|g_t^s\|_2^2]$$

- informally, this says the accumulated squared bias of gradient estimates (w.r.t. batch gradients) can be controlled by the accumulated variance
Proof of Lemma 12.4

Key step:

**Lemma 12.5**

\[
\mathbb{E}
\left[
\left\| \nabla F(x_s^t) - g_s^t \right\|_2^2
\right]\leq \sum_{k=1}^{t}
\mathbb{E}
\left[
\left\| g_s^k - g_s^{k-1} \right\|_2^2
\right]
\]

- convert the bias of gradient estimates to the differences of consecutive gradient estimates (a consequence of the smoothness and the recursive formula of \( g_s^t \))
Proof of Lemma 12.4 (cont.)

From Lemma 12.5, it suffices to connect \( \{\|g_t^s - g_t^{s-1}\|_2\} \) with \( \{\|g_t^s\|_2\} \):

\[
\|g_t^s - g_t^{s-1}\|_2^2 = \|\nabla f_i^t(x_t^s) - \nabla f_i^t(x_t^{s-1})\|_2^2 \leq L^2 \|x_t^s - x_t^{s-1}\|_2^2
\]

\[
= \eta^2 L^2 \|g_t^{s-1}\|_2^2
\]

Invoking Lemma 12.5 then gives

\[
\mathbb{E}\left[\|\nabla F(x_t^s) - g_t^s\|_2^2\right] \leq \sum_{k=1}^{t} \mathbb{E}\left[\|g_k^s - g_k^{s-1}\|_2^2\right] \leq \eta^2 L^2 \sum_{k=1}^{t} \mathbb{E}\left[\|g_k^{s-1}\|_2^2\right]
\]

Summing over \( t = 0, \cdots, m \), we obtain

\[
\sum_{t=0}^{m} \mathbb{E}\left[\|\nabla F(x_t^s) - g_t^s\|_2^2\] \leq \eta^2 L^2 m \sum_{t=0}^{m-1} \mathbb{E}\left[\|g_t^s\|_2^2\]
\]

which establishes Lemma 12.4 if \( \eta \lesssim \frac{1}{L \sqrt{m}} \)
Proof of Lemma 12.5

Since this lemma only concerns a single epoch, we shall drop the dependency on $s$ for simplicity. Let $\mathcal{F}_k$ contain all info up to $x^k$ and $g^{k-1}$, then

$$
\mathbb{E} \left[ \left\| \nabla F(x^k) - g^k \right\|_2^2 \mid \mathcal{F}_k \right]
= \mathbb{E} \left[ \left\| \nabla F(x^{k-1}) - g^{k-1} + \left( \nabla F(x^k) - \nabla F(x^{k-1}) \right) - \left( g^k - g^{k-1} \right) \right\|_2^2 \mid \mathcal{F}_k \right]
= \left\| \nabla F(x^{k-1}) - g^{k-1} \right\|_2^2 + \left\| \nabla F(x^k) - \nabla F(x^{k-1}) \right\|_2^2 + \mathbb{E} \left[ \left\| g^k - g^{k-1} \right\|_2^2 \mid \mathcal{F}_k \right]
+ 2 \left\langle \nabla F(x^{k-1}) - g^{k-1}, \nabla F(x^k) - \nabla F(x^{k-1}) \right\rangle
- 2 \left\langle \nabla F(x^{k-1}) - g^{k-1}, \mathbb{E} \left[ g^k - g^{k-1} \mid \mathcal{F}_k \right] \right\rangle
- 2 \left\langle \nabla F(x^k) - \nabla F(x^{k-1}), \mathbb{E} \left[ g^k - g^{k-1} \mid \mathcal{F}_k \right] \right\rangle
= \left\| \nabla F(x^{k-1}) - g^{k-1} \right\|_2^2 - \left\| \nabla F(x^k) - \nabla F(x^{k-1}) \right\|_2^2 + \mathbb{E} \left[ \left\| g^k - g^{k-1} \right\|_2^2 \mid \mathcal{F}_k \right]
$$

(exercise)

Since $\nabla F(x^0) = g^0$. Sum over $k = 1, \ldots, t$ to obtain

$$
\mathbb{E} \left[ \left\| \nabla F(x^k) - g^k \right\|_2^2 \right] = \sum_{k=1}^{t} \mathbb{E} \left[ \left\| g^k - g^{k-1} \right\|_2^2 \right] - \sum_{k=1}^{t} \left\| \nabla F(x^k) - \nabla F(x^{k-1}) \right\|_2^2 \leq 0; \text{ done!}
$$
Stochastic dual coordinate ascent (SDCA) — a dual perspective
A class of finite-sum optimization

\begin{equation}
\text{minimize}_{x \in \mathbb{R}^d} \quad F(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) + \frac{\mu}{2} \|x\|^2_2
\end{equation}

• $f_i$: convex and $L$-smooth
The dual problem of (12.8)

\[
\maximize_{\nu} \quad D(\nu) = \frac{1}{n} \sum_{i=1}^{n} -f_i^*(-\nu_i) - \frac{\mu}{2} \left\| \frac{1}{\mu n} \sum_{i=1}^{n} \nu_i \right\|_2^2
\]  (12.9)

• a primal-dual relation

\[
x(\nu) = \frac{1}{\mu n} \sum_{i=1}^{n} \nu_i
\]  (12.10)
Derivation of the dual formulation

\[
\min_x \frac{1}{n} \sum_{i=1}^{n} f_i(x) + \frac{\mu}{2} \|x\|_2^2
\]

\[\iff\]
\[
\min_{x,\{z_i\}} \frac{1}{n} \sum_{i=1}^{n} f_i(z_i) + \frac{\mu}{2} \|x\|_2^2 \quad \text{s.t. } z_i = x
\]

\[\iff\]
\[
\max_{\{\nu_i\}} \min_{x,\{z_i\}} \frac{1}{n} \sum_{i=1}^{n} f_i(z_i) + \frac{\mu}{2} \|x\|_2^2 + \frac{1}{n} \sum_{i=1}^{n} \langle \nu_i, z_i - x \rangle
\]

Lagrangian

\[\iff\]
\[
\max_{\{\nu_i\}} \min_x \frac{1}{n} \sum_{i=1}^{n} -f_i^*(-\nu_i) + \frac{\mu}{2} \|x\|_2^2 - \frac{1}{n} \sum_{i=1}^{n} \langle \nu_i, x \rangle
\]

conjugate: \(f_i^*(\nu) := \max_z \{\langle \nu, z \rangle - f_i(z)\}\)

\[\iff\]
\[
\max_{\{\nu_i\}} \frac{1}{n} \sum_{i=1}^{n} -f_i^*(-\nu_i) - \frac{\mu}{2} \left\| \frac{1}{\mu n} \sum_{i=1}^{n} \nu_i \right\|_2^2
\]

optimal \(x = \frac{1}{\mu n} \sum_i \nu_i\)

Variance reduction 12-43
Randomized coordinate ascent on dual problem

— Shalev-Shwartz, Zhang ’13

• randomized coordinate ascent: at each iteration, randomly pick one dual (block) coordinate \( \nu_{i_t} \) of (12.9) to optimize

• maintain the primal-dual relation (12.10)

\[
x^t = \frac{1}{\mu n} \sum_{i=1}^{n} \nu^t_i
\]  

(12.11)
Stochastic dual coordinate ascent (SDCA)

Algorithm 12.5 SDCA for finite-sum optimization

1: initialize $\mathbf{x}^0 = \frac{1}{\mu n} \sum_{i=1}^{n} \nu_i^0$
2: for $t = 0, 1, \ldots$ do
3: // choose a random coordinate to optimize
4: choose $i_t$ uniformly from $\{1, \ldots, n\}$
5: $\Delta^t \leftarrow \arg \max_{\Delta} -\frac{1}{n} f_{i_t}^* \left( -\nu_{i_t}^t - \Delta \right) - \frac{\mu}{2} \| \mathbf{x}^t + \frac{1}{\mu n} \Delta \|^2$
   \[\text{find the optimal step with all } \{\nu_i^t\}_{i \neq i_t} \text{ fixed}\]
6: $\nu_{i_t}^{t+1} \leftarrow \nu_{i_t}^t + \Delta^t \mathbb{1}\{i = i_t\}$ \hspace{1cm} (1 ≤ $i$ ≤ $n$)
   \[\text{update only the } i_t^{\text{th}} \text{ coordinate}\]
7: $\mathbf{x}^{t+1} \leftarrow \mathbf{x}^t + \frac{1}{\mu n} \Delta^t$ \hspace{1cm} // based on (12.11)

Variance reduction
A variant of SDCA without duality

SDCA might not be applicable if the conjugate functions are difficult to evaluate

This calls for a dual-free version of SDCA
A variant of SDCA without duality

— S. Shalev-Shwartz ’16

Algorithm 12.6 SDCA without duality

1: initialize $x^0 = \frac{1}{\mu n} \sum_{i=1}^{n} \nu^0_i$
2: for $t = 0, 1, \ldots$ do
3: // choose a random coordinate to optimize
4: choose $i_t$ uniformly from $\{1, \ldots, n\}$
5: $\Delta^t \leftarrow -\eta \mu n (\nabla f_{i_t}(x^t) + \nu^t_{i_t})$
6: $\nu^t_{i+1} \leftarrow \nu^t_i + \Delta^t 1\{i = i_t\}$ (1 ≤ $i$ ≤ $n$) update only the $i^{th}$ coordinate
7: $x^{t+1} \leftarrow x^t + \frac{1}{\mu n} \Delta^t$ // based on (12.11)
A variant of SDCA without duality

A little intuition

- the optimality condition requires (check!)

\[ \nu_i^* = -\nabla f_i(x^*), \quad \forall i \]  \hspace{1cm} (12.12)

- with a modified update rule, one has

\[ \nu_{i_t}^{t+1} \leftarrow (1 - \eta \mu n) \nu_{i_t}^t + \eta \mu n(-\nabla f_{i_t}(x_t)) \]


\text{cvx combination of current dual iterate and gradient component}

— when it converges, it will satisfy (12.12)
The SDCA (without duality) update rule reads:

\[ x^{t+1} = x^t - \eta \left( \nabla f_i(x^t) + \nu_{it} \right) \]

:= \boldsymbol{g}^t

It is straightforward to verify that \( \boldsymbol{g}^t \) is an unbiased gradient estimate

\[
\mathbb{E}[\boldsymbol{g}^t] = \mathbb{E}[\nabla f_i(x^t)] + \mathbb{E}[\nu_{it}] = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x^t) + \frac{1}{n} \sum_{i=1}^{n} \nu_i = \nabla F(x^t) = \mu x^t
\]
SDCA as variance-reduced SGD

The SDCA (without duality) update rule reads:

$$x^{t+1} = x^t - \eta \left( \nabla f_{i_t}(x^t) + \nu_{i_t}^t \right)$$

The variance of $$\|g^t\|_2$$ goes to 0 as we converge to the optimizer

$$\mathbb{E}[\|g^t\|_2^2] = \mathbb{E}[\|\nu_{i_t}^t - \nu_{i_t}^* + \nu_{i_t}^* + \nabla f_{i_t}(x^t)\|_2^2]$$

$$\leq 2 \mathbb{E}[\|\nu_{i_t}^t - \nu_{i_t}^*\|_2^2] + 2 \mathbb{E}[\|\nu_{i_t}^* + \nabla f_{i_t}(x^t)\|_2^2]$$

$$\rightarrow 0 \text{ as } t \rightarrow \infty$$

$$\leq \|w^t - w^*\|_2^2 \text{ (Shalev-Shwartz '16)}$$
Convergence guarantees of SDCA

Theorem 12.6 (informal, Shalev-Shwartz ’16)

Assume each $f_i$ is convex and $L$-smooth, and set $\eta = \frac{1}{L + \mu n}$. Then it takes SDCA (without duality) $O\left( (n + \frac{L}{\mu}) \log \frac{1}{\varepsilon} \right)$ iterations to yield $\varepsilon$ accuracy.

- the same computational complexity as SVRG
- storage complexity: $O(nd)$ (needs to store $\{v_i\}_{1 \leq i \leq n}$)
Reference


[12] "Optimal finite-Sum smooth non-convex optimization with SARAH,”
L. Nguyen, M. vanDijk, D. Phan, P. Nguyen, T. Weng, J. Kalagnanam,

minimization,” S. Shalev-Shwartz, T. Zhang, Journal of Machine
Learning Research, 2013.

[14] "SDCA without duality, regularization, and individual convexity,”
S. Shalev-Shwartz, ICML, 2016.

[15] "Optimization methods for large-scale machine learning,” L. Bottou,