Variance reduction for stochastic gradient methods

Yuxin Chen

Wharton Statistics & Data Science, Fall 2023

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

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^d} \qquad F(\boldsymbol{x}) := \frac{1}{n}\sum_{i=1}^{n} \underbrace{f_i(\boldsymbol{x})}_{\text{loss for } \underbrace{i\text{th sample}}_{(\boldsymbol{a}_i, y_i)}} \quad + \underbrace{\psi(\boldsymbol{x})}_{\text{regularizer}}$$

common task in machine learning

- linear regression: $f_i(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{a}_i^\top \boldsymbol{x} - y_i)^2$, $\psi(\boldsymbol{x}) = 0$
- logistic regression: $f_i(\boldsymbol{x}) = \log(1 + e^{-y_i \boldsymbol{a}_i^\top \boldsymbol{x}})$, $\psi(\boldsymbol{x}) = 0$
- Lasso: $f_i(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{a}_i^\top \boldsymbol{x} - y_i)^2$, $\psi(\boldsymbol{x}) = \lambda\|\boldsymbol{x}\|_1$
- SVM: $f_i(\boldsymbol{x}) = \max\{0, 1 - y_i \boldsymbol{a}_i^\top \boldsymbol{x}\}$, $\psi(\boldsymbol{x}) = \frac{\lambda}{2}\|\boldsymbol{x}\|_2^2$
- ...

Stochastic gradient descent (SGD)

Algorithm 12.1 Stochastic gradient descent (SGD)

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

As we have shown in the last lecture

- large stepsizes poorly suppress variability of stochastic gradients

$$\Longrightarrow \quad \text{SGD with } \eta_t \asymp 1 \text{ tends to oscillate around global mins}$$

- 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 $F(x^t)$
- $\mathbb{E}[\|g^t\|_2^2] \leq \sigma_{\mathrm{g}}^2 + c_{\mathrm{g}} \|\nabla F(x^t)\|_2^2$
- $F(\cdot)$: μ-strongly convex; L-smooth

From the last lecture, we know

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

Recall: SGD theory with fixed stepsizes

$$\mathbb{E}[F(\boldsymbol{x}^t) - F(\boldsymbol{x}^*)] \leq \frac{\eta L \sigma_{\mathrm{g}}^2}{2\mu} + (1 - \eta\mu)^t \big(F(\boldsymbol{x}^0) - F(\boldsymbol{x}^*)\big)$$

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

A simple idea

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

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

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

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

answer: find some zero-mean $\boldsymbol{v}^t$ that is positively correlated with $\nabla f_{i_t}(\boldsymbol{x}^t)$ (i.e. $\langle \boldsymbol{v}^t, \nabla f_{i_t}(\boldsymbol{x}^t)\rangle > 0$) (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 $\boldsymbol{v}^t$ 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)

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^d} \qquad F(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x})$$

- f_i: convex and L-smooth
- F: μ-strongly convex
- $\kappa := L/\mu$: condition number

Stochastic variance reduced gradient (SVRG)

— *Johnson, Zhang '13*

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

$$\underbrace{\nabla f_{i_t}(\boldsymbol{x}^t) - \nabla f_{i_t}(\boldsymbol{x}^{\text{old}})}_{\to \boldsymbol{0} \text{ if } \boldsymbol{x}^t \approx \boldsymbol{x}^{\text{old}}} + \underbrace{\nabla F(\boldsymbol{x}^{\text{old}})}_{\to \boldsymbol{0} \text{ if } \boldsymbol{x}^{\text{old}} \approx \boldsymbol{x}^*} \qquad \text{with } i_t \sim \text{Unif}(1, \cdots, n)$$

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

Stochastic variance reduced gradient (SVRG)

- operate in epochs
- in the s^{th} epoch
  - **very beginning**: take a snapshot $\boldsymbol{x}_s^{\text{old}}$ of the current iterate, and compute the *batch* gradient $\nabla F(\boldsymbol{x}_s^{\text{old}})$
  - **inner loop**: use the snapshot point to help reduce variance

$$\boldsymbol{x}_s^{t+1} = \boldsymbol{x}_s^t - \eta\big\{\nabla f_{i_t}(\boldsymbol{x}_s^t) - \nabla f_{i_t}(\boldsymbol{x}_s^{\text{old}}) + \nabla F(\boldsymbol{x}_s^{\text{old}})\big\}$$

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

1: **for** $s = 1, 2, \ldots$ **do**

2: $\quad \boldsymbol{x}_s^{\mathsf{old}} \leftarrow \boldsymbol{x}_{s-1}^m$, and compute $\underbrace{\nabla F(\boldsymbol{x}_s^{\mathsf{old}})}_{\text{batch gradient}}$ // update snapshot

3: $\quad$ initialize $\boldsymbol{x}_s^0 \leftarrow \boldsymbol{x}_s^{\mathsf{old}}$

4: $\quad$ **for** $\underbrace{t = 0, \ldots, m-1}_{\text{each epoch contains } m \text{ iterations}}$ **do**

5: $\quad\quad$ choose i_t uniformly from $\{1, \ldots, n\}$, and

$$\boldsymbol{x}_s^{t+1} = \boldsymbol{x}_s^t - \eta \{ \underbrace{\nabla f_{i_t}(\boldsymbol{x}_s^t) - \nabla f_{i_t}(\boldsymbol{x}_s^{\mathsf{old}})}_{\text{stochastic gradient}} + \nabla F(\boldsymbol{x}_s^{\mathsf{old}}) \}$$

Remark

- constant stepsize η
- 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 μ-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

$$\mathbb{E}[F(\boldsymbol{x}_s^{\mathsf{old}}) - F(\boldsymbol{x}^*)] \le \rho^s [F(\boldsymbol{x}_0^{\mathsf{old}}) - F(\boldsymbol{x}^*)]$$

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

$$\Longrightarrow \quad O(\log \tfrac{1}{\varepsilon}) \text{ epochs to attain } \varepsilon \text{ accuracy}$$

Convergence analysis of SVRG

Theorem 12.1

Assume each f_i is convex and L-smooth, and F is μ-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

$$\mathbb{E}[F(\boldsymbol{x}_s^{\mathsf{old}}) - F(\boldsymbol{x}^*)] \leq \rho^s [F(\boldsymbol{x}_0^{\mathsf{old}}) - F(\boldsymbol{x}^*)]$$

- **total computational cost:**

$$\underbrace{(m+n)}_{\#\text{ grad computation per epoch}} \log\frac{1}{\varepsilon} \asymp \underbrace{(n+\kappa)\log\frac{1}{\varepsilon}}_{\text{if } m \asymp \max\{n,\kappa\}}$$

Proof of Theorem 12.1

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

$$\boldsymbol{x}_{s+1}^{\mathsf{old}} = \boldsymbol{x}_s^j \qquad \underbrace{\textit{with } j \sim \mathsf{Unif}(0, \cdots, m-1)}_{\text{rather than } j=m} \tag{12.1}$$

The interested reader is referred to Tan et al. '16 for the proof of the original version

Proof of Theorem 12.1

Let $\boldsymbol{g}_s^t := \nabla f_{i_t}(\boldsymbol{x}_s^t) - \nabla f_{i_t}(\boldsymbol{x}_s^{\mathsf{old}}) + \nabla F(\boldsymbol{x}_s^{\mathsf{old}})$ for simplicity. As usual, conditional on everything prior to $\boldsymbol{x}_s^{t+1}$, one has

$$
\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{x}_s^{t+1} - \boldsymbol{x}^*\|_2^2\right] &= \mathbb{E}\left[\|\boldsymbol{x}_s^t - \eta \boldsymbol{g}_s^t - \boldsymbol{x}^*\|_2^2\right] \\
&= \|\boldsymbol{x}_s^t - \boldsymbol{x}^*\|_2^2 - 2\eta(\boldsymbol{x}_s^t - \boldsymbol{x}^*)^\top \mathbb{E}[\boldsymbol{g}_s^t] + \eta^2 \mathbb{E}\left[\|\boldsymbol{g}_s^t\|_2^2\right] \\
&\leq \|\boldsymbol{x}_s^t - \boldsymbol{x}^*\|_2^2 - 2\eta(\boldsymbol{x}_s^t - \boldsymbol{x}^*)^\top \underbrace{\nabla F(\boldsymbol{x}_s^t)}_{\text{since } \boldsymbol{g}_s^t \text{ is an unbiased estimate of } \nabla F(\boldsymbol{x}_s^t)} + \eta^2 \mathbb{E}\left[\|\boldsymbol{g}_s^t\|_2^2\right] \\
&\leq \|\boldsymbol{x}_s^t - \boldsymbol{x}^*\|_2^2 - \underbrace{2\eta\big(F(\boldsymbol{x}_s^t) - F(\boldsymbol{x}^*)\big)}_{\text{by convexity}} + \eta^2 \mathbb{E}\left[\|\boldsymbol{g}_s^t\|_2^2\right] \qquad (12.2)
\end{aligned}
$$

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

Proof of Theorem 12.1

main pillar: control $\mathbb{E}[\|\boldsymbol{g}_s^t\|_2^2]$ via ...

Lemma 12.2

$$\mathbb{E}\big[\|\boldsymbol{g}_s^t\|_2^2\big] \leq 4L\big[F(\boldsymbol{x}_s^t) - F(\boldsymbol{x}^*) + F(\boldsymbol{x}_s^{\mathsf{old}}) - F(\boldsymbol{x}^*)\big]$$

this means if $\boldsymbol{x}_s^t \approx \boldsymbol{x}_s^{\mathsf{old}} \approx \boldsymbol{x}^*$, then $\mathbb{E}\big[\|\boldsymbol{g}_s^t\|_2^2\big] \approx 0$ (reduced variance)

Proof of Theorem 12.1

main pillar: control $\mathbb{E}\big[\|\boldsymbol{g}_s^t\|_2^2\big]$ via ...

Lemma 12.2

$$\mathbb{E}\big[\|\boldsymbol{g}_s^t\|_2^2\big] \le 4L\big[F(\boldsymbol{x}_s^t) - F(\boldsymbol{x}^*) + F(\boldsymbol{x}_s^{\mathsf{old}}) - F(\boldsymbol{x}^*)\big]$$

this allows one to obtain: conditional on everything prior to $\boldsymbol{x}_s^{t+1}$,

$$\begin{aligned}
\mathbb{E}\big[\|\boldsymbol{x}_s^{t+1} - \boldsymbol{x}^*\|_2^2\big] &\le \text{(12.2)} \\
&\le \|\boldsymbol{x}_s^t - \boldsymbol{x}^*\|_2^2 - 2\eta[F(\boldsymbol{x}_s^t) - F(\boldsymbol{x}^*)] \\
&\qquad + 4L\eta^2\big[F(\boldsymbol{x}_s^t) - F(\boldsymbol{x}^*) + F(\boldsymbol{x}_s^{\mathsf{old}}) - F(\boldsymbol{x}^*)\big] \\
&= \|\boldsymbol{x}_s^t - \boldsymbol{x}^*\|_2^2 - 2\eta(1 - 2L\eta)[F(\boldsymbol{x}_s^t) - F(\boldsymbol{x}^*)] \\
&\qquad + 4L\eta^2\big[F(\boldsymbol{x}_s^{\mathsf{old}}) - F(\boldsymbol{x}^*)\big] \qquad (12.3)
\end{aligned}$$

Proof of Theorem 12.1 (cont.)

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

$$
\begin{aligned}
&2\eta(1-2L\eta)m\,\mathbb{E}\big[F(\boldsymbol{x}_{s+1}^{\mathsf{old}})-F(\boldsymbol{x}^*)\big] \\
&=2\eta(1-2L\eta)\sum_{t=0}^{m-1}\mathbb{E}\big[F(\boldsymbol{x}_s^t)-F(\boldsymbol{x}^*)\big] \qquad \text{by (12.1)} \\
&\le \underbrace{\mathbb{E}\big[\|\boldsymbol{x}_{s+1}^m-\boldsymbol{x}^*\|_2^2\big]}_{\ge 0}+2\eta(1-2L\eta)\sum_{t=0}^{m-1}\mathbb{E}\big[F(\boldsymbol{x}_s^t)-F(\boldsymbol{x}^*)\big] \\
&\le \mathbb{E}\big[\|\boldsymbol{x}_{s+1}^0-\boldsymbol{x}^*\|_2^2\big]+4Lm\eta^2[F(\boldsymbol{x}_s^{\mathsf{old}})-F(\boldsymbol{x}^*)] \qquad \text{(apply (12.3) recursively)} \\
&=\mathbb{E}\big[\|\boldsymbol{x}_s^{\mathsf{old}}-\boldsymbol{x}^*\|_2^2\big]+4Lm\eta^2\mathbb{E}\big[F(\boldsymbol{x}_s^{\mathsf{old}})-F(\boldsymbol{x}^*)\big] \\
&\le \tfrac{2}{\mu}\mathbb{E}\big[F(\boldsymbol{x}_s^{\mathsf{old}})-F(\boldsymbol{x}^*)\big]+4Lm\eta^2\mathbb{E}\big[F(\boldsymbol{x}_s^{\mathsf{old}})-F(\boldsymbol{x}^*)\big] \qquad \text{(strong convexity)} \\
&=\Big(\tfrac{2}{\mu}+4Lm\eta^2\Big)\mathbb{E}[F(\boldsymbol{x}_s^{\mathsf{old}})-F(\boldsymbol{x}^*)]
\end{aligned}
$$

Proof of Theorem 12.1 (cont.)

Consequently,

$$\begin{aligned}
&\mathbb{E}\big[F(\boldsymbol{x}_{s+1}^{\mathsf{old}}) - F(\boldsymbol{x}^*)\big] \\
&\qquad \leq \frac{\frac{2}{\mu} + 4Lm\eta^2}{2\eta(1-2L\eta)m}\mathbb{E}[F(\boldsymbol{x}_s^{\mathsf{old}}) - F(\boldsymbol{x}^*)] \\
&\qquad = \underbrace{\left(\frac{1}{\mu\eta(1-2L\eta)m} + \frac{2L\eta}{1-2L\eta}\right)}_{=\rho}\mathbb{E}[F(\boldsymbol{x}_s^{\mathsf{old}}) - F(\boldsymbol{x}^*)]
\end{aligned}$$

Applying this bound recursively establishes the theorem.

Proof of Lemma 12.2

$$
\begin{aligned}
&\mathbb{E}\big[\|\nabla f_{i_t}(\boldsymbol{x}_s^t) - \nabla f_{i_t}(\boldsymbol{x}_s^{\text{old}}) + \nabla F(\boldsymbol{x}_s^{\text{old}})\|_2^2\big] \\
&= \mathbb{E}\big[\|\nabla f_{i_t}(\boldsymbol{x}_s^t) - \nabla f_{i_t}(\boldsymbol{x}^*) - \big(\nabla f_{i_t}(\boldsymbol{x}_s^{\text{old}}) - \nabla f_{i_t}(\boldsymbol{x}^*) - \nabla F(\boldsymbol{x}_s^{\text{old}})\big)\|_2^2\big] \\
&\leq 2\mathbb{E}\big[\|\nabla f_{i_t}(\boldsymbol{x}_s^t) - \nabla f_{i_t}(\boldsymbol{x}^*)\|_2^2\big] + 2\mathbb{E}\big[\|\nabla f_{i_t}(\boldsymbol{x}_s^{\text{old}}) - \nabla f_{i_t}(\boldsymbol{x}^*) - \nabla F(\boldsymbol{x}_s^{\text{old}})\|_2^2\big] \\
&= 2\mathbb{E}\big[\|\nabla f_{i_t}(\boldsymbol{x}_s^t) - \nabla f_{i_t}(\boldsymbol{x}^*)\|_2^2\big] \\
&\qquad + 2\mathbb{E}\big[\|\nabla f_{i_t}(\boldsymbol{x}_s^{\text{old}}) - \nabla f_{i_t}(\boldsymbol{x}^*) - \underbrace{\mathbb{E}[\nabla f_{i_t}(\boldsymbol{x}_s^{\text{old}}) - \nabla f_{i_t}(\boldsymbol{x}^*)]}_{\text{since } \mathbb{E}[\nabla f_{i_t}(\boldsymbol{x}^*)] = \nabla F(\boldsymbol{x}^*) = \boldsymbol{0}}\|_2^2\big] \\
&\leq 2\mathbb{E}\big[\|\nabla f_{i_t}(\boldsymbol{x}_s^t) - \nabla f_{i_t}(\boldsymbol{x}^*)\|_2^2\big] + 2\mathbb{E}\big[\|\nabla f_{i_t}(\boldsymbol{x}_s^{\text{old}}) - \nabla f_{i_t}(\boldsymbol{x}^*)\|_2^2\big] \\
&\leq 4L[F(\boldsymbol{x}_s^t) - F(\boldsymbol{x}^*) + F(\boldsymbol{x}_s^{\text{old}}) - F(\boldsymbol{x}^*)]
\end{aligned}
$$

where the last inequality would hold if we could justify

$$
\underbrace{\frac{1}{n}\sum_{i=1}^{n}\big\|\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{x}^*)\big\|_2^2 \leq 2L\big[F(\boldsymbol{x}) - F(\boldsymbol{x}^*)\big]}_{\text{relies on both smoothness and convexity of } f_i} \tag{12.4}
$$

Proof of Lemma 12.2 (cont.)

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

$$\underbrace{\frac{1}{2L}\big\|\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{x}^*)\big\|_2^2 \le f_i(\boldsymbol{x}) - f_i(\boldsymbol{x}^*) - \nabla f_i(\boldsymbol{x}^*)^\top(\boldsymbol{x} - \boldsymbol{x}^*)}_{\text{an equivalent characterization of } L\text{-smoothness}}$$

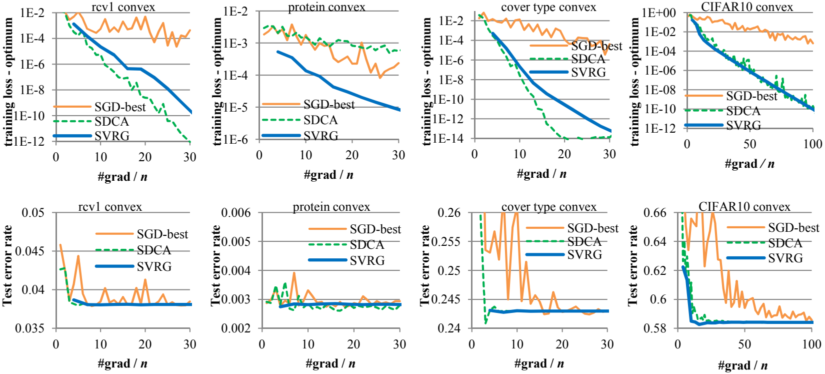
Summing over all i and recognizing that $\nabla F(\boldsymbol{x}^*) = \boldsymbol{0}$ yield

$$\begin{aligned}\frac{1}{2L}\sum_{i=1}^{n}\big\|\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{x}^*)\big\|_2^2 &\le nF(\boldsymbol{x}) - nF(\boldsymbol{x}^*) - n\big(\nabla F(\boldsymbol{x}^\star)\big)^\top(\boldsymbol{x} - \boldsymbol{x}^*)\\ &= nF(\boldsymbol{x}) - nF(\boldsymbol{x}^*)\end{aligned}$$

as claimed

Numerical example: logistic regression

— *Johnson, Zhang '13*



ℓ_2-regularized logistic regression on CIFAR-10

Comparisons with GD and SGD

	SVRG	GD	SGD
comp. cost	$(n+\kappa)\log\frac{1}{\varepsilon}$	$n\kappa\log\frac{1}{\varepsilon}$	$\frac{\kappa^2}{\varepsilon}$ (practically often $\frac{\kappa}{\varepsilon}$)

Proximal extension

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^d} \qquad \underbrace{\frac{1}{n}\sum_{i=1}^{n} f_i\left(\boldsymbol{x}\right)}_{=:F(\boldsymbol{x})} + \psi(\boldsymbol{x})$$

- f_i: convex and L-smooth
- F: μ-strongly convex
- $\kappa := L/\mu$: condition number
- ψ: potentially non-smooth

Proximal extension (Xiao, Zhang '14)

Algorithm 12.3 Prox-SVRG for finite-sum optimization

for $s = 1, 2, \ldots$ **do**

$\quad \boldsymbol{x}_s^{\mathsf{old}} \leftarrow \boldsymbol{x}_{s-1}^m$, and compute $\underbrace{\nabla F(\boldsymbol{x}_s^{\mathsf{old}})}_{\text{batch gradient}}$ // update snapshot

$\quad$ initialize $\boldsymbol{x}_s^0 \leftarrow \boldsymbol{x}_s^{\mathsf{old}}$

$\quad$ **for** $\underbrace{t = 0, \ldots, m-1}_{\text{each epoch contains } m \text{ iterations}}$ **do**

$\quad\quad$ choose i_t uniformly from $\{1, \ldots, n\}$, and

$$\boldsymbol{x}_s^{t+1} = \mathsf{prox}_{\eta\psi}\Big(\boldsymbol{x}_s^t - \eta\{\underbrace{\nabla f_{i_t}(\boldsymbol{x}_s^t) - \nabla f_{i_t}(\boldsymbol{x}_s^{\mathsf{old}})}_{\text{stochastic gradient}} + \nabla F(\boldsymbol{x}_s^{\mathsf{old}})\}\Big)$$

Stochastic recursive gradient algorithm (SARAH)

Nonconvex and smooth problems

$$\text{minimize}_{\boldsymbol{x} \in \mathbb{R}^d} \qquad F(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x})$$

- f_i: L-smooth, potentially nonconvex

Recursive stochastic gradient estimates

— *Nguyen, Liu, Scheinberg, Takac '17*

key idea: recursive / adaptive updates of $\underbrace{\text{gradient estimates}}_{\text{stochastic}}$

$$\begin{aligned} \boldsymbol{g}^t &= \nabla f_{i_t}(\boldsymbol{x}^t) - \nabla f_{i_t}(\boldsymbol{x}^{t-1}) + \boldsymbol{g}^{t-1} \\ \boldsymbol{x}^{t+1} &= \boldsymbol{x}^t - \eta \boldsymbol{g}^t \end{aligned} \tag{12.5}$$

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

$$\text{(SVRG)} \quad \boldsymbol{g}^t = \nabla f_{i_t}(\boldsymbol{x}^t) - \nabla f_{i_t}(\boldsymbol{x}^{\mathsf{old}}) + \nabla F(\boldsymbol{x}^{\mathsf{old}})$$

Restarting gradient estimate every epoch

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

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

solution: reset $\boldsymbol{g}^t$ every few iterations to calibrate with the true batch gradient

Bias of gradient estimates

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

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

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

$$\mathbb{E}[\boldsymbol{g}^t] = \mathbb{E}\big[\nabla F(\boldsymbol{x}^t)\big]$$

StochAstic Recursive grAdient algoritHm

Algorithm 12.4 SARAH (Nguyen et al. '17)

1: **for** $s = 1, 2, \ldots, S$ **do**

2: $\quad \boldsymbol{x}_s^0 \leftarrow \boldsymbol{x}_{s-1}^{m+1}$, and compute $\underbrace{\boldsymbol{g}_s^0 = \nabla F(\boldsymbol{x}_s^0)}_{\text{batch gradient}}$ // restart $\boldsymbol{g}$ anew

3: $\quad \boldsymbol{x}_s^1 = \boldsymbol{x}_s^0 - \eta \boldsymbol{g}_s^0$

4: $\quad$ **for** $t = 1, \ldots, m$ **do**

5: $\quad\quad$ choose i_t uniformly from $\{1, \ldots, n\}$

6: $\quad\quad \boldsymbol{g}_s^t = \underbrace{\nabla f_{i_t}(\boldsymbol{x}_s^t) - \nabla f_{i_t}(\boldsymbol{x}_s^{t-1})}_{\text{stochastic gradient}} + \boldsymbol{g}_s^{t-1}$

7: $\quad\quad \boldsymbol{x}_s^{t+1} = \boldsymbol{x}_s^t - \eta \boldsymbol{g}_s^t$

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

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

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

$$O\left(n + \frac{L\sqrt{n}}{\varepsilon^2}\right) \qquad \text{(setting } m \asymp n, \eta \asymp \tfrac{1}{L\sqrt{m}}\text{)}$$

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

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

- 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?)

$$\mathbb{E}[F(\boldsymbol{x}_s^{m+1})] \leq \mathbb{E}[F(\boldsymbol{x}_s^0)] - \frac{\eta}{2}\sum_{t=0}^{m}\mathbb{E}[\|\nabla F(\boldsymbol{x}_s^t)\|_2^2] \tag{12.6}$$

We will then focus on estalibshing (12.6)

Proof of Theorem 12.3 (cont.)

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

$$\mathbb{E}\big[F(\boldsymbol{x}_s^{t+1})\big] \le \mathbb{E}\big[F(\boldsymbol{x}_s^t)\big] - \eta \mathbb{E}\big[\nabla F(\boldsymbol{x}_s^t)^\top \boldsymbol{g}_s^t\big] + \frac{L\eta^2}{2}\mathbb{E}\big[\big\|\boldsymbol{g}_s^t\big\|_2^2\big] \qquad (12.7)$$

Since $\boldsymbol{g}_s^t$ is not an unbiased estimate of $\nabla F(\boldsymbol{x}_s^t)$, we first decouple

$$2\mathbb{E}\big[\nabla F(\boldsymbol{x}_s^t)^\top \boldsymbol{g}_s^t\big] = \underbrace{\mathbb{E}\big[\big\|\nabla F(\boldsymbol{x}_s^t)\big\|_2^2\big]}_{\text{desired gradient estimate}} + \underbrace{\mathbb{E}\big[\big\|\boldsymbol{g}_s^t\big\|_2^2\big]}_{\text{variance}} - \underbrace{\mathbb{E}\big[\big\|\nabla F(\boldsymbol{x}_s^t) - \boldsymbol{g}_s^t\big\|_2^2\big]}_{\text{squared bias of gradient estimate}}$$

Substitution into (12.7) with straightforward algebra gives

$$\begin{aligned}\mathbb{E}\big[F(\boldsymbol{x}_s^{t+1})\big] \le \mathbb{E}\big[F(\boldsymbol{x}_s^t)\big] - \frac{\eta}{2}\mathbb{E}\big[\big\|\nabla F(\boldsymbol{x}_s^t)\big\|_2^2\big] + \frac{\eta}{2}\mathbb{E}\big[\big\|\nabla F(\boldsymbol{x}_s^t) - \boldsymbol{g}_s^t\big\|_2^2\big] \\ - \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right)\mathbb{E}\big[\big\|\boldsymbol{g}_s^t\big\|_2^2\big]\end{aligned}$$

Proof of Theorem 12.3 (cont.)

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

$$\mathbb{E}\big[F(\boldsymbol{x}_s^{m+1})\big] \leq \mathbb{E}\big[F(\boldsymbol{x}_s^{0})\big] - \frac{\eta}{2}\sum\nolimits_{t=0}^{m}\mathbb{E}\big[\big\|\nabla F(\boldsymbol{x}_s^{t})\big\|_2^2\big]$$
$$+ \frac{\eta}{2}\Big\{\sum\nolimits_{t=0}^{m}\mathbb{E}\big[\big\|\nabla F(\boldsymbol{x}_s^{t}) - \boldsymbol{g}_s^t\big\|_2^2\big] - \underbrace{(1 - L\eta)}_{\geq 1/2}\sum\nolimits_{t=0}^{m}\mathbb{E}\big[\big\|\boldsymbol{g}_s^t\big\|_2^2\big]\Big\}$$

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 η, the epoch length m cannot be too large)

$$\sum\nolimits_{t=0}^{m}\underbrace{\mathbb{E}\Big[\big\|\nabla F(\boldsymbol{x}_s^{t}) - \boldsymbol{g}_s^t\big\|_2^2\Big]}_{\text{squared bias of gradient estimate}} \leq \frac{1}{2}\sum\nolimits_{t=0}^{m}\underbrace{\mathbb{E}\Big[\big\|\boldsymbol{g}_s^t\big\|_2^2\Big]}_{\text{variance}}$$

- 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

$$\mathbb{E}\Big[\|\nabla F(\boldsymbol{x}_s^t) - \boldsymbol{g}_s^t\|_2^2\Big] \leq \sum\nolimits_{k=1}^{t} \mathbb{E}\Big[\|\boldsymbol{g}_s^k - \boldsymbol{g}_s^{k-1}\|_2^2\Big]$$

- convert the bias of gradient estimates to the differences of consecutive gradient estimates (a consequence of the smoothness and the recursive formula of $\boldsymbol{g}_s^t$)

Proof of Lemma 12.4 (cont.)

From Lemma 12.5, it suffices to connect $\{\|\boldsymbol{g}_s^t - \boldsymbol{g}_s^{t-1}\|_2\}$ with $\{\|\boldsymbol{g}_s^t\|_2\}$:

$$\begin{aligned}\big\|\boldsymbol{g}_s^t - \boldsymbol{g}_s^{t-1}\big\|_2^2 &\overset{(12.5)}{=} \big\|\nabla f_{i_t}(\boldsymbol{x}_s^t) - \nabla f_{i_t}(\boldsymbol{x}_s^{t-1})\big\|_2^2 \overset{\text{smoothness}}{\leq} L^2\big\|\boldsymbol{x}_s^t - \boldsymbol{x}_s^{t-1}\big\|_2^2 \\ &= \eta^2 L^2 \big\|\boldsymbol{g}_s^{t-1}\big\|_2^2\end{aligned}$$

Invoking Lemma 12.5 then gives

$$\mathbb{E}\Big[\big\|\nabla F(\boldsymbol{x}_s^t) - \boldsymbol{g}_s^t\big\|_2^2\Big] \leq \sum\nolimits_{k=1}^{t} \mathbb{E}\Big[\big\|\boldsymbol{g}_s^k - \boldsymbol{g}_s^{k-1}\big\|_2^2\Big] \leq \eta^2 L^2 \sum\nolimits_{k=1}^{t} \mathbb{E}\Big[\big\|\boldsymbol{g}_s^{k-1}\big\|_2^2\Big]$$

Summing over $t = 0, \cdots, m$, we obtain

$$\sum\nolimits_{t=0}^{m} \mathbb{E}\Big[\big\|\nabla F(\boldsymbol{x}_s^t) - \boldsymbol{g}_s^t\big\|_2^2\Big] \leq \eta^2 L^2 m \sum\nolimits_{t=0}^{m-1} \mathbb{E}\Big[\big\|\boldsymbol{g}_s^t\big\|_2^2\Big]$$

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 $\boldsymbol{x}^k$ and $\boldsymbol{g}^{k-1}$, then

$$
\begin{aligned}
&\mathbb{E}\Big[\big\|\nabla F(\boldsymbol{x}^k)-\boldsymbol{g}^k\big\|_2^2 \mid \mathcal{F}_k\Big] \\
&= \mathbb{E}\Big[\Big\|\nabla F(\boldsymbol{x}^{k-1})-\boldsymbol{g}^{k-1}+\big(\nabla F(\boldsymbol{x}^k)-\nabla F(\boldsymbol{x}^{k-1})\big)-\big(\boldsymbol{g}^k-\boldsymbol{g}^{k-1}\big)\Big\|_2^2 \mid \mathcal{F}_k\Big] \\
&= \big\|\nabla F(\boldsymbol{x}^{k-1})-\boldsymbol{g}^{k-1}\big\|_2^2+\big\|\nabla F(\boldsymbol{x}^k)-\nabla F(\boldsymbol{x}^{k-1})\big\|_2^2+\mathbb{E}\Big[\big\|\boldsymbol{g}^k-\boldsymbol{g}^{k-1}\big\|_2^2 \mid \mathcal{F}_k\Big] \\
&\quad + 2\Big\langle \nabla F(\boldsymbol{x}^{k-1})-\boldsymbol{g}^{k-1}, \nabla F(\boldsymbol{x}^k)-\nabla F(\boldsymbol{x}^{k-1})\Big\rangle \\
&\quad - 2\Big\langle \nabla F(\boldsymbol{x}^{k-1})-\boldsymbol{g}^{k-1}, \mathbb{E}\big[\boldsymbol{g}^k-\boldsymbol{g}^{k-1} \mid \mathcal{F}_k\big]\Big\rangle \\
&\quad - 2\Big\langle \nabla F(\boldsymbol{x}^k)-\nabla F(\boldsymbol{x}^{k-1}), \mathbb{E}\big[\boldsymbol{g}^k-\boldsymbol{g}^{k-1} \mid \mathcal{F}_k\big]\Big\rangle \\
&\overset{\text{(exercise)}}{=} \big\|\nabla F(\boldsymbol{x}^{k-1})-\boldsymbol{g}^{k-1}\big\|_2^2-\big\|\nabla F(\boldsymbol{x}^k)-\nabla F(\boldsymbol{x}^{k-1})\big\|_2^2+\mathbb{E}\Big[\big\|\boldsymbol{g}^k-\boldsymbol{g}^{k-1}\big\|_2^2 \mid \mathcal{F}_k\Big]
\end{aligned}
$$

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

$$
\mathbb{E}\Big[\big\|\nabla F(\boldsymbol{x}^k)-\boldsymbol{g}^k\big\|_2^2\Big] = \sum_{k=1}^{t}\mathbb{E}\Big[\big\|\boldsymbol{g}^k-\boldsymbol{g}^{k-1}\big\|_2^2\Big] \underbrace{-\sum_{k=1}^{t}\big\|\nabla F(\boldsymbol{x}^k)-\nabla F(\boldsymbol{x}^{k-1})\big\|_2^2}_{\leq 0;\ \text{done!}}
$$

Stochastic dual coordinate ascent (SDCA)

— a dual perspective

A class of finite-sum optimization

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^d} \quad F(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{x}\|_2^2 \qquad (12.8)$$

- f_i: convex and L-smooth

Dual formulation

The dual problem of (12.8)

$$\text{maximize}_{\boldsymbol{v}} \quad D(\boldsymbol{\nu}) = \frac{1}{n}\sum_{i=1}^{n} -f_i^*(-\boldsymbol{v}_i) - \frac{\mu}{2}\left\|\frac{1}{\mu n}\sum_{i=1}^{n}\boldsymbol{\nu}_i\right\|_2^2 \qquad (12.9)$$

- a primal-dual relation

$$\boldsymbol{x}(\boldsymbol{\nu}) = \frac{1}{\mu n}\sum_{i=1}^{n}\boldsymbol{\nu}_i \qquad (12.10)$$

Derivation of the dual formulation

$$\min_{\boldsymbol{x}} \quad \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{x}\|_2^2$$

$$\Longleftrightarrow \qquad \min_{\boldsymbol{x},\{\boldsymbol{z}_i\}} \quad \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{z}_i) + \frac{\mu}{2}\|\boldsymbol{x}\|_2^2 \qquad \text{s.t. } \boldsymbol{z}_i = \boldsymbol{x}$$

$$\Longleftrightarrow \qquad \max_{\{\boldsymbol{\nu}_i\}} \; \min_{\boldsymbol{x},\{\boldsymbol{z}_i\}} \quad \underbrace{\frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{z}_i) + \frac{\mu}{2}\|\boldsymbol{x}\|_2^2 + \frac{1}{n}\sum_{i=1}^{n}\langle \boldsymbol{\nu}_i, \boldsymbol{z}_i - \boldsymbol{x}\rangle}_{\text{Lagrangian}}$$

$$\Longleftrightarrow \qquad \max_{\{\boldsymbol{\nu}_i\}} \; \min_{\boldsymbol{x}} \quad \underbrace{\frac{1}{n}\sum_{i=1}^{n} -f_i^*(-\boldsymbol{\nu}_i)}_{\text{conjugate: } f_i^*(\boldsymbol{\nu}) := \max_{\boldsymbol{z}}\{\langle \boldsymbol{\nu}, \boldsymbol{z}\rangle - f_i(\boldsymbol{z})\}} + \frac{\mu}{2}\|\boldsymbol{x}\|_2^2 - \frac{1}{n}\sum_{i=1}^{n}\langle \boldsymbol{\nu}_i, \boldsymbol{x}\rangle$$

$$\Longleftrightarrow \qquad \max_{\{\boldsymbol{\nu}_i\}} \quad \frac{1}{n}\sum_{i=1}^{n} -f_i^*(-\boldsymbol{\nu}_i) - \frac{\mu}{2}\Big\| \underbrace{\frac{1}{\mu n}\sum_{i=1}^{n}\boldsymbol{\nu}_i}_{\text{optimal } \boldsymbol{x} = \frac{1}{\mu n}\sum_i \boldsymbol{\nu}_i} \Big\|_2^2$$

Randomized coordinate ascent on dual problem

— *Shalev-Shwartz, Zhang '13*

- **randomized coordinate ascent:** at each iteration, randomly pick one dual (block) coordinate $\boldsymbol{\nu}_{i_t}$ of (12.9) to optimize
- **maintain the primal-dual relation** (12.10)

$$\boldsymbol{x}^t = \frac{1}{\mu n} \sum_{i=1}^{n} \boldsymbol{\nu}_i^t \qquad (12.11)$$

Stochastic dual coordinate ascent (SDCA)

Algorithm 12.5 SDCA for finite-sum optimization

initialize $\boldsymbol{x}^0 = \frac{1}{\mu n}\sum_{i=1}^n \boldsymbol{\nu}_i^0$
for $t = 0, 1, \ldots$ **do**
    // choose a random coordinate to optimize
    choose i_t uniformly from $\{1, \ldots, n\}$
    $\boldsymbol{\Delta}^t \leftarrow \underbrace{\arg\max_{\boldsymbol{\Delta}} -\frac{1}{n} f_{i_t}^*\big(-\boldsymbol{\nu}_{i_t}^t - \boldsymbol{\Delta}\big) - \frac{\mu}{2}\big\|\boldsymbol{x}^t + \frac{1}{\mu n}\boldsymbol{\Delta}\big\|_2^2}_{\text{find the optimal step with all } \{\boldsymbol{\nu}_i^t\}_{i:i\neq i_t} \text{ fixed}}$
    $\boldsymbol{\nu}_i^{t+1} \leftarrow \underbrace{\boldsymbol{\nu}_i^t + \boldsymbol{\Delta}^t \mathbb{1}\{i = i_t\}}_{\text{update only the } i_t^{\text{th}} \text{ coordinate}} \qquad (1 \leq i \leq n)$
    $\boldsymbol{x}^{t+1} \leftarrow \boldsymbol{x}^t + \frac{1}{\mu n}\boldsymbol{\Delta}^t$ // based on (12.11)

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

initialize $\boldsymbol{x}^0 = \frac{1}{\mu n}\sum_{i=1}^n \boldsymbol{\nu}_i^0$
for $t = 0, 1, \ldots$ **do**
  // choose a random coordinate to optimize
  choose i_t uniformly from $\{1, \ldots, n\}$
  $\boldsymbol{\Delta}^t \leftarrow -\eta\mu n(\nabla f_{i_t}(\boldsymbol{x}^t) + \boldsymbol{\nu}_{i_t}^t)$
  $\boldsymbol{\nu}_i^{t+1} \leftarrow \underbrace{\boldsymbol{\nu}_i^t + \boldsymbol{\Delta}^t \mathbb{1}\{i = i_t\}}_{\text{update only the } i_t^{\text{th}} \text{ coordinate}} \qquad (1 \leq i \leq n)$
  $\boldsymbol{x}^{t+1} \leftarrow \boldsymbol{x}^t + \frac{1}{\mu n}\boldsymbol{\Delta}^t$ // based on (12.11)

A variant of SDCA without duality

A little intuition

- the optimality condition requires (check!)

$$\boldsymbol{\nu}_i^* = -\nabla f_i(\boldsymbol{x}^*), \qquad \forall i \tag{12.12}$$

- with a modified update rule, one has

$$\boldsymbol{\nu}_{i_t}^{t+1} \;\leftarrow\; \underbrace{(1-\eta\mu n)\boldsymbol{\nu}_{i_t}^{t} + \eta\mu n\big(-\nabla f_{i_t}(\boldsymbol{x}^t)\big)}_{\text{cvx combination of current dual iterate and gradient component}}$$

— when it converges, it will satisfy (12.12)

SDCA as SGD

The SDCA (without duality) update rule reads:

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta\big(\underbrace{\nabla f_{i_t}(\boldsymbol{x}^t) + \boldsymbol{\nu}_{i_t}^t}_{:=\boldsymbol{g}^t}\big)$$

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

$$\mathbb{E}[\boldsymbol{g}^t] = \mathbb{E}[\nabla f_{i_t}(\boldsymbol{x}^t)] + \mathbb{E}[\boldsymbol{\nu}_{i_t}^t] = \frac{1}{n}\sum_{i=1}^n \nabla f_i(\boldsymbol{x}^t) + \underbrace{\frac{1}{n}\sum_{i=1}^n \boldsymbol{\nu}_i^t}_{=\mu\boldsymbol{x}^t} = \nabla F(\boldsymbol{x}^t)$$

SDCA as variance-reducedSGD

The SDCA (without duality) update rule reads:

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta\big(\underbrace{\nabla f_{i_t}(\boldsymbol{x}^t) + \boldsymbol{\nu}_{i_t}^t}_{:=\boldsymbol{g}^t}\big)$$

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

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{g}^t\|_2^2] &= \mathbb{E}[\|\boldsymbol{\nu}_{i_t}^t - \boldsymbol{\nu}_{i_t}^* + \boldsymbol{\nu}_{i_t}^* + \nabla f_{i_t}(\boldsymbol{x}^t)\|_2^2] \\ &\leq 2\underbrace{\mathbb{E}[\|\boldsymbol{\nu}_{i_t}^t - \boldsymbol{\nu}_{i_t}^*\|_2^2]}_{\to\ 0 \text{ as } t\to\infty} + 2\underbrace{\mathbb{E}[\|\boldsymbol{\nu}_{i_t}^* + \nabla f_{i_t}(\boldsymbol{x}^t)\|_2^2]}_{\leq \|\boldsymbol{w}^t - \boldsymbol{w}^*\|_2^2 \text{ (Shalev-Shwartz '16)}}\end{aligned}$$

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) $O\big((n + \frac{L}{\mu}) \log \frac{1}{\varepsilon}\big)$ iterations to yield ε accuracy

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

Reference

- "*Recent advances in stochastic convex and non-convex optimization*," Z. Allen-Zhu, *ICML Tutorial*, 2017.
- "*Accelerating stochastic gradient descent using predictive variance reduction*," R. Johnson, T. Zhang, *NIPS*, 2013.
- "*Barzilai-Borwein step size for stochastic gradient descent*," C. Tan, S. Ma, Y.H. Dai, Y. Qian, *NIPS*, 2016.
- "*A proximal stochastic gradient method with progressive variance reduction*," L. Xiao, T. Zhang, *SIAM Journal on Optimization*, 2014.
- "*Minimizing finite sums with the stochastic average gradient*," M. Schmidt, N. Le Roux, F. Bach, *Mathematical Programming*, 2013.
- "*SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives*," A. Defazio, F. Bach, and S. Lacoste-Julien, *NIPS*, 2014.

Reference

- "*Variance reduction for faster non-convex optimization*," Z. Allen-Zhu, E. Hazan, *ICML*, 2016.
- "*Katyusha: The first direct acceleration of stochastic gradient methods*," Z. Allen-Zhu, *STOC*, 2017.
- "*SARAH: A novel method for machine learning problems using stochastic recursive gradient*," L. Nguyen, J. Liu, K. Scheinberg, M. Takac, *ICML*, 2017.
- "*Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator*," C. Fang, C. Li, Z. Lin, T. Zhang, *NIPS*, 2018.
- "*SpiderBoost and momentum: Faster variance reduction algorithms*," Z. Wang, K. Ji, Y. Zhou, Y. Liang, V. Tarokh, *NIPS*, 2019.

Reference

- "*Optimal finite-Sum smooth non-convex optimization with SARAH*," L. Nguyen, M. vanDijk, D. Phan, P. Nguyen, T. Weng, J. Kalagnanam, arXiv:1901.07648, 2019.
- "*Stochastic dual coordinate ascent methods for regularized loss minimization*," S. Shalev-Shwartz, T. Zhang, *Journal of Machine Learning Research*, 2013.
- "*SDCA without duality, regularization, and individual convexity*," S. Shalev-Shwartz, ICML, 2016.
- "*Optimization methods for large-scale machine learning*," L. Bottou, F. Curtis, J. Nocedal, 2016.