Optimization for Machine Learning

Stochastic Variance Reduced Gradient Methods

Lecturer: Robert M. Gower

TELECOM ParisTech

28th of April to 5th of May 2020, Cornell mini-lecture series, online

References for this class

O. Sebbouh, N. Gazagnadou, S. Jelassi, F. Bach, R. M. G. **Towards closing the gap between the theory and practice of SVRG,** Neurips 2019.

M. Schmidt, N. Le Roux, F. Bach (2016), Mathematical Programming **Minimizing Finite Sums with the Stochastic Average Gradient.**

RMG, P. Richtárik and Francis Bach (2018) **Stochastic quasi-gradient methods: variance reduction via Jacobian sketching**

EXE: variance_reduced_exe + convergence_prob_exe

Optimization Sum of Terms

A Datum Function

$$f_i(w) := \ell\left(h_w(x^i), y^i\right) + \lambda R(w)$$

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \ell\left(h_w(x^i), y^i\right) + \lambda R(w) &= \frac{1}{n}\sum_{i=1}^{n} \left(\ell\left(h_w(x^i), y^i\right) + \lambda R(w)\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} f_i(w)\end{aligned}$$

Finite Sum Training Problem

$$\min_{w \in \mathbf{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(w) =: f(w)$$

Issue with variance of SGD

Complexity / Convergence

Theorem

If f is μ–str. convex, f_i is convex, L_i–smooth, $\alpha \in [0, \frac{1}{2L_{\max}}]$

then the iterates of the SGD satisfy

$\sigma^2 := \mathbb{E}_j[||\nabla f_j(w^*)||_2^2]$

$$\mathbb{E}\left[||w^t - w^*||_2^2\right] \leq (1-\alpha\mu)^t ||w^0 - w^*||_2^2 + \frac{2\alpha}{\mu}\sigma^2$$

Complexity / Convergence

Theorem

If f is μ–str. convex, f_i is convex, L_i–smooth, $\alpha \in [0, \frac{1}{2L_{\max}}]$ then the iterates of the SGD satisfy

$\sigma^2 := \mathbb{E}_j[||\nabla f_j(w^*)||_2^2]$

$$\mathbb{E}\left[||w^t - w^*||_2^2\right] \leq (1-\alpha\mu)^t ||w^0 - w^*||_2^2 + \frac{2\alpha}{\mu}\sigma^2$$

This stops SGD from naturally converging

Complexity / Convergence

Theorem

If f is μ–str. convex, f_i is convex, L_i–smooth, $\alpha \in [0, \frac{1}{2L_{\max}}]$ then the iterates of the SGD satisfy

$\sigma^2 := \mathbb{E}_j[||\nabla f_j(w^*)||_2^2]$

$$\mathbb{E}\left[||w^t - w^*||_2^2\right] \leq (1-\alpha\mu)^t ||w^0 - w^*||_2^2 + \frac{2\alpha}{\mu}\sigma^2$$

Where did this term come from ?

This stops SGD from naturally converging

Proof:

$$\begin{aligned}||w^{t+1} - w^*||_2^2 &= ||w^t - w^* - \alpha \nabla f_j(w^t)||_2^2 \\ &= ||w^t - w^*||_2^2 - 2\alpha\langle \nabla f_j(w^t), w^t - w^* \rangle + \alpha^2 ||\nabla f_j(w^t)||_2^2.\end{aligned}$$

Proof:

$$
\begin{aligned}
||w^{t+1} - w^*||_2^2 &= ||w^t - w^* - \alpha \nabla f_j(w^t)||_2^2 \\
&= ||w^t - w^*||_2^2 - 2\alpha \langle \nabla f_j(w^t), w^t - w^* \rangle + \alpha^2 ||\nabla f_j(w^t)||_2^2.
\end{aligned}
$$

Taking expectation conditioned on respect to w^t

Proof:

$$\begin{aligned} ||w^{t+1} - w^*||_2^2 &= ||w^t - w^* - \alpha \nabla f_j(w^t)||_2^2 \\ &= ||w^t - w^*||_2^2 - 2\alpha \langle \nabla f_j(w^t), w^t - w^* \rangle + \alpha^2 ||\nabla f_j(w^t)||_2^2. \end{aligned}$$

Taking expectation conditioned on respect to w^t

$$\begin{aligned} \mathbb{E}_j \left[||w^{t+1} - w^*||_2^2 \right] &= ||w^t - w^*||_2^2 - 2\alpha \langle \nabla f(w^t), w^t - w^* \rangle + \alpha^2 \mathbb{E}_j \left[||\nabla f_j(w^t)||_2^2 \right] \\ &\leq (1 - \alpha\mu)||w^t - w^*||_2^2 - 2\alpha(f(w^t) - f(w^*)) + \alpha^2 \mathbb{E}_j \left[||\nabla f_j(w^t)||_2^2 \right] \end{aligned}$$

Proof:

$$\begin{aligned} ||w^{t+1} - w^*||_2^2 &= ||w^t - w^* - \alpha \nabla f_j(w^t)||_2^2 \\ &= ||w^t - w^*||_2^2 - 2\alpha\langle \nabla f_j(w^t), w^t - w^*\rangle + \alpha^2 ||\nabla f_j(w^t)||_2^2. \end{aligned}$$

Taking expectation conditioned on respect to w^t

$\mathbb{E}[\nabla f_j(w)] = \nabla f(w)$

$$\begin{aligned} \mathbb{E}_j\left[||w^{t+1} - w^*||_2^2\right] &= ||w^t - w^*||_2^2 - 2\alpha\langle \nabla f(w^t), w^t - w^*\rangle + \alpha^2 \mathbb{E}_j\left[||\nabla f_j(w^t)||_2^2\right] \\ &\leq (1-\alpha\mu)||w^t - w^*||_2^2 - 2\alpha(f(w^t) - f(w^*)) + \alpha^2 \mathbb{E}_j\left[||\nabla f_j(w^t)||_2^2\right] \end{aligned}$$

quasi strong conv

Proof:

$$
\begin{aligned}
||w^{t+1} - w^*||_2^2 &= ||w^t - w^* - \alpha \nabla f_j(w^t)||_2^2 \\
&= ||w^t - w^*||_2^2 - 2\alpha \langle \nabla f_j(w^t), w^t - w^* \rangle + \alpha^2 ||\nabla f_j(w^t)||_2^2.
\end{aligned}
$$

Taking expectation conditioned on respect to w^t

$\mathbb{E}[\nabla f_j(w)] = \nabla f(w)$

$$
\begin{aligned}
\mathbb{E}_j \left[||w^{t+1} - w^*||_2^2 \right] &= ||w^t - w^*||_2^2 - 2\alpha \langle \nabla f(w^t), w^t - w^* \rangle + \alpha^2 \mathbb{E}_j \left[||\nabla f_j(w^t)||_2^2 \right] \\
&\leq (1 - \alpha\mu)||w^t - w^*||_2^2 - 2\alpha (f(w^t) - f(w^*)) + \alpha^2 \mathbb{E}_j \left[||\nabla f_j(w^t)||_2^2 \right]
\end{aligned}
$$

quasi strong conv

Proof:

$$\begin{aligned} \|w^{t+1} - w^*\|_2^2 &= \|w^t - w^* - \alpha \nabla f_j(w^t)\|_2^2 \\ &= \|w^t - w^*\|_2^2 - 2\alpha \langle \nabla f_j(w^t), w^t - w^* \rangle + \alpha^2 \|\nabla f_j(w^t)\|_2^2. \end{aligned}$$

Taking expectation conditioned on respect to w^t

$\mathbb{E}[\nabla f_j(w)] = \nabla f(w)$

$$\mathbb{E}_j\left[\|w^{t+1} - w^*\|_2^2\right] = \|w^t - w^*\|_2^2 - 2\alpha \langle \nabla f(w^t), w^t - w^* \rangle + \alpha^2 \mathbb{E}_j\left[\|\nabla f_j(w^t)\|_2^2\right]$$

quasi strong conv

$$\leq (1 - \alpha\mu)\|w^t - w^*\|_2^2 - 2\alpha(f(w^t) - f(w^*)) + \alpha^2 \mathbb{E}_j\left[\|\nabla f_j(w^t)\|_2^2\right]$$

$$\mathbb{E}_j\left[\|\nabla f_j(w^t)\|_2^2\right] \leq 2\mathbb{E}_j\left[\|\nabla f_j(w^t) - \nabla f_j(w^*)\|_2^2\right] + 2\mathbb{E}_j\left[\|\nabla f_j(w^*)\|_2^2\right]$$

f_i is cvx and $L_{\max}$–smooth

$$\leq 4L_{\max}(f(w) - f(w^*)) + 2\sigma^2$$

Proof:

$$\begin{aligned} ||w^{t+1} - w^*||_2^2 &= ||w^t - w^* - \alpha \nabla f_j(w^t)||_2^2 \\ &= ||w^t - w^*||_2^2 - 2\alpha \langle \nabla f_j(w^t), w^t - w^* \rangle + \alpha^2 ||\nabla f_j(w^t)||_2^2. \end{aligned}$$

Taking expectation conditioned on respect to w^t

$\mathbb{E}[\nabla f_j(w)] = \nabla f(w)$

$$\begin{aligned} \mathbb{E}_j \left[||w^{t+1} - w^*||_2^2 \right] &= ||w^t - w^*||_2^2 - 2\alpha \langle \nabla f(w^t), w^t - w^* \rangle + \alpha^2 \mathbb{E}_j \left[||\nabla f_j(w^t)||_2^2 \right] \\ &\leq (1 - \alpha\mu) ||w^t - w^*||_2^2 - 2\alpha (f(w^t) - f(w^*)) + \alpha^2 \mathbb{E}_j \left[||\nabla f_j(w^t)||_2^2 \right] \end{aligned}$$

quasi strong conv

$$\begin{aligned} \mathbb{E}_j \left[||\nabla f_j(w^t)||_2^2 \right] &\leq 2\mathbb{E}_j \left[||\nabla f_j(w^t) - \nabla f_j(w^*)||_2^2 \right] + 2\mathbb{E}_j \left[||\nabla f_j(w^*)||_2^2 \right] \\ &\leq 4 L_{\max} (f(w) - f(w^*)) + 2\sigma^2 \end{aligned}$$

f_i is cvx and $L_{\max}$–smooth

Proof:

$$
\begin{aligned}
||w^{t+1} - w^*||_2^2 &= ||w^t - w^* - \alpha \nabla f_j(w^t)||_2^2 \\
&= ||w^t - w^*||_2^2 - 2\alpha \langle \nabla f_j(w^t), w^t - w^* \rangle + \alpha^2 ||\nabla f_j(w^t)||_2^2.
\end{aligned}
$$

Taking expectation conditioned on respect to w^t

$\mathbb{E}[\nabla f_j(w)] = \nabla f(w)$

$$
\begin{aligned}
\mathbb{E}_j\left[||w^{t+1} - w^*||_2^2\right] &= ||w^t - w^*||_2^2 - 2\alpha \langle \nabla f(w^t), w^t - w^* \rangle + \alpha^2 \mathbb{E}_j\left[||\nabla f_j(w^t)||_2^2\right] \\
&\le (1-\alpha\mu)||w^t - w^*||_2^2 - 2\alpha (f(w^t) - f(w^*)) + \alpha^2 \mathbb{E}_j\left[||\nabla f_j(w^t)||_2^2\right]
\end{aligned}
$$

quasi strong conv

$$
\begin{aligned}
\mathbb{E}_j\left[||\nabla f_j(w^t)||_2^2\right] &\le 2\mathbb{E}_j\left[||\nabla f_j(w^t) - \nabla f_j(w^*)||_2^2\right] + 2\mathbb{E}_j\left[||\nabla f_j(w^*)||_2^2\right] \\
&\le 4L_{\max}(f(w) - f(w^*)) + 2\sigma^2
\end{aligned}
$$

f_i is cvx and $L_{\max}$–smooth

$$
\begin{aligned}
\mathbb{E}_j\left[||w^{t+1} - w^*||_2^2\right] &\le (1-\alpha\mu)||w^t - w^*||_2^2 + 2\gamma(2\gamma L_{\max} - 1)(f(w) - f(w^*)) + 2\alpha^2\sigma^2 \\
&\le (1-\alpha\mu)||w^t - w^*||_2^2 + 2\alpha^2\sigma^2
\end{aligned}
$$

$\alpha \le \dfrac{1}{2L_{\max}}$

Proof:

$$
\begin{aligned}
||w^{t+1} - w^*||_2^2 &= ||w^t - w^* - \alpha \nabla f_j(w^t)||_2^2 \\
&= ||w^t - w^*||_2^2 - 2\alpha \langle \nabla f_j(w^t), w^t - w^* \rangle + \alpha^2 ||\nabla f_j(w^t)||_2^2.
\end{aligned}
$$

Taking expectation conditioned on respect to w^t

$\mathbb{E}[\nabla f_j(w)] = \nabla f(w)$

$$
\mathbb{E}_j \left[||w^{t+1} - w^*||_2^2 \right] = ||w^t - w^*||_2^2 - 2\alpha \langle \nabla f(w^t), w^t - w^* \rangle + \alpha^2 \mathbb{E}_j \left[||\nabla f_j(w^t)||_2^2 \right]
$$

quasi strong conv

$$
\leq (1 - \alpha\mu)||w^t - w^*||_2^2 - 2\alpha (f(w^t) - f(w^*)) + \alpha^2 \mathbb{E}_j \left[||\nabla f_j(w^t)||_2^2 \right]
$$

$$
\mathbb{E}_j \left[||\nabla f_j(w^t)||_2^2 \right] \leq 2\mathbb{E}_j \left[||\nabla f_j(w^t) - \nabla f_j(w^*)||_2^2 \right] + 2\mathbb{E}_j \left[||\nabla f_j(w^*)||_2^2 \right]
$$

f_i is cvx and $L_{\max}$–smooth

$$
\leq 4L_{\max}(f(w) - f(w^*)) + 2\sigma^2
$$

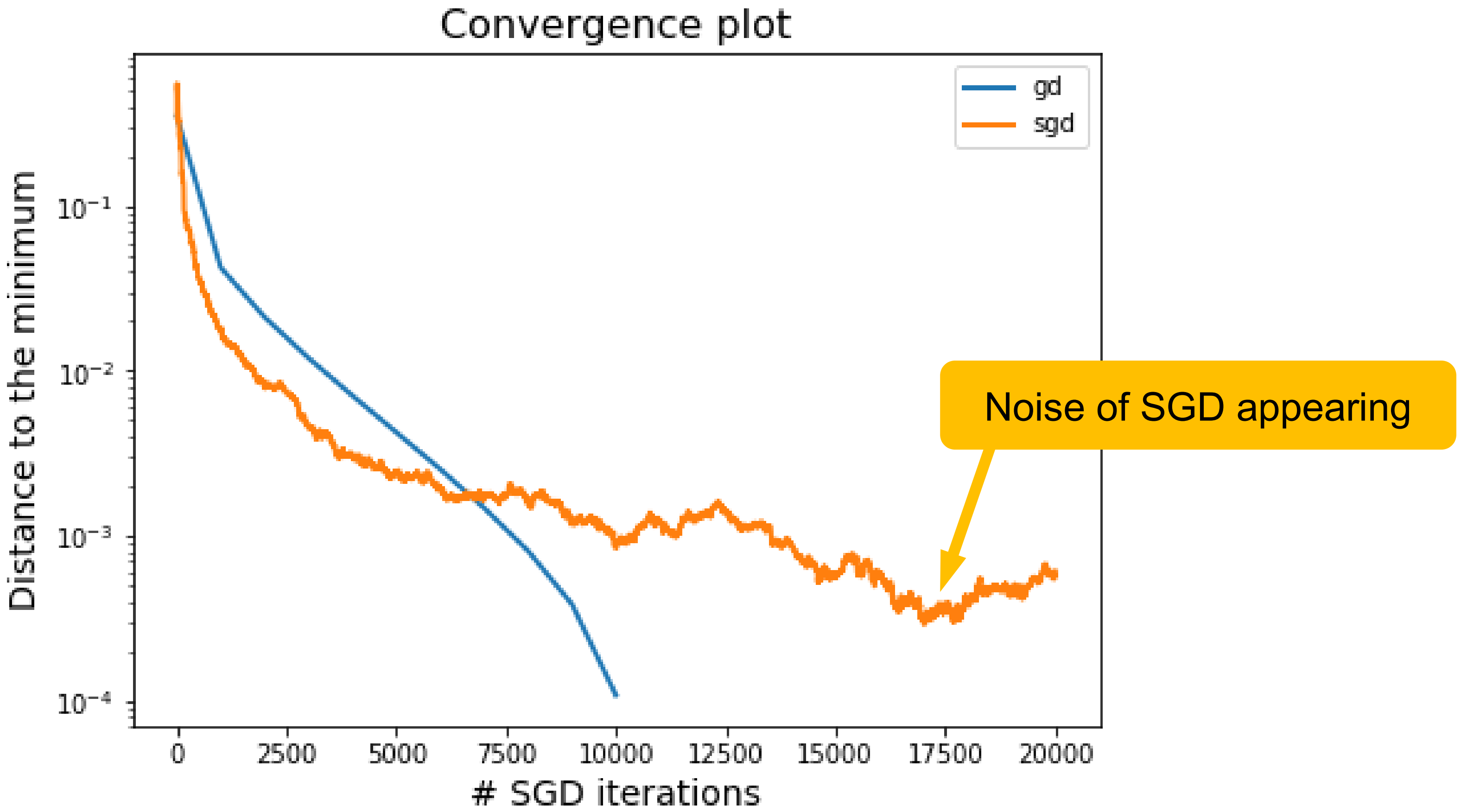
$$
\mathbb{E}_j \left[||w^{t+1} - w^*||_2^2 \right] \leq (1 - \alpha\mu)||w^t - w^*||_2^2 + 2\gamma(2\gamma L_{\max} - 1)(f(w) - f(w^*)) + 2\alpha^2\sigma^2
$$

$\alpha \leq \dfrac{1}{2L_{\max}}$

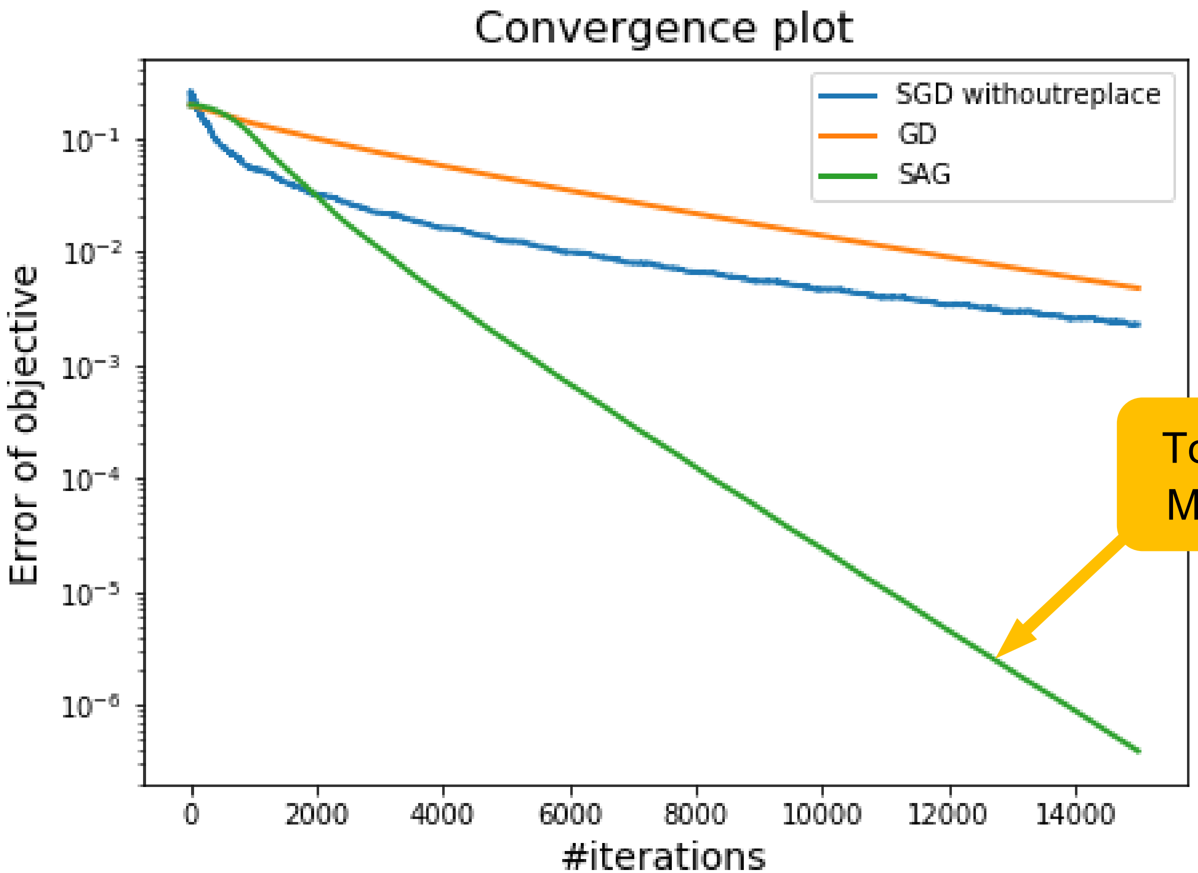
$$
\leq (1 - \alpha\mu)||w^t - w^*||_2^2 + 2\alpha^2\sigma^2
$$

Proof follows by expanding recurrence and summing up

SGD initially fast, slow later



Can we get best of both?



Today we learn about Methods like this one

Stochastic variance reduced methods

Build an Estimate of the Gradient

Instead of using directly $\nabla f_j(w^t) \approx \nabla f(w^t)$
Use $\nabla f_j(w^t)$ to update estimate $g_t \approx \nabla f(w^t)$

$$w^{t+1} = w^t - \gamma g^t$$

We would like gradient estimate such that:

Good estimate

$$g^t \approx \nabla f(w^t)$$

Converges in *L2*

$$\mathbb{E}_t ||g^t||_2^2 \underset{w^t \to w^*}{\longrightarrow} 0$$

Build an Estimate of the Gradient

Instead of using directly $\nabla f_j(w^t) \approx \nabla f(w^t)$
Use $\nabla f_j(w^t)$ to update estimate $g_t \approx \nabla f(w^t)$

$$w^{t+1} = w^t - \gamma g^t$$

We would like gradient estimate such that:

Good estimate

$$g^t \approx \nabla f(w^t)$$

Converges in *L2*

$$\mathbb{E}_t ||g^t||_2^2 \underset{w^t \to w^*}{\longrightarrow} 0$$

Build an Estimate of the Gradient

Instead of using directly $\nabla f_j(w^t) \approx \nabla f(w^t)$
Use $\nabla f_j(w^t)$ to update estimate $g_t \approx \nabla f(w^t)$

$$w^{t+1} = w^t - \gamma g^t$$

Typically unbiased
$\mathbf{E}[g^t] = \nabla f(w^t)$

We would like gradient estimate such that:

Good estimate

$$g^t \approx \nabla f(w^t)$$

Converges in *L2*

$$\mathbb{E}_t||g^t||_2^2 \underset{w^t \to w^*}{\longrightarrow} 0$$

Build an Estimate of the Gradient

Instead of using directly $\nabla f_j(w^t) \approx \nabla f(w^t)$
Use $\nabla f_j(w^t)$ to update estimate $g_t \approx \nabla f(w^t)$

$$w^{t+1} = w^t - \gamma g^t$$

We would like gradient estimate such that:

Good estimate

$$g^t \approx \nabla f(w^t)$$

Typically unbiased
$\mathbf{E}[g^t] = \nabla f(w^t)$

Converges in *L2*

$$\mathbb{E}_t ||g^t||_2^2 \xrightarrow[w^t \to w^*]{} 0$$

Solves SGD problem
$\mathbb{E}_j \left[||\nabla f_j(w^t)||_2^2\right]$

High Level Proof when $\mathbf{E}[g^t] = \nabla f(w^t)$:

$$\begin{aligned} \|w^{t+1} - w^*\|_2^2 &= \|w^t - w^* - \gamma g^t\|_2^2 \\ &= \|w^t - w^*\|_2^2 - 2\gamma\langle g^t, w^t - w^*\rangle + \gamma^2\|g^t\|_2^2. \end{aligned}$$

Taking expectation conditioned on respect to w^t

$\mathbb{E}[\nabla f_j(w)] = \nabla f(w)$

quasi strong conv

$$\begin{aligned} \mathbb{E}_t\left[\|w^{t+1} - w^*\|_2^2\right] &= \|w^t - w^*\|_2^2 - 2\gamma\langle \nabla f(w^t), w^t - w^*\rangle + \gamma^2\mathbb{E}_t\left[\|g^t\|_2^2\right] \\ &\leq (1-\gamma\mu)\|w^t - w^*\|_2^2 - 2\gamma(f(w^t) - f(w^*)) + \gamma^2\mathbb{E}_t\left[\|g^t\|_2^2\right] \end{aligned}$$

High Level Proof when $\mathbf{E}[g^t] = \nabla f(w^t)$:

$$\begin{aligned} ||w^{t+1} - w^*||_2^2 &= ||w^t - w^* - \gamma g^t||_2^2 \\ &= ||w^t - w^*||_2^2 - 2\gamma \langle g^t, w^t - w^* \rangle + \gamma^2 ||g^t||_2^2. \end{aligned}$$

Taking expectation conditioned on respect to w^t

$\mathbb{E}[\nabla f_j(w)] = \nabla f(w)$

$$\begin{aligned} \mathbb{E}_t\left[||w^{t+1} - w^*||_2^2\right] &= ||w^t - w^*||_2^2 - 2\gamma \langle \nabla f(w^t), w^t - w^* \rangle + \gamma^2 \mathbb{E}_t\left[||g^t||_2^2\right] \\ &\leq (1-\gamma\mu)||w^t - w^*||_2^2 - 2\gamma(f(w^t) - f(w^*)) + \gamma^2 \mathbb{E}_t\left[||g^t||_2^2\right] \end{aligned}$$

quasi strong conv

Converge to 0 as $w^t \rightarrow w^*$

High Level Proof when $\mathbf{E}[g^t] = \nabla f(w^t)$:

$$
\begin{aligned}
||w^{t+1} - w^*||_2^2 &= ||w^t - w^* - \gamma g^t||_2^2 \\
&= ||w^t - w^*||_2^2 - 2\gamma\langle g^t, w^t - w^*\rangle + \gamma^2||g^t||_2^2.
\end{aligned}
$$

Taking expectation conditioned on respect to w^t

$\mathbb{E}[\nabla f_j(w)] = \nabla f(w)$

$$
\begin{aligned}
\mathbb{E}_t\left[||w^{t+1} - w^*||_2^2\right] &= ||w^t - w^*||_2^2 - 2\gamma\langle \nabla f(w^t), w^t - w^*\rangle + \gamma^2\mathbb{E}_t\left[||g^t||_2^2\right] \\
&\leq (1-\gamma\mu)||w^t - w^*||_2^2 - 2\gamma(f(w^t) - f(w^*)) + \gamma^2\mathbb{E}_t\left[||g^t||_2^2\right]
\end{aligned}
$$

quasi strong conv

Converge to 0 as $w^t \to w^*$

What exactly should g^t be?

Controlled Stochastic Reformulation

Covariate functions:

$$z_i : w \mapsto z_i(w) \in \mathbb{R}, \quad \text{for } i = 1, \ldots, n$$

$$\frac{1}{n}\sum_{i=1}^{n} f_i(w) = \mathbb{E}[f_i(w)] = \mathbb{E}[f_i(w)] - \mathbb{E}[z_i(w)] + \mathbb{E}[z_i(w)]$$

$i \sim \frac{1}{n}$

Controlled Stochastic Reformulation

Covariate functions:

$$z_i \,:\, w \;\mapsto\; z_i(w) \;\in\; \mathbb{R}, \qquad \text{for } i = 1, \ldots, n$$

Cancel out

$$\frac{1}{n}\sum_{i=1}^{n} f_i(w) \;=\; \mathbb{E}[f_{\color{red}{i}}(w)] \;=\; \mathbb{E}[f_{\color{red}{i}}(w)] - \mathbb{E}[z_{\color{red}{i}}(w)] + \mathbb{E}[z_{\color{red}{i}}(w)]$$

$i \sim \frac{1}{n}$

Controlled Stochastic Reformulation

Covariate functions:

$$z_i : w \mapsto z_i(w) \in \mathbb{R}, \quad \text{for } i = 1, \ldots, n$$

Cancel out

$$\frac{1}{n}\sum_{i=1}^{n} f_i(w) = \mathbb{E}[f_i(w)] = \mathbb{E}[f_i(w)] - \mathbb{E}[z_i(w)] + \mathbb{E}[z_i(w)]$$

$$i \sim \frac{1}{n} \qquad = \mathbb{E}\left[f_i(w) - z_i(w) + \mathbb{E}[z_i(w)]\right]$$

Controlled Stochastic Reformulation

Covariate functions:

$$z_i : w \mapsto z_i(w) \in \mathbb{R}, \quad \text{for } i = 1, \ldots, n$$

Cancel out

$$\frac{1}{n}\sum_{i=1}^{n} f_i(w) = \mathbb{E}[f_i(w)] = \mathbb{E}[f_i(w)] - \mathbb{E}[z_i(w)] + \mathbb{E}[z_i(w)]$$

$i \sim \frac{1}{n}$

$$= \mathbb{E}\left[f_i(w) - z_i(w) + \mathbb{E}[z_i(w)]\right]$$

Original finite sum problem

$$\min_{w\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(w)$$

Controlled Stochastic Reformulation

$$\min_{w\in\mathbb{R}^d} \mathbb{E}\left[f_i(w) - z_i(w) + \mathbb{E}[z_i(w)]\right]$$

Use covariates to **control the variance**

Variance reduction as SGD

$$\min_{w \in \mathbb{R}^d} \mathbb{E}\left[f_i(w) - z_i(w) + \mathbb{E}[z_i(w)]\right]$$

Variance reduction as SGD

$$\min_{w\in\mathbb{R}^d} \mathbb{E}\left[f_i(w) - z_i(w) + \mathbb{E}[z_i(w)]\right]$$

$$\text{Sample } i \sim \frac{1}{n}$$

$$w^{t+1} = w^t - \gamma g_i(w^t)$$

Variance reduction as SGD

$$\min_{w\in\mathbb{R}^d} \mathbb{E}\left[f_i(w) - z_i(w) + \mathbb{E}[z_i(w)]\right]$$

Sample $i \sim \frac{1}{n}$

$$w^{t+1} = w^t - \gamma g_i(w^t)$$

$$g_i(w) := \nabla f_i(w) - \nabla z_i(w) + \mathbb{E}[\nabla z_i(w)]$$

Variance reduction as SGD

$$\min_{w\in\mathbb{R}^d} \mathbb{E}\left[f_i(w) - z_i(w) + \mathbb{E}[z_i(w)]\right]$$

Sample $i \sim \frac{1}{n}$

$$w^{t+1} = w^t - \gamma g_i(w^t)$$

By design we have that
$\mathbb{E}[g_i(w^t)] = \nabla f(w^t)$

$$g_i(w) := \nabla f_i(w) - \nabla z_i(w) + \mathbb{E}[\nabla z_i(w)]$$

Variance reduction as SGD

$$\min_{w\in\mathbb{R}^d} \mathbb{E}\left[f_i(w) - z_i(w) + \mathbb{E}[z_i(w)]\right]$$

Sample $i \sim \frac{1}{n}$

$$w^{t+1} = w^t - \gamma g_i(w^t)$$

By design we have that $\mathbb{E}[g_i(w^t)] = \nabla f(w^t)$

$$g_i(w) := \nabla f_i(w) - \nabla z_i(w) + \mathbb{E}[\nabla z_i(w)]$$

How to choose $z_i(w)$?

Noise of covariate estimate

$$\text{Sample } i \sim \frac{1}{n}$$

$$w^{t+1} = w^t - \gamma g_i(w^t)$$

$$\begin{aligned}
\mathbb{E}_i[\|g_i(w)\|^2] &= \mathbb{E}_i[\|\nabla f_i(w) - \nabla z_i(w) + \mathbb{E}[\nabla z_i(w)]\|^2] \\
&= \mathbb{E}_i[\|\nabla f_i(w) - \nabla z_i(w) + \mathbb{E}[\nabla z_i(w) - \nabla f(w)] + \nabla f(w)\|^2] \\
&\leq 2\mathbb{E}_i[\|\nabla f_i(w) - \nabla z_i(w) + \mathbb{E}[\nabla z_i(w) - \nabla f(w)]\|^2 + 2\|\nabla f(w)\|^2 \\
&\leq 2\mathbb{E}_i[\|\nabla f_i(w) - \nabla z_i(w)\|^2 + 2\|\nabla f(w)\|^2
\end{aligned}$$

Noise of covariate estimate

$$\text{Sample } i \sim \frac{1}{n}$$
$$w^{t+1} = w^t - \gamma g_i(w^t)$$

$\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$

$$\begin{aligned}
\mathbb{E}_i[\|g_i(w)\|^2] &= \mathbb{E}_i[\|\nabla f_i(w) - \nabla z_i(w) + \mathbb{E}[\nabla z_i(w)]\|^2] \\
&= \mathbb{E}_i[\|\nabla f_i(w) - \nabla z_i(w) + \mathbb{E}[\nabla z_i(w) - \nabla f(w)] + \nabla f(w)\|^2] \\
&\leq 2\mathbb{E}_i[\|\nabla f_i(w) - \nabla z_i(w) + \mathbb{E}[\nabla z_i(w) - \nabla f(w)]\|^2 + 2\|\nabla f(w)\|^2 \\
&\leq 2\mathbb{E}_i[\|\nabla f_i(w) - \nabla z_i(w)\|^2 + 2\|\nabla f(w)\|^2
\end{aligned}$$

Noise of covariate estimate

$$\text{Sample } i \sim \frac{1}{n}$$
$$w^{t+1} = w^t - \gamma g_i(w^t)$$

$$\begin{aligned}
\mathbb{E}_i[\|g_i(w)\|^2] &= \mathbb{E}_i[\|\nabla f_i(w) - \nabla z_i(w) + \mathbb{E}[\nabla z_i(w)]\|^2] \\
&= \mathbb{E}_i[\|\nabla f_i(w) - \nabla z_i(w) + \mathbb{E}[\nabla z_i(w) - \nabla f(w)] + \nabla f(w)\|^2] \\
&\leq 2\mathbb{E}_i[\|\nabla f_i(w) - \nabla z_i(w) + \mathbb{E}[\nabla z_i(w) - \nabla f(w)]\|^2 + 2\|\nabla f(w)\|^2 \\
&\leq 2\mathbb{E}_i[\|\nabla f_i(w) - \nabla z_i(w)\|^2 + 2\|\nabla f(w)\|^2
\end{aligned}$$

$\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$

$\mathbb{E}[\|X - E[X]\|^2] \leq \mathbb{E}[\|X\|^2$

where$X := \nabla f_i(w) - \nabla z_i(w)$

Noise of covariate estimate

Sample $i \sim \frac{1}{n}$

$$w^{t+1} = w^t - \gamma g_i(w^t)$$

$$\begin{aligned}
\mathbb{E}_i[\|g_i(w)\|^2] &= \mathbb{E}_i[\|\nabla f_i(w) - \nabla z_i(w) + \mathbb{E}[\nabla z_i(w)]\|^2] \\
&= \mathbb{E}_i[\|\nabla f_i(w) - \nabla z_i(w) + \mathbb{E}[\nabla z_i(w) - \nabla f(w)] + \nabla f(w)\|^2] \\
&\leq 2\mathbb{E}_i[\|\nabla f_i(w) - \nabla z_i(w) + \mathbb{E}[\nabla z_i(w) - \nabla f(w)]\|^2 + 2\|\nabla f(w)\|^2 \\
&\leq 2\mathbb{E}_i[\|\nabla f_i(w) - \nabla z_i(w)\|^2 + 2\|\nabla f(w)\|^2
\end{aligned}$$

$\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$

$\mathbb{E}[\|X - E[X]\|^2] \leq \mathbb{E}[\|X\|^2$

where $X := \nabla f_i(w) - \nabla z_i(w)$

Converge to 0 as $w^t \to w^*$

Noise of covariate estimate

Sample $i \sim \frac{1}{n}$

$$w^{t+1} = w^t - \gamma g_i(w^t)$$

$$\begin{aligned}
\mathbb{E}_i[\|g_i(w)\|^2] &= \mathbb{E}_i[\|\nabla f_i(w) - \nabla z_i(w) + \mathbb{E}[\nabla z_i(w)]\|^2] \\
&= \mathbb{E}_i[\|\nabla f_i(w) - \nabla z_i(w) + \mathbb{E}[\nabla z_i(w) - \nabla f(w)] + \nabla f(w)\|^2] \\
&\leq 2\mathbb{E}_i[\|\nabla f_i(w) - \nabla z_i(w) + \mathbb{E}[\nabla z_i(w) - \nabla f(w)]\|^2 + 2\|\nabla f(w)\|^2 \\
&\leq 2\mathbb{E}_i[\|\nabla f_i(w) - \nabla z_i(w)\|^2 + 2\|\nabla f(w)\|^2
\end{aligned}$$

$\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$

$\mathbb{E}[\|X - E[X]\|^2] \leq \mathbb{E}[\|X\|^2$

where$X := \nabla f_i(w) - \nabla z_i(w)$

$\nabla z_i(w) \approx \nabla f_i(w)$

Converge to 0 as $w^t \to w^*$

Choosing the covariate as a linear approximation

We would like: $\nabla z_i(w) \approx \nabla f_i(w)$

Choosing the covariate as a linear approximation

We would like:

$$\nabla z_i(w) \approx \nabla f_i(w)$$

Choosing the covariate as a linear approximation

We would like:

$$\nabla z_i(w) \approx \nabla f_i(w)$$

Expensive to compute for all i

Choosing the covariate as a linear approximation

We would like:

$$\nabla z_i(w) \approx \nabla f_i(w)$$

Expensive to compute for all *i*

Use snapshot: $\nabla z_i(w) = \nabla f_i(\tilde{w})$

Reference point. Rarely update

Choosing the covariate as a linear approximation

We would like:

$$\nabla z_i(w) \approx \nabla f_i(w)$$

Expensive to compute for all *i*

Use snapshot: $\nabla z_i(w) = \nabla f_i(\tilde{w})$

Reference point. Rarely update

$$\text{If } f_i(w) \text{ is } L_{\max}\text{–smooth}$$

$$\Downarrow$$

$$\|\nabla f_i(w) - \nabla f_i(\tilde{w})\| \leq L_{\max}\|w - \tilde{w}\|$$

Choosing the covariate as a linear approximation

We would like:

$$\nabla z_i(w) \approx \nabla f_i(w)$$

Expensive to compute for all *i*

Use snapshot: $\nabla z_i(w) = \nabla f_i(\tilde{w})$

Reference point. Rarely update

If $f_i(w)$ is $L_{\max}$–smooth

$$\|\nabla f_i(w) - \nabla f_i(\tilde{w})\| \leq L_{\max}\|w - \tilde{w}\|$$

$$\mathbb{E}_i[\|g_i(w)\|^2] \quad \leq \quad \mathbb{E}_i[\|w - \tilde{w}\|^2 + 2\|\nabla f(w)\|^2$$

Choosing the covariate as a linear approximation

We would like: $\nabla z_i(w) \approx \nabla f_i(w)$

Expensive to compute for all i

Use snapshot: $\nabla z_i(w) = \nabla f_i(\tilde{w})$

Reference point. Rarely update

If $f_i(w)$ is $L_{\max}$–smooth

$$\|\nabla f_i(w) - \nabla f_i(\tilde{w})\| \leq L_{\max}\|w - \tilde{w}\|$$

But update frequently enough to control noise

$$\mathbb{E}_i[\|g_i(w)\|^2] \quad \leq \quad \mathbb{E}_i[\|w - \tilde{w}\|^2 + 2\|\nabla f(w)\|^2$$

SVRG: Stochastic Variance reduced method gradient

Johnson & Zhang, 2013 NIPS

$$w^{t+1} = w^t - \gamma g_i(w^t)$$

Reference point $\tilde{w} \in \mathbb{R}^d$

Sample $\nabla f_i(w^t), \quad$ i.i.d sample with prob $\frac{1}{n}$

Grad. estimate $g_i(w^t) = \nabla f_i(w^t) - \nabla f_i(\tilde{w}) + \nabla f(\tilde{w})$

$\nabla z_i(w^t) = \nabla f_i(\tilde{w})$

$\mathbb{E}[\nabla z_i(w^t)]$

free-SVRG: Stochastic Variance Reduced Gradients

Jonhson & Zhang NIPS 2013

Sebbouh, et. al 2019 Neurips 2019

Set $\tilde{w}^0 = 0 = x_0^m$, choose $\gamma > 0, m \in \mathbb{N}$,

$\alpha_t > 0$ with $\sum_{t=0}^{m-1} \alpha_t = 1$

for $s = 1, 2, \ldots, T$

$x_s^0 = x_{s-1}^m$

for $t = 0, 1, 2, \ldots, m-1$

i.i.d sample $i \sim \frac{1}{n}$

$g^t = \nabla f_i(x_s^t) - \nabla f_i(\tilde{w}^{s-1}) + \nabla f(\tilde{w}^{s-1})$

$x_s^{t+1} = x_s^t - \gamma g^t$

$\tilde{w}^{s+1} = \sum_{t=0}^{m-1} \alpha_t x_s^t$

Output $\tilde{w}^{T+1}$

Most iterates cost *O(1)*

Tune inner loop size *m*

free-SVRG: Stochastic Variance Reduced Gradients

Jonhson & Zhang NIPS 2013

Sebbouh, et. al 2019 Neurips 2019

Set $\tilde{w}^0 = 0 = x_0^m$, choose $\gamma > 0, m \in \mathbb{N}$,

$\alpha_t > 0$ with $\sum_{t=0}^{m-1} \alpha_t = 1$

for $s = 1, 2, \ldots, T$

$x_s^0 = x_{s-1}^m$

for $t = 0, 1, 2, \ldots, m-1$

i.i.d sample $i \sim \frac{1}{n}$

$g^t = \nabla f_i(x_s^t) - \nabla f_i(\tilde{w}^{s-1}) + \nabla f(\tilde{w}^{s-1})$

$x_s^{t+1} = x_s^t - \gamma g^t$

$\tilde{w}^{s+1} = \sum_{t=0}^{m-1} \alpha_t x_s^t$

Output $\tilde{w}^{T+1}$

Adding indices in *t* and *s*

Most iterates cost *O(1)*

Tune inner loop size *m*

free-SVRG: Stochastic Variance Reduced Gradients

Jonhson & Zhang NIPS 2013

Sebbouh, et. al 2019 Neurips 2019

Set $\tilde{w}^0 = 0 = x_0^m$, choose $\gamma > 0, m \in \mathbb{N}$,

$\alpha_t > 0$ with $\sum_{t=0}^{m-1} \alpha_t = 1$

for $s = 1, 2, \ldots, T$

$x_s^0 = x_{s-1}^m$

for $t = 0, 1, 2, \ldots, m-1$

i.i.d sample $i \sim \frac{1}{n}$

$g^t = \nabla f_i(x_s^t) - \nabla f_i(\tilde{w}^{s-1}) + \nabla f(\tilde{w}^{s-1})$

$x_s^{t+1} = x_s^t - \gamma g^t$

$\tilde{w}^{s+1} = \sum_{t=0}^{m-1} \alpha_t x_s^t$

Output $\tilde{w}^{T+1}$

Adding indices in *t* and *s*

Reference point is an average of inner iterates

Most iterates cost *O(1)*

Tune inner loop size *m*

SAGA: Stochastic Average Gradient

Defazio, Bach, & Lacoste-Julien, 2014 NIPs

$$w^{t+1} = w^t - \gamma g_i(w^t)$$

Sample

$$\nabla f_i(w^t), \qquad \text{i.i.d sample with prob } \frac{1}{n}$$

Grad. estimate

$$g_i(w^t) = \nabla f_i(w^t) - \nabla f_i(w^{t_i}) + \frac{1}{n}\sum_{j=1}^{n} \nabla f_j(w^{t_j})$$

Store grad.

$$\nabla f_i(w^{t_i}) = \nabla f_i(w^t)$$

SAGA: Stochastic Average Gradient

Defazio, Bach, & Lacoste-Julien, 2014 NIPs

$$w^{t+1} = w^t - \gamma g_i(w^t)$$

Sample

$$\nabla f_i(w^t), \quad \text{i.i.d sample with prob } \frac{1}{n}$$

Grad. estimate

$$g_i(w^t) = \nabla f_i(w^t) - \nabla f_i(w^{t_i}) + \frac{1}{n}\sum_{j=1}^{n} \nabla f_j(w^{t_j})$$

$$\nabla z_i(w^t) = \nabla f_i(w^{t_i})$$

Store grad.

$$\nabla f_i(w^{t_i}) = \nabla f_i(w^t)$$

SAGA: Stochastic Average Gradient

Defazio, Bach, & Lacoste-Julien, 2014 NIPs

$$w^{t+1} = w^t - \gamma g_i(w^t)$$

Sample

$$\nabla f_i(w^t), \qquad \text{i.i.d sample with prob } \frac{1}{n}$$

Grad. estimate

$$g_i(w^t) = \nabla f_i(w^t) - \nabla f_i(w^{t_i}) + \frac{1}{n}\sum_{j=1}^{n} \nabla f_j(w^{t_j})$$

$$\nabla z_i(w^t) = \nabla f_i(w^{t_i})$$

$$\mathbb{E}[\nabla z_i(w^t)]$$

Store grad.

$$\nabla f_i(w^{t_i}) = \nabla f_i(w^t)$$

SAGA: Stochastic Average Gradient

Defazio, Bach, & Lacoste-Julien, 2014 NIPs

$$w^{t+1} = w^t - \gamma g_i(w^t)$$

Sample

$$\nabla f_i(w^t), \qquad \text{i.i.d sample with prob } \frac{1}{n}$$

Grad. estimate

$$g_i(w^t) = \nabla f_i(w^t) - \nabla f_i(w^{t_i}) + \frac{1}{n}\sum_{j=1}^{n} \nabla f_j(w^{t_j})$$

$$z_i(w) = f_i(w^{t_i}) + \langle \nabla f_i(w^{t_i}), w - w^{t_i} \rangle$$

$\nabla z_i(w^t) = \nabla f_i(w^{t_i})$

$\mathbb{E}[\nabla z_i(w^t)]$

Store grad.

$$\nabla f_i(w^{t_i}) = \nabla f_i(w^t)$$

SAGA: Stochastic Average Gradient

Set $w^0 = 0, g_i = \nabla f_i(w^0),$ for $i = 1\ldots, n$

Choose $\gamma > 0$

for $t = 0, 1, 2, \ldots, T-1$

  sample $i \in \{1, \ldots, n\}$

  $g^t = \nabla f_i(w^t) - g_i + \frac{1}{n}\sum_{j=1}^{n} g_j$

  $w^{t+1} = w^t - \gamma g^t$

  $g_i = \nabla f_i(w^t)$

Output w^T

No inner loop, rolling update

Stores a $d \times n$ matrix

SAG: Stochastic Average Gradient (Biased version)

M. Schmidt, N. Le Roux, F. Bach (2016), Math prog

$$w^{t+1} = w^t - \gamma g_i(w^t)$$

Sample $\nabla f_i(w^t), \quad \text{i.i.d sample with prob } \frac{1}{n}$

Grad. estimate $g_i(w^t) = \frac{1}{n}\sum_{j=1}^{n} \nabla f_j(w^{t_j})$

$$\mathbb{E}[g^t] \neq \nabla f(w^t)$$

$$\cancel{g_i(w) := \nabla f_i(w) - \nabla z_i(w) + \mathbb{E}[\nabla z_i(w)]}$$

Store grad. $\nabla f_i(w^{t_i}) = \nabla f_i(w^t)$

SAG: Stochastic Average Gradient

Set $w^0 = 0, g_i = \nabla f_i(w^0),$ for $i = 1, \ldots, n$

Choose $\gamma > 0$

for $t = 0, 1, 2, \ldots, T-1$

  sample $i \in \{1, \ldots, n\}$

  $g_i = \nabla f_i(w^t)$ (update grad)

  $g^t = \frac{1}{n} \sum_{j=1}^{n} g_j$

  $w^{t+1} = w^t - \gamma g^t$

Output w^T

Very easy to implement

Stores a $d \times n$ matrix

SAG: Stochastic Average Gradient

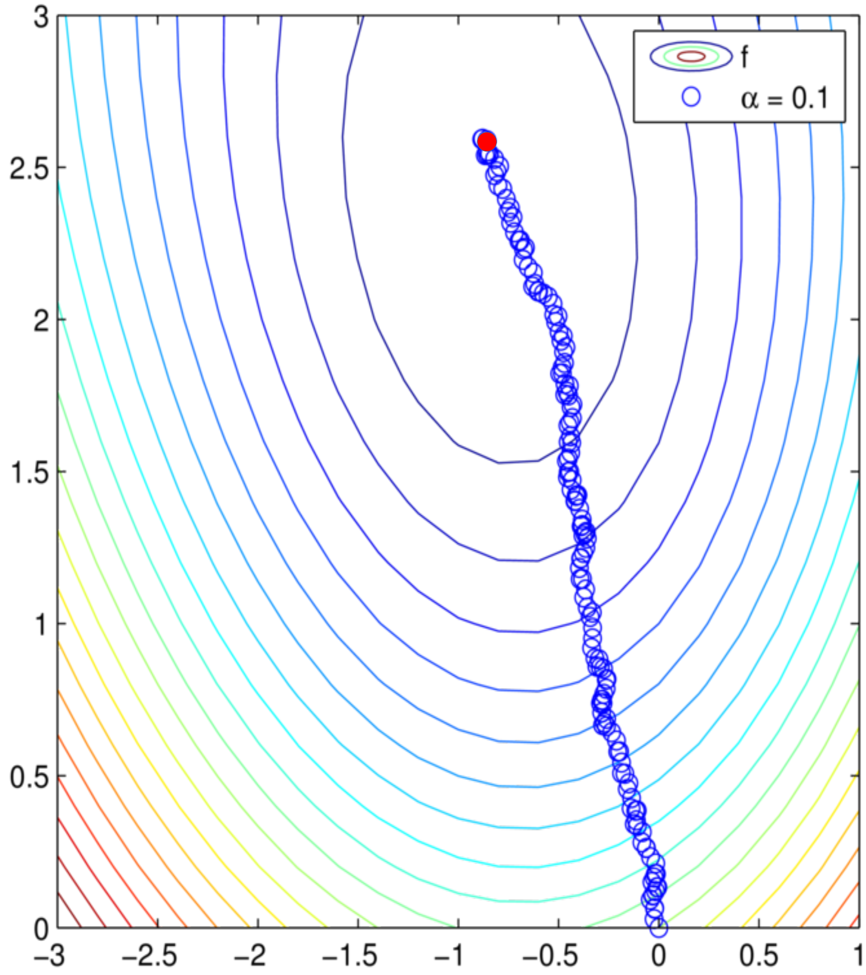
$$
\begin{array}{l}
\text{Set } w^0 = 0, g_i = \nabla f_i(w^0), \text{ for } i = 1, \ldots, n \\
\text{Choose } \gamma > 0 \\
\text{for } t = 0, 1, 2, \ldots, T-1 \\
\qquad \text{sample } i \in \{1, \ldots, n\} \\
\qquad g_i = \nabla f_i(w^t) \qquad \text{(update grad)} \\
\qquad g^t = \frac{1}{n} \sum_{j=1}^{n} g_j \\
\qquad w^{t+1} = w^t - \gamma g^t \\
\text{Output } w^T
\end{array}
$$

Very easy to implement

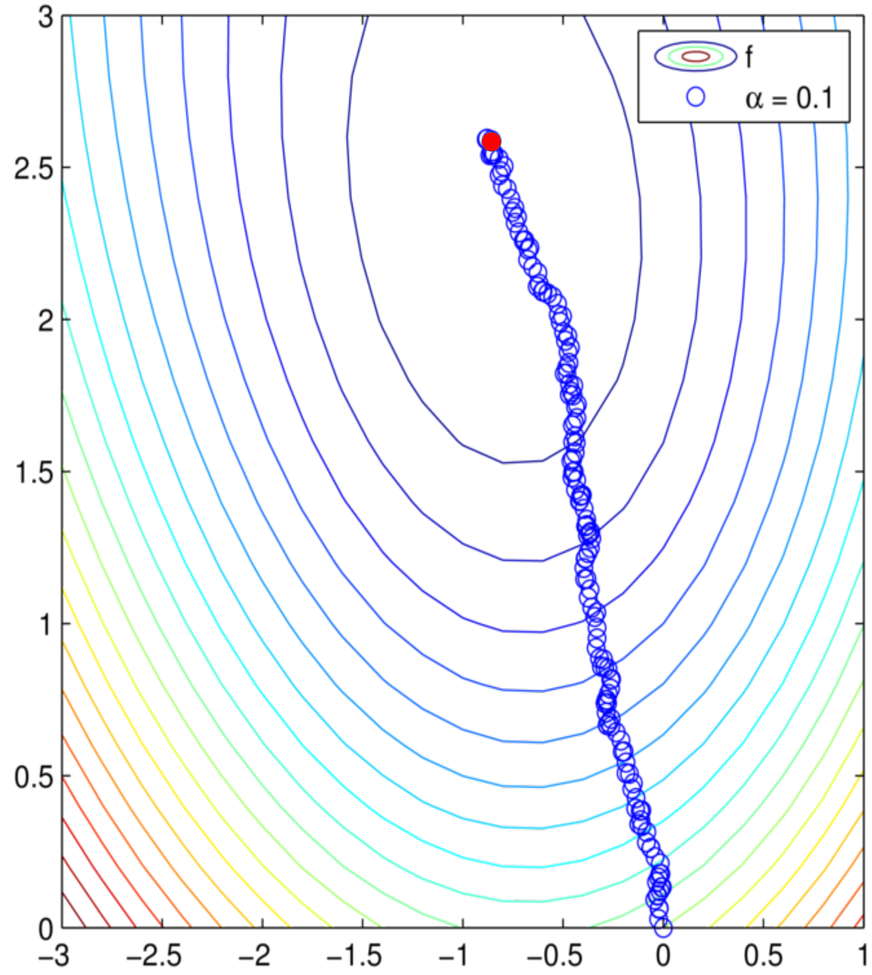
Stores a $d \times n$ matrix

EXE: Introduce a variable $G = (1/n) \sum_{j=1} g_j$. Re-write the SAG algorithm so G is updated efficiently at each iteration.

The Stochastic Average Gradient



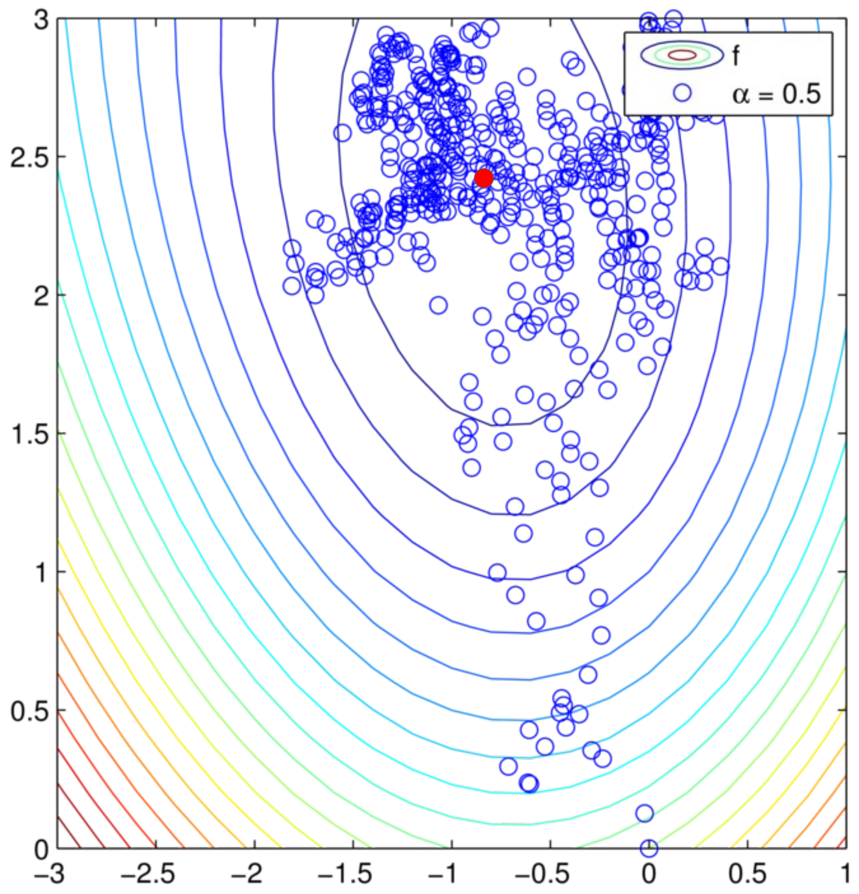
The Stochastic Average Gradient



How to prove this converges? Is this the only option?

Stochastic Gradient Descent

$\alpha = 0.5$



Convergence Theorems

Assumptions for Convergence

Strong Convexity

$$f(w) \geq f(y) + \langle \nabla f(y), w - y \rangle + \frac{\mu}{2} ||w - y||_2^2$$

Smoothness + convexity

$$f_i(w) \leq f_i(y) + \langle \nabla f_i(y), w - y \rangle + \frac{L_i}{2} ||w - y||_2^2$$

$$f_i(w) \geq f_i(y) + \langle \nabla f_i(y), w - y \rangle \qquad \text{for } i = 1, \ldots, n$$

$$L_{\max} := \max_{i=1,\ldots,n} L_i$$

Convergence SAG

Theorem SAG

If $f(w)$ is μ–strongly convex, $f_i(w)$ is cvx & $L_{\max}$–smooth and $\alpha = 1/(16L_{\max})$ then

$$\mathbb{E}\left[||w^t - w^*||_2^2\right] \leq \left(1 - \min\left\{\frac{1}{8n}, \frac{\mu}{16L_{\max}}\right\}\right)^t C_0$$

where $C_0 = \frac{3}{2}(f(w^0) - f(w^*)) + \frac{4L_{\max}}{n}||w^0 - w^*||_2^2 \geq 0$

A practical convergence result!

Because of biased gradients, difficult proof that relies on computer assisted steps

M. Schmidt, N. Le Roux, F. Bach (2016)
Mathematical Programming
Minimizing Finite Sums with the Stochastic Average Gradient.

Convergence SAGA

Theorem SAGA

If $f(w)$ is μ–strongly convex, $f_i(w)$ is cvx & $L_{\max}$–smooth and $\alpha = 1/(3L_{\max})$ then

$$\mathbb{E}\left[||w^t - w^*||_2^2\right] \leq \left(1 - \min\left\{\frac{1}{4n}, \frac{\mu}{3L_{\max}}\right\}\right)^t C_0$$

where $C_0 = \dfrac{2n}{3L_{\max}}(f(w^0) - f(w^*)) + ||w^0 - w^*||_2^2 \geq 0$

An even more practical convergence result!

Much easier prove due to unbiased estimate

A. Defazio, F. Bach and J. Lacoste-Julien (2014) NIPS, **SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives.**

free-SVRG: Stochastic Variance Reduced Gradients

Jonhson & Zhang NIPS 2013

Sebbouh, et. al 2019 Neurips 2019

Set $\tilde{w}^0 = 0 = x_0^m$, choose $\gamma > 0, m \in \mathbb{N}$,

$\alpha_t > 0$ with $\sum_{t=0}^{m-1} \alpha_t = 1$

for $s = 1, 2, \ldots, T$

$x_s^0 = x_{s-1}^m$

for $t = 0, 1, 2, \ldots, m-1$

i.i.d sample $i \sim \frac{1}{n}$

$g^t = \nabla f_i(x_s^t) - \nabla f_i(\tilde{w}_{s-1}) + \nabla f(\tilde{w}_{s-1})$

$x_s^{t+1} = x_s^t - \gamma g^t$

$\tilde{w}^{s+1} = \sum_{t=0}^{m-1} \alpha_t x_s^t$

Output $\tilde{w}^{T+1}$

Adding indices in *k* and *t*

Most iterates cost *O(1)*

Tune inner loop size *m*

free-SVRG: Stochastic Variance Reduced Gradients

Jonhson & Zhang NIPS 2013

Sebbouh, et. al 2019 Neurips 2019

Set $\tilde{w}^0 = 0 = x_0^m$, choose $\gamma > 0, m \in \mathbb{N}$,

$\alpha_t > 0$ with $\sum_{t=0}^{m-1} \alpha_t = 1$

for $s = 1, 2, \ldots, T$

$x_s^0 = x_{s-1}^m$

for $t = 0, 1, 2, \ldots, m-1$

i.i.d sample $i \sim \frac{1}{n}$

$g^t = \nabla f_i(x_s^t) - \nabla f_i(\tilde{w}_{s-1}) + \nabla f(\tilde{w}_{s-1})$

$x_s^{t+1} = x_s^t - \gamma g^t$

$\tilde{w}^{s+1} = \sum_{t=0}^{m-1} \alpha_t x_s^t$

Output $\tilde{w}^{T+1}$

Adding indices in *k* and *t*

$$\alpha_k = \frac{(1-\gamma\mu)^{m-1-t}}{\sum_{i=0}^{m-1}(1-\gamma\mu)^{m-1-i}}$$

Most iterates cost *O(1)*

Tune inner loop size *m*

Convergence Theorem for SVRG

Theorem

If $f(w)$ is μ–strongly convex, $f_i(w)$ is $L_{\max}$–smooth

$$\Psi(x, \tilde{w}) := \|x - w^*\|^2 + cnst \times \left(f(\tilde{w}) - f(w^*)\right)$$

where $cnst := 8L_{\max}\gamma^2 \sum_{i=1}^{m-1} (1-\gamma\mu)^i$

Sebbouh, Gazagnadou, Jelassi, Bach, G, 2019

Convergence Theorem for SVRG

Theorem

If $f(w)$ is μ–strongly convex, $f_i(w)$ is $L_{\max}$–smooth

$$\Psi(x, \tilde{w}) := \|x - w^*\|^2 + cnst \times \left(f(\tilde{w}) - f(w^*)\right)$$

If $\gamma \leq \frac{1}{6L_{\max}}$ then

$$\mathbb{E}[\Psi(x_s^m, \tilde{w}_s)] \leq \max\left\{(1-\gamma\mu)^m, \frac{1}{2}\right\}^t \Psi(x_0^0, \tilde{w}_0)$$

where $cnst := 8L_{\max}\gamma^2 \sum_{i=1}^{m-1} (1-\gamma\mu)^i$

Sebbouh, Gazagnadou, Jelassi, Bach, G, 2019

Convergence Theorem for SVRG

Theorem

If $f(w)$ is μ–strongly convex, $f_i(w)$ is $L_{\max}$–smooth

$$\Psi(x, \tilde{w}) := \|x - w^*\|^2 + cnst \times \left(f(\tilde{w}) - f(w^*)\right)$$

If $\gamma \leq \frac{1}{6L_{\max}}$ then

$$\Rightarrow \quad \mathbb{E}[\Psi(x_s^m, \tilde{w}_s)] \;\leq\; \max\left\{(1-\gamma\mu)^m, \; \frac{1}{2}\right\}^t \Psi(x_0^0, \tilde{w}_0)$$

where $cnst := 8L_{\max}\gamma^2 \sum_{i=1}^{m-1} (1-\gamma\mu)^i$

Free to choose the number of inner iterates *m*

Sebbouh, Gazagnadou, Jelassi, Bach, G, 2019

Convergence Theorem for SVRG

Theorem

If $f(w)$ is μ–strongly convex, $f_i(w)$ is $L_{\max}$–smooth

$$\Psi(x,\tilde{w}) := \|x - w^*\|^2 + cnst \times \left(f(\tilde{w}) - f(w^*)\right)$$

If $\gamma \leq \frac{1}{6L_{\max}}$ then

$$\Rightarrow \quad \mathbb{E}[\Psi(x_s^m, \tilde{w}_s)] \leq \max\left\{(1-\gamma\mu)^m, \frac{1}{2}\right\}^t \Psi(x_0^0, \tilde{w}_0)$$

where $cnst := 8L_{\max}\gamma^2 \sum_{i=1}^{m-1}(1-\gamma\mu)^i$

Free to choose the number of inner iterates *m*

Corollary If $\gamma = 1/6L_{\max}$ and $m = n$

$$t = O\left(\frac{6}{m}\frac{L_{\max}}{\mu}\right)\log\left(\frac{1}{\epsilon}\right) \quad \Rightarrow \quad \frac{\mathbb{E}[\|x_t^m - w^*\|^2]}{\Psi(x_0^0, \tilde{w}^0)} \leq \epsilon$$

Sebbouh, Gazagnadou, Jelassi, Bach, G, 2019

Comparisons in total complexity for strongly convex

Approximate solution

$$\mathbb{E}[f(w^T)] - f(w^*) \leq \epsilon \quad \text{or} \quad \mathbb{E}\|w^t - w^*\|^2 \leq \epsilon$$

SGD

$$O\left(\frac{1}{\epsilon}\right)$$

Gradient descent

$$O\left(\frac{nL}{\mu}\log\left(\frac{1}{\epsilon}\right)\right)$$

SVRG/SAGA/SAG

$$O\left(\left(n + \frac{L_{\max}}{\mu}\right)\log\left(\frac{1}{\epsilon}\right)\right)$$

Variance reduction faster than GD when $L \geq \mu + L_{\max}/n$

How did I get these complexity results from the convergence results?

Section 1.3.5, R.M. Gower, Ph.d thesis: Sketch and Project: Randomized Iterative Methods for Linear Systems and Inverting Matrices University of Edinburgh, 2016

Practicals implementation of SAG for Linear Classifiers

Finite Sum Training Problem

$$\min_{w\in\mathbf{R}^d} \frac{1}{n}\sum_{i=1}^{n} \ell\left(\langle w, x^i\rangle, y^i\right) + \frac{\lambda}{2}||w||_2^2$$

L2 regularizor + linear hypothesis

Practicals implementation of SAG for Linear Classifiers

L2 regularizor + linear hypothesis

Finite Sum Training Problem

$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell\left(\langle w, x^i \rangle, y^i\right) + \frac{\lambda}{2}||w||_2^2$$

$$\nabla f_i(w) = \ell'(\langle w, x^i \rangle, y^i) x^i + \lambda w$$

Practicals implementation of SAG for Linear Classifiers

L2 regularizor + linear hypothesis

Finite Sum Training Problem

$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell\left(\langle w, x^i \rangle, y^i\right) + \frac{\lambda}{2}||w||_2^2$$

$$\nabla f_i(w) = \underbrace{\ell'(\langle w, x^i \rangle, y^i)}_{\text{Nonlinear in } w} x^i + \underbrace{\lambda w}_{\text{Linear in } w}$$

Practicals implementation of SAG for Linear Classifiers

L2 regularizor + linear hypothesis

Finite Sum Training Problem

$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell\left(\langle w, x^i \rangle, y^i\right) + \frac{\lambda}{2} ||w||_2^2$$

$$\nabla f_i(w) = \underbrace{\ell'(\langle w, x^i \rangle, y^i)}_{\text{Nonlinear in } w} x^i + \underbrace{\lambda w}_{\text{Linear in } w}$$

Reduce Storage to *O(n)*

Only store real number

$$\beta_i = \ell'(\langle w^{t_i}, x^i \rangle, y^i)$$

Stoch. gradient estimate

$$\nabla f_i(w^{t_i}) = \beta_i x^i + \lambda w^t$$

Full gradient estimate

$$g^t = \frac{1}{n} \sum_{j=1}^{n} \beta_j x_j + \lambda w^t$$

Proving Convergence of SVRG

Proof:

$$\begin{aligned} ||x_s^{t+1} - w^*||_2^2 &= ||x_s^t - w^* - \gamma g^t||_2^2 \\ &= ||x_s^t - w^*||_2^2 - 2\gamma\langle g^t, x_s^t - w^*\rangle + \gamma^2||g^t||_2^2. \end{aligned}$$

Taking expectation with respect to *j*

Unbiased estimator

$$\begin{aligned} \mathbb{E}_j\left[||x_s^{t+1} - w^*||_2^2\right] &= ||x_s^t - w^*||_2^2 - 2\gamma\langle \nabla f(x_s^t), x_s^t - w^*\rangle + \gamma^2\mathbb{E}_j\left[||g^t||_2^2\right] \\ &\overset{\text{str. conv.}}{\leq} (1-\mu\gamma)||x_s^t - w^*||_2^2 - 2\gamma(f(x_s^t) - f(w^*)) + \gamma^2\mathbb{E}_j\left[||g^t||_2^2\right] \end{aligned}$$

Need to bound this!

$$\boxed{\mathbb{E}_j\left[||g^t||_2^2\right]}$$

Smoothness Consequences I

Smoothness

$$f(w) \leq f(y) + \langle \nabla f(y), w - y \rangle + \frac{L}{2}||w - y||_2^2, \quad \text{for } i = 1, \dots, n$$

EXE: Lemma 1

$$f(y - \frac{1}{L}\nabla f(y)) - f(y) \leq -\frac{1}{2L}||\nabla f(y)||_2^2, \quad \forall y.$$

Proof:

Substituting $w = y - \dfrac{1}{L}\nabla f(y)$ into the smoothness inequality gives

$$\begin{aligned} f(y - \tfrac{1}{L}\nabla f(y)) - f(y) \quad &\leq \quad \langle \nabla f(y), -\frac{1}{L}\nabla f(y)\rangle + \frac{L}{2}|| - \frac{1}{L}\nabla f(y)||_2^2 \\ &= \quad -\tfrac{1}{2L}||\nabla f(y)||_2^2. \quad \blacksquare \end{aligned}$$

Smoothness Consequences II

Smoothness

$$f_i(w) \leq f_i(y) + \langle \nabla f_i(y), w - y\rangle + \frac{L_i}{2}||w-y||_2^2, \quad \text{for } i = 1,\ldots,n$$

EXE: Lemma 2

$$\mathbb{E}[||\nabla f_i(w) - \nabla f_i(w^*)||_2^2] \leq 2L_{\max}(f(w) - f(w^*))$$

Proof: Let $g_i(w) = f_i(w) - f_i(w^*) - \langle \nabla f_i(w^*), w - w^*\rangle$ which is L_i–smooth.

Smoothness Consequences II

Smoothness

$$f_i(w) \leq f_i(y) + \langle \nabla f_i(y), w - y \rangle + \frac{L_i}{2}||w - y||_2^2, \quad \text{for } i = 1, \ldots, n$$

EXE: Lemma 2

$$\mathbb{E}[||\nabla f_i(w) - \nabla f_i(w^*)||_2^2] \leq 2L_{\max}(f(w) - f(w^*))$$

Proof: Let $g_i(w) = f_i(w) - f_i(w^*) - \langle \nabla f_i(w^*), w - w^* \rangle$ which is L_i–smooth.

Smoothness Consequences II

Smoothness

$$f_i(w) \leq f_i(y) + \langle \nabla f_i(y), w - y \rangle + \frac{L_i}{2}||w-y||_2^2, \quad \text{for } i = 1, \ldots, n$$

EXE: Lemma 2

$$\mathbb{E}[||\nabla f_i(w) - \nabla f_i(w^*)||_2^2] \leq 2L_{\max}(f(w) - f(w^*))$$

Proof: Let $g_i(w) = f_i(w) - f_i(w^*) - \langle \nabla f_i(w^*), w - w^* \rangle$ which is L_i–smooth.

Convexity of $f_i(w) \Rightarrow g_i(w) \geq 0$ for all w. From Lemma 1 we have

$$g_i(w) \geq g_i(w) - g_i(w - \frac{1}{L_i}\nabla g_i(w)) \underbrace{\geq}_{\text{Lemma 1}} \frac{1}{2L_i}||\nabla g_i(w)||_2^2 \geq \frac{1}{2L_{\max}}||\nabla g_i(w)||_2^2$$

Inserting definition of $g_i(w)$ we have

$$\frac{1}{2L_{\max}}||\nabla f_i(w) - \nabla f_i(w^*)||_2^2 \leq f_i(w) - f_i(w^*) - \langle \nabla f_i(w^*), w - w^* \rangle$$

Result follows by taking expectation of i. ∎

Bounding gradient estimate

$$g^t = \nabla f_i(x^t) - \nabla f_i(\tilde{w}) + \nabla f(\tilde{w})$$

EXE: Lemma 3

$$\mathbb{E}[||g^t||_2^2] \leq 4L_{\max}(f(x^t) - f(w^*)) + 4L_{\max}(f(\tilde{w}) - f(w^*))$$

Proof: Hint: use $||a+b||_2^2 \leq 2||a||_2^2 + 2||b||_2^2$ and Lemma 2

Bounding gradient estimate

$g^t = \nabla f_i(x^t) - \nabla f_i(\tilde{w}) + \nabla f(\tilde{w})$

EXE: Lemma 3

$$\mathbb{E}[||g^t||_2^2] \leq 4L_{\max}(f(x^t) - f(w^*)) + 4L_{\max}(f(\tilde{w}) - f(w^*))$$

Proof: Hint: use $||a+b||_2^2 \leq 2||a||_2^2 + 2||b||_2^2$ and Lemma 2

Bounding gradient estimate

$g^t = \nabla f_i(x^t) - \nabla f_i(\tilde{w}) + \nabla f(\tilde{w})$

EXE: Lemma 3

$$\mathbb{E}[||g^t||_2^2] \leq 4L_{\max}(f(x^t) - f(w^*)) + 4L_{\max}(f(\tilde{w}) - f(w^*))$$

Proof: Hint: use $||a+b||_2^2 \leq 2||a||_2^2 + 2||b||_2^2$ and Lemma 2

$$\begin{aligned}
\mathbb{E}_j[||g^t||_2^2] &= \mathbb{E}_j[||\nabla f_i(x^t) - \nabla f_i(w^*) + \nabla f_i(w^*) - \nabla f_i(\tilde{w}) + \nabla f(\tilde{w})||_2^2] \\
&\leq 2\mathbb{E}_j[||\nabla f_i(x^t) - \nabla f_i(w^*)||_2^2] + 2\mathbb{E}_j[||\nabla f_i(w^*) - \nabla f_i(\tilde{w}) + \nabla f(\tilde{w})||_2^2] \\
&\leq 2\mathbb{E}_j[||\nabla f_i(x^t) - \nabla f_i(w^*)||_2^2] + 2\mathbb{E}_j[||\nabla f_i(w^*) - \nabla f_i(\tilde{w})||_2^2] \\
&= 4L_{\max}\left(f(x^t) - f(w^*) + f(\tilde{w}) - f(w^*)\right) \qquad \blacksquare
\end{aligned}$$

Lemma 2

Bounding gradient estimate

$g^t = \nabla f_i(x^t) - \nabla f_i(\tilde{w}) + \nabla f(\tilde{w})$

EXE: Lemma 3

$$\mathbb{E}[||g^t||_2^2] \leq 4L_{\max}(f(x^t) - f(w^*)) + 4L_{\max}(f(\tilde{w}) - f(w^*))$$

Proof: Hint: use $||a+b||_2^2 \leq 2||a||_2^2 + 2||b||_2^2$ and Lemma 2

$$\begin{aligned}
\mathbb{E}_j[||g^t||_2^2] &= \mathbb{E}_j[||\nabla f_i(x^t) - \nabla f_i(w^*) + \nabla f_i(w^*) - \nabla f_i(\tilde{w}) + \nabla f(\tilde{w})||_2^2] \\
&\leq 2\mathbb{E}_j[||\nabla f_i(x^t) - \nabla f_i(w^*)||_2^2] + 2\mathbb{E}_j[||\nabla f_i(w^*) - \nabla f_i(\tilde{w}) + \nabla f(\tilde{w})||_2^2] \\
&\leq 2\mathbb{E}_j[||\nabla f_i(x^t) - \nabla f_i(w^*)||_2^2] + 2\mathbb{E}_j[||\nabla f_i(w^*) - \nabla f_i(\tilde{w})||_2^2] \\
&= 4L_{\max}\left(f(x^t) - f(w^*) + f(\tilde{w}) - f(w^*)\right) \qquad \blacksquare
\end{aligned}$$

Lemma 2

Bounding gradient estimate

$g^t = \nabla f_i(x^t) - \nabla f_i(\tilde{w}) + \nabla f(\tilde{w})$

EXE: Lemma 3

$$\mathbb{E}[||g^t||_2^2] \leq 4L_{\max}(f(x^t) - f(w^*)) + 4L_{\max}(f(\tilde{w}) - f(w^*))$$

Proof: Hint: use $||a+b||_2^2 \leq 2||a||_2^2 + 2||b||_2^2$ and Lemma 2

$$\begin{aligned}
\mathbb{E}_j[||g^t||_2^2] &= \mathbb{E}_j[||\nabla f_i(x^t) - \nabla f_i(w^*) + \nabla f_i(w^*) - \nabla f_i(\tilde{w}) + \nabla f(\tilde{w})||_2^2] \\
&\leq 2\mathbb{E}_j[||\nabla f_i(x^t) - \nabla f_i(w^*)||_2^2] + 2\mathbb{E}_j[||\nabla f_i(w^*) - \nabla f_i(\tilde{w}) + \nabla f(\tilde{w})||_2^2] \\
&\leq 2\mathbb{E}_j[||\nabla f_i(x^t) - \nabla f_i(w^*)||_2^2] + 2\mathbb{E}_j[||\nabla f_i(w^*) - \nabla f_i(\tilde{w})||_2^2] \\
&= 4L_{\max}\left(f(x^t) - f(w^*) + f(\tilde{w}) - f(w^*)\right) \qquad \blacksquare
\end{aligned}$$

Lemma 2

Where we used in the first inequality that $\mathbb{E}[||X - \mathbb{E}X||_2^2] \leq \mathbb{E}[||X||_2^2]$ with $X = \nabla f_i(w^*) - \nabla f_i(\tilde{w})$ thus $\mathbb{E}[X] = -\nabla f(\tilde{w})$

Proof:

$$\begin{aligned} ||x_s^{t+1} - w^*||_2^2 &= ||x_s^t - w^* - \gamma g^t||_2^2 \\ &= ||x_s^t - w^*||_2^2 - 2\gamma\langle g^t, x_s^t - w^*\rangle + \gamma^2||g^t||_2^2. \end{aligned}$$

Taking expectation with respect to *j*

Unbiased estimator

$$\begin{aligned} \mathbb{E}_j\left[||x_s^{t+1} - w^*||_2^2\right] &= ||x_s^t - w^*||_2^2 - 2\gamma\langle \nabla f(x_s^t), x_s^t - w^*\rangle + \gamma^2\mathbb{E}_j\left[||g^t||_2^2\right] \\ &\overset{\text{str. conv.}}{\leq} (1-\mu\gamma)||x_s^t - w^*||_2^2 - 2\gamma(f(x_s^t) - f(w^*)) + \gamma^2\mathbb{E}_j\left[||g^t||_2^2\right] \end{aligned}$$

Need to bound this!

$$\boxed{\mathbb{E}_j\left[||g^t||_2^2\right]}$$

Lemma 3 $g^t = \nabla f_{i}(x_s^t) - \nabla f_{i}(\tilde{w}_{s-1}) + \nabla f(\tilde{w}_{s-1})$

$$\mathbb{E}_j[||g^t||_2^2] \leq 4L_{\max}(f(x_s^t) - f(w^*)) + 4L_{\max}(f(\tilde{w}_{s-1}) - f(w^*))$$

Proof:

$$\begin{aligned}||x_s^{t+1} - w^*||_2^2 &= ||x_s^t - w^* - \gamma g^t||_2^2 \\ &= ||x_s^t - w^*||_2^2 - 2\gamma\langle g^t, x_s^t - w^*\rangle + \gamma^2||g^t||_2^2.\end{aligned}$$

Unbiased estimator

Taking expectation with respect to *j*

$$\begin{aligned}\mathbb{E}_j\left[||x_s^{t+1} - w^*||_2^2\right] &= ||x_s^t - w^*||_2^2 - 2\gamma\langle \nabla f(x_s^t), x_s^t - w^*\rangle + \gamma^2\mathbb{E}_j\left[||g^t||_2^2\right] \\ &\overset{\text{str. conv.}}{\leq} (1-\mu\gamma)||x_s^t - w^*||_2^2 - 2\gamma(f(x_s^t) - f(w^*)) + \gamma^2\mathbb{E}_j\left[||g^t||_2^2\right] \\ &\leq (1-\mu\gamma)||x_s^t - w^*||_2^2 - 2\gamma(1 - 2\gamma L_{\max})(f(x_s^t) - f(w^*)) \\ &\quad +4\gamma^2 L_{\max}(f(w_{s-1}) - f(w^*))\end{aligned}$$

Proof:

$$\begin{aligned} ||x_s^{t+1} - w^*||_2^2 &= ||x_s^t - w^* - \gamma g^t||_2^2 \\ &= ||x_s^t - w^*||_2^2 - 2\gamma\langle g^t, x_s^t - w^*\rangle + \gamma^2 ||g^t||_2^2. \end{aligned}$$

Unbiased estimator

Taking expectation with respect to *j*

$$\begin{aligned} \mathbb{E}_j\left[||x_s^{t+1} - w^*||_2^2\right] &= ||x_s^t - w^*||_2^2 - 2\gamma\langle \nabla f(x_s^t), x_s^t - w^*\rangle + \gamma^2 \mathbb{E}_j\left[||g^t||_2^2\right] \\ &\overset{\text{str. conv.}}{\leq} (1-\mu\gamma)||x_s^t - w^*||_2^2 - 2\gamma(f(x_s^t) - f(w^*)) + \gamma^2 \mathbb{E}_j\left[||g^t||_2^2\right] \\ &\leq (1-\mu\gamma)||x_s^t - w^*||_2^2 - 2\gamma(1 - 2\gamma L_{\max})(f(x_s^t) - f(w^*)) \\ &\quad + 4\gamma^2 L_{\max}(f(w_{s-1}) - f(w^*)) \end{aligned}$$

Taking expectation and iterating from $t = 0, \ldots, m-1$

Proof:

$$
\begin{aligned}
||x_s^{t+1} - w^*||_2^2 &= ||x_s^t - w^* - \gamma g^t||_2^2 \\
&= ||x_s^t - w^*||_2^2 - 2\gamma\langle g^t, x_s^t - w^*\rangle + \gamma^2||g^t||_2^2.
\end{aligned}
$$

Unbiased estimator

Taking expectation with respect to *j*

$$
\begin{aligned}
\mathbb{E}_j\left[||x_s^{t+1} - w^*||_2^2\right] &= ||x_s^t - w^*||_2^2 - 2\gamma\langle \nabla f(x_s^t), x_s^t - w^*\rangle + \gamma^2\mathbb{E}_j\left[||g^t||_2^2\right] \\
&\overset{\text{str. conv.}}{\leq} (1-\mu\gamma)||x_s^t - w^*||_2^2 - 2\gamma(f(x_s^t) - f(w^*)) + \gamma^2\mathbb{E}_j\left[||g^t||_2^2\right] \\
&\leq (1-\mu\gamma)||x_s^t - w^*||_2^2 - 2\gamma(1 - 2\gamma L_{\max})(f(x_s^t) - f(w^*)) \\
&\quad +4\gamma^2 L_{\max}(f(w_{s-1}) - f(w^*))
\end{aligned}
$$

Taking expectation and iterating from *t = 0, … , m-1*

$$
\begin{aligned}
\mathbb{E}_j\left[||x_s^m - w^*||_2^2\right] \leq\ & (1-\mu\gamma)^m||x_s^0 - w^*||_2^2 \\
& -2\gamma(1-2\gamma L_{\max})S_m\sum_{t=0}^{m-1}\alpha_t(f(x_s^t) - f(w^*)) \\
+\ & 4S_m\gamma^2 L_{\max}(f(w_{s-1}) - f(w^*))
\end{aligned}
$$

$\alpha_t := (1-\mu\gamma)^{m-1-t}$

Proof:

$$
\begin{aligned}
||x_s^{t+1} - w^*||_2^2 &= ||x_s^t - w^* - \gamma g^t||_2^2 \\
&= ||x_s^t - w^*||_2^2 - 2\gamma\langle g^t, x_s^t - w^*\rangle + \gamma^2 ||g^t||_2^2.
\end{aligned}
$$

Unbiased estimator

Taking expectation with respect to *j*

$$
\begin{aligned}
\mathbb{E}_j\left[||x_s^{t+1} - w^*||_2^2\right] &= ||x_s^t - w^*||_2^2 - 2\gamma\langle \nabla f(x_s^t), x_s^t - w^*\rangle + \gamma^2 \mathbb{E}_j\left[||g^t||_2^2\right] \\
&\overset{\text{str. conv.}}{\leq} (1-\mu\gamma)||x_s^t - w^*||_2^2 - 2\gamma(f(x_s^t) - f(w^*)) + \gamma^2 \mathbb{E}_j\left[||g^t||_2^2\right] \\
&\leq (1-\mu\gamma)||x_s^t - w^*||_2^2 - 2\gamma(1 - 2\gamma L_{\max})(f(x_s^t) - f(w^*)) \\
&\quad +4\gamma^2 L_{\max}(f(w_{s-1}) - f(w^*))
\end{aligned}
$$

Taking expectation and iterating from *t = 0, … , m-1*

$$
\begin{aligned}
\mathbb{E}_j\left[||x_s^m - w^*||_2^2\right] &\leq (1-\mu\gamma)^m ||x_s^0 - w^*||_2^2 \\
&\quad -2\gamma(1-2\gamma L_{\max}) S_m \sum_{t=0}^{m-1} \alpha_t (f(x_s^t) - f(w^*)) \\
&\quad + 4 S_m \gamma^2 L_{\max}(f(w_{s-1}) - f(w^*))
\end{aligned}
$$

$\alpha_t := (1-\mu\gamma)^{m-1-t}$

$S_m := \sum_{t=0}^{m-1} \alpha_t$

Proof:

$$
\begin{aligned}
||x_s^{t+1} - w^*||_2^2 &= ||x_s^t - w^* - \gamma g^t||_2^2 \\
&= ||x_s^t - w^*||_2^2 - 2\gamma\langle g^t, x_s^t - w^*\rangle + \gamma^2||g^t||_2^2.
\end{aligned}
$$

Unbiased estimator

Taking expectation with respect to *j*

$$
\begin{aligned}
\mathbb{E}_j\left[||x_s^{t+1} - w^*||_2^2\right] &= ||x_s^t - w^*||_2^2 - 2\gamma\langle \nabla f(x_s^t), x_s^t - w^*\rangle + \gamma^2\mathbb{E}_j\left[||g^t||_2^2\right] \\
&\overset{\text{str. conv.}}{\leq} (1-\mu\gamma)||x_s^t - w^*||_2^2 - 2\gamma(f(x_s^t) - f(w^*)) + \gamma^2\mathbb{E}_j\left[||g^t||_2^2\right] \\
&\leq (1-\mu\gamma)||x_s^t - w^*||_2^2 - 2\gamma(1 - 2\gamma L_{\max})(f(x_s^t) - f(w^*)) \\
&\quad +4\gamma^2 L_{\max}(f(w_{s-1}) - f(w^*))
\end{aligned}
$$

Taking expectation and iterating from *t = 0, … , m-1*

$$
\begin{aligned}
\mathbb{E}_j\left[||x_s^m - w^*||_2^2\right] \leq\ & (1-\mu\gamma)^m||x_s^0 - w^*||_2^2 \\
& -2\gamma(1 - 2\gamma L_{\max})S_m \sum_{t=0}^{m-1} \alpha_t(f(x_s^t) - f(w^*)) \\
+\ & 4S_m\gamma^2 L_{\max}(f(w_{s-1}) - f(w^*))
\end{aligned}
$$

$\alpha_t := (1-\mu\gamma)^{m-1-t}$

$S_m := \sum_{t=0}^{m-1} \alpha_t$

Rest on the board

Take for home Variance Reduction

- Variance reduced methods use only **one stochastic gradient per iteration** and converge linearly on strongly convex functions
- Choice of **fixed stepsize** possible
- **SAGA** only needs to know the smoothness parameter to work, but requires storing *n* past stochastic gradients
- **SVRG** only has *O(d)* storage, but requires full gradient computations every so often. Has an extra “number of inner iterations” parameter to tune