Optimization for Data Science

Stochastic Variance Reduced Gradient Methods

Robert M. Gower

References for this class

Section 6.3:

Sébastien Bubeck (2015)
Foundations and Trends
Convex Optimization: Algorithms and Complexity

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

How to transform convergence results into iteration complexity

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

Solving the Finite Sum Training Problem

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

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

SGD shrinking stepsize

SGD 1.0: Descreasing stepsize

Set $w^0 = 0$, choose $\alpha > 0$, $\alpha_t = \frac{\alpha}{\sqrt{t+1}}$,

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

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

$w^{t+1} = w^t - \alpha_t \nabla f_j(w^t)$

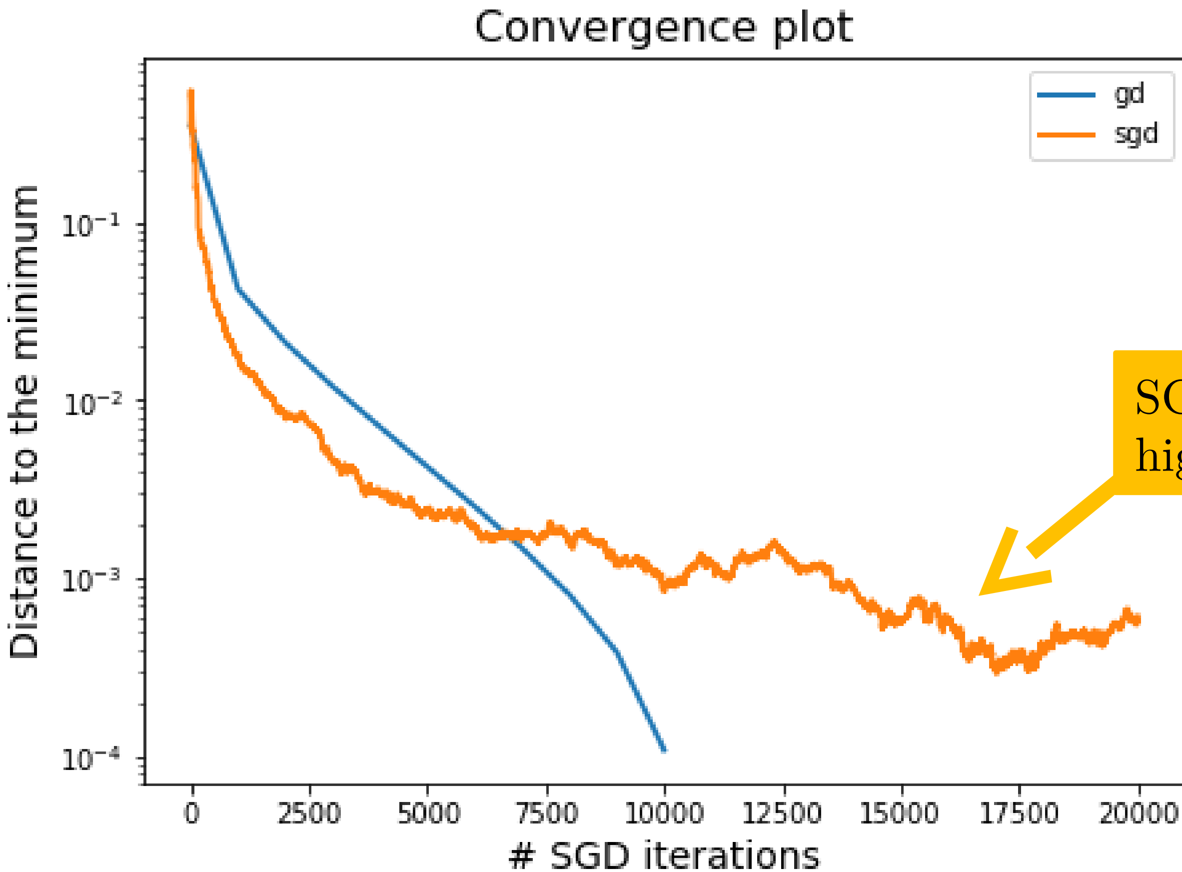
Output w^T

Convergence for Strongly Convex

- $f(w)$ is λ - strongly convex
- Subgradients bounded

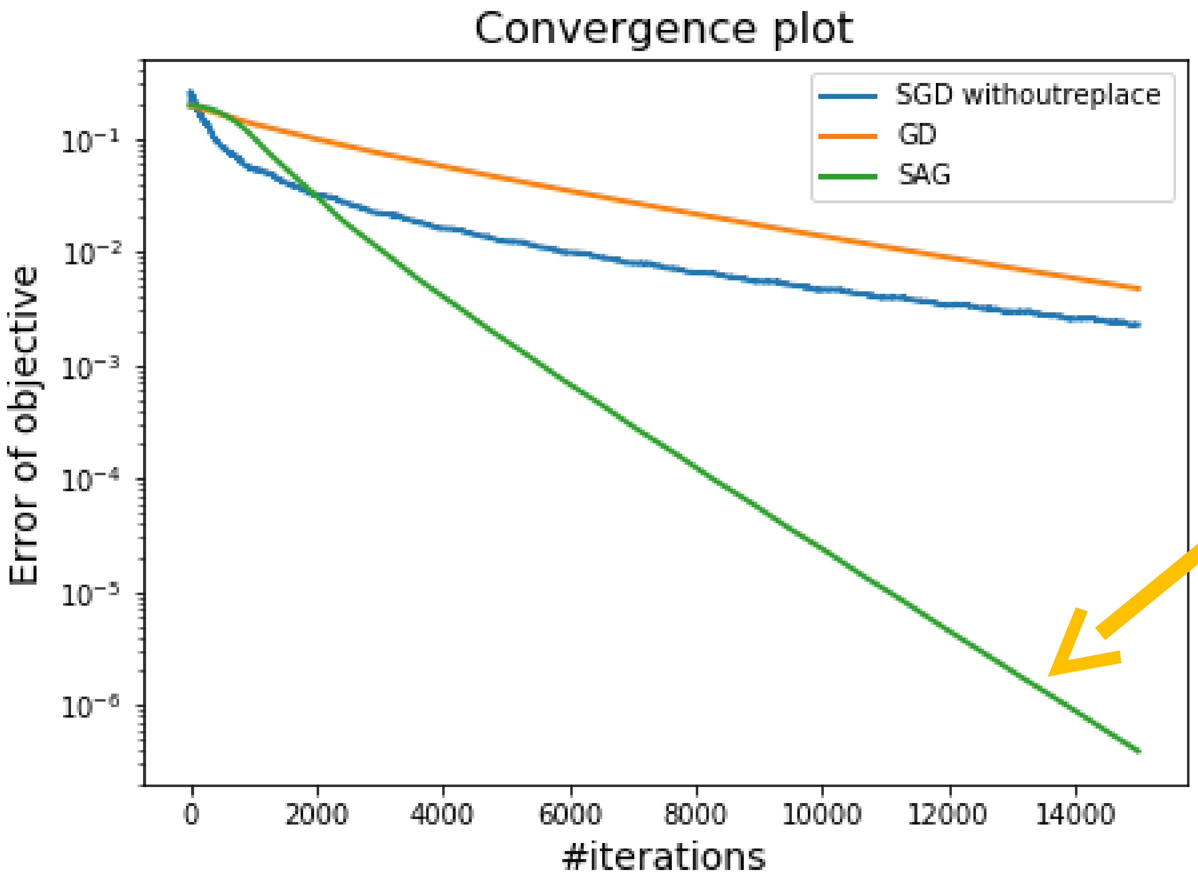
$$\alpha_t = O\left(\frac{1}{\lambda t}\right) \quad \Rightarrow \quad \mathbb{E}[f(w^T)] - f(w^*) = O\left(\frac{1}{\lambda T}\right)$$

SGD initially fast, slow later



SGD suffers from high variance

Can we get best of both?



Today we learn about methods like this one

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

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

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

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

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

We would like gradient estimate such that:

Similar

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

Converges in *L2*

$$\mathbb{E}||g^t - \nabla f(w^t)||_2^2 \underset{w^t \to w^*}{\longrightarrow} 0$$

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

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

We would like gradient estimate such that:

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

Similar

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

Converges in *L2*

$$\mathbb{E}||g^t - \nabla f(w^t)||_2^2 \underset{w^t \to w^*}{\longrightarrow} 0$$

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

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

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

We would like gradient estimate such that:

Similar

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

Solves problem of
$\mathbb{E}||\nabla f_j(w)||_2^2 \leq B^2$

Converges in *L2*

$$\mathbb{E}||g^t - \nabla f(w^t)||_2^2 \underset{w^t \to w^*}{\longrightarrow} 0$$

Covariates

Let x and z be random variables. We say that x and z are covariates if:

$$\mathrm{cov}(x, z) \geq 0$$

Variance Reduced Estimate:

$$x_z = x - z + \mathbb{E}[z]$$

Covariates

$$\mathrm{cov}(x, z) := \mathbb{E}[(x - \mathbb{E}[x])(z - \mathbb{E}[z])]$$

Let x and z be random variables. We say that x and z are covariates if:

$$\mathrm{cov}(x, z) \geq 0$$

Variance Reduced Estimate:

$$x_z = x - z + \mathbb{E}[z]$$

EXE:

1. Show that $\mathbb{E}[x_z] = \mathbb{E}[x]$
2. $\mathbb{VAR}[x_z] = \mathbb{E}[(x_z - \mathbb{E}[x_z])^2] = ?$
3. When is $\mathbb{VAR}[x_z] \leq \mathbb{VAR}[x]$

Covariates

$$\mathrm{cov}(x, z) := \mathbb{E}[(x - \mathbb{E}[x])(z - \mathbb{E}[z])]$$

Let x and z be random variables. We say that x and z are covariates if:

$$\mathrm{cov}(x, z) \geq 0$$

Variance Reduced Estimate:

$$x_z = x - z + \mathbb{E}[z]$$

EXE:

1. Show that $\mathbb{E}[x_z] = \mathbb{E}[x]$
2. $\mathbb{VAR}[x_z] = \mathbb{E}[(x_z - \mathbb{E}[x_z])^2] = ?$
3. When is $\mathbb{VAR}[x_z] \leq \mathbb{VAR}[x]$

$$\begin{aligned}
\mathbb{E}[(x_z - \mathbb{E}[x_z])^2] &= \mathbb{E}[(x - \mathbb{E}[x] - (z - \mathbb{E}[z]))^2] \\
&= \mathbb{E}[(x - \mathbb{E}[x])^2] - 2\mathbb{E}[(x - \mathbb{E}[x])(z - \mathbb{E}[z])] \\
&\quad + \mathbb{E}[(z - \mathbb{E}[z])^2] \\
&= \mathbb{VAR}[x] - 2\mathrm{cov}(x, z) + \mathbb{VAR}[z]
\end{aligned}$$

SVRG: Stochastic Variance Reduced Gradients

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

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

Sample $\quad \nabla f_i(w^t), \quad i \in \{1, \ldots, n\}$ uniformly

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

$$x_z = \quad x \quad - \quad z \quad + \quad \mathbb{E}[z]$$

SVRG: Stochastic Variance Reduced Gradients

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

$\tilde{w}^0 = w^0$

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

  calculate $\nabla f(\tilde{w}^t)$

  $w^0 = \tilde{w}^t$

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

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

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

    $w^{k+1} = w^k - \alpha g^k$

  Option I: $\tilde{w}^{t+1} = w^m$

  Option II: $\tilde{w}^{t+1} = \frac{1}{m}\sum_{i=0}^{m-1} w^i$

Output $\tilde{w}^T$

Freeze reference point for m iterations

SAGA: Stochastic Average Gradient unbiased version

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

Sample

$$\nabla f_i(w^t), \quad i \in \{1, \ldots, n\} \text{ uniformly}$$

grad estimate

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

$$x_z = \quad x \quad - \quad z \quad + \quad \mathbb{E}[z]$$

Store gradient

$$\nabla f_i(w_i^t) = \nabla f_i(w^t), \quad \nabla f_i(w_j^{t+1}) = \nabla f_i(w_j^t)$$

$$\forall j \neq i$$

SAGA: Stochastic Average Gradient

Set $w^0 = 0, w_i^0 = w^0,$ for $i = 1 \ldots, n,$ choose $\alpha > 0$
Setup table $[\nabla f_1(w_1^0), \ldots, \nabla f_n(w_n^0)] \in R^{d \times n}$
for $t = 0, 1, 2, \ldots, T-1$

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

$\quad g^t = \nabla f_i(w^t) - \nabla f_i(w_i^t) + \frac{1}{n}\sum_{j=1}^n \nabla f_j(w_j^t)$

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

$\quad \nabla f_i(w_i^{t+1}) = \nabla f_i(w^t)$ $\qquad$ (update grad)

$\quad \nabla f_j(w_j^{t+1}) = \nabla f_j(w_j^t)$ $\qquad \forall j \neq i$

Output w^T

$d \times n$

SAGA: Stochastic Average Gradient

Set $w^0 = 0, w_i^0 = w^0,$ for $i = 1\ldots, n,$ choose $\alpha > 0$

Setup table $[\nabla f_1(w_1^0), \ldots, \nabla f_n(w_n^0)] \in R^{d\times n}$

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

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

$\quad$ $g^t = \nabla f_i(w^t) - \nabla f_i(w_i^t) + \frac{1}{n}\sum_{j=1}^n \nabla f_j(w_j^t)$

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

$\quad$ $\nabla f_i(w_i^{t+1}) = \nabla f_i(w^t)$ $\qquad$ (update grad)

$\quad$ $\nabla f_j(w_j^{t+1}) = \nabla f_j(w_j^t)$ $\qquad$ $\forall j \neq i$

Output w^T

No inner loop, rolling update

Stores a $d\times n$ matrix

SAG: Stochastic Average Gradient (Biased version)

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

Sample

$$\nabla f_i(w^t), \quad i \in \{1, \ldots, n\} \text{ uniformly}$$

Store gradient

$$\nabla f_i(w_i^t) = \nabla f_i(w^t), \quad \nabla f_i(w_j^{t+1}) = \nabla f_i(w_j^t)$$

$$\forall j \neq i$$

grad estimate

$$g^t = \frac{1}{n} \sum_{j=1}^{n} \nabla f_j(w^{t_j})$$

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

$$\cancel{x_z = x - z + \mathbb{E}[z]}$$

SAGA: Stochastic Average Gradient

Set $w^0 = 0, w_i^0 = w^0,$ for $i = 1, \ldots, n,$ choose $\alpha > 0$
Setup table $[\nabla f_1(w_1^0), \ldots, \nabla f_n(w_n^0)] \in R^{d \times n}$
for $t = 0, 1, 2, \ldots, T-1$

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

$\quad$ $\nabla f_i(w_i^{t+1}) = \nabla f_i(w^t) \qquad$ (update grad)

$\quad$ $\nabla f_j(w_j^{t+1}) = \nabla f_j(w_j^t) \qquad \forall j \neq i$

$\quad$ $g^t = \frac{1}{n} \sum_{j=1}^n \nabla f_j(w_j^{t+1})$

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

Output w^T

$d \times n$

SAGA: Stochastic Average Gradient

Set $w^0 = 0, w_i^0 = w^0,$ for $i = 1, \ldots, n,$ choose $\alpha > 0$

Setup table $[\nabla f_1(w_1^0), \ldots, \nabla f_n(w_n^0)] \in R^{d\times n}$

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

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

$$\nabla f_i(w_i^{t+1}) = \nabla f_i(w^t) \qquad \text{(update grad)}$$

$$\nabla f_j(w_j^{t+1}) = \nabla f_j(w_j^t) \qquad \forall j \neq i$$

$$g^t = \tfrac{1}{n}\textstyle\sum_{j=1}^n \nabla f_j(w_j^{t+1})$$

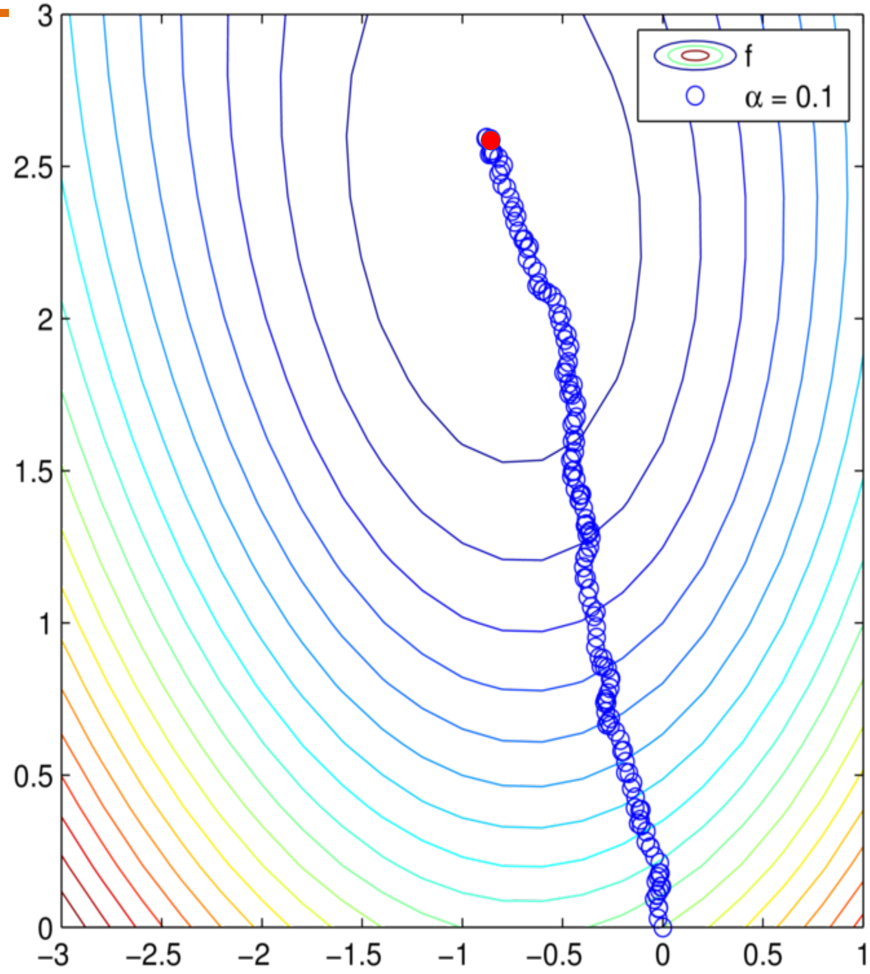
$$w^{t+1} = w^t - \alpha g^t$$

Output w^T

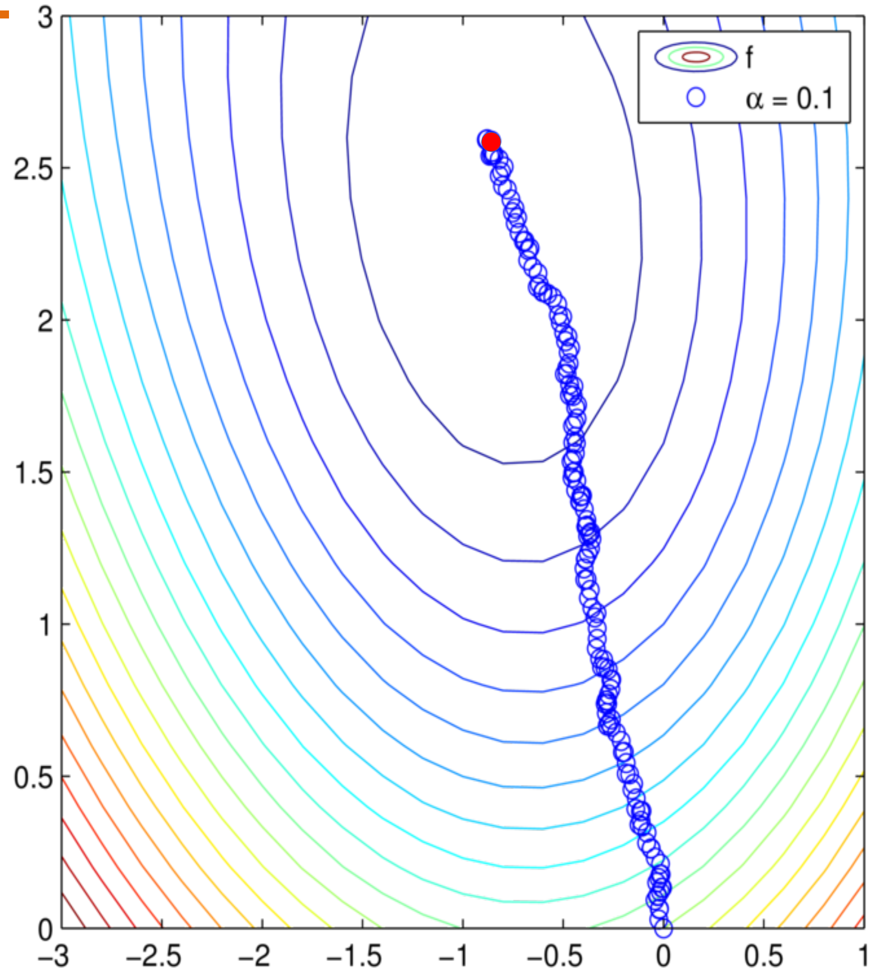
Very easy to implement

Stores a $d \times n$ matrix

The Stochastic Average Gradient

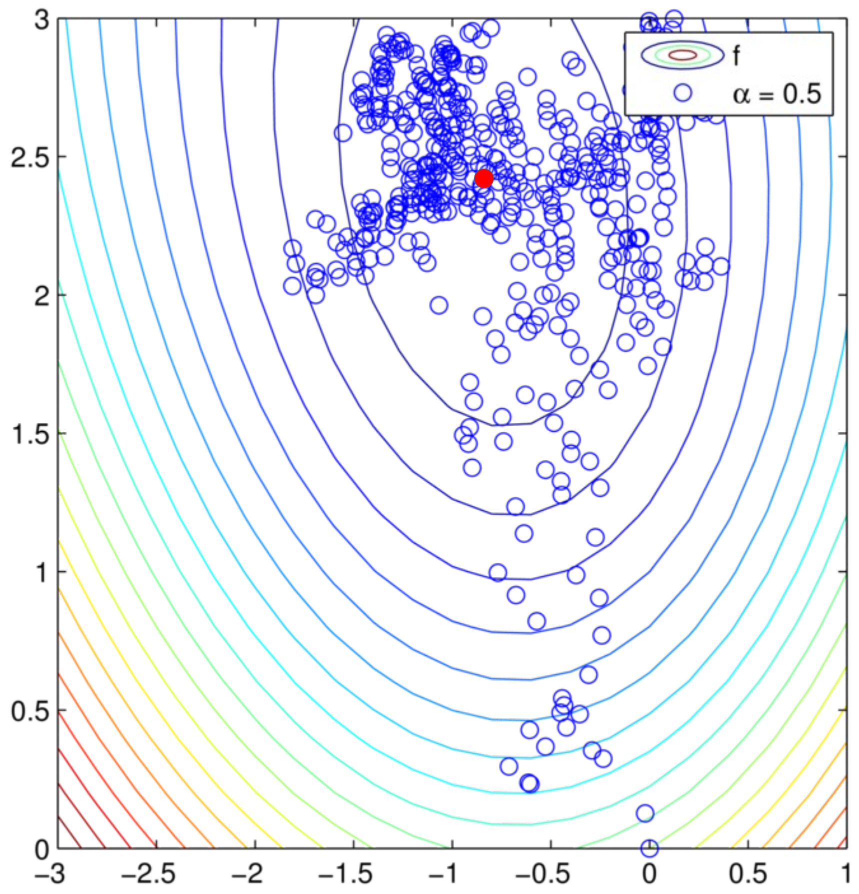


The Stochastic Average Gradient



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

Stochastic Gradient Descent
α =0.5



Proving Convergence

Assumptions for Convergence

Strong Convexity

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

Smoothness + convexity

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

for $i = 1, \ldots, n$

EXE: Calculate L_i and $L_{\max} := \max_{i=1,\ldots,n} L_i$ for

1. $f(w) = \frac{1}{2}||Xw - y||_2^2 + \frac{\lambda}{2}||w||_2^2,$ where $X \in \mathbb{R}^{n \times d}$
2. $f(w) = \frac{1}{n}\sum_{i=1}^{n} \ln(1 + e^{-y_i \langle w, x_i \rangle}) + \frac{\lambda}{2}||w||_2^2$

Assumptions for Convergence

EXE: Calculate L_i for

1. $$f(w) = \frac{1}{2}||Aw - y||_2^2 + \frac{\lambda}{2}||w||_2^2$$

Assumptions for Convergence

EXE: Calculate L_i for

1. $$f(w) = \frac{1}{2}||Aw - y||_2^2 + \frac{\lambda}{2}||w||_2^2$$

1. $$f(w) = \frac{1}{2}||Aw - y||_2^2 + \frac{\lambda}{2}||w||_2^2 = \frac{1}{n}\sum_{i=1}^{n}(\frac{n}{2}(A_{i:}^{\top}w - y_i)^2 + \frac{\lambda}{2}||w||_2^2)$$
$$= \frac{1}{n}\sum_{i=1}^{n} f_i(w)$$

Assumptions for Convergence

EXE: Calculate L_i for

1. $$f(w) = \frac{1}{2}||Aw - y||_2^2 + \frac{\lambda}{2}||w||_2^2$$

$$1. \quad f(w) = \frac{1}{2}||Aw - y||_2^2 + \frac{\lambda}{2}||w||_2^2 = \frac{1}{n}\sum_{i=1}^{n}(\frac{n}{2}(A_{i:}^\top w - y_i)^2 + \frac{\lambda}{2}||w||_2^2)$$

$$= \frac{1}{n}\sum_{i=1}^{n} f_i(w)$$

$$\nabla^2 f_i(w) = nA_{i:}A_{i:}^\top + \lambda \quad \preceq \quad (n||A_{i:}||_2^2 + \lambda)I \quad = \quad L_i \ I$$

Assumptions for Convergence

EXE: Calculate L_{i} for

$$2. \quad f(w) = \frac{1}{n}\sum_{i=1}^{n} \ln(1 + e^{-y_i\langle w, a_i\rangle}) + \frac{\lambda}{2}||w||_2^2$$

Assumptions for Convergence

EXE: Calculate L_i for

$$2. \quad f(w) = \frac{1}{n}\sum_{i=1}^{n} \ln(1 + e^{-y_i\langle w, a_i\rangle}) + \frac{\lambda}{2}||w||_2^2$$

$$2. \quad f_i(w) = \ln(1 + e^{-y_i\langle w, a_i\rangle}) + \frac{\lambda}{2}||w||_2^2,$$

Assumptions for Convergence

EXE: Calculate L_i for

$$2. \quad f(w) = \frac{1}{n}\sum_{i=1}^{n} \ln(1+e^{-y_i\langle w, a_i\rangle}) + \frac{\lambda}{2}||w||_2^2$$

$$2. \quad f_i(w) = \ln(1+e^{-y_i\langle w, a_i\rangle}) + \frac{\lambda}{2}||w||_2^2,$$

$$\nabla f_i(w) = \frac{-y_i a_i e^{-y_i\langle w, a_i\rangle}}{1+e^{-y_i\langle w, a_i\rangle}} + \lambda w$$

$$\nabla^2 f_i(w) = a_i a_i^\top \left(\frac{(1+e^{-y_i\langle w, a_i\rangle})e^{-y_i\langle w, a_i\rangle}}{(1+e^{-y_i\langle w, a_i\rangle})^2} - \frac{e^{-2y_i\langle w, a_i\rangle}}{(1+e^{-y_i\langle w, a_i\rangle})^2} \right) + \lambda I$$

$$= a_i a_i^\top \frac{e^{-y_i\langle w, a_i\rangle}}{(1+e^{-y_i\langle w, a_i\rangle})^2} + \lambda I \quad \preceq \quad \left(\frac{||a_i||_2^2}{4} + \lambda \right) I = L_i \, I$$

Assumptions for Convergence

EXE: Let $f(w)$ be L–smooth and $f_i(w)$ be L_i–smooth for $i = 1, \ldots, n$.

Show that

$$L \quad \leq \quad \frac{1}{n}\sum_{i=1}^{n} L_i \quad \leq \quad L_{\max} := \max_{i=1,\ldots,n} L_i$$

Proof: From definition of $f_i(w)$ smoothness

Assumptions for Convergence

EXE: Let $f(w)$ be L–smooth and $f_i(w)$ be L_i–smooth for $i = 1, \ldots, n$.

Show that

$$L \quad \leq \quad \frac{1}{n}\sum_{i=1}^{n} L_i \quad \leq \quad L_{\max} := \max_{i=1,\ldots,n} L_i$$

Proof: From definition of $f_i(w)$ smoothness

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

$$= f(y) + \langle \nabla f(y), x - y\rangle + \frac{1}{2n}\sum_{i=1}^{n} L_i ||w - y||_2^2$$

Convergence SVRG

Theorem

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

If $\alpha = 1/10L_{\max}$ and $m = 20L_{\max}/\lambda$ then

$$\mathbb{E}[f(\tilde{w}^t)] - f(w^*) \quad \leq \quad \left(\frac{7}{8}\right)^t (f(\tilde{w}^0) - f(w^*))$$

Need $O(L_{\max}/\lambda)$ inner iterations to have linear convergence

In practice use $\alpha = 1/L_{\max}, \ m = n$

Johnson, R. & Zhang, T. **Accelerating Stochastic Gradient Descent using Predictive Variance Reduction,** NIPS 2013

Proof:

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

Taking expectation with respect to j

Unbiased estimator

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

Must control this!

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

Smoothness Consequences I

Smoothness

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

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

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

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

EXE: Lemma 3

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

Proof: Hint: use $||a+b||_2^2 \leq 2||a||_2^2 + 2||b||_2^2$ and 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 $X = \nabla f_i(w^*) - \nabla f_i(\tilde{w}^t)$ thus $\mathbb{E}[X] = -\nabla f(\tilde{w}^t)$

Bounding gradient estimate

EXE: Lemma 3

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

Proof: Hint: use $||a+b||_2^2 \leq 2||a||_2^2 + 2||b||_2^2$ and 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 $X = \nabla f_i(w^*) - \nabla f_i(\tilde{w}^t)$ thus $\mathbb{E}[X] = -\nabla f(\tilde{w}^t)$

Bounding gradient estimate

EXE: Lemma 3

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

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

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

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

Proof:

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

Taking expectation with respect to j

Unbiased estimator

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

Must control this!

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

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

Proof (continued I):

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

Taking expectation with respect to j

Unbiased estimator

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

Proof (continued I):

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

Taking expectation with respect to j

Unbiased estimator

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

Taking expectation and summing from $k = 0, \ldots, m-1$ gives

$$
\begin{aligned}
\mathbb{E}\left[||w^m - w^*||_2^2\right] &\leq \mathbb{E}\left[||w^0 - w^*||_2^2\right] - 2\alpha(1 - 2\alpha L_{\max})\mathbb{E}[\textstyle\sum_{k=0}^{m-1}(f(w^k) - f(w^*))] \\
&\quad +4m\alpha^2 L_{\max}\mathbb{E}\left[f(\tilde{w}^t) - f(w^*)\right]
\end{aligned}
$$

Proof (continued II):

$$\mathbb{E}\left[||w^m - w^*||_2^2\right] \leq \mathbb{E}\left[||w^0 - w^*||_2^2\right] - 2\alpha(1 - 2\alpha L_{\max})\mathbb{E}[\textstyle\sum_{k=0}^{m-1}(f(w^k) - f(w^*))]$$

$$+4m\alpha^2 L_{\max}\mathbb{E}\left[f(\tilde{w}^t) - f(w^*)\right]$$

Proof (continued II):

$$\begin{aligned}\mathbb{E}\left[||w^m - w^*||_2^2\right] &\leq \mathbb{E}\left[||w^0 - w^*||_2^2\right] - 2\alpha(1 - 2\alpha L_{\max})\mathbb{E}[\textstyle\sum_{k=0}^{m-1}(f(w^k) - f(w^*))] \\ &\quad +4m\alpha^2 L_{\max}\mathbb{E}\left[f(\tilde{w}^t) - f(w^*)\right]\end{aligned}$$

Re-arranging and using strong convexity $f(\tilde{w}^t) - f(w^*) \geq \dfrac{\lambda}{2}||\tilde{w}^t - w^*||_2^2$

Proof (continued II):

$$\mathbb{E}\left[||w^m - w^*||_2^2\right] \leq \mathbb{E}\left[||w^0 - w^*||_2^2\right] - 2\alpha(1-2\alpha L_{\max})\mathbb{E}[\textstyle\sum_{k=0}^{m-1}(f(w^k) - f(w^*))]$$
$$+4m\alpha^2 L_{\max}\mathbb{E}\left[f(\tilde{w}^t) - f(w^*)\right]$$

Re-arranging and using strong convexity $f(\tilde{w}^t) - f(w^*) \geq \dfrac{\lambda}{2}||\tilde{w}^t - w^*||_2^2$

$$2\alpha(1-2\alpha L_{\max})\mathbb{E}[\textstyle\sum_{k=0}^{m-1}(f(w^k) - f(w^*))] \leq \mathbb{E}\left[||w^0 - w^*||_2^2\right] - \mathbb{E}\left[||w^m - w^*||_2^2\right]$$
$$+4m\alpha^2 L_{\max}\mathbb{E}\left[f(\tilde{w}^t) - f(w^*)\right]$$
$$\leq 2(2m\alpha^2 L_{\max} - \lambda^{-1})\mathbb{E}\left[f(\tilde{w}^t) - f(w^*)\right]$$

$\boxed{w^0 = \tilde{w}^t}$

Proof (continued II):

$$\mathbb{E}\left[||w^m - w^*||_2^2\right] \leq \mathbb{E}\left[||w^0 - w^*||_2^2\right] - 2\alpha(1-2\alpha L_{\max})\mathbb{E}[\textstyle\sum_{k=0}^{m-1}(f(w^k)-f(w^*))]$$
$$+4m\alpha^2 L_{\max}\mathbb{E}\left[f(\tilde{w}^t)-f(w^*)\right]$$

Re-arranging and using strong convexity $f(\tilde{w}^t) - f(w^*) \geq \frac{\lambda}{2}||\tilde{w}^t - w^*||_2^2$

$$2\alpha(1-2\alpha L_{\max})\mathbb{E}[\textstyle\sum_{k=0}^{m-1}(f(w^k)-f(w^*))] \leq \mathbb{E}\left[||w^0 - w^*||_2^2\right] - \mathbb{E}\left[||w^m - w^*||_2^2\right]$$

$w^0 = \tilde{w}^t$

$$+4m\alpha^2 L_{\max}\mathbb{E}\left[f(\tilde{w}^t)-f(w^*)\right]$$
$$\leq 2(2m\alpha^2 L_{\max} - \lambda^{-1})\mathbb{E}\left[f(\tilde{w}^t)-f(w^*)\right]$$

Re-arranging again

$$\mathbb{E}[(f(\sum_{k=0}^{m-1}\frac{w^k}{m}) - f(w^*))] \leq \mathbb{E}[\frac{1}{m}\sum_{k=0}^{m-1}(f(w^k)-f(w^*))]$$

Jensen's inequality

$$\leq \left(\frac{2\alpha L_{\max}}{1-2\alpha L_{\max}} + \frac{1}{\lambda\alpha(1-2\alpha L_{\max})m}\right)\mathbb{E}\left[f(\tilde{w}^t)-f(w^*)\right]$$

Now plug in values $\alpha = 1/(10L_{\max})$ and $m = 20L_{\max}/\lambda$ ■

Proof (continued II):

$$\mathbb{E}\left[||w^m - w^*||_2^2\right] \leq \mathbb{E}\left[||w^0 - w^*||_2^2\right] - 2\alpha(1-2\alpha L_{\max})\mathbb{E}[\textstyle\sum_{k=0}^{m-1}(f(w^k) - f(w^*))]$$
$$+4m\alpha^2 L_{\max}\mathbb{E}\left[f(\tilde{w}^t) - f(w^*)\right]$$

Re-arranging and using strong convexity $f(\tilde{w}^t) - f(w^*) \geq \frac{\lambda}{2}||\tilde{w}^t - w^*||_2^2$

$$2\alpha(1-2\alpha L_{\max})\mathbb{E}[\textstyle\sum_{k=0}^{m-1}(f(w^k) - f(w^*))] \leq \mathbb{E}\left[||w^0 - w^*||_2^2\right] - \mathbb{E}\left[||w^m - w^*||_2^2\right]$$

$w^0 = \tilde{w}^t$

$$+4m\alpha^2 L_{\max}\mathbb{E}\left[f(\tilde{w}^t) - f(w^*)\right]$$
$$\leq 2(2m\alpha^2 L_{\max} - \lambda^{-1})\mathbb{E}\left[f(\tilde{w}^t) - f(w^*)\right]$$

Re-arranging again

$$\mathbb{E}[(f(\sum_{k=0}^{m-1}\frac{w^k}{m}) - f(w^*))] \leq \mathbb{E}[\frac{1}{m}\sum_{k=0}^{m-1}(f(w^k) - f(w^*))]$$

Jensen's inequality

$$\leq \left(\underbrace{\frac{2\alpha L_{\max}}{1-2\alpha L_{\max}} + \frac{1}{\lambda\alpha(1-2\alpha L_{\max})m}}_{=7/8}\right)\mathbb{E}\left[f(\tilde{w}^t) - f(w^*)\right]$$

Now plug in values $\alpha = 1/(10L_{\max})$ and $m = 20L_{\max}/\lambda$ ∎

Convergence SAG

Theorem SAG

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

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

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

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 λ–strongly convex, $f_i(w)$ is $L_{\max}$–smooth and $\alpha = 1/(3L_{\max})$ then

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

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 proof due to unbiased gradients

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

Comparisons in complexity for strongly convex

Approximate solution

$$\mathbb{E}[f(w^T)] - f(w^*) \leq \epsilon$$

SGD

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

Gradient descent

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

SVRG/SAGA/SAG

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

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

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

Finite Sum Training Problem

L2 regularizor + linear hypothesis

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

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

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

$$\beta_i = \ell'(\langle w_i^t, x^i\rangle, y^i)$$

Stoch. gradient estimate

$$\nabla f_i(w_i^t) = \beta_i x^i + \lambda w^t$$

Full gradient estimate

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

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