Introduction to Machine Learning and Stochastic Optimization

Robert M. Gower

ENS
ÉCOLE NORMALE SUPÉRIEURE

Inria
INVENTEURS DU MONDE NUMÉRIQUE

cost
EUROPEAN COOPERATION IN SCIENCE AND TECHNOLOGY

Spring School on Optimization and Data Science,
Novi Saad, March 2017

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

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

The Training Problem

Solving the *training problem*:

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

Reference method: Gradient descent

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

Gradient Descent Algorithm

Set $w^0 = 0$, choose $\alpha > 0$.

for $t = 1, 2, 3, \ldots, T$

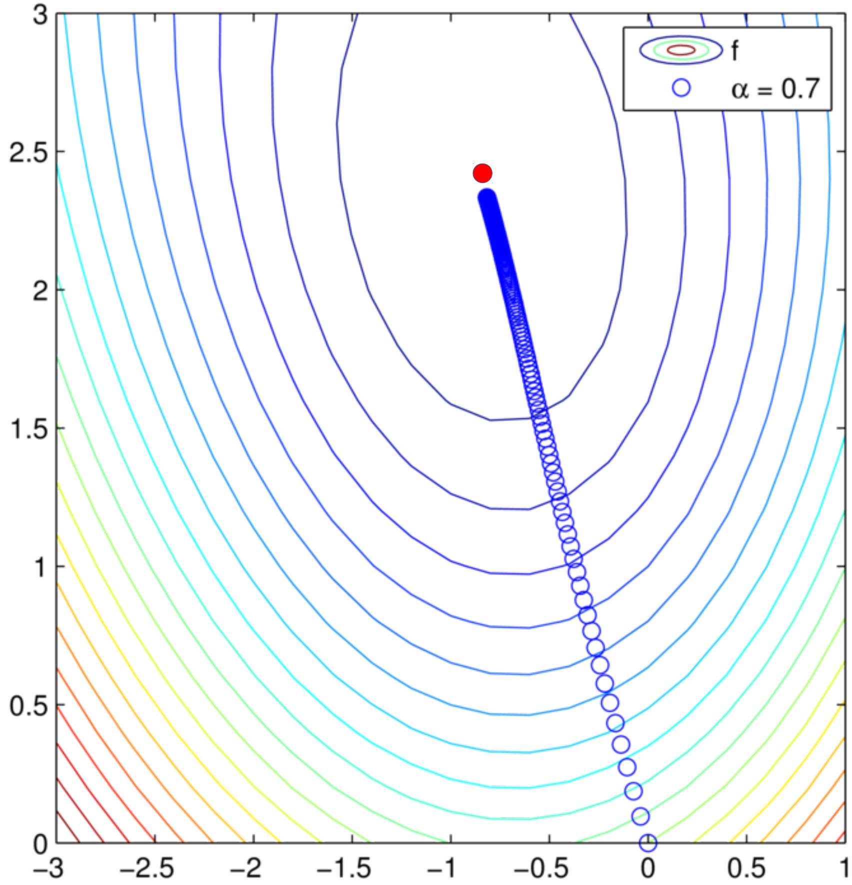
$w^{t+1} = w^t - \frac{\alpha}{n} \sum_{i=1}^{n} \nabla f_i(w^t)$

Output w^{T+1}

Gradient Descent Example

A Logistic Regression problem using the fourclass labelled data from LIBSVM

$(n,\ d) = (862{,}2)$

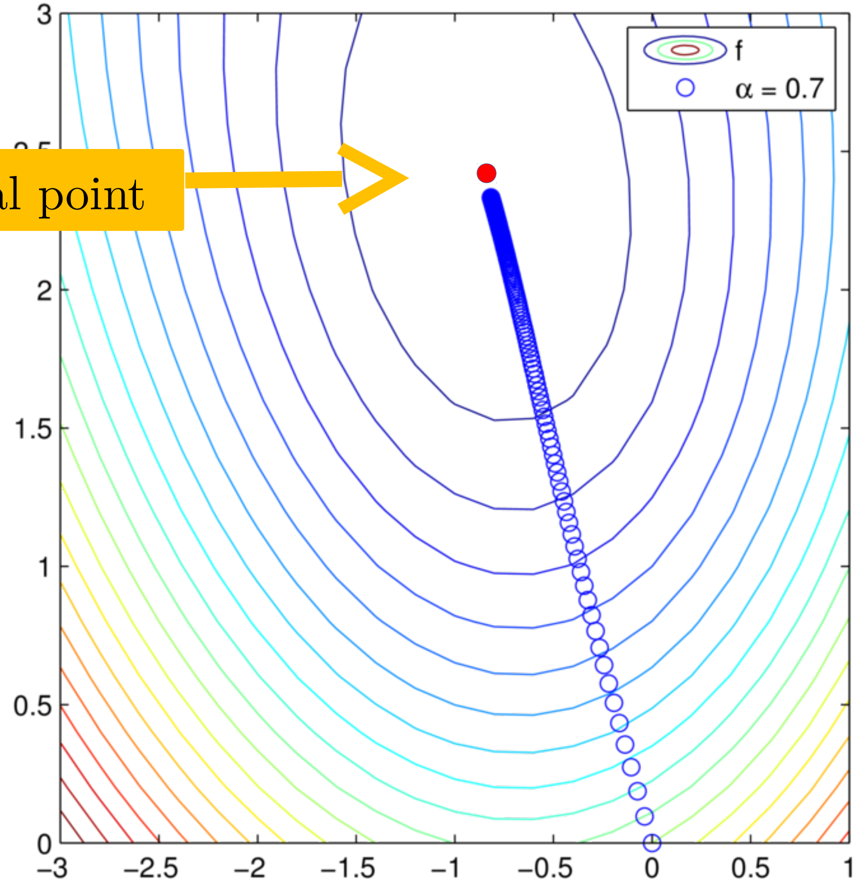


Gradient Descent Example

Optimal point

A Logistic Regression problem using the fourclass labelled data from LIBSVM

$(n,\ d) = (862{,}2)$



The Training Problem

Solving the *training problem*:

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

Problem with Gradient Descent:
Each iteration requires computing a gradient $\nabla f_i(w)$ for each data point. One gradient for each cat on the internet!

Gradient Descent Algorithm

Set $w^0 = 0$, choose $\alpha > 0$.
for $t = 1, 2, 3, \ldots, T$

$$w^{t+1} = w^t - \frac{\alpha}{n} \sum_{i=1}^{n} \nabla f_i(w^t)$$

Output w^{T+1}

Stochastic Gradient Descent

Is it possible to design a method that uses only the gradient of a **single** data function $f_i(w)$ at each iteration?

Stochastic Gradient Descent

Is it possible to design a method that uses only the gradient of a **single** data function $f_i(w)$ at each iteration?

Unbiased Estimate

Let j be a random index sampled from $\{1, \ldots, n\}$ selected uniformly at random. Then

$$\mathbb{E}_j\left[\nabla f_j(w)\right] = \frac{1}{n}\sum \nabla f_i(w) = \nabla f(w)$$

Stochastic Gradient Descent

Is it possible to design a method that uses only the gradient of a **single** data function $f_i(w)$ at each iteration?

Unbiased Estimate

Let j be a random index sampled from $\{1, \ldots, n\}$ selected uniformly at random. Then

$$\mathbb{E}_j\left[\nabla f_j(w)\right] = \frac{1}{n}\sum \nabla f_i(w) = \nabla f(w)$$

Use $\nabla f_j(w) \approx \nabla f(w)$

Stochastic Gradient Descent

Stochastic Gradient Descent Algorithm

Set $w^0 = 0$, choose $\alpha > 0$.

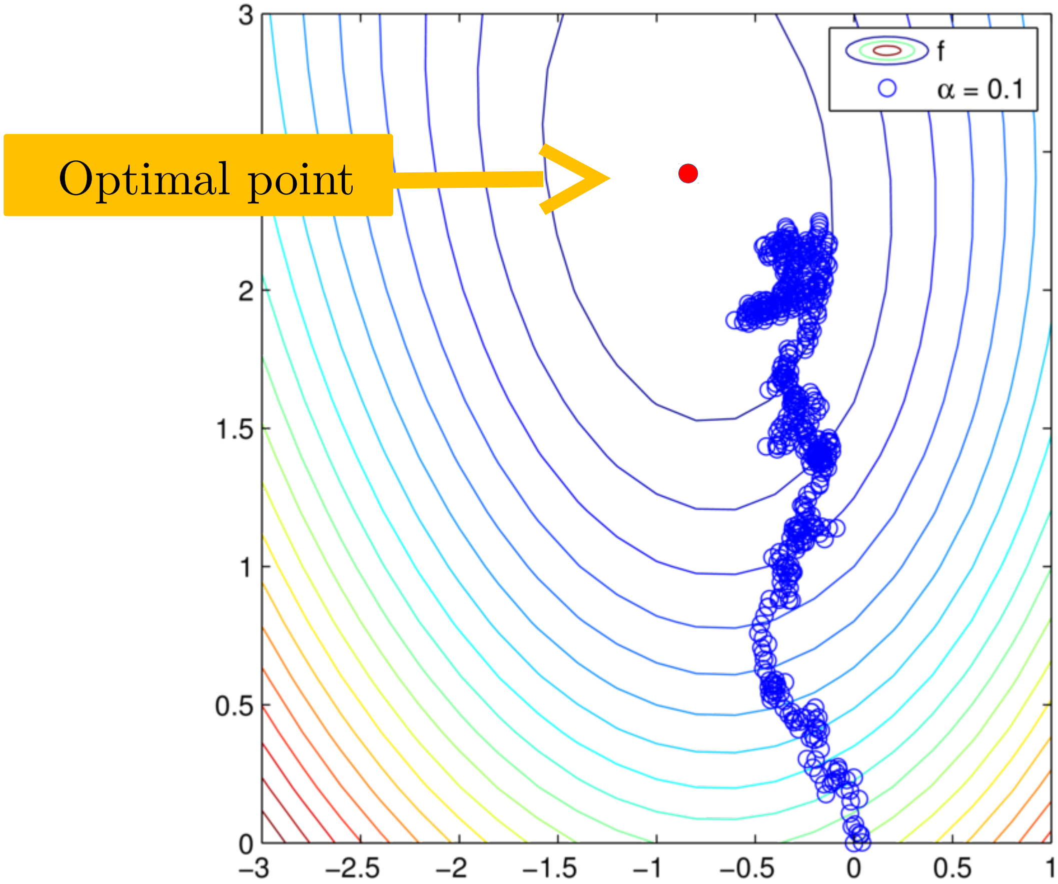
for $t = 1, 2, 3, \ldots, T$

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

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

Output w^{T+1}

Stochastic Gradient Descent



Assumptions for Convergence

Strong Convexity

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

$$2\langle \nabla f(w), w - w^* \rangle \geq \lambda ||w - w^*||_2^2$$

EXE: Using that

$$\frac{\sigma_{\min}(A)^2}{2} ||w - y||_2^2 \leq \frac{1}{2} ||A(w - y)||_2^2$$

Show that

$$\frac{1}{2} ||Aw - b||_2^2 \geq \frac{1}{2} ||Ay - b||_2^2 + \langle A^\top (Ay - b), w - y \rangle + \frac{\sigma_{\min}(A)^2}{2} ||w - y||_2^2$$

Assumptions for Convergence

Strong Convexity

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

Often the same as the regularization parameter

$$2\langle \nabla f(w), w - w^* \rangle \geq \lambda ||w - w^*||_2^2$$

EXE: Using that

$$\frac{\sigma_{\min}(A)^2}{2}||w - y||_2^2 \leq \frac{1}{2}||A(w - y)||_2^2$$

Show that

$$\frac{1}{2}||Aw - b||_2^2 \geq \frac{1}{2}||Ay - b||_2^2 + \langle A^\top (Ay - b), w - y \rangle + \frac{\sigma_{\min}(A)^2}{2}||w - y||_2^2$$

Assumptions for Convergence

Often the same as the regularization parameter

Strong Convexity

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

$$2\langle \nabla f(w), w - w^* \rangle \geq \lambda ||w - w^*||_2^2$$

Strong convexity parameter!

EXE: **Using that**

$$\frac{\sigma_{\min}(A)^2}{2}||w - y||_2^2 \leq \frac{1}{2}||A(w - y)||_2^2$$

Show that

$$\frac{1}{2}||Aw - b||_2^2 \geq \frac{1}{2}||Ay - b||_2^2 + \langle A^\top (Ay - b), w - y \rangle + \frac{\sigma_{\min}(A)^2}{2}||w - y||_2^2$$

Assumptions for Convergence

Often the same as the regularization parameter

Strong Convexity

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

$$2\langle \nabla f(w), w - w^* \rangle \geq \lambda ||w - w^*||_2^2$$

Strong convexity parameter!

EXE: **Using that**

$$\frac{\sigma_{\min}(A)^2}{2} ||w - y||_2^2 \leq \frac{1}{2} ||A(w - y)||_2^2$$

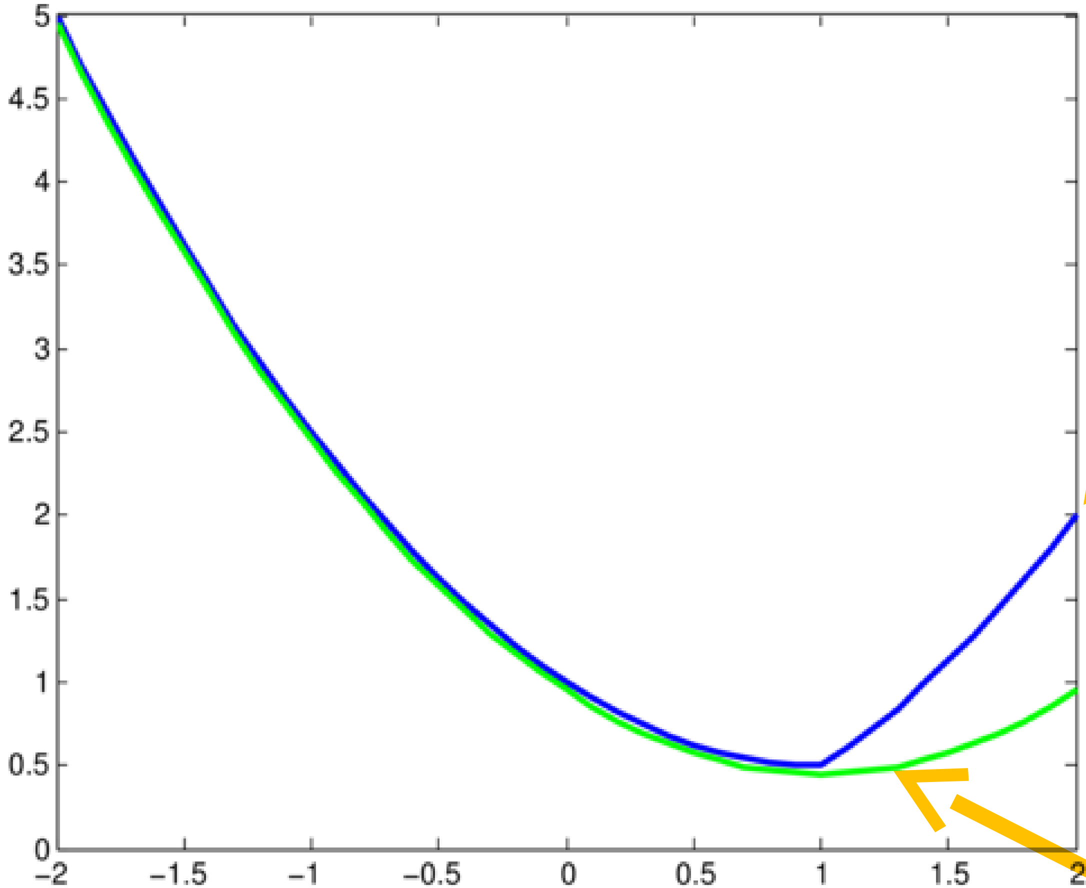
Show that

$$\frac{1}{2}||Aw - b||_2^2 \geq \frac{1}{2}||Ay - b||_2^2 + \langle A^\top (Ay - b), w - y \rangle + \frac{\sigma_{\min}(A)^2}{2} ||w - y||_2^2$$

Expected Bounded Stochastic Gradients

$$\mathbb{E}\left[||\nabla f_j(w^t)||_2^2 \right] \leq B^2, \text{ for all iterates } w^t \text{ of SGD}$$

Example of Strong Convexity



Hinge loss + L2

$$\max\{0, 1 - x\} + \frac{1}{2}||x||_2^2$$

Quadratic lower bound

Complexity / Convergence

Theorem

If $\frac{1}{\lambda} \geq \alpha > 0$ then the iterates of the SGD method satisfy

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

Shows that $\alpha \approx \frac{1}{\lambda}$

Shows that $\alpha \approx 0$

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 with respect to j

Unbiased estimator

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

Strong conv.

Bounded Stoch grad

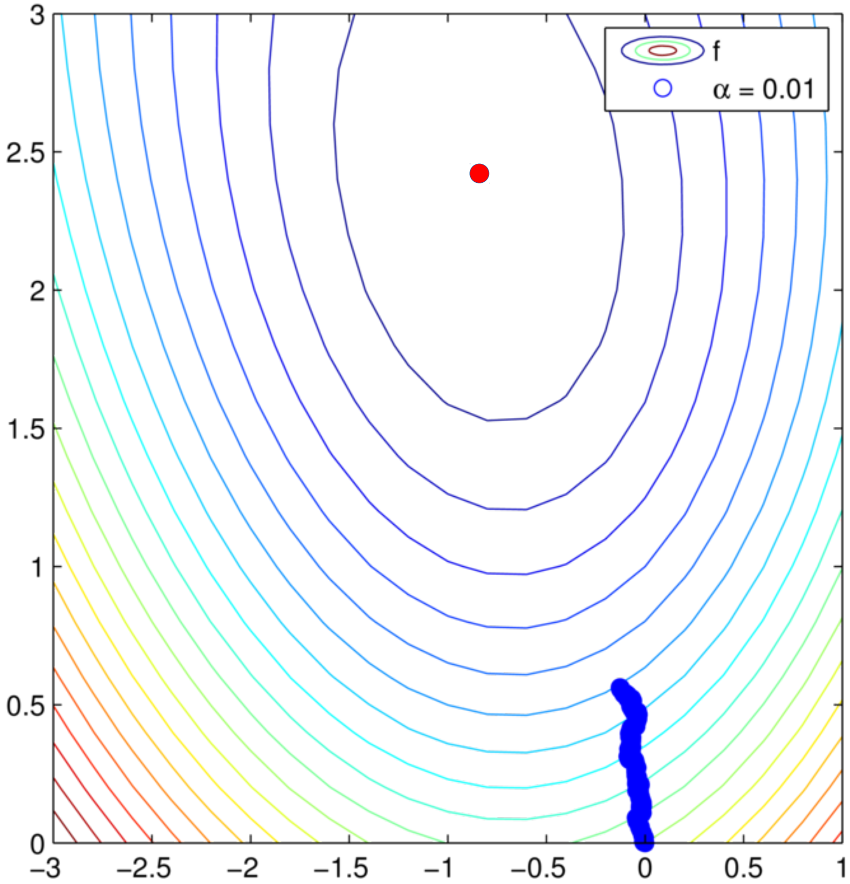
Taking total expectation

$$\begin{aligned}\mathbb{E}\left[||w^{t+1}-w^*||_2^2\right] &\leq (1-\alpha\lambda)\mathbb{E}\left[||w^t-w^*||_2^2\right]+\alpha^2B^2\\ &= (1-\alpha\lambda)^{t+1}||w^0-w^*||_2^2+\textstyle\sum_{i=0}^t(1-\alpha\lambda)^i\alpha^2B^2\end{aligned}$$

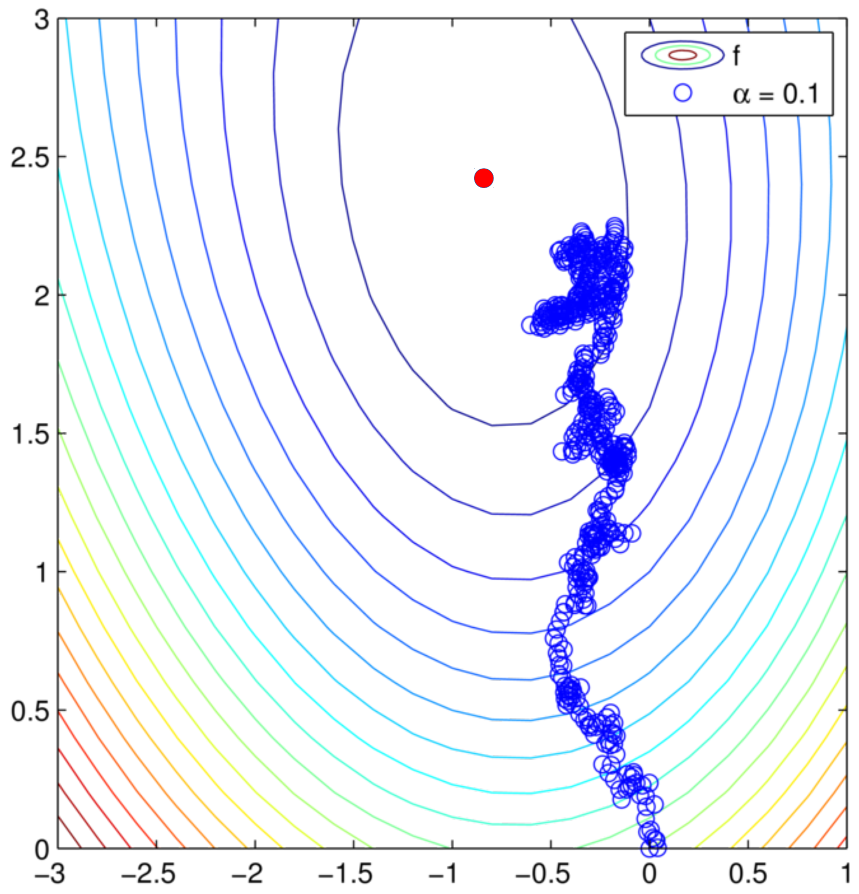
Using the geometric series sum $\displaystyle\sum_{i=0}^t(1-\alpha\lambda)^i=\frac{1-(1-\alpha\mu)^{t+1}}{\alpha\lambda}\leq\frac{1}{\alpha\lambda}$

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

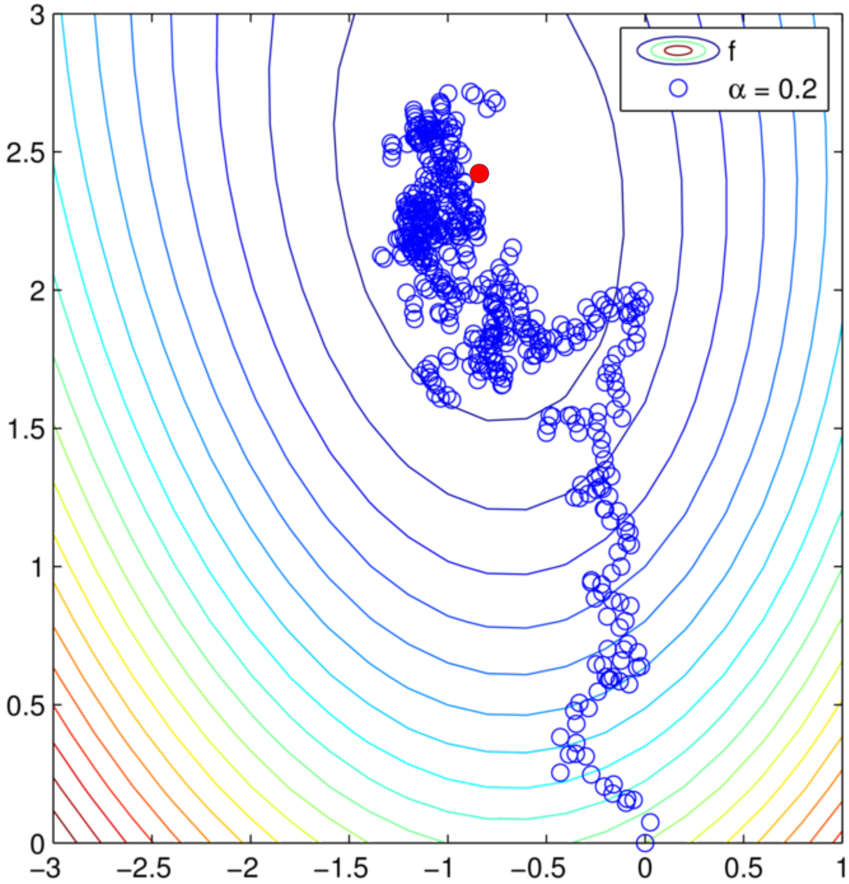
Stochastic Gradient Descent
α =0.01



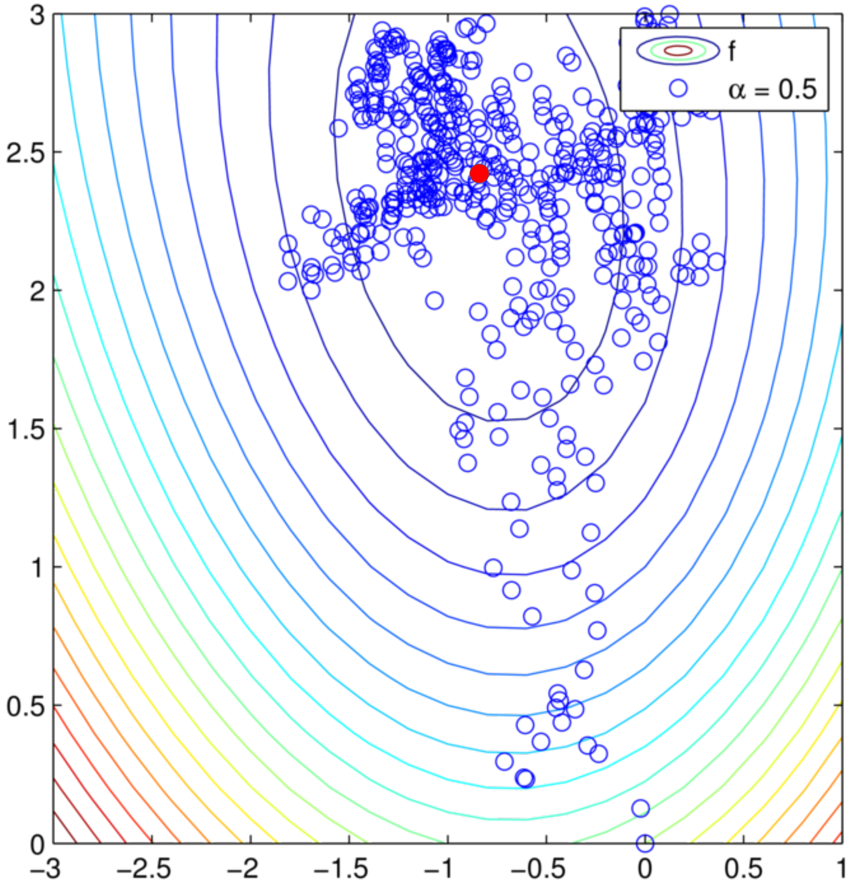
Stochastic Gradient Descent
α =0.1



Stochastic Gradient Descent
α =0.2



Stochastic Gradient Descent
α =0.5



Complexity / Convergence

Theorem (Shrinking stepsize)

If $\alpha_t = \frac{1}{t\lambda}$ then the iterates of the SGD method satisfy

$$\mathbb{E}\left[||w^t - w^*||_2^2\right] \leq \frac{4B^2}{t}$$

Stochastic Gradient Descent Algorithm

Set $w^0 = 0, \alpha_t = \frac{1}{t\lambda}$.
for $t = 1, 2, 3, \ldots, T$
    Sor $j \in \{1, \ldots, n\}$
    $w^{t+1} = w^t - \alpha_t \nabla f_j(w^t)$
Output w^{T+1}

Complexity / Convergence

Theorem (Shrinking stepsize)

If $\alpha_t = \frac{1}{t\lambda}$ then the iterates of the SGD method satisfy

$$\mathbb{E}\left[||w^t - w^*||_2^2\right] \leq \frac{4B^2}{t}$$

Sublinear convergence

Stochastic Gradient Descent Algorithm

Set $w^0 = 0, \alpha_t = \frac{1}{t\lambda}$.
for $t = 1, 2, 3, \ldots, T$
Sor $j \in \{1, \ldots, n\}$
$w^{t+1} = w^t - \alpha_t \nabla f_j(w^t)$
Output w^{T+1}

Complexity / Convergence

Theorem (Shrinking stepsize)

If $\alpha_t = \frac{1}{t\lambda}$ then the iterates of the SGD method satisfy

$$\mathbb{E}\left[||w^t - w^*||_2^2\right] \leq \frac{4B^2}{t}$$

Sublinear convergence

Stochastic Gradient Descent Algorithm

Set $w^0 = 0, \alpha_t = \frac{1}{t\lambda}$.

for $t = 1, 2, 3, \ldots, T$

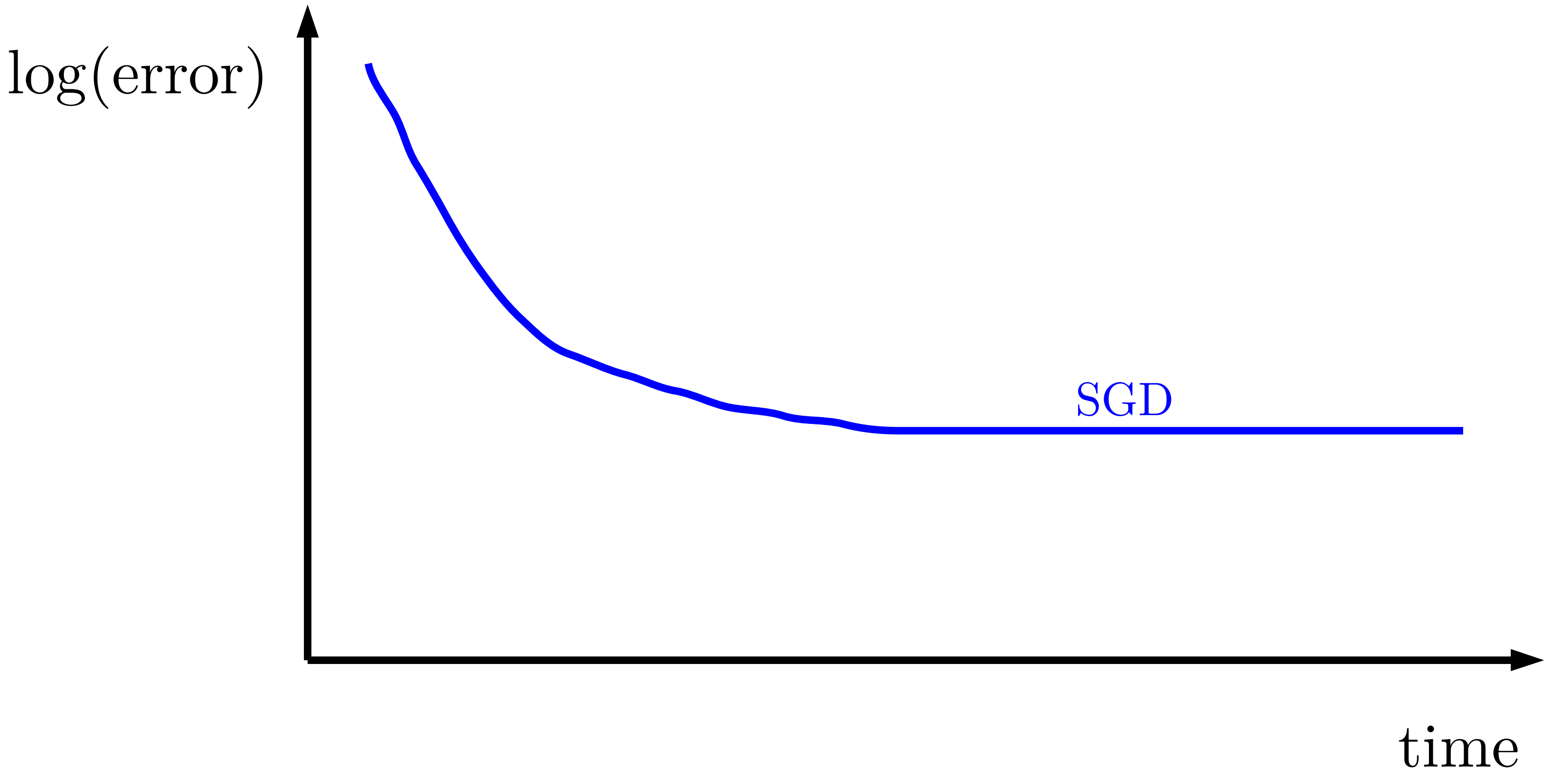
Sor $j \in \{1, \ldots, n\}$

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

Output w^{T+1}

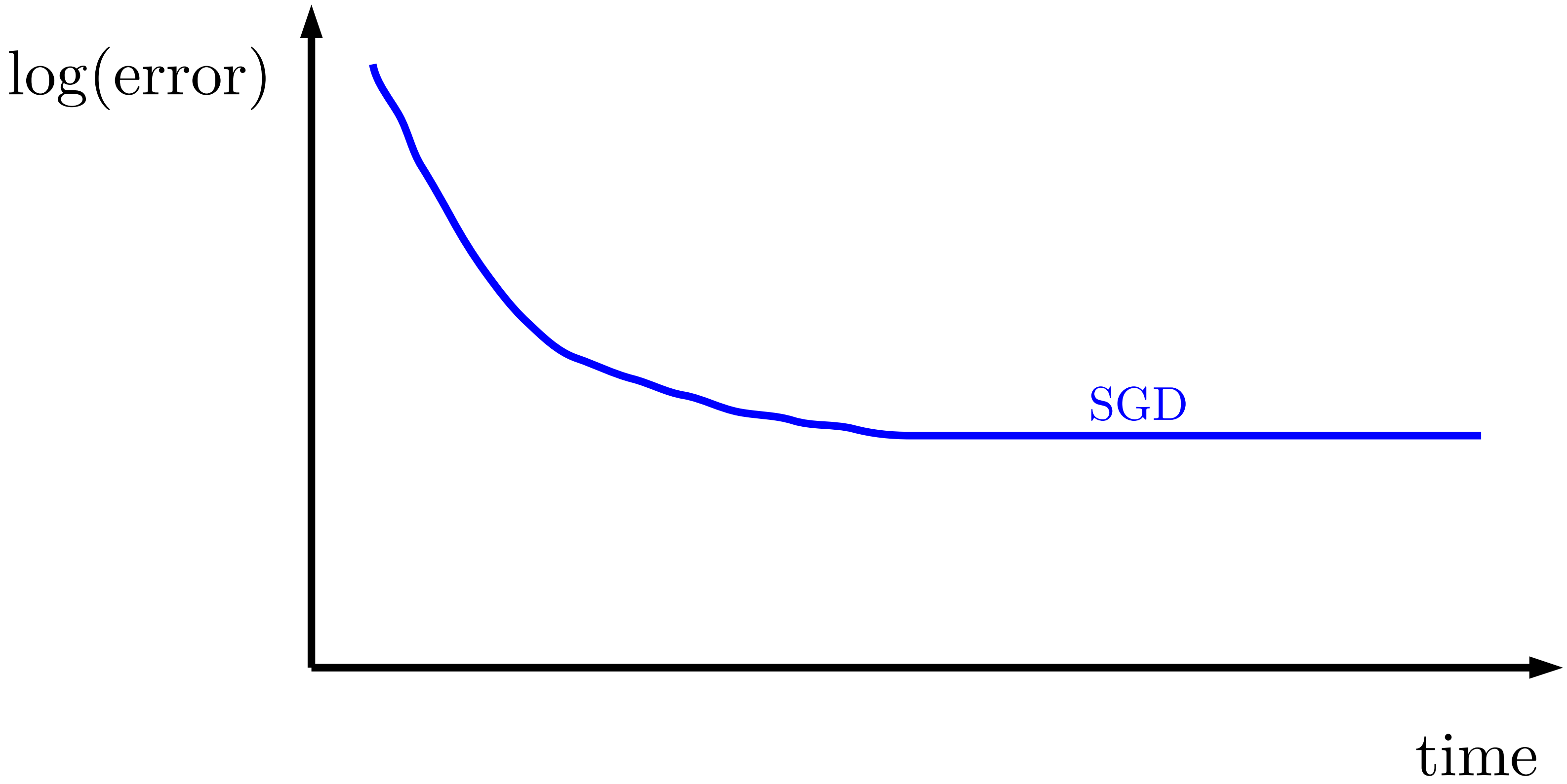
Shrinking Stepsize

Comparison SGD vs GD



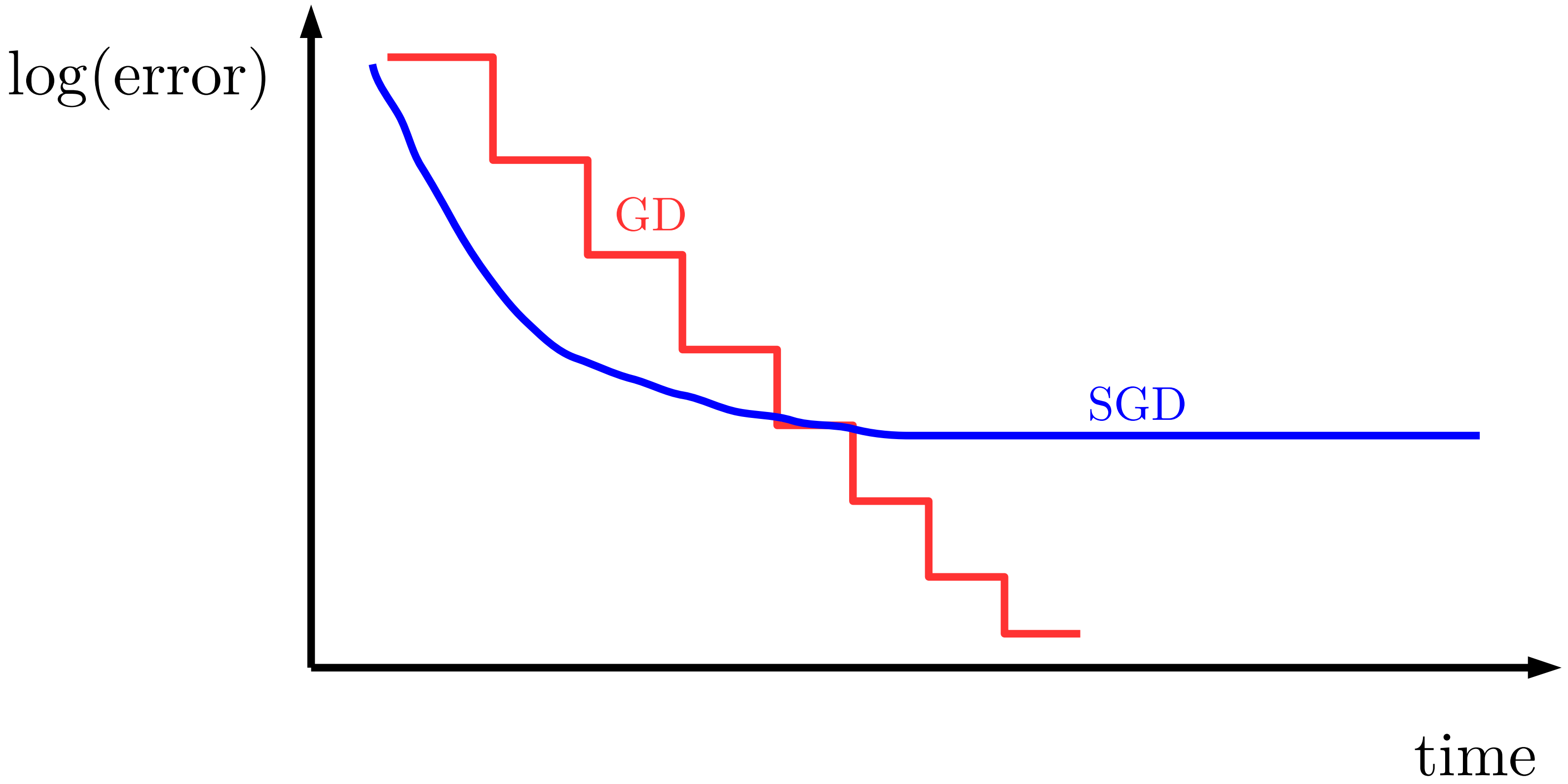
M. Schmidt, N. Le Roux, F. Bach (2016)
Mathematical Programming
Minimizing Finite Sums with the Stochastic Average Gradient.

Comparison SGD vs GD



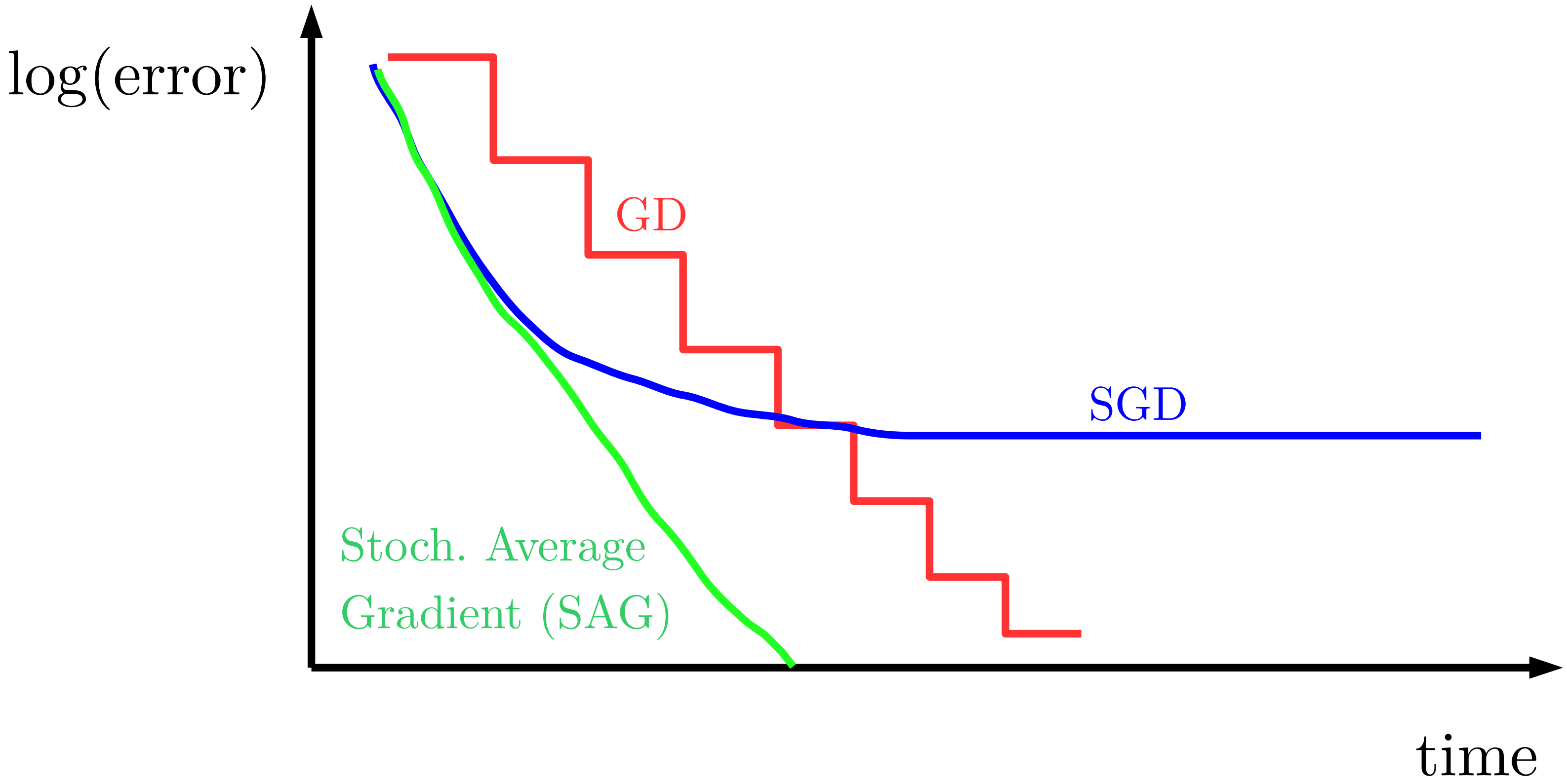
M. Schmidt, N. Le Roux, F. Bach (2016)
Mathematical Programming
Minimizing Finite Sums with the Stochastic Average Gradient.

Comparison SGD vs GD



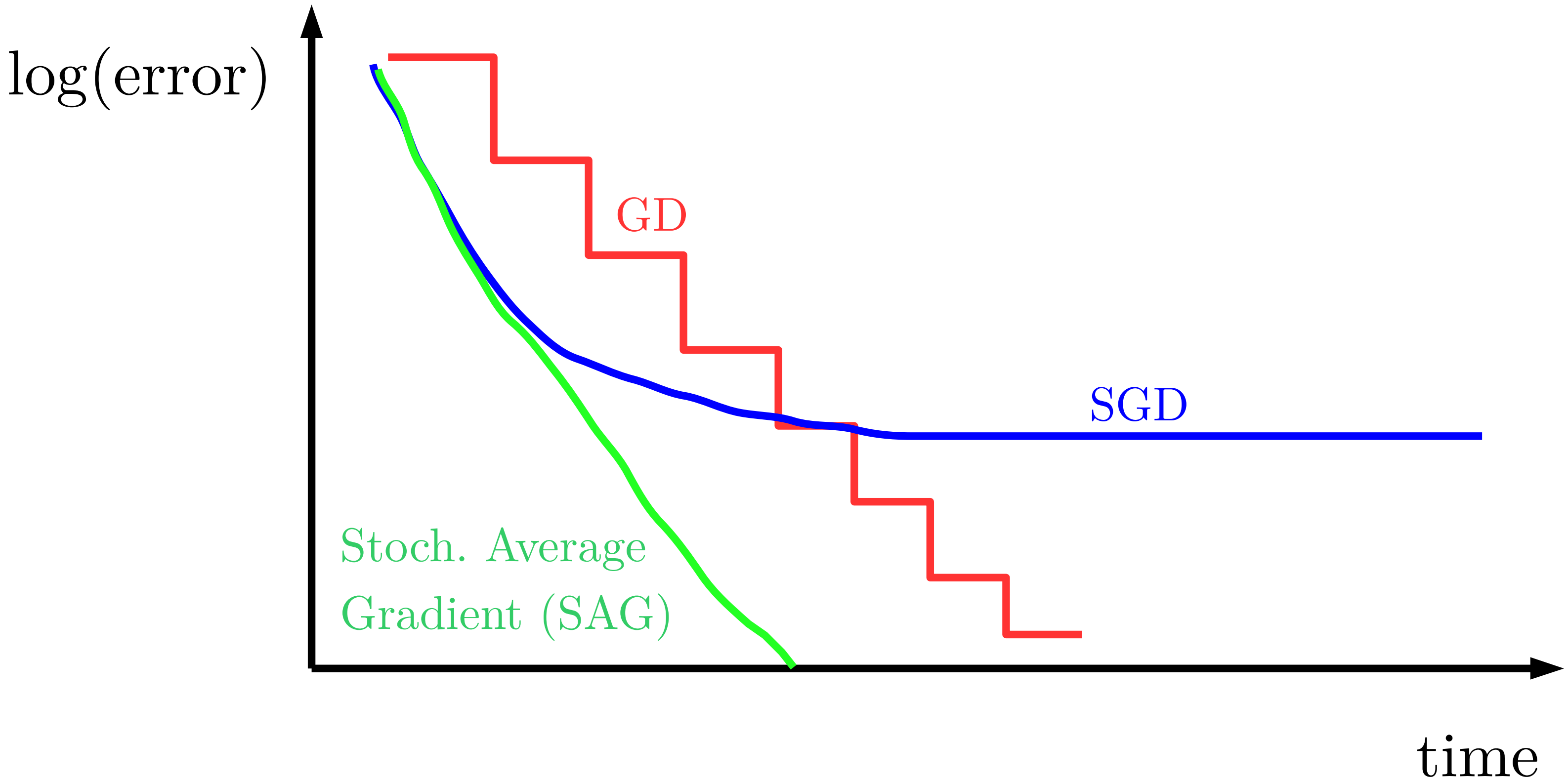
M. Schmidt, N. Le Roux, F. Bach (2016)
Mathematical Programming
Minimizing Finite Sums with the Stochastic Average Gradient.

Comparison SGD vs GD



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

Comparison SGD vs GD



Maybe just an unbiased estimate is not enough.

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

Variance reduced methods through Sketching

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:

Unbiased

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

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:

Unbiased

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

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

Converges in *L2*

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

Example: The Stochastic Average Gradient

Maintain $J^t \approx [\nabla f_1(w^t), \dots, \nabla f_n(w^t)]$ and iterate

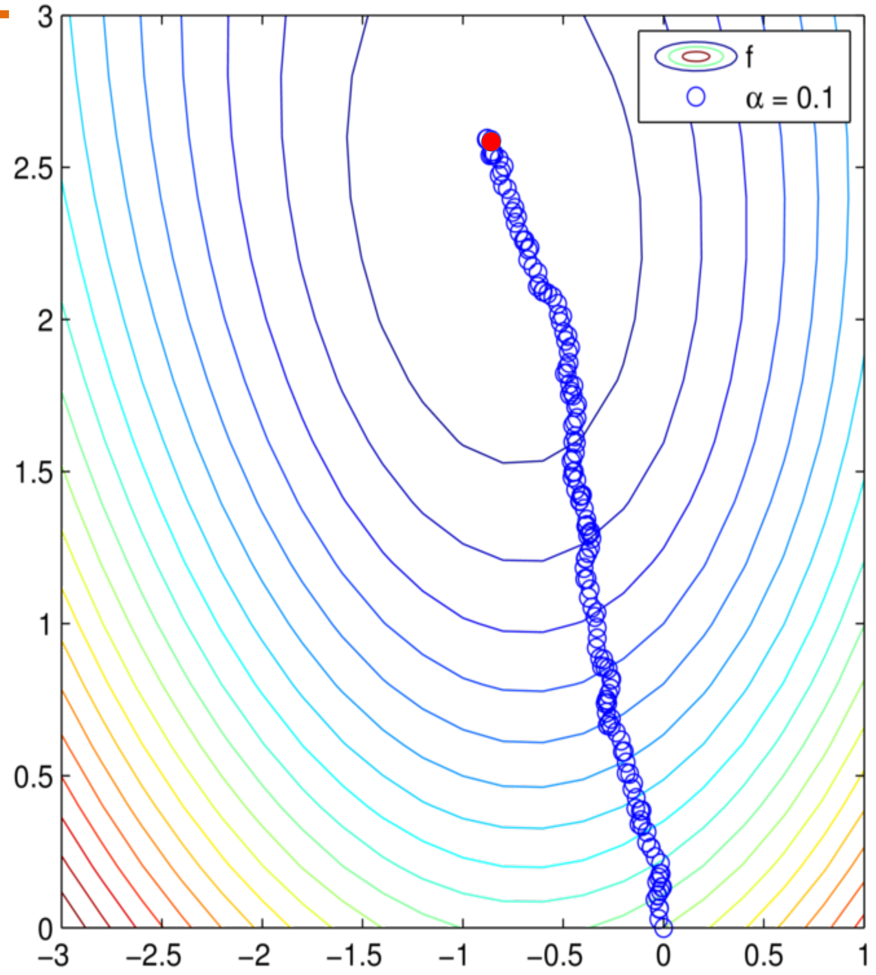
$$w^{t+1} = w^t - \frac{\alpha}{n} \sum_{i=1}^{n} J_i^t = w^t - \alpha g^t$$

Update J_i^t's by sampling $j \in \{1, \dots, n\}$ uniformly at random and setting:

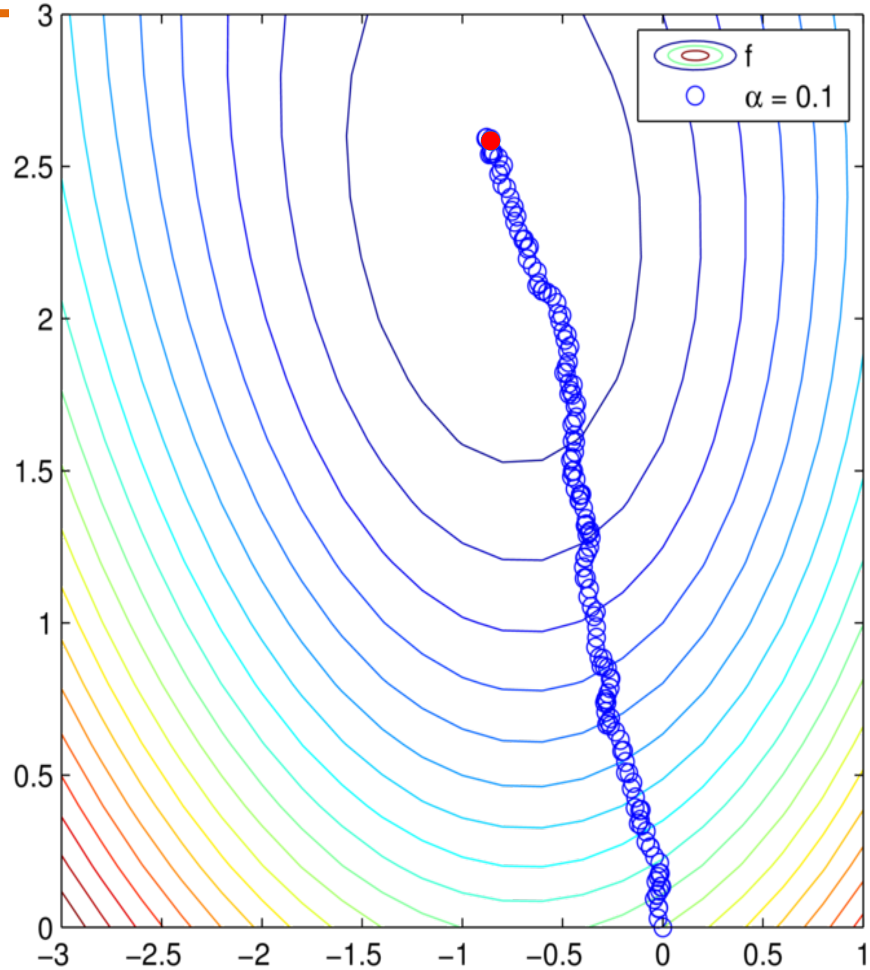
$$J_i^t = \begin{cases} J_i^t = \nabla f_i(w^t) & \text{if } i = j \\ J_i^t = J_i^{t-1} & \text{if } i \neq j \end{cases}$$

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

The Stochastic Average Gradient



The Stochastic Average Gradient



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

Introducing the Jacobian

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

$$F(w) \stackrel{\text{def}}{=} (f_1(w), \ldots, f_n(w))$$

$$DF(w) = (\nabla f_1(w), \ldots, \nabla f_n(w))$$

Introducing the Jacobian

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

$$F(w) \overset{\text{def}}{=} (f_1(w), \ldots, f_n(w))$$

$$DF(w) = (\nabla f_1(w), \ldots, \nabla f_n(w))$$

$$\nabla f(w) = \frac{1}{n} DF(w) \mathbf{1}, \quad \text{where } \mathbf{1}^\top = (1, 1, \ldots, 1) \in \mathbf{R}^n$$

Introducing the Jacobian

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

$$F(w) \stackrel{\text{def}}{=} (f_1(w), \ldots, f_n(w))$$

$$DF(w) = (\nabla f_1(w), \ldots, \nabla f_n(w))$$

$$\nabla f(w) = \frac{1}{n} DF(w) \mathbf{1}, \quad \text{where } \mathbf{1}^\top = (1, 1, \ldots, 1) \in \mathbf{R}^n$$

$\nabla f(w)$ is a *dense* linear meassurement of $DF(w)$

The Stochastic Average Gradient

Maintain $J^t \approx [\nabla f_1(w^t), \ldots, \nabla f_n(w^t)] = DF(w^t)$ and iterate

$$w^{t+1} = w^t - \frac{\alpha}{n} \sum_{i=1}^{n} J_i^t$$

Update J_i^t's by sampling $j \in \{1, \ldots, n\}$ uniformly at random and setting:

$$J_i^t = \begin{cases} J_i^t = \nabla f_i(w^t) & \text{if } i = j \\ J_i^t = J_i^{t-1} & \text{if } i \neq j \end{cases}$$

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

The Stochastic Average Gradient

Maintain $J^t \approx [\nabla f_1(w^t), \ldots, \nabla f_n(w^t)] = DF(w^t)$ and iterate

$$w^{t+1} = w^t - \frac{\alpha}{n}\sum_{i=1}^{n} J_i^t$$

Estimate of $\frac{1}{n}DF(w^t)\mathbf{1}$

Update J_i^t's by sampling $j \in \{1, \ldots, n\}$ uniformly at random and setting:

$$J_i^t = \begin{cases} J_i^t = \nabla f_i(w^t) & \text{if } i = j \\ J_i^t = J_i^{t-1} & \text{if } i \neq j \end{cases}$$

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

The Stochastic Average Gradient

Maintain $J^t \approx [\nabla f_1(w^t), \ldots, \nabla f_n(w^t)] = DF(w^t)$ and iterate

$$w^{t+1} = w^t - \frac{\alpha}{n}\sum_{i=1}^{n} J_i^t$$

Estimate of $\frac{1}{n} DF(w^t)\mathbf{1}$

Update J_i^t's by sampling $j \in \{1, \ldots, n\}$ uniformly at random and setting:

Stoch. Linear Measurement $DF(w^t)e_j$

$$J_i^t = \begin{cases} J_i^t = \nabla f_i(w^t) & \text{if } i = j \\ J_i^t = J_i^{t-1} & \text{if } i \neq j \end{cases}$$

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

Stochastic Sparse Sketches

Sparse Stochastic Matrix

$$S \in \mathbf{R}^{n\times\tau} \text{ a sparse matrix and } \tau \ll d$$

$$S \sim \mathcal{D} \text{ fixed distribution}$$

Stochastic Sketch

$$DF(w)S = \sum_{i=1}^{\tau} DF(w)S_{:i}$$

Stochastic Sparse Sketches

Sparse Stochastic Matrix

$$S \in \mathbf{R}^{n\times\tau} \text{ a sparse matrix and } \tau \ll d$$

$$S \sim \mathcal{D} \text{ fixed distribution}$$

Stochastic Sketch

$$DF(w)S = \sum_{i=1}^{\tau} DF(w)S_{:i}$$

Eg: SGD Sketch

$$S = e_j \in \mathbf{R}^d \text{ the } j\text{th unit coordinate vector}$$

$$\text{with } \mathbb{P}(S = e_j) = \frac{1}{n}$$

$$DF(x)S = \nabla f_j(w)$$

Stochastic Sparse Sketches

Eg: Mini-batch SGD Sketch

$S = I_C \in \mathbf{R}^{n \times \tau}$ where $C \subset \{1, \ldots, n\}$

$$DF(w)S = [\nabla f_{C_1}(w), \ldots, \nabla f_{C_\tau}(w)]$$

Exe. $\tau = 3, n = 6, \quad S = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ and $DF(w)S = [\nabla f_1(w), \nabla f_4(w), \nabla f_6(w)]$

Many examples: Sparse Rademacher matrices, sampling with replacement, nonuniform...etc

A Jacobian Based Method

Maintain Jacobian Estimate

$$J^{t-1} \approx DF(w^{t-1})$$

+

Sample Stochastic Sketch

$$S \sim \mathcal{D}$$

$$DF(w^t)S$$

A Jacobian Based Method

Maintain Jacobian Estimate

$$J^{t-1} \approx DF(w^{t-1})$$

Sample Stochastic Sketch

$$S \sim \mathcal{D}$$

$$DF(w^t)S$$

? ?

Improved Guess

$$J^t \approx DF(w^t)$$

A Jacobian Based Method

Jacobian Sketching Algorithm

Set $\alpha > 0, w^1 = 0, J^0 \in \mathbb{R}^{d\times n}$
For $t = 1, \ldots, T$
  Sample $S \sim \mathcal{D}$
  Calculate Sketch $DF(w^t)S$
  Update J^t using $DF(w^t)S$ and J^{t-1}
  Calculate $g^t = \frac{1}{n} J^t \mathbf{1}$
  Step $w^{t+1} = w^t - \alpha g^t$.

A Jacobian Based Method

Jacobian Sketching Algorithm

Set $\alpha > 0, w^1 = 0, J^0 \in \mathbb{R}^{d \times n}$
For $t = 1, \ldots, T$
  Sample $S \sim \mathcal{D}$
  Calculate Sketch $DF(w^t)S$
  Update J^t using $DF(w^t)S$ and J^{t-1}
  Calculate $g^t = \frac{1}{n} J^t \mathbf{1}$
  Step $w^{t+1} = w^t - \alpha g^t$.

$\approx \frac{1}{n} DF(w)\mathbf{1}$

A Jacobian Based Method

Jacobian Sketching Algorithm

Set $\alpha > 0, w^1 = 0, J^0 \in \mathbb{R}^{d\times n}$
For $t = 1, \ldots, T$
  Sample $S \sim \mathcal{D}$
  Calculate Sketch $DF(w^t)S$
? Update J^t using $DF(w^t)S$ and J^{t-1} ?
  Calculate $g^t = \frac{1}{n} J^t \mathbf{1}$
  Step $w^{t+1} = w^t - \alpha g^t$.

$\approx \frac{1}{n} DF(w)\mathbf{1}$

Updating the Jacobian Estimate: Sketch and project

$$J^t = DF(w^t)$$

Updating the Jacobian Estimate: Sketch and project

$$J^t S \;=\; DF(w^t)S, \qquad S \sim \mathcal{D}$$

Updating the Jacobian Estimate: Sketch and project

$$J^t = \arg\min_{J \in \mathbb{R}^{d \times n}} ||J - J^{t-1}||_F^2$$

$$J^t S \;=\; DF(w^t)S, \qquad S \sim \mathcal{D}$$

Updating the Jacobian Estimate: Sketch and project

Sketch and Project the Jacobian

$$J^t = \arg\min_{J \in \mathbb{R}^{d\times n}} ||J - J^{t-1}||_F^2$$

$$J^t S \ = \ DF(w^t)S, \qquad S \sim \mathcal{D}$$

Updating the Jacobian Estimate: Sketch and project

Sketch and Project the Jacobian

$$J^t = \arg \min_{J \in \mathbb{R}^{d\times n}} ||J - J^{t-1}||_F^2$$

$$J^t S \;=\; DF(w^t)S, \qquad S \sim \mathcal{D}$$

RMG and Peter Richtarik (2015)
Randomized iterative methods for linear systems
SIAM Journal on Matrix Analysis and Applications 36(4)

Exercise

$$J^t = \arg \min_{J \in \mathbb{R}^{d \times n}} ||J - J^{t-1}||_F^2$$
$$\text{subject to} \quad JS = DF(w^t)S$$

Show that the solution J^t is given by

Solution: $$J^t = J^{t-1} - (J^{t-1} - DF(w^t))S(S^\top S)^{-1}S^\top$$

Proof: The Lagrangian is given by

Exercise

$$J^t = \arg \min_{J \in \mathbb{R}^{d \times n}} ||J - J^{t-1}||_F^2$$
$$\text{subject to} \quad JS = DF(w^t)S$$

Show that the solution J^t is given by

Solution: $J^t = J^{t-1} - (J^{t-1} - DF(w^t))S(S^\top S)^{-1}S^\top$

Proof: The Lagrangian is given by

$$\begin{aligned} L(J, Y) \quad &:= \tfrac{1}{2}||J - J^{t-1}||_F^2 + \langle Y, (DF^t - J)S \rangle \\ &= \tfrac{1}{2}||J - J^{t-1}||_F^2 + \langle YS^\top, DF^t - J \rangle \end{aligned}$$

Exercise

$$J^t = \arg \min_{J \in \mathbb{R}^{d \times n}} ||J - J^{t-1}||_F^2$$
$$\text{subject to} \quad JS = DF(w^t)S$$

Show that the solution J^t is given by

Solution: $J^t = J^{t-1} - (J^{t-1} - DF(w^t))S(S^\top S)^{-1}S^\top$

Proof: The Lagrangian is given by

$$\begin{aligned} L(J, Y) \quad &:= \tfrac{1}{2}||J - J^{t-1}||_F^2 + \langle Y, (DF^t - J)S \rangle \\ &= \tfrac{1}{2}||J - J^{t-1}||_F^2 + \langle YS^\top, DF^t - J \rangle \end{aligned}$$

Differentiating in J and setting to zero: $\quad YS^\top = J - J^{t-1} \qquad (1)$

Exercise

$$
\begin{aligned}
J^t = \arg\min_{J \in \mathbb{R}^{d\times n}} & \; ||J - J^{t-1}||_F^2 \\
\text{subject to} & \quad JS = DF(w^t)S
\end{aligned}
$$

Show that the solution J^t is given by

Solution: $$J^t = J^{t-1} - (J^{t-1} - DF(w^t))S(S^\top S)^{-1}S^\top$$

Proof: The Lagrangian is given by

$$
\begin{aligned}
L(J, Y) \quad & := \tfrac{1}{2}||J - J^{t-1}||_F^2 + \langle Y, (DF^t - J)S \rangle \\
& = \tfrac{1}{2}||J - J^{t-1}||_F^2 + \langle YS^\top, DF^t - J \rangle
\end{aligned}
$$

Differentiating in J and setting to zero: $$YS^\top = J - J^{t-1} \qquad (1)$$

Right multiplying by $S(S^\top S)^{-1}$ gives : $$Y = (DF^t - J^{t-1})S(S^\top S)^{-1} \qquad (2)$$

Exercise

$$J^t = \arg\min_{J \in \mathbb{R}^{d \times n}} ||J - J^{t-1}||_F^2$$

$$\text{subject to} \quad JS = DF(w^t)S$$

Show that the solution J^t is given by

Solution: $J^t = J^{t-1} - (J^{t-1} - DF(w^t))S(S^\top S)^{-1}S^\top$

Proof: The Lagrangian is given by

$$\begin{aligned} L(J, Y) \quad &:= \tfrac{1}{2}||J - J^{t-1}||_F^2 + \langle Y, (DF^t - J)S\rangle \\ &= \tfrac{1}{2}||J - J^{t-1}||_F^2 + \langle YS^\top, DF^t - J\rangle \end{aligned}$$

Differentiating in J and setting to zero: $\quad YS^\top = J - J^{t-1} \quad (1)$

Right multiplying by $S(S^\top S)^{-1}$ gives : $\quad Y = (DF^t - J^{t-1})S(S^\top S)^{-1} \quad (2)$

Substituting (1) into (2) gives the solution.

Sketch and project the Jacobian

$$J^t = \arg \min_{J \in \mathbb{R}^{d \times n}} ||J - J^{t-1}||_F^2$$
$$\text{subject to} \quad JS = DF(w^t)S$$

Solution:

$$J^t = J^{t-1} - (J^{t-1} - DF(w^t))S(S^\top S)^{-1}S^\top$$

$$g^t = \frac{1}{n}J^{t-1}\mathbf{1} - \frac{\eta}{n}(J^{t-1} - DF(w^t))S(S^\top S)^{-1}S^\top \mathbf{1}$$

Sketch and project the Jacobian

$$
\begin{aligned}
J^t = \arg \min_{J \in \mathbb{R}^{d\times n}} & \; ||J - J^{t-1}||_F^2 \\
\text{subject to} & \quad JS = DF(w^t)S
\end{aligned}
$$

Solution:

$$J^t = J^{t-1} - (J^{t-1} - DF(w^t))S(S^\top S)^{-1}S^\top$$

$$g^t = \frac{1}{n}J^{t-1}\mathbf{1} - \frac{\eta}{n}(J^{t-1} - DF(w^t))S(S^\top S)^{-1}S^\top\mathbf{1}$$

If $\eta = 1$ then $\mathrm{g}^t = \frac{1}{n}J^t\mathbf{1}$

Sketch and project the Jacobian

$$J^t = \arg\min_{J \in \mathbb{R}^{d\times n}} ||J - J^{t-1}||^2_{F(W)}$$

$$\text{subject to} \quad JS = DF(w^t)S$$

Solution:

$$J^t = J^{t-1} - (J^{t-1} - DF(w^t))S(S^\top W^{-1} S)^{-1} S^\top W^{-1}$$

$$g^t = \frac{1}{n} J^{t-1}\mathbf{1} - \frac{\eta}{n}(J^{t-1} - DF(w^t))S(S^\top W^{-1} S)^{-1} S^\top W^{-1}\mathbf{1}$$

Sketch and project the Jacobian

$$J^t = \arg \min_{J \in \mathbb{R}^{d \times n}} ||J - J^{t-1}||_F^2$$
$$\text{subject to} \quad JS = DF(w^t)S$$

Solution:

$$J^t = J^{t-1} - (J^{t-1} - DF(w^t))S(S^\top S)^{-1}S^\top$$

$$g^t = \frac{1}{n}J^{t-1}\mathbf{1} - \frac{\eta}{n}(J^{t-1} - DF(w^t))S(S^\top S)^{-1}S^\top \mathbf{1}$$

Sketch and project the Jacobian

$$J^t = \arg \min_{J \in \mathbb{R}^{d \times n}} ||J - J^{t-1}||_F^2$$
$$\text{subject to} \quad JS = DF(w^t)S$$

Solution:

$$J^t = J^{t-1} - (J^{t-1} - DF(w^t))\boxed{S(S^\top S)^{-1}S^\top \; =: P_S}$$

$$g^t = \frac{1}{n} J^{t-1}\mathbf{1} - \frac{\eta}{n}(J^{t-1} - DF(w^t))S(S^\top S)^{-1}S^\top \mathbf{1}$$

Unbiased Condition

Lemma. If $(\frac{1}{\eta}, \mathbf{1})$ is an eigenpair of $\mathbb{E}[P_S]$ then

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

consequently g^t is an unbiased estimator.

Proof: $g^t = g^{t-1} - \frac{\eta}{n}(J^{t-1} - DF(w^t))S(S^\top S)^{-1}S^\top \mathbf{1}$

Unbiased Condition

Lemma. If $(\frac{1}{\eta}, \mathbf{1})$ is an eigenpair of $\mathbb{E}[P_S]$ then

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

consequently g^t is an unbiased estimator.

Proof: $$g^t = g^{t-1} - \frac{\eta}{n}(J^{t-1} - DF(w^t))S(S^\top S)^{-1}S^\top \mathbf{1}$$

Unbiased Condition

Lemma. If $(\frac{1}{\eta}, \mathbf{1})$ is an eigenpair of $\mathbb{E}[P_S]$ then

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

consequently g^t is an unbiased estimator.

Proof:

$$g^t = g^{t-1} - \frac{\eta}{n}(J^{t-1} - DF(w^t))S(S^\top S)^{-1}S^\top \mathbf{1}$$

$$\begin{aligned}
\mathbb{E}_S[g^t] &= \frac{1}{n}J^{t-1}\mathbf{1} - \frac{\eta}{n}(J^{t-1} - DF(w^t))\mathbb{E}_S\underbrace{[S(S^\top S)^{-1}S^\top]}_{P_S}\mathbf{1} \\
&= \frac{1}{n}J^{t-1}\mathbf{1} - \frac{\eta}{n\eta}(J^{t-1} - DF(w^t))\mathbf{1} \\
&= \cancel{\frac{1}{n}J^{t-1}}\mathbf{1} - \cancel{\frac{1}{n}J^{t-1}}\mathbf{1} + \frac{1}{n}DF(w^t))\mathbf{1} \quad = \quad \nabla f(w^t)
\end{aligned}$$

Exercise

Let $\mathbb{P}[S = e_i] = \dfrac{1}{n}$ for $i = 1, \ldots, n$. Show that

$$\mathbb{E}[P_S]\mathbf{1} = \mathbb{E}[S(S^\top S)^{-1}S^\top]\mathbf{1} = \frac{1}{n}\mathbf{1}$$

Proof:

Exercise

Let $\mathbb{P}[S = e_i] = \dfrac{1}{n}$ for $i = 1, \ldots, n$. Show that

$$\boxed{\mathbb{E}[P_S]\mathbf{1} = \mathbb{E}[S(S^\top S)^{-1}S^\top]\mathbf{1} = \frac{1}{n}\mathbf{1}}$$

Proof:

Exercise

Let $\mathbb{P}[S = e_i] = \dfrac{1}{n}$ for $i = 1, \ldots, n$. Show that

$$\boxed{\mathbb{E}[P_S]\mathbf{1} = \mathbb{E}[S(S^\top S)^{-1}S^\top]\mathbf{1} = \frac{1}{n}\mathbf{1}}$$

Proof:

$$\begin{aligned}
\mathbb{E}[S(S^\top S)^{-1}S^\top]\mathbf{1} &= \textstyle\sum_{i=1}^n \frac{1}{n} \frac{e_i e_i^\top}{e_i^\top e_i} \\
&= \frac{1}{n}\sum_{i=1}^n e_i e_i^\top \mathbf{1} \\
&= \frac{1}{n} I \mathbf{1} = \frac{1}{n}\mathbf{1}
\end{aligned}$$

A Jacobian Based Method

Archetype Jacobian Sketching Algorithm

Choose distribution $\mathcal{D}$ and unbiased $\eta > 0$
Set $\alpha > 0, w^1 = 0, J^0 \in \mathbb{R}^{d \times n}$
For $t = 1, \ldots, T$
Sample $S \sim \mathcal{D}$
Calculate Sketch $DF(w^t)S$
Update $J^t = J^{t-1} - (J^{t-1} - DF(w^t))S(S^\top S)^{-1}S^\top$
Calculate $g^t = \frac{1}{n} J^{t-1}\mathbf{1} - \frac{\eta}{n}(J^{t-1} - DF(w^t))S(S^\top S)^{-1}S^\top \mathbf{1}$
Step $w^{t+1} = w^t - \alpha g^t$

A Jacobian Based Method

Archetype Jacobian Sketching Algorithm

Choose distribution $\mathcal{D}$ and unbiased $\eta > 0$

Set $\alpha > 0, w^1 = 0, J^0 \in \mathbb{R}^{d\times n}$

For $t = 1, \ldots, T$

Sample $S \sim \mathcal{D}$

Calculate Sketch $DF(w^t)S$

Update $J^t = J^{t-1} - (J^{t-1} - DF(w^t))S(S^\top S)^{-1}S^\top$

Calculate $g^t = \frac{1}{n}J^{t-1}\mathbf{1} - \frac{\eta}{n}(J^{t-1} - DF(w^t))S(S^\top S)^{-1}S^\top\mathbf{1}$

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

Looks expensive and complicated. Investigate

Example: minibatch-SAGA

Homework:

$$\mathbb{E}[P_S]\mathbf{1} = \frac{\tau}{n}\mathbf{1}$$

Let $C \subset \{1, \ldots, n\}$ with $|C| = \tau$ and $\mathbb{P}[S = I_C] = \frac{1}{\binom{n}{\tau}}$

Exe. $\tau = 3, n = 6, \quad S = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ and $DF(w)S = [\nabla f_1(w), \nabla f_4(w), \nabla f_6(w)]$

Example: minibatch-SAGA

Homework:

$$\mathbb{E}[P_S]\mathbf{1} = \frac{\tau}{n}\mathbf{1}$$

Let $C \subset \{1, \ldots, n\}$ with $|C| = \tau$ and $\mathbb{P}[S = I_C] = \frac{1}{\binom{n}{\tau}}$

Exe. $\tau = 3, n = 6, \quad S = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ and $DF(w)S = [\nabla f_1(w), \nabla f_4(w), \nabla f_6(w)]$

Jacobain update

$$J_j^t = \begin{cases} \nabla f_j(w^t) & \text{if } j \in C, \\ J_j^{t-1} & \text{if } j \neq C. \end{cases}$$

Gradiant estimate

$$g^t = \frac{1}{n} J^{t-1}\mathbf{1} - \frac{1}{\tau}\sum_{j \in C}(J_j^{t-1} - \nabla f_j(w^t))$$

Proving Convergence of Variance reduced methods