Optimization for Machine Learning

Stochastic Gradient Methods

Lecturer: Robert M. Gower

Master IASD: AI Systems and Data Science, 2019

Core Info

- **Where**: ENS: 07/11 amphi Langevin, 03/12 U209, 05/12 amphi Langevin.

- **Online:** Teaching materials for these 3 classes:
  https://gowerrobert.github.io/

- **Google docs with course info:** Can also be found on
  https://gowerrobert.github.io/

Outline of my three classes

- 07/11/19 Foundations and the empirical risk problem, revision probability, SGD (Stochastic Gradient Descent) for ridge regression
- 03/12/19 (**TODAY**) SGD for convex optimization. Theory, variants including averaging, decreasing stepsizes and momentum.
- 05/12/19 Lab on SGD and variants **BRING LAPTOPS!**

Solving the Finite Sum Training Problem

Recap

Training Problem

$$\min_{w \in \mathbf{R}^d} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell\left(h_w(x^i), y^i\right)}_{L(w) \,=\, loss} + \lambda R(w) =: f(w)$$

General methods

$\min f(w)$

- Gradient Descent

Two parts

$\min L(w) + \lambda R(w)$

- Proximal gradient (ISTA)
- Fast proximal gradient (FISTA)

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

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

Can we use this sum structure?

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

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

Output w^T

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 $t = 0, 1, 2, \ldots, T$
$$w^{t+1} = w^t - \frac{\alpha}{n} \sum_{i=1}^{n} \nabla f_i(w^t)$$
Output w^T

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

$$\mathbb{E}_j[\nabla f_j(w)] = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(w) = \nabla f(w)$$

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 $\{1, \dots, n\}$ selected uniformly at random. Then

$$\mathbb{E}_j[\nabla f_j(w)] = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(w) = \nabla f(w)$$

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

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

$$\mathbb{E}_j[\nabla f_j(w)] = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(w) = \nabla f(w)$$

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

EXE: Let $\sum_{i=1}^{n} p_i = 1$ and $j \sim p_j$. Show $\mathbb{E}[\nabla f_j(w)/(np_j)] = \nabla f(w)$

Stochastic Gradient Descent

SGD 0.0 Constant stepsize

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

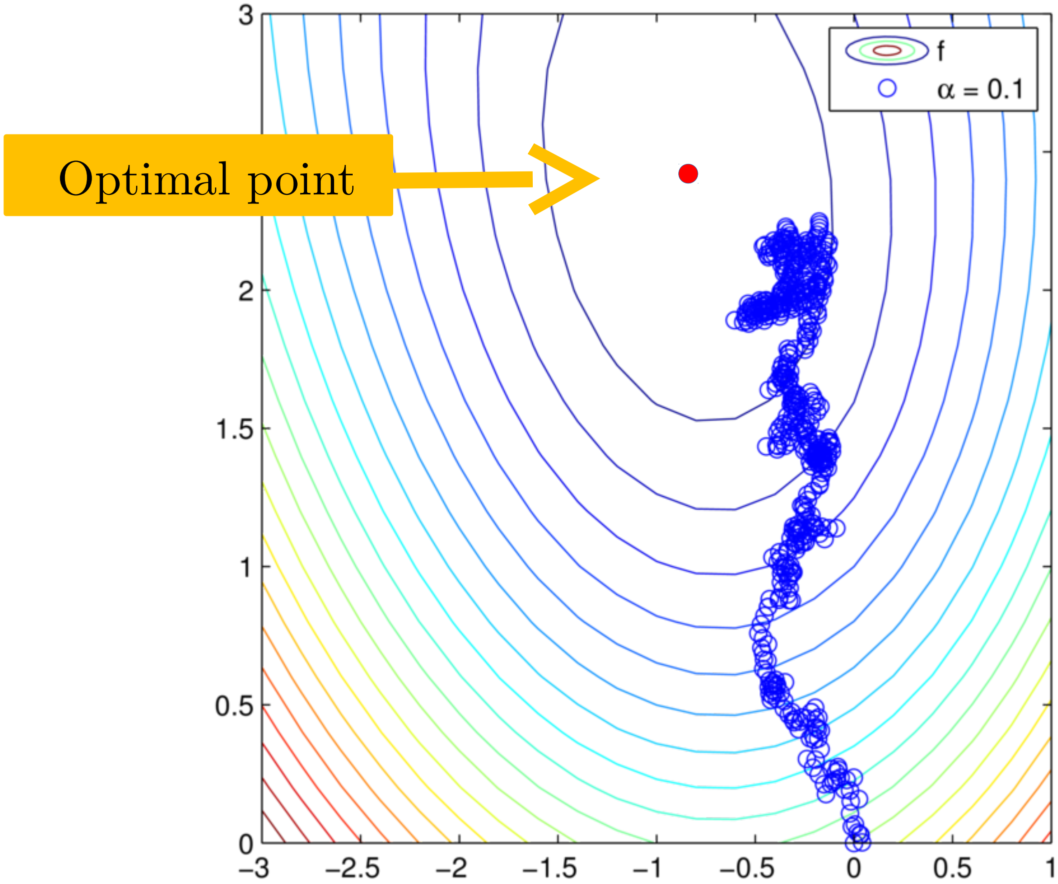
for $t = 0, 1, 2, \ldots, T-1$

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

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

Output w^T

Stochastic Gradient Descent



Assumptions for Convergence

Strong Convexity

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

$$y = w^*$$

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

Assumptions for Convergence

Strong Convexity

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

$$\Big\Downarrow \; y = w^*$$

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

Assumptions for Convergence

Strong Convexity

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

$$y = w^*$$

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

Expected Bounded Stochastic Gradients

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

Assumptions for Convergence

Strong Convexity

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

$$y = w^*$$

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

Expected Bounded Stochastic Gradients

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

Complexity / Convergence

Theorem

If $0 < \alpha \leq \frac{1}{\lambda}$ then the iterates of the SGD 0.0 method satisfy

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

EXE: Do exercises on convergence of random sequences.

Complexity / Convergence

Theorem

If $0 < \alpha \leq \frac{1}{\lambda}$ then the iterates of the SGD 0.0 method satisfy

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

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

EXE: Do exercises on convergence of random sequences.

Complexity / Convergence

Theorem

If $0 < \alpha \leq \frac{1}{\lambda}$ then the iterates of the SGD 0.0 method satisfy

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

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

Shows that $\alpha \approx 0$

EXE: Do exercises on convergence of random sequences.

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

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

Strong conv.

Bounded Stoch grad

Taking total expectation

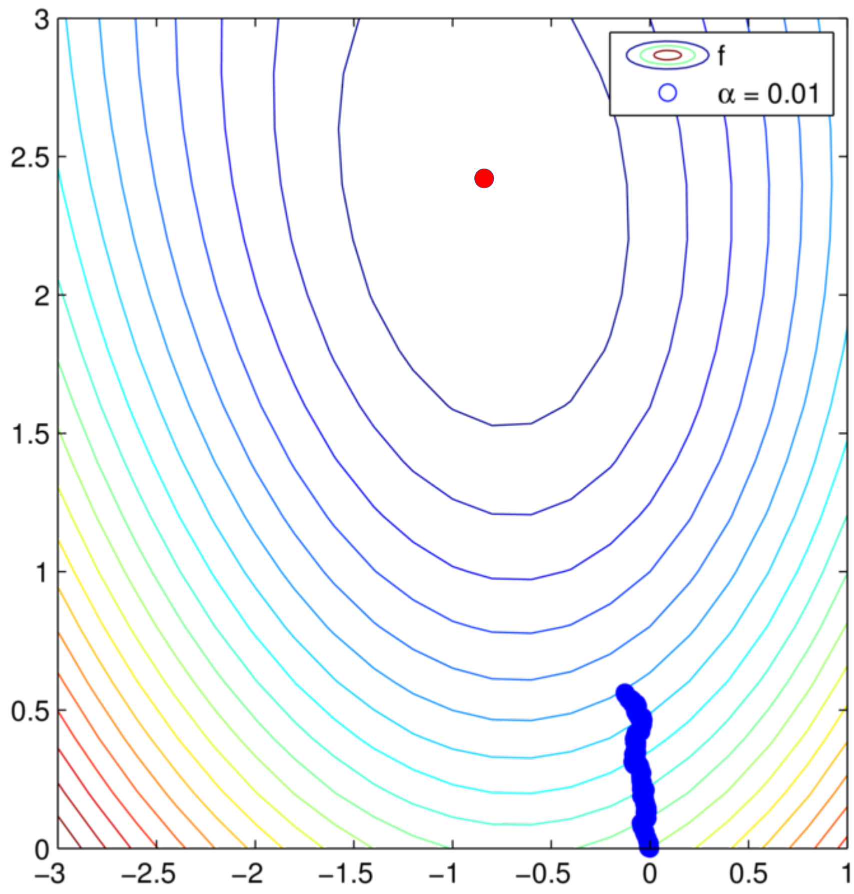
$$
\begin{aligned}
\mathbb{E} \left[||w^{t+1} - w^*||_2^2 \right] &\leq (1 - \alpha\lambda) \mathbb{E} \left[||w^t - w^*||_2^2 \right] + \alpha^2 B^2 \\
&= (1 - \alpha\lambda)^{t+1} ||w^0 - w^*||_2^2 + \textstyle\sum_{i=0}^{t} (1 - \alpha\lambda)^i \alpha^2 B^2
\end{aligned}
$$

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

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

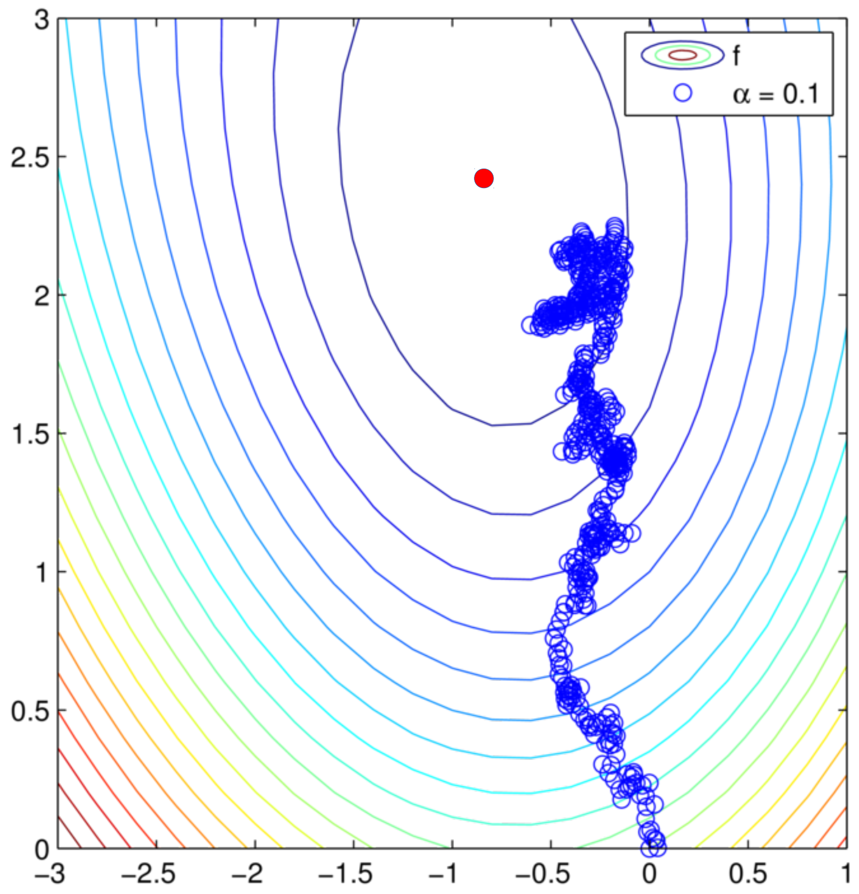
Stochastic Gradient Descent

α =0.01



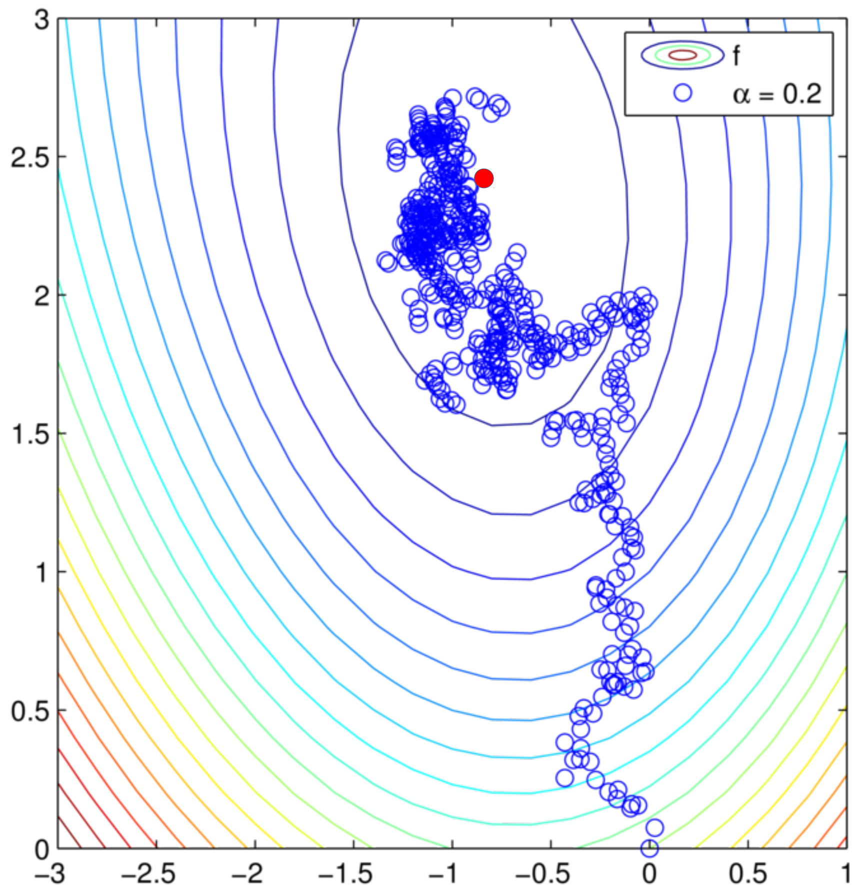
Stochastic Gradient Descent

α =0.1



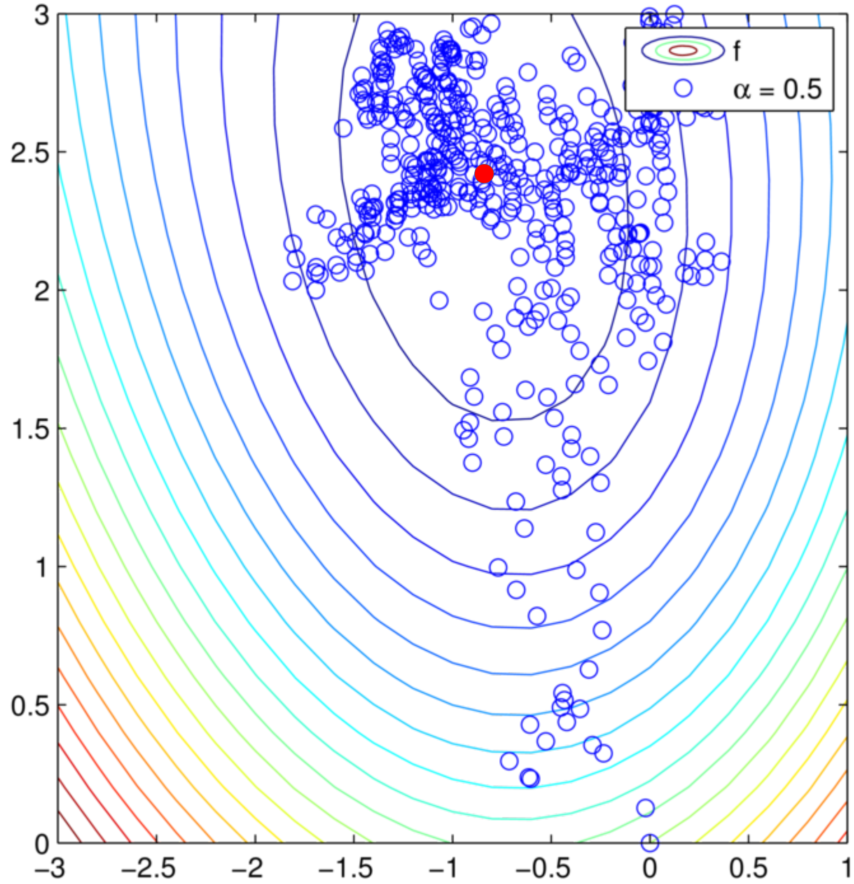
Stochastic Gradient Descent

α =0.2



Stochastic Gradient Descent

α =0.5



Assumptions for Convergence

Strong Convexity

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

$$\Downarrow \; y = w^*$$

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

Expected Bounded Stochastic Gradients

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

Assumptions for Convergence

Strong Convexity

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

$$\Big\Downarrow \; y = w^*$$

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

Expected Bounded Stochastic Gradients

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

Assumptions for Convergence

Strong Convexity

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

$$\Downarrow \; y = w^*$$

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

~~**Expected Bounded Stochastic Gradients**~~

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

Assumptions for Convergence

Strong Convexity

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

$$\Big\downarrow \; y = w^*$$

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

~~**Expected Bounded Stochastic Gradients**~~

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

EXE:
Let $A \in \mathbb{R}^{n \times d}$, $f_j(w) = (A_{j:}w - b_j)^2$. $\quad \max\limits_{w} \mathbb{E}_{j \sim \frac{1}{n}}[\|\nabla f_j(w)\|^2] = ?$

EXE:
Let $A \in \mathbb{R}^{n\times d}$, $f_j(w) = (A_{j:}w - b_j)^2$. $\quad \max_w \mathbb{E}_{j\sim\frac{1}{n}}[\|\nabla f_j(w)\|^2] \; = ?$

Proof: $\max_w \mathbb{E}_{j\sim\frac{1}{n}}[\|\nabla f_j(w)\|^2] \; = \; \infty$, indeed since

$$\begin{aligned}
\|\nabla f_j(w)\|^2 &= 4\|A_{j:}^\top(A_{j:}w - b_j)\|^2 \\
&= 4\|A_{j:}\|^2(A_{j:}w - b_j)^2 \\
&= 4(\hat{A}_{j:}w - \hat{b}_j)^2 \qquad \text{where } \hat{A}_{j:} := A_{j:}\|A_{j:}\|, \quad \hat{b}_j := b_j\|A_{j:}\|
\end{aligned}$$

Taking expectation

$$\mathbb{E}_{j\sim\frac{1}{n}}\|\nabla f_j(w)\|^2 = \frac{1}{n}\sum_{j=1}^n 4(\hat{A}_{j:}w - \hat{b}_j)^2 \; = \frac{1}{n}\|\hat{A}w - \hat{b}\|^2$$

$$\lim_{w\to\infty} \|\hat{A}w - b\|^2 \; = \; \infty$$

Realistic assumptions for Convergence

Strongly quasi-convexity

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

Each f_i is convex and L_i smooth

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

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

Definition: Gradient Noise

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

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

Assumptions for Convergence

EXE: Calculate the L_i's and $L_{\max}$ for

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

HINT: A twice differentiable f_i is L_i - smooth if and only if

$$\nabla^2 f_i(w) \preceq L_i I \quad \Leftrightarrow \quad v^\top \nabla^2 f_i(w) v \leq L_i \|v\|^2, \forall v$$

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

Assumptions for Convergence

EXE: Calculate the L_i's and $L_{\max}$ for

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

HINT: A twice differentiable f_i is L_i - smooth if and only if

$$\nabla^2 f_i(w) \preceq L_i\, I \quad \Leftrightarrow \quad v^\top \nabla^2 f_i(w) v \leq L_i \|v\|^2, \forall v$$

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

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

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

Assumptions for Convergence

EXE: Calculate the L_i's and $L_{\max}$ for

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

HINT: A twice differentiable f_i is L_i - smooth if and only if

$$\nabla^2 f_i(w) \preceq L_i\, I \quad \Leftrightarrow \quad v^\top \nabla^2 f_i(w) v \leq L_i \|v\|^2, \forall v$$

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

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

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

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

Assumptions for Convergence

EXE: Calculate the L_i's and $L_{\max}$ for

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

HINT: A twice differentiable f_i is L_i - smooth if and only if

$$\nabla^2 f_i(w) \preceq L_i\, I \quad \Leftrightarrow \quad v^\top \nabla^2 f_i(w) v \leq L_i \|v\|^2, \forall v$$

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

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

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

$$L_{\max} \; = \; \max_{i=1,\ldots,n}(||A_{i:}||_2^2 + \lambda) \; = \; \max_{i=1,\ldots,n}||A_{i:}||_2^2 + \lambda$$

Assumptions for Convergence

EXE: Calculate the L_i's and $L_{\max}$ for

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

Assumptions for Convergence

EXE: Calculate the L_i's and $L_{\max}$ 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 the L_i's and $L_{\max}$ for

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

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

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

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

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

Relationship between smoothness constants

EXE: Let f be differentiable and convex. Show that $f(w)$ is L–smooth with

$$L = \max_{w \in \mathbb{R}^d} \lambda_{\max}(\nabla^2 f(w))$$

Thus $f_i(w)$ is L_i–smooth with $L_i = \max_{w \in \mathbb{R}^d} \lambda_{\max}(\nabla^2 f_i(w))$ show that

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

Relationship between smoothness constants

EXE: Let f be differentiable and convex. Show that $f(w)$ is L–smooth with

$$L = \max_{w\in\mathbb{R}^d} \lambda_{\max}(\nabla^2 f(w))$$

Thus $f_i(w)$ is L_i–smooth with $L_i = \max_{w\in\mathbb{R}^d} \lambda_{\max}(\nabla^2 f_i(w))$ 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 the Hessian definition of smoothness

$$\nabla^2 f(w) \quad \preceq \quad \lambda_{\max}(\nabla^2 f(w))I \quad \preceq \quad \max_{w\in\mathbb{R}^d} \lambda_{\max}(\nabla^2 f(w))I$$

Furthermore

$$\lambda_{\max}(\nabla^2 f(w)) = \lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n \nabla^2 f_i(w)\right) \leq \frac{1}{n}\sum_{i=1}^n \lambda_{\max}(\nabla^2 f_i(w)) \leq \frac{1}{n}\sum_{i=1}^n L_i$$

Which follows since the largest eigenvalue function is convex over psd matrices. Now take the max over w, *then* max over i.

Complexity / Convergence

Theorem.

Let f be μ–strongly quasi-convex and f_i be L_i–smooth.

If $0 < \alpha \leq \frac{1}{2L_{\max}}$ then the iterates of the SGD 0.0 satisfy

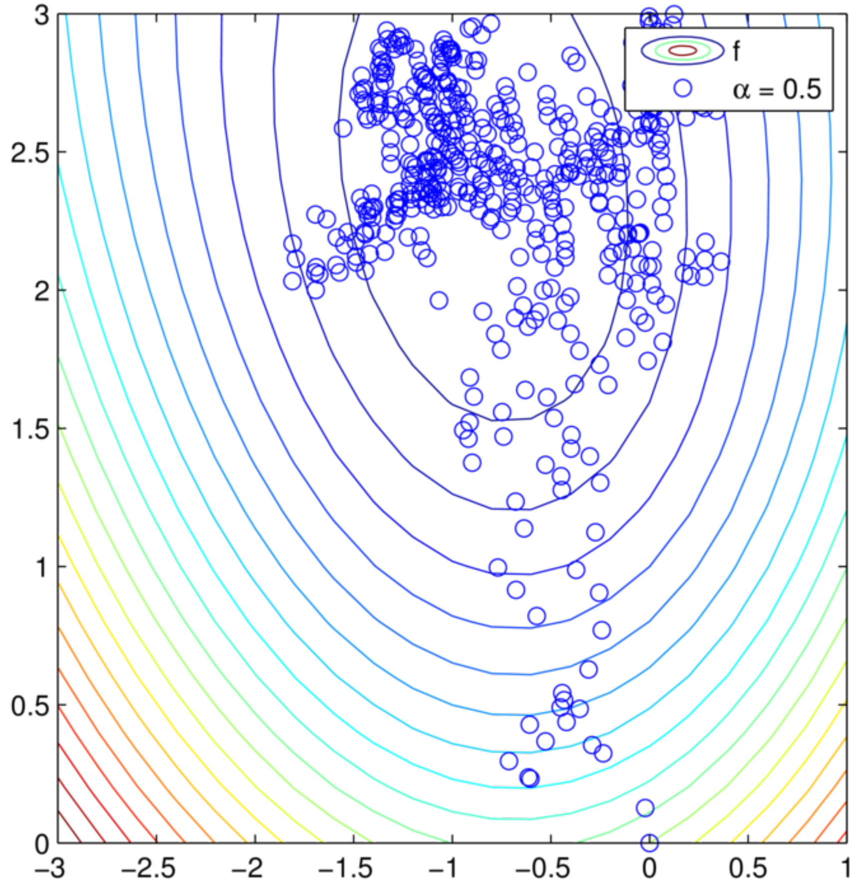
$$\mathbb{E}\left[||w^t - w^*||_2^2\right] \leq (1-\alpha\mu)^t ||w^0 - w^*||_2^2 + \frac{2\alpha}{\mu}\sigma^2$$

EXE: The steps of the proof are given in the SGD_proof exercise list for homework!

RMG, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, P. Richtarik (2019) ICML 2019
SGD: General Analysis and Improved Rates.

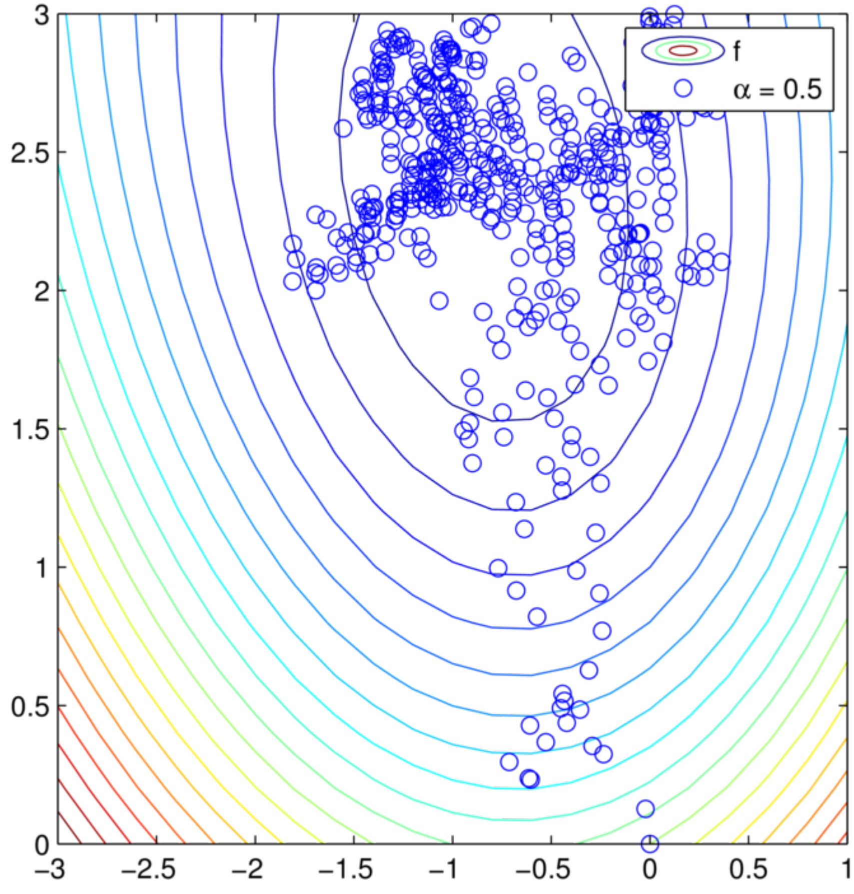
Stochastic Gradient Descent

$\alpha = 0.5$



Stochastic Gradient Descent

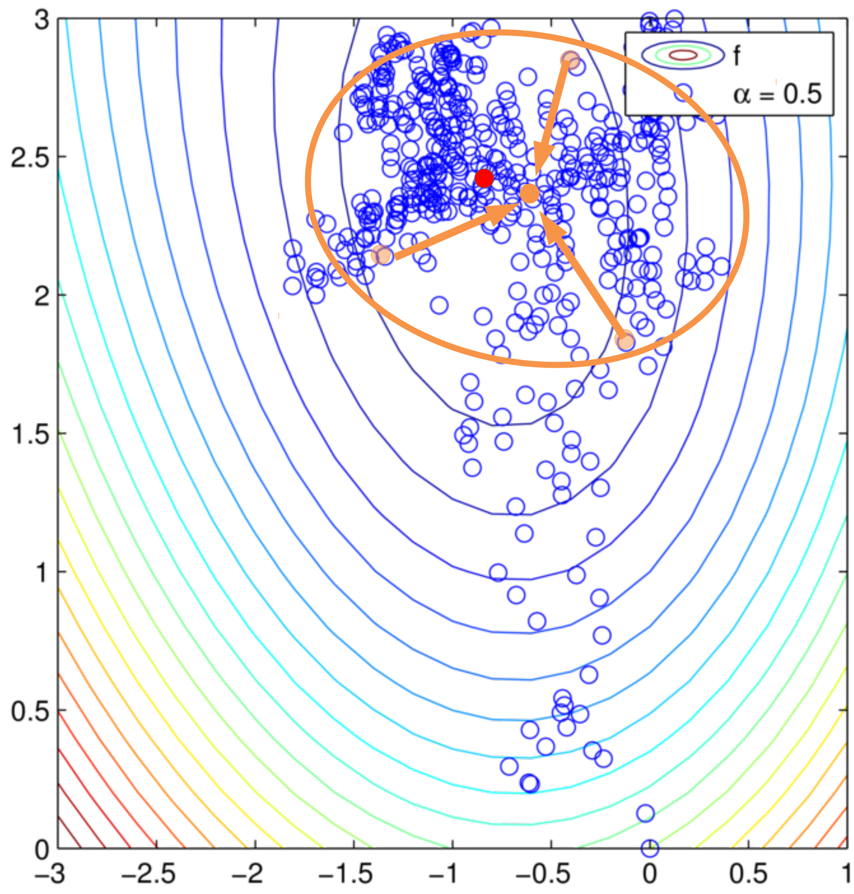
α =0.5



1) Start with big steps and end with smaller steps

Stochastic Gradient Descent

α =0.5



1) Start with big steps and end with smaller steps

2) Try averaging the points

SGD shrinking stepsize

SGD 1.0: Descreasing stepsize

Set $w^0 = 0$

Choose $\alpha_t > 0$, $\alpha_t \to 0$, $\sum_{t=0}^{\infty} \alpha_t = \infty$

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

Shrinking Stepsize

SGD shrinking stepsize

SGD 1.0: Descreasing stepsize

Set $w^0 = 0$

Choose $\alpha_t > 0$, $\alpha_t \to 0$, $\sum_{t=0}^{\infty} \alpha_t = \infty$

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

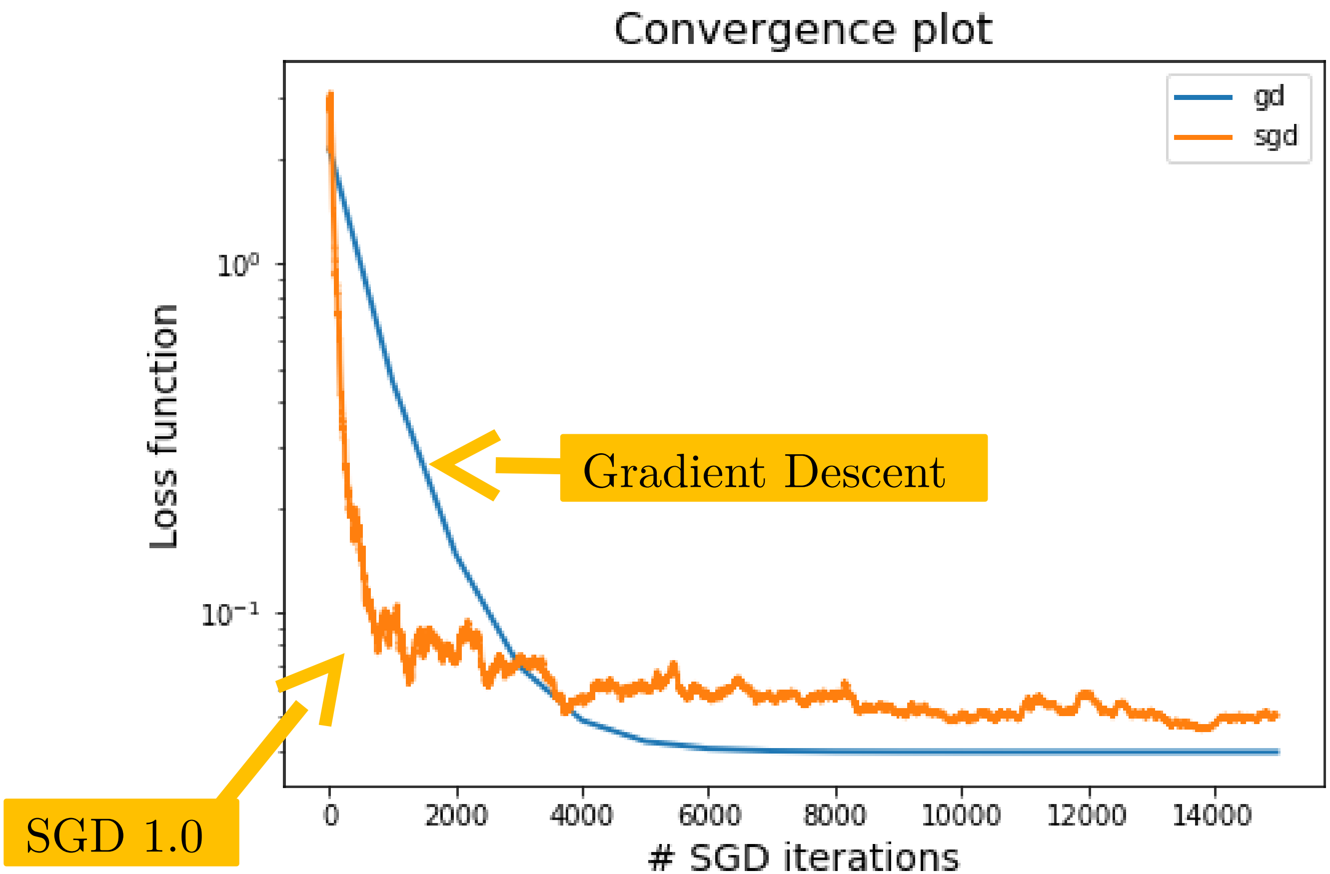
Shrinking Stepsize

How should we sample j ?

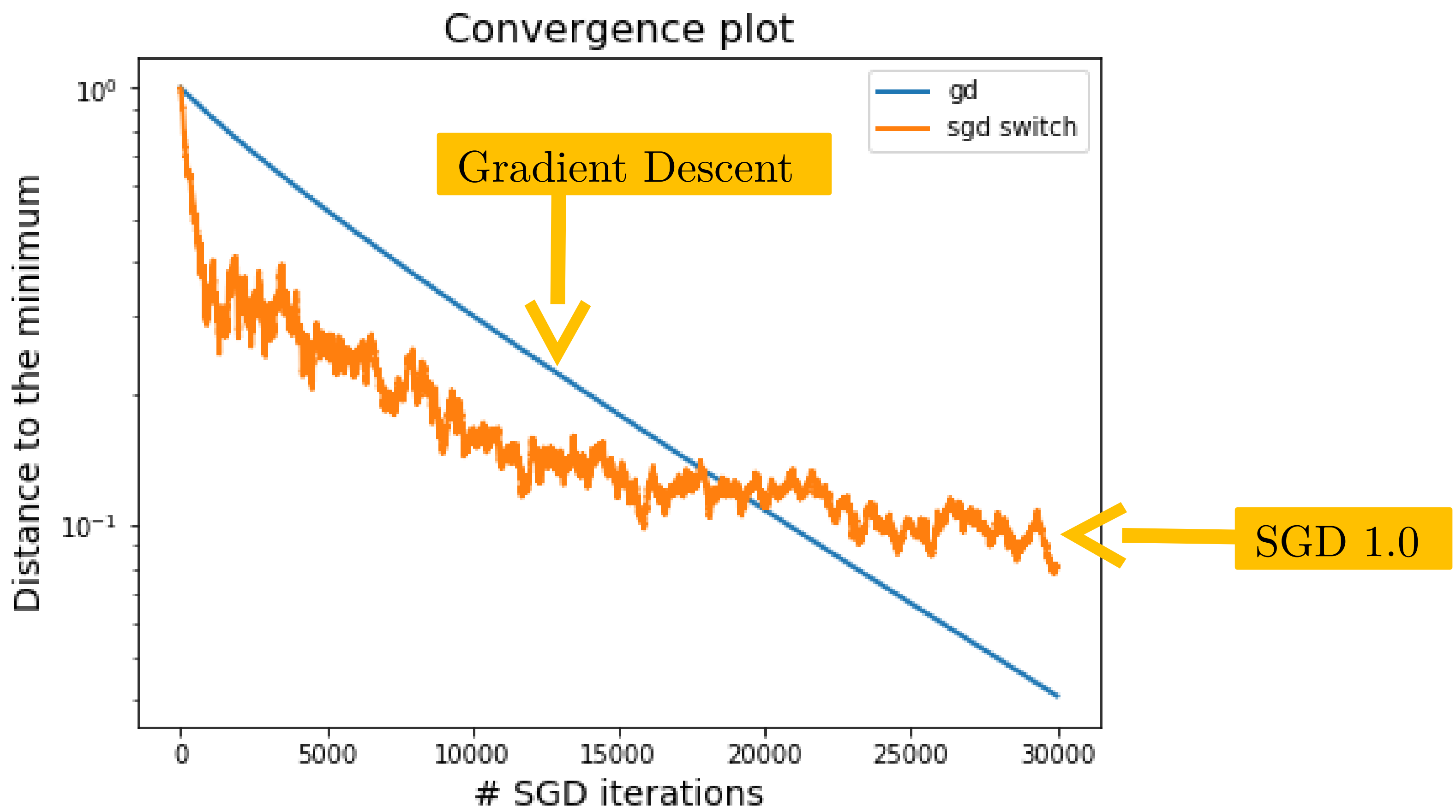
How fast $\alpha_t \to 0$?

Does this converge?

SGD with shrinking stepsize
Compared with Gradient Descent



SGD with shrinking stepsize
Compared with Gradient Descent



Complexity / Convergence

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

Theorem for shrinking stepsizes

Let f be μ–strongly quasi-convex and f_i be L_i–smooth.
Let $\mathcal{K} := L_{\max}/\mu$ and let

$$\alpha^t = \begin{cases} \frac{1}{2L_{\max}} & \text{for} \quad t \leq 4\lceil \mathcal{K} \rceil \\ \frac{2t+1}{(t+1)^2\mu} & \text{for} \quad t > 4\lceil \mathcal{K} \rceil. \end{cases}$$

If $t \geq 4\lceil \mathcal{K} \rceil$, then SGD 1.0 satifies

$$\mathbb{E}\|w^t - w^*\|^2 \leq \frac{\sigma^2}{\mu^2}\frac{8}{t} + \frac{16}{e^2}\frac{\lceil \mathcal{K} \rceil^2}{t^2}\|w^0 - w^*\|^2$$

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

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

Complexity / Convergence

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

Theorem for shrinking stepsizes

Let f be μ–strongly quasi-convex and f_i be L_i–smooth.
Let $\mathcal{K} := L_{\max}/\mu$ and let

$$\alpha^t = \begin{cases} \frac{1}{2L_{\max}} & \text{for} \quad t \leq 4\lceil\mathcal{K}\rceil \\ \frac{2t+1}{(t+1)^2\mu} & \text{for} \quad t > 4\lceil\mathcal{K}\rceil. \end{cases}$$

$\alpha^t = O(1/(t+1))$

If $t \geq 4\lceil\mathcal{K}\rceil$, then SGD 1.0 satifies

$$\mathbb{E}\|w^t - w^*\|^2 \leq \frac{\sigma^2}{\mu^2}\frac{8}{t} + \frac{16}{e^2}\frac{\lceil\mathcal{K}\rceil^2}{t^2}\|w^0 - w^*\|^2$$

$O\left(\frac{1}{t}\right)$ $\Rightarrow$ Iteration complexity $O\left(\frac{1}{\epsilon}\right)$

Complexity / Convergence

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

Theorem for shrinking stepsizes

Let f be μ–strongly quasi-convex and f_i be L_i–smooth.
Let $\mathcal{K} := L_{\max}/\mu$ and let

$$\alpha^t = \begin{cases} \frac{1}{2L_{\max}} & \text{for} \quad t \leq 4\lceil \mathcal{K} \rceil \\ \frac{2t+1}{(t+1)^2\mu} & \text{for} \quad t > 4\lceil \mathcal{K} \rceil. \end{cases}$$

$\alpha^t = O(1/(t+1))$

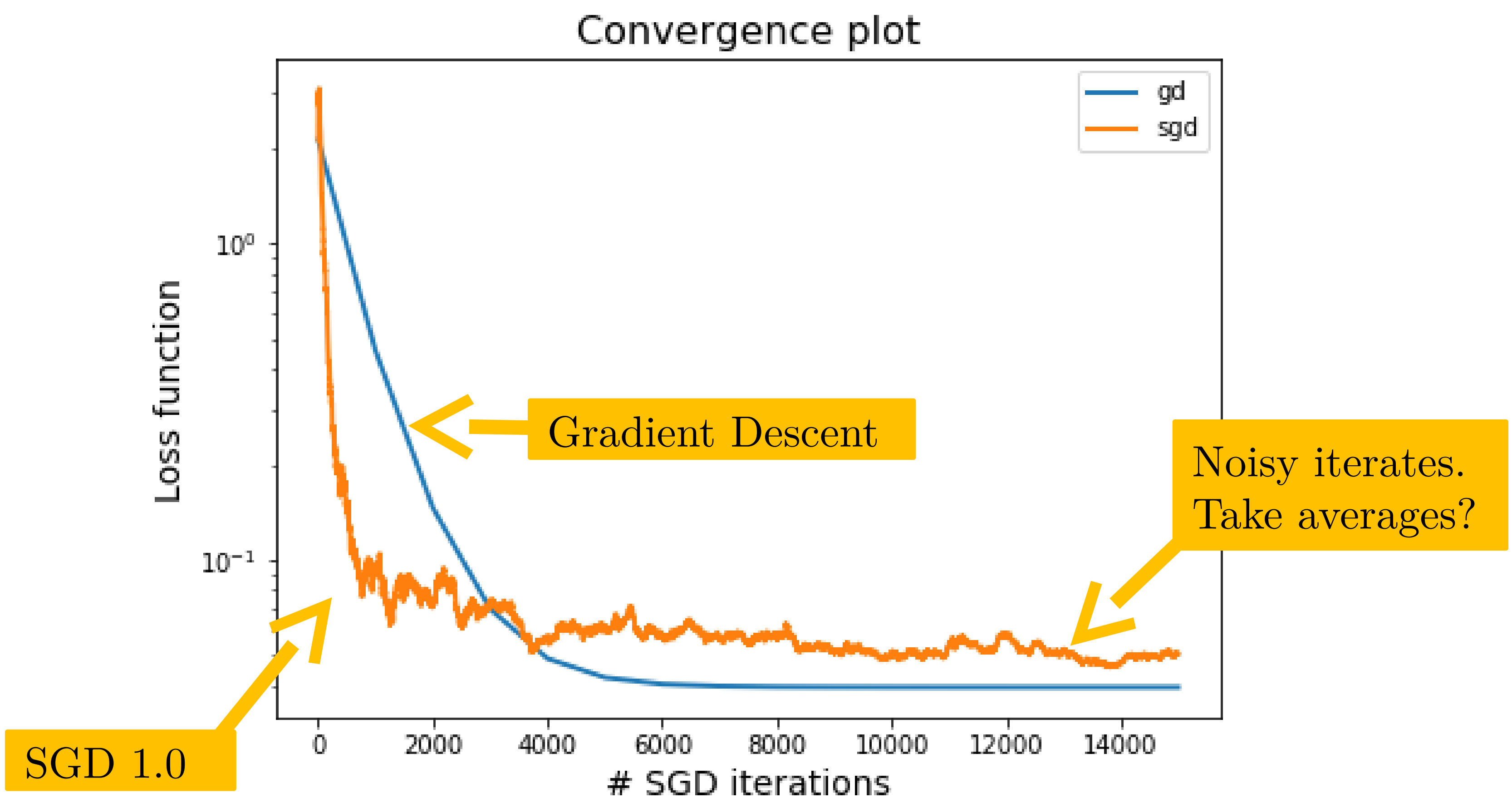
If $t \geq 4\lceil \mathcal{K} \rceil$, then SGD 1.0 satifies

$$\mathbb{E}\|w^t - w^*\|^2 \leq \frac{\sigma^2}{\mu^2}\frac{8}{t} + \frac{16}{e^2}\frac{\lceil \mathcal{K} \rceil^2}{t^2}\|w^0 - w^*\|^2$$

$O\left(\frac{1}{t}\right)$ → Iteration complexity $O\left(\frac{1}{\epsilon}\right)$

In practice $\alpha^t = C/(t+1)$ or $\alpha^t = C/\sqrt{t+1}$ where C is tuned

Stochastic Gradient Descent
Compared with Gradient Descent



SGD with (late start) averaging

SGDA 1.1

Set $w^0 = 0$

Choose $\alpha_t > 0,\ \alpha_t \rightarrow 0,\ \sum_{t=0}^{\infty} \alpha_t = \infty$

Choose averaging start $s_0 \in \mathbb{N}$

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

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

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

  if $t > s_0$

    $\overline{w} = \frac{1}{t-s_0} \sum_{i=s_0}^{t} w^t$

  else: $\overline{w} = w$

Output $\overline{w}$

B. T. Polyak and A. B. Juditsky, SIAM Journal on Control and Optimization (1992)
Acceleration of stochastic approximation by averaging

SGD with (late start) averaging

SGDA 1.1

Set $w^0 = 0$

Choose $\alpha_t > 0,\ \alpha_t \rightarrow 0,\ \sum_{t=0}^{\infty} \alpha_t = \infty$

Choose averaging start $s_0 \in \mathbb{N}$

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

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

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

  if $t > s_0$

    $\overline{w} = \frac{1}{t - s_0} \sum_{i=s_0}^{t} w^t$

  else: $\overline{w} = w$

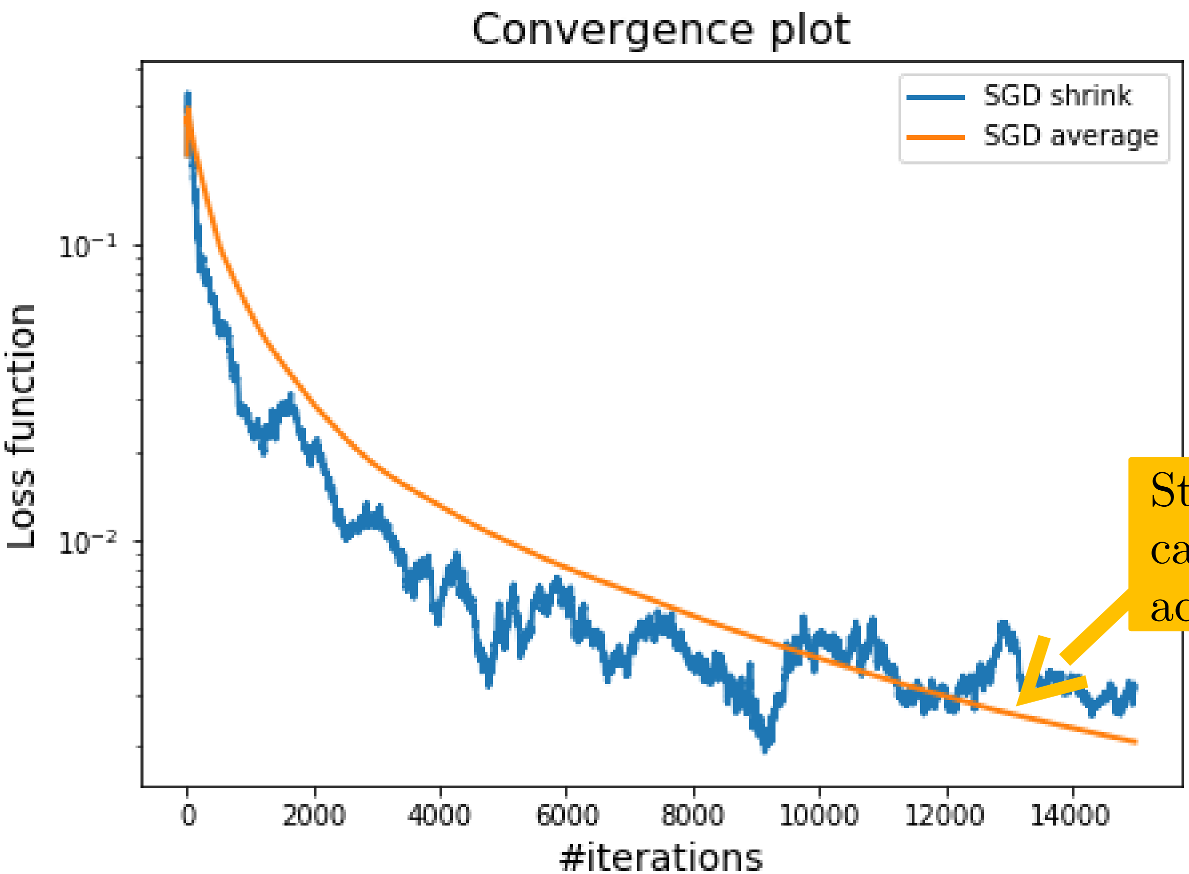
Output $\overline{w}$

This is not efficient. How to make this efficient?

B. T. Polyak and A. B. Juditsky, SIAM Journal on Control and Optimization (1992)
Acceleration of stochastic approximation by averaging

Stochastic Gradient Descent

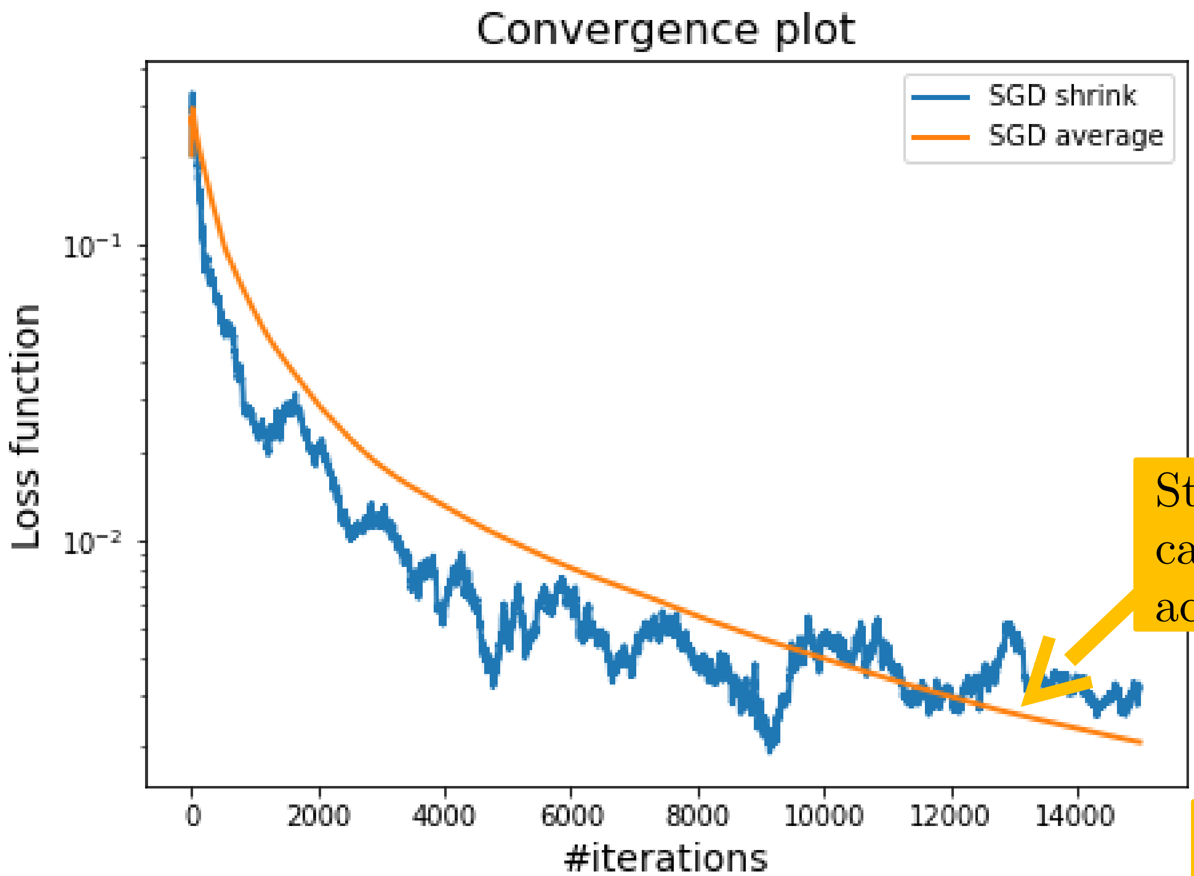
With and without averaging



Starts slow, but can reach higher accuracy

Stochastic Gradient Descent

With and without averaging

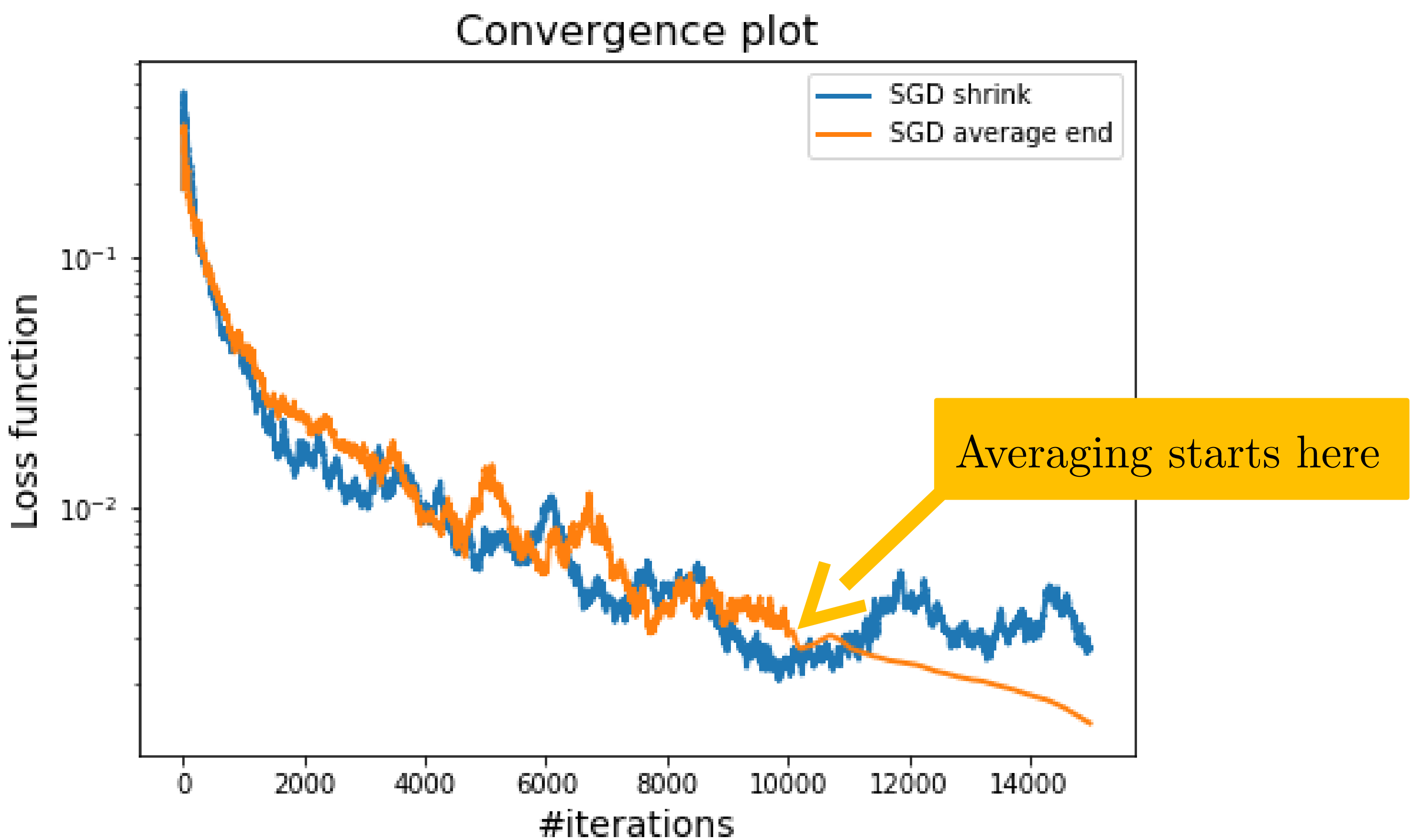


Starts slow, but can reach higher accuracy

Only use averaging towards the end?

Stochastic Gradient Descent

Averaging the last few iterates



Comparison GD and SGD for strongly convex

	SGD	**GD**
Iteration complexity	$O\left(\frac{1}{\epsilon}\right)$	$O\left(\log\left(\frac{1}{\epsilon}\right)\right)$

Comparison GD and SGD for strongly convex

	SGD	**GD**
Iteration complexity	$O\left(\frac{1}{\epsilon}\right)$	$O\left(\log\left(\frac{1}{\epsilon}\right)\right)$
Cost of an iteration	$O\left(1\right)$	$O\left(n\right)$

Comparison GD and SGD for strongly convex

	SGD	**GD**
Iteration complexity	$O\left(\frac{1}{\epsilon}\right)$	$O\left(\log\left(\frac{1}{\epsilon}\right)\right)$
Cost of an iteration	$O(1)$	$O(n)$
Total complexity*	$O\left(\frac{1}{\epsilon}\right)$	$O\left(n\log\left(\frac{1}{\epsilon}\right)\right)$

Comparison GD and SGD for strongly convex

	SGD	**GD**
Iteration complexity	$O\left(\frac{1}{\epsilon}\right)$	$O\left(\log\left(\frac{1}{\epsilon}\right)\right)$
Cost of an iteration	$O(1)$	$O(n)$
Total complexity*	$O\left(\frac{1}{\epsilon}\right)$	$O\left(n\log\left(\frac{1}{\epsilon}\right)\right)$

*Total complexity $=$ (Iteration complexity) $\times$ (Cost of an iteration)

Comparison GD and SGD for strongly convex

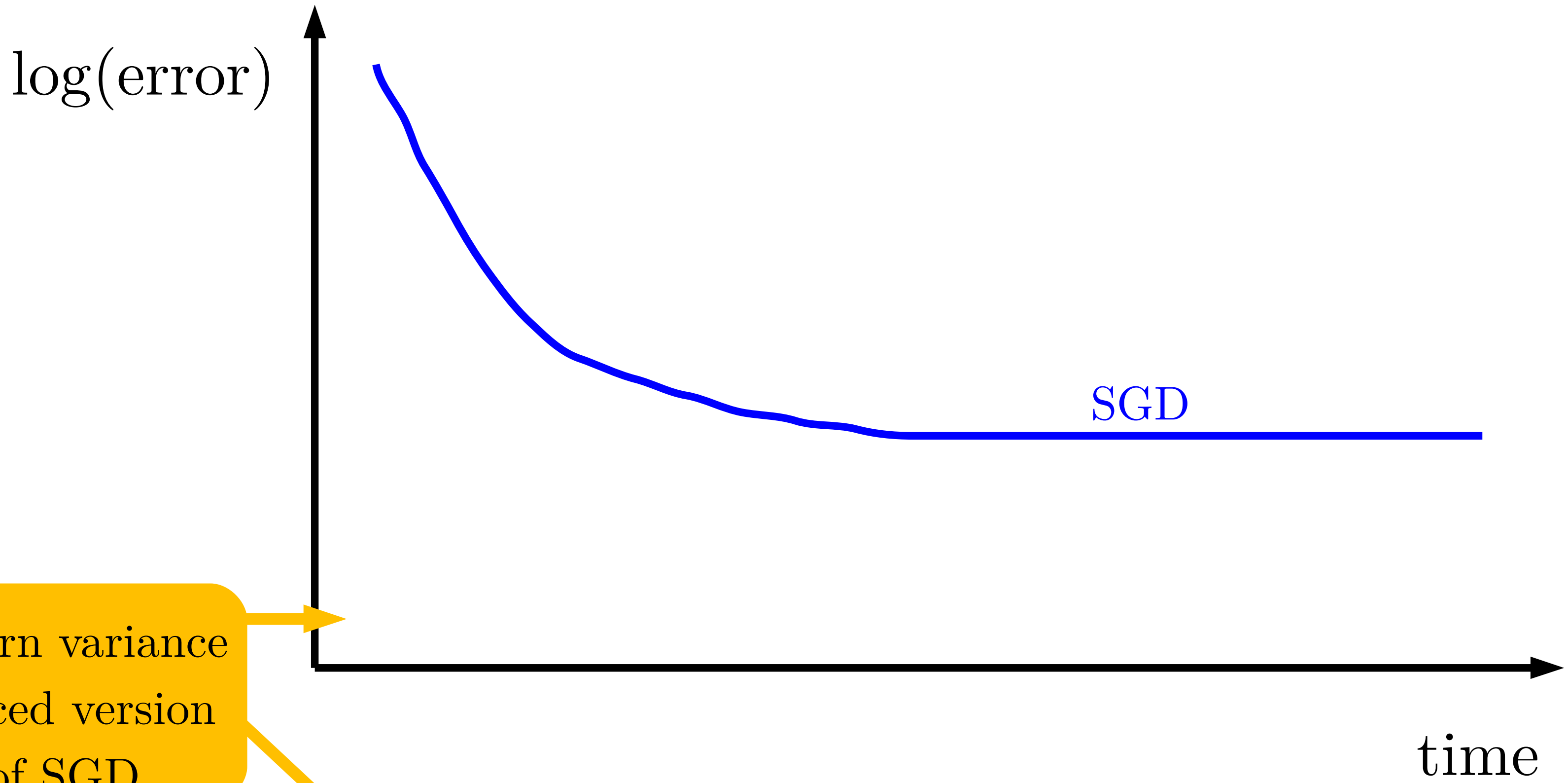
	SGD	**GD**
Iteration complexity	$O\left(\frac{1}{\epsilon}\right)$	$O\left(\log\left(\frac{1}{\epsilon}\right)\right)$
Cost of an iteration	$O(1)$	$O(n)$
Total complexity*	$O\left(\frac{1}{\epsilon}\right)$	$O\left(n\log\left(\frac{1}{\epsilon}\right)\right)$

What happens if ϵ is small?

What happens if n is big?

*Total complexity $=$ (Iteration complexity) $\times$ (Cost of an iteration)

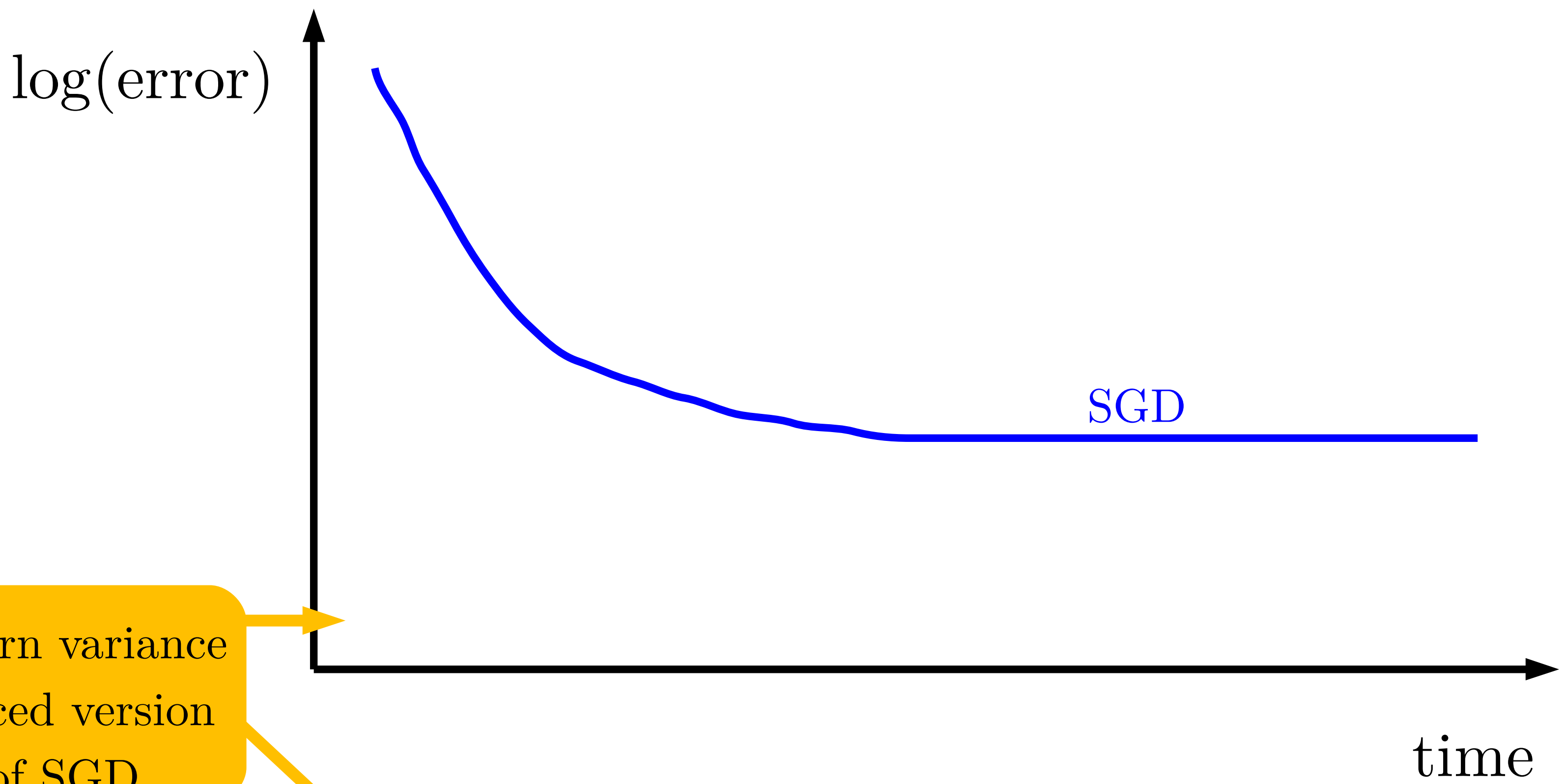
Comparison SGD vs GD



Modern variance reduced version of SGD

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

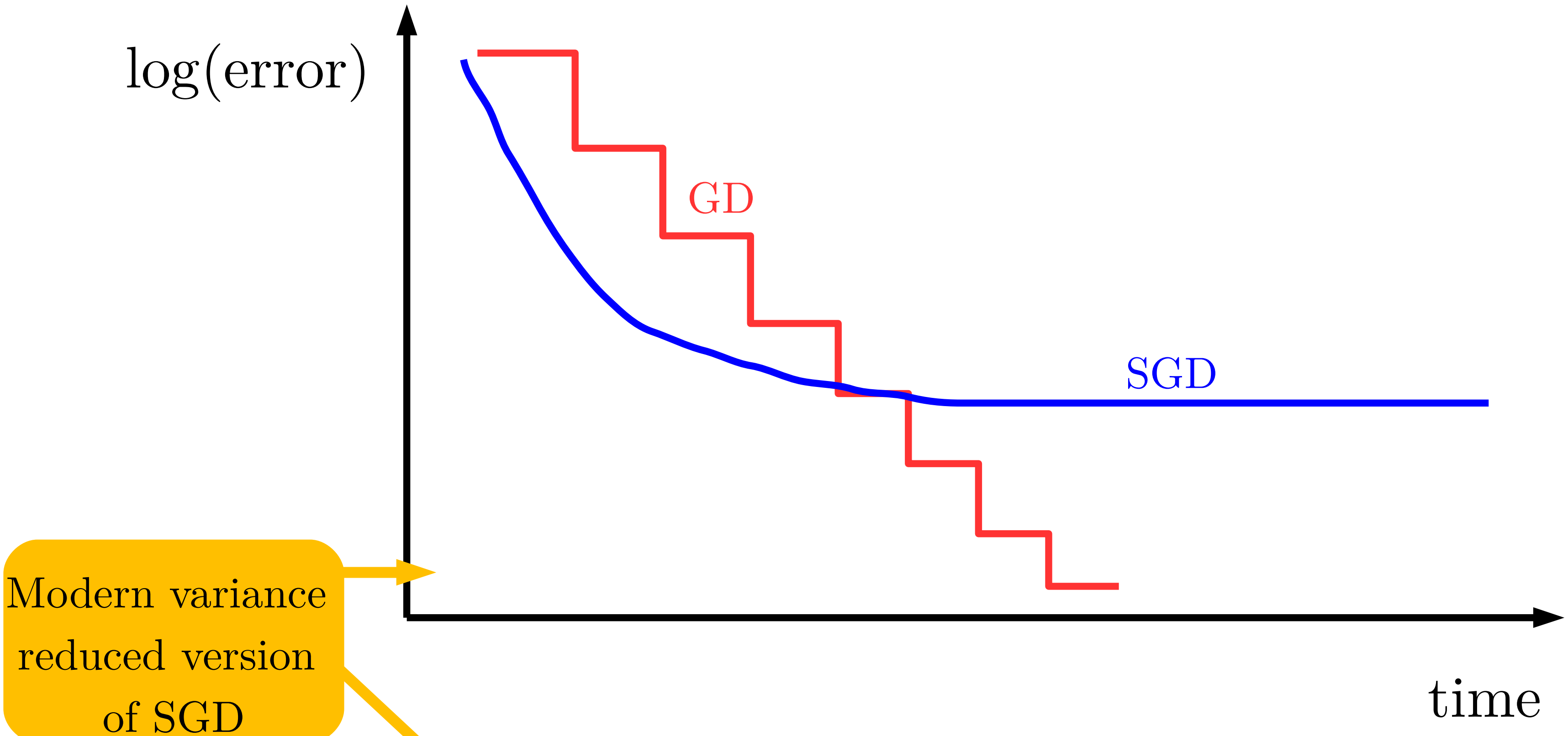
Comparison SGD vs GD



Modern variance reduced version of SGD

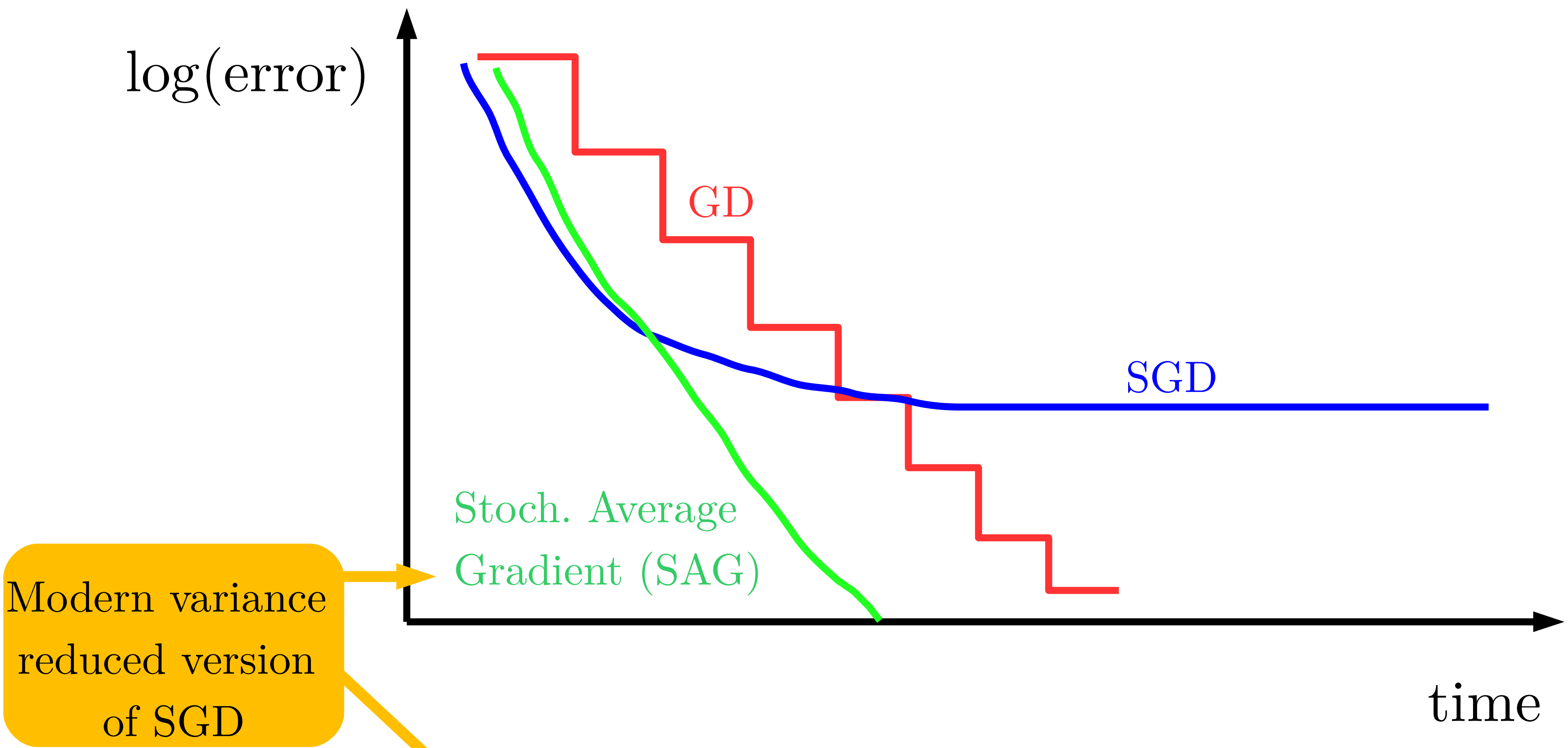
M. Schmidt, N. Le Roux, F. Bach (2016)
Mathematical Programming
Minimizing Finite Sums with the Stochastic Average Gradient.

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.

20 min tea time break?

Practical SGD for Sparse Data

Lazy SGD updates for Sparse Data

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

Assume each data point x^i is s-sparse, how many operations does each SGD step cost?

Lazy SGD updates for Sparse Data

L2 regularizor + linear hypothesis

Finite Sum Training Problem

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

Assume each data point x^i is s-sparse, how many operations does each SGD step cost?

$$\begin{aligned} w^{t+1} &= w^t - \alpha_t \left(\ell'(\langle w^t, x^i \rangle, y^i) x^i + \lambda w^t\right) \\ &= (1 - \lambda \alpha_t) w^t - \alpha_t \ell'(\langle w^t, x^i \rangle, y^i) x^i \end{aligned}$$

Lazy SGD updates for Sparse Data

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

Assume each data point x^i is s-sparse, how many operations does each SGD step cost?

$$\begin{aligned} w^{t+1} &= w^t - \alpha_t\left(\ell'(\langle w^t, x^i\rangle, y^i)x^i + \lambda w^t\right) \\ &= \underbrace{(1-\lambda\alpha_t)w^t}_{\text{Rescaling } O(d)} - \underbrace{\alpha_t \ell'(\langle w^t, x^i\rangle, y^i)x^i}_{\text{Addition sparse vector } O(s)} \end{aligned}$$

Rescaling $O(d)$ + Addition sparse vector $O(s)$ = $O(d)$

Lazy SGD updates for Sparse Data

SGD step

$$w^{t+1} = (1-\lambda\alpha_t)w^t - \alpha_t \ell'(\langle w^t, x^i\rangle, y^i)x^i$$

EXE: re-write the iterates using $w^t = \beta_t z^t$ where $\beta_t \in \mathbb{R},\; z^t \in \mathbb{R}^d$

Can you update β_t and z^t so that each iteration is $O(s)$?

Lazy SGD updates for Sparse Data

SGD step

$$w^{t+1} = (1-\lambda\alpha_t)w^t - \alpha_t \ell'(\langle w^t, x^i\rangle, y^i)x^i$$

EXE: re-write the iterates using $w^t = \beta_t z^t$ where $\beta_t \in \mathbb{R}$, $z^t \in \mathbb{R}^d$

Can you update β_t and z^t so that each iteration is $O(s)$?

$$\begin{aligned}
\beta_{t+1} z^{t+1} &= (1-\lambda\alpha_t)\beta_t z^t - \alpha_t \ell'(\beta_t\langle z^t, x^i\rangle, y^i)x^i \\
&= (1-\lambda\alpha_t)\beta_t \left(z^t - \frac{\alpha_t \ell'(\beta_t\langle z^t, x^i\rangle, y^i)}{(1-\lambda\alpha_t)\beta_t} x^i \right)
\end{aligned}$$

Lazy SGD updates for Sparse Data

SGD step

$$w^{t+1} = (1-\lambda\alpha_t)w^t - \alpha_t \ell'(\langle w^t, x^i\rangle, y^i)x^i$$

EXE: re-write the iterates using $w^t = \beta_t z^t$ where $\beta_t \in \mathbb{R},\ z^t \in \mathbb{R}^d$

Can you update β_t and z^t so that each iteration is $O(s)$?

$$\begin{aligned}\beta_{t+1}z^{t+1} &= (1-\lambda\alpha_t)\beta_t z^t - \alpha_t \ell'(\beta_t\langle z^t, x^i\rangle, y^i)x^i \\ &= \underbrace{(1-\lambda\alpha_t)\beta_t}_{\beta_{t+1}} \left(\underbrace{z^t - \frac{\alpha_t \ell'(\beta_t\langle z^t, x^i\rangle, y^i)}{(1-\lambda\alpha_t)\beta_t}x^i}_{z^{t+1}}\right)\end{aligned}$$

$$\beta_{t+1} = (1-\lambda\alpha_t)\beta_t, \qquad z^{t+1} = z^t - \frac{\alpha_t \ell'(\beta_t\langle z^t, x^i\rangle, y^i)}{(1-\lambda\alpha_t)\beta_t}x^i$$

Lazy SGD updates for Sparse Data

SGD step

$$w^{t+1} = (1-\lambda\alpha_t)w^t - \alpha_t \ell'(\langle w^t, x^i\rangle, y^i)x^i$$

EXE: re-write the iterates using $w^t = \beta_t z^t$ where $\beta_t \in \mathbb{R},\ z^t \in \mathbb{R}^d$

Can you update β_t and z^t so that each iteration is $O(s)$?

$$\begin{aligned}\beta_{t+1}z^{t+1} &= (1-\lambda\alpha_t)\beta_t z^t - \alpha_t \ell'(\beta_t\langle z^t, x^i\rangle, y^i)x^i \\ &= \underbrace{(1-\lambda\alpha_t)\beta_t}_{\beta_{t+1}} \underbrace{\left(z^t - \frac{\alpha_t \ell'(\beta_t\langle z^t, x^i\rangle, y^i)}{(1-\lambda\alpha_t)\beta_t}x^i\right)}_{z^{t+1}}\end{aligned}$$

$O(1)$ scaling +
$O(s)$ sparse add
$= O(s)$ update

$$\beta_{t+1} = (1-\lambda\alpha_t)\beta_t, \qquad z^{t+1} = z^t - \frac{\alpha_t \ell'(\beta_t\langle z^t, x^i\rangle, y^i)}{(1-\lambda\alpha_t)\beta_t}x^i$$

Momentum

Issue with Gradient Descent

Solving the *training problem*: $$\min_{w\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(w) =: f(w)$$

Baseline method: Gradient Descent (GD)

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

Step size/
Learning rate

Why GD and the the Issues

Local rate of change

$$\Delta(d) \;:=\; \lim_{s \to 0^+} \frac{f(x + ds) - f(x)}{s}$$

Max local rate

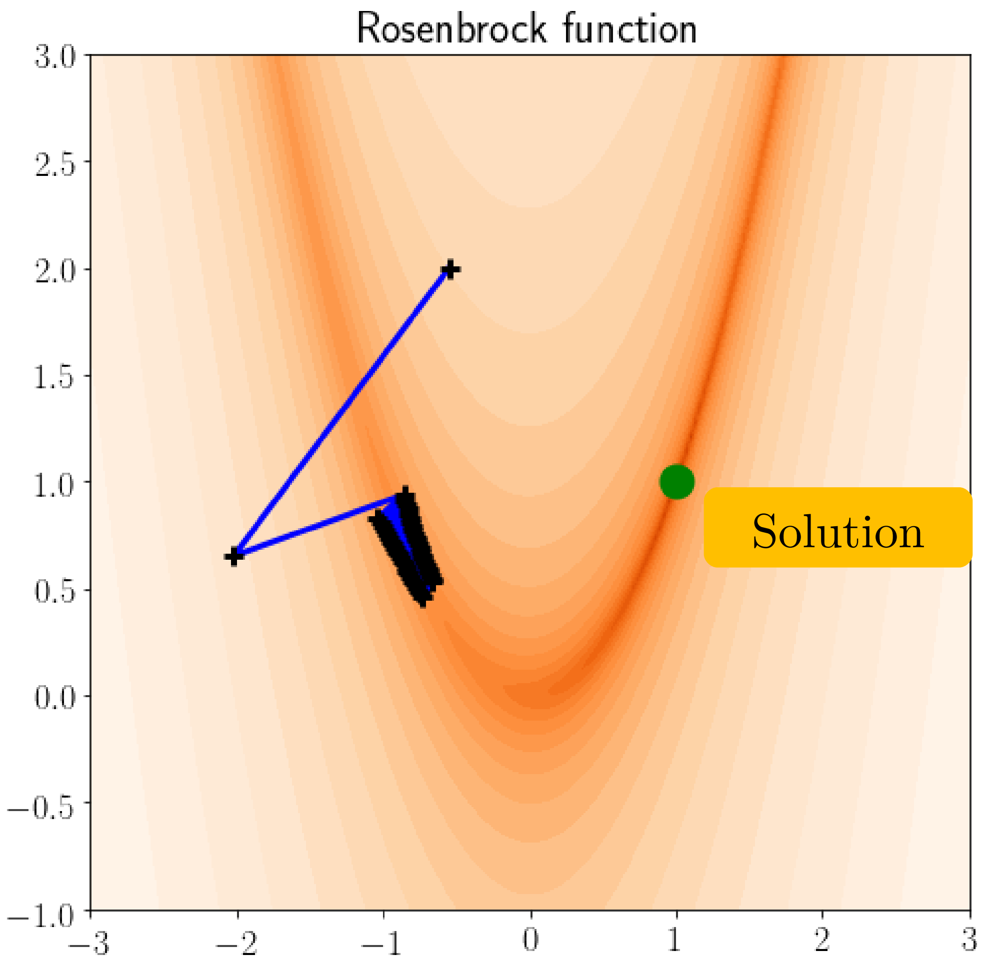
$$\frac{\nabla f(w^t)}{\|\nabla f(w^t)\|} \;:=\; \max_{w \in \mathbb{R}^d} \Delta(d)$$

$$\text{subject to} \quad \|d\| \;=\; 1$$

GD is the "steepest descent"

Issue with Gradient Descent

$$f(x_1, x_2) = 100(x_1 - x_2^2)^2 + (1 - x_2)^2$$



Get's stuck in "flat" valleys

Give momentum to keep going

Adding some Momentum to GD

Heavey Ball Method:

$$w^{t+1} = w^t - \gamma \nabla f(w^t) + \underbrace{\beta(w^t - w^{t-1})}_{\text{Adds “Inertia” to update}}$$

Adding some Momentum to GD

Heavey Ball Method:

$$w^{t+1} = w^t - \gamma \nabla f(w^t) + \underbrace{\beta(w^t - w^{t-1})}_{\text{Adds "Inertia" to update}}$$

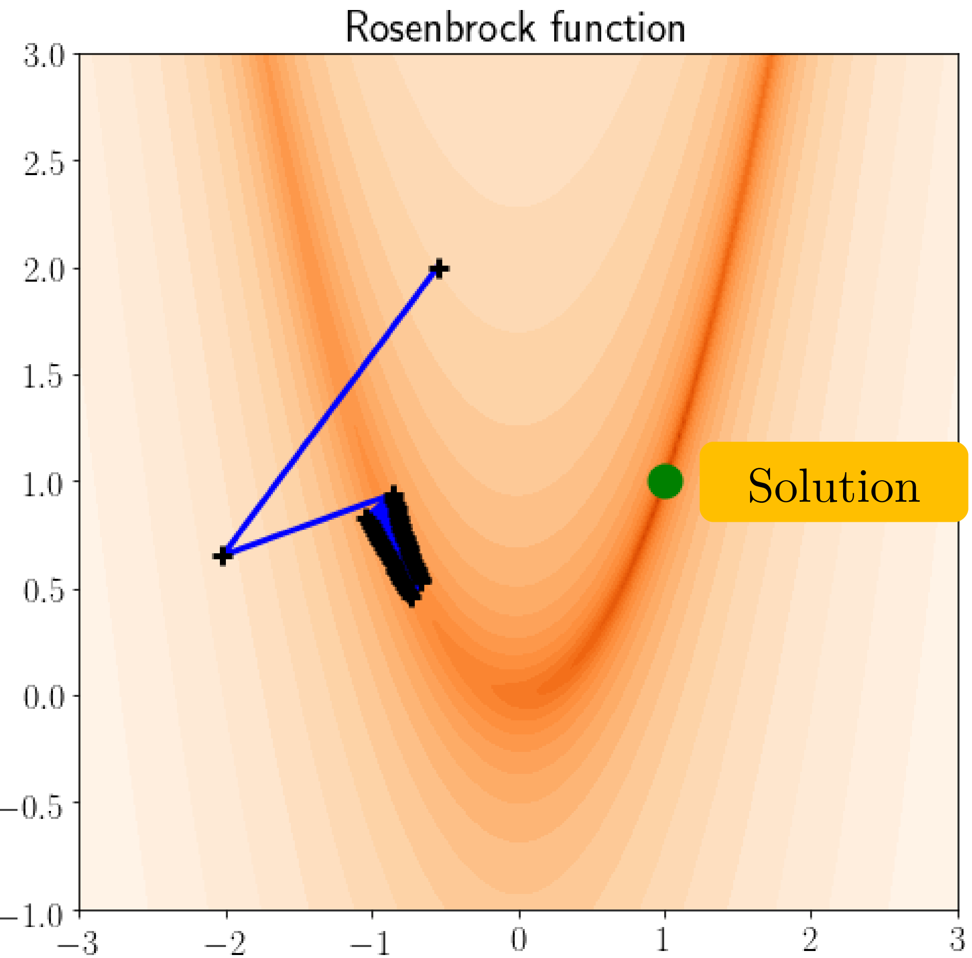
GD with momentum (GDm):

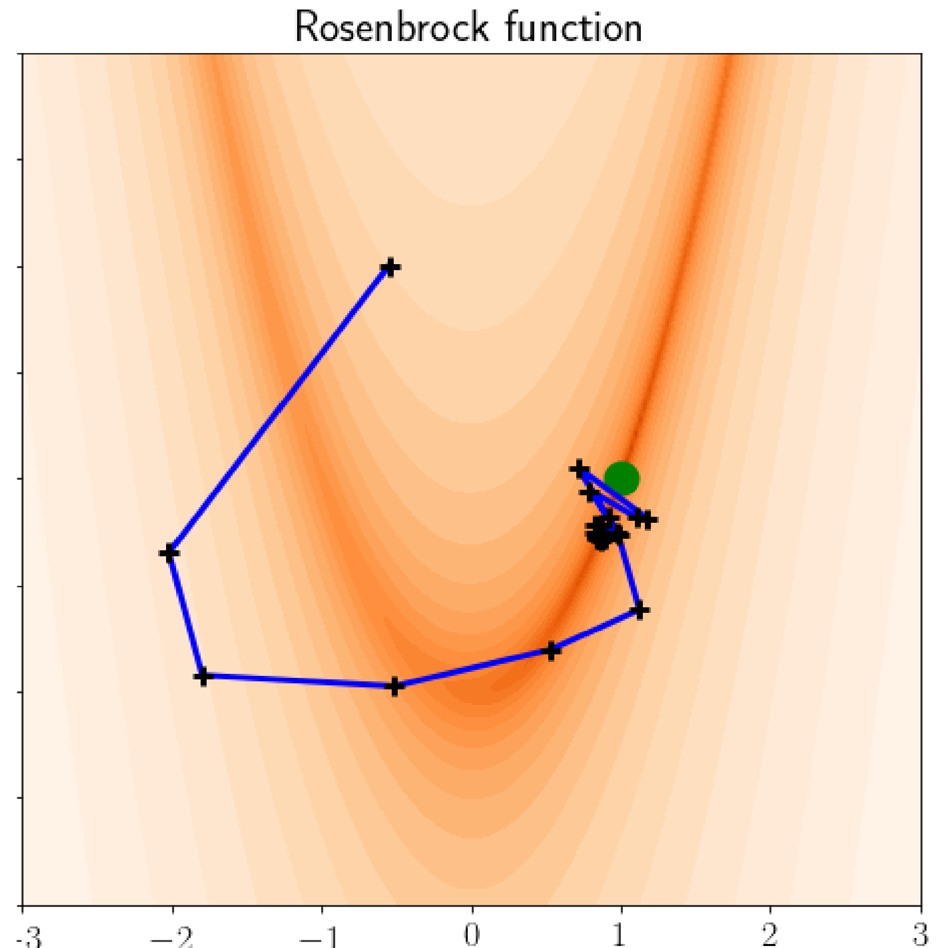
Adds "Momentum" to update

$$m^t = \beta\, m^{t-1} + \nabla f(w^t)$$

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

Issue with Gradient Descent





Get's stuck in "flat" valleys

Give momentum to keep going

GDm and Heavy Ball Equivalence

GD with momentum:

$$
\begin{aligned}
m^t &= \beta\, m^{t-1} + \nabla f(w^t) \\
w^{t+1} &= w^t - \gamma\, m^t
\end{aligned}
$$

GDm and Heavy Ball Equivalence

GD with momentum:

$$m^t = \beta\, m^{t-1} + \nabla f(w^t)$$
$$w^{t+1} = w^t - \gamma\, m^t$$

$$\begin{aligned} w^{t+1} &= w^t - \gamma\, m^t \\ &= w^t - \gamma\left(\beta m^{t-1} + \nabla f(w^t)\right) \\ &= w^t - \gamma\,\nabla f(w^t) - \gamma\beta\, m^{t-1} \\ &= w^t - \gamma\,\nabla f(w^t) + \tfrac{\gamma\beta}{\gamma}\left(w^t - w^{t-1}\right) \end{aligned}$$

GDm and Heavy Ball Equivalence

GD with momentum:

$$m^t = \beta\, m^{t-1} + \nabla f(w^t)$$
$$w^{t+1} = w^t - \gamma\, m^t$$

$$m^{t-1} = -\frac{1}{\gamma}(w^t - w^{t-1})$$

$$\begin{aligned}
w^{t+1} &= w^t - \gamma\, m^t \\
&= w^t - \gamma\,(\beta m^{t-1} + \nabla f(w^t)) \\
&= w^t - \gamma\,\nabla f(w^t) - \gamma\beta\, m^{t-1} \\
&= w^t - \gamma\,\nabla f(w^t) + \frac{\gamma\beta}{\gamma}\,\overbrace{(w^t - w^{t-1})}
\end{aligned}$$

GDm and Heavy Ball Equivalence

GD with momentum:

$$m^t = \beta\, m^{t-1} + \nabla f(w^t)$$
$$w^{t+1} = w^t - \gamma\, m^t$$

$$m^{t-1} = -\frac{1}{\gamma}(w^t - w^{t-1})$$

$$\begin{aligned} w^{t+1} &= w^t - \gamma\, m^t \\ &= w^t - \gamma\,(\beta m^{t-1} + \nabla f(w^t)) \\ &= w^t - \gamma\,\nabla f(w^t) - \gamma\beta\, m^{t-1} \\ &= w^t - \gamma\,\nabla f(w^t) + \frac{\gamma\beta}{\gamma}\,(w^t - w^{t-1}) \end{aligned}$$

$$w^{t+1} = w^t - \gamma\,\nabla f(w^t) + \beta(w^t - w^{t-1})$$

GDm and Heavy Ball Equivalence

GD with momentum:

$$m^t = \beta\, m^{t-1} + \nabla f(w^t)$$
$$w^{t+1} = w^t - \gamma\, m^t$$

$$m^{t-1} = -\frac{1}{\gamma}(w^t - w^{t-1})$$

$$\begin{aligned} w^{t+1} &= w^t - \gamma\, m^t \\ &= w^t - \gamma\, (\beta m^{t-1} + \nabla f(w^t)) \\ &= w^t - \gamma\, \nabla f(w^t) - \gamma\beta\, m^{t-1} \\ &= w^t - \gamma\, \nabla f(w^t) + \frac{\gamma\beta}{\gamma}\, \overbrace{(w^t - w^{t-1})} \end{aligned}$$

Heavey Ball Method:

$$w^{t+1} = w^t - \gamma\, \nabla f(w^t) + \beta(w^t - w^{t-1})$$

Convergence of Gradient Descent with Momentum

Polyak 1964

Theorem Let f be μ–strongly convex and L–smooth, that is

$$\mu I \preceq \nabla^2 f(w) \preceq LI, \quad \forall w \in \mathbb{R}^d$$

stepsize

If $\gamma = \dfrac{4}{(\sqrt{L}+\sqrt{\mu})^2}$ and $\beta = \dfrac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$ then SGDm converges

momentum parameter

$$\Rightarrow \quad \|w^t - w^*\| \leq \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^t \|w^0 - w^*\|$$

$\kappa := L/\mu$

Convergence of Gradient Descent with Momentum

Polyak 1964

Theorem Let f be μ–strongly convex and L–smooth, that is

stepsize

$$\mu I \quad \preceq \quad \nabla^2 f(w) \quad \preceq \quad LI, \quad \forall w \in \mathbb{R}^d$$

If $\gamma = \dfrac{4}{(\sqrt{L}+\sqrt{\mu})^2}$ and $\beta = \dfrac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$ then SGDm converges

momentum parameter

$$\Rightarrow \quad \|w^t - w^*\| \;\leq\; \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^t \|w^0 - w^*\|$$

$\kappa := L/\mu$

Corollary $t \;\geq\; \dfrac{1}{\sqrt{\kappa}+1}\log\left(\dfrac{1}{\epsilon}\right) \quad \Rightarrow \quad \dfrac{\|w^t - w^*\|}{\|w^0 - w^*\|} \leq \epsilon$

Proof sketch: GDm convergence

Fundamental Theorem of Calculus

$$\int_{s=0}^{1} \nabla^2 f(\underbrace{w_s}_{w_s := w^* + s(w^t - w^*)}) ds (w^t - w^*) = \nabla f(w^t) - \nabla f(w^*) = \nabla f(w^t)$$

Proof sketch: GDm convergence

Fundamental Theorem of Calculus

$$\int_{s=0}^{1} \nabla^2 f(\underbrace{w_s}_{w_s := w^* + s(w^t - w^*)})ds(w^t - w^*) = \nabla f(w^t) - \nabla f(w^*) = \nabla f(w^t)$$

$$\begin{aligned}
w^{t+1} - w^* &= w^t - w^* - \gamma \nabla f(w^t) + \beta(w^t - w^{t-1}) \qquad +w^* - w^* \\
&= \left(I - \gamma \int_{s=0}^{1} \nabla^2 f(w^s)\right)(w^t - w^*) + \beta(w^t - w^{t-1}) \\
&= \left((1+\beta)I - \gamma \int_{s=0}^{1} \nabla^2 f(w^s)\right)(w^t - w^*) - \beta(w^{t-1} - w^*)
\end{aligned}$$

Proof sketch: GDm convergence

Fundamental Theorem of Calculus

$$\int_{s=0}^{1} \nabla^2 f(\underbrace{w_s}_{w_s := w^* + s(w^t - w^*)})ds(w^t - w^*) = \nabla f(w^t) - \nabla f(w^*) = \nabla f(w^t)$$

$$\begin{aligned} w^{t+1} - w^* &= w^t - w^* - \gamma \nabla f(w^t) + \beta(w^t - w^{t-1}) \quad +w^* - w^* \\ &= \left(I - \gamma \int_{s=0}^{1} \nabla^2 f(w^s)\right)(w^t - w^*) + \beta(w^t - w^{t-1}) \\ &= \left(\underbrace{(1+\beta)I - \gamma \int_{s=0}^{1} \nabla^2 f(w^s)}_{=: A_s}\right)(w^t - w^*) - \beta(w^{t-1} - w^*) \end{aligned}$$

Proof sketch: GDm convergence

Fundamental Theorem of Calculus

$$\int_{s=0}^{1} \nabla^2 f(\underbrace{w_s}_{w_s := w^* + s(w^t - w^*)}) ds (w^t - w^*) = \nabla f(w^t) - \nabla f(w^*) = \nabla f(w^t)$$

$$\begin{aligned}
w^{t+1} - w^* &= w^t - w^* - \gamma \nabla f(w^t) + \beta(w^t - w^{t-1}) \quad +w^* - w^* \\
&= \left(I - \gamma \int_{s=0}^{1} \nabla^2 f(w^s)\right) (w^t - w^*) + \beta(w^t - w^{t-1}) \\
&= \Big(\underbrace{(1+\beta)I - \gamma \int_{s=0}^{1} \nabla^2 f(w^s)}_{=: A_s}\Big) (w^t - w^*) - \beta(w^{t-1} - w^*) \\
&= A_s(w^t - w^*) - \beta(w^{t-1} - w^*)
\end{aligned}$$

Proof sketch: GDm convergence

Fundamental Theorem of Calculus

$$\int_{s=0}^{1} \nabla^2 f(\underbrace{w_s}_{w_s := w^* + s(w^t - w^*)})ds(w^t - w^*) = \nabla f(w^t) - \nabla f(w^*) = \nabla f(w^t)$$

$$\begin{aligned}
w^{t+1} - w^* &= w^t - w^* - \gamma \nabla f(w^t) + \beta(w^t - w^{t-1}) \quad +w^* - w^* \\
&= \left(I - \gamma \int_{s=0}^{1} \nabla^2 f(w^s)\right)(w^t - w^*) + \beta(w^t - w^{t-1}) \\
&= \left(\underbrace{(1+\beta)I - \gamma \int_{s=0}^{1} \nabla^2 f(w^s)}_{=: A_s}\right)(w^t - w^*) - \beta(w^{t-1} - w^*) \\
&= A_s(w^t - w^*) - \beta(w^{t-1} - w^*)
\end{aligned}$$

Depends on past. Difficult recurrence

Proof: Convergence of Heavy Ball

$$z^{t+1} = \begin{bmatrix} w^{t+1} - w^* \\ w^t - w^* \end{bmatrix} \in \mathbb{R}^{2d}$$

Proof: Convergence of Heavy Ball

$$z^{t+1} = \begin{bmatrix} w^{t+1} - w^* \\ w^t - w^* \end{bmatrix} \in \mathbb{R}^{2d}$$

$$z^{t+1} = \begin{bmatrix} w^{t+1} - w^* \\ w^t - w^* \end{bmatrix} = \begin{bmatrix} A_s(w^t - w^*) - \beta(w^{t-1} - w^*) \\ w^t - w^* \end{bmatrix}$$

Proof: Convergence of Heavy Ball

$$z^{t+1} = \begin{bmatrix} w^{t+1} - w^* \\ w^t - w^* \end{bmatrix} \in \mathbb{R}^{2d}$$

$$z^{t+1} = \begin{bmatrix} w^{t+1} - w^* \\ w^t - w^* \end{bmatrix} = \begin{bmatrix} A_s(w^t - w^*) - \beta(w^{t-1} - w^*) \\ w^t - w^* \end{bmatrix}$$

$$= \begin{bmatrix} A_s & -I\beta \\ I & 0 \end{bmatrix} \begin{bmatrix} w^t - w^* \\ w^{t-1} - w^* \end{bmatrix}$$

Proof: Convergence of Heavy Ball

$$z^{t+1} = \begin{bmatrix} w^{t+1} - w^* \\ w^t - w^* \end{bmatrix} \in \mathbb{R}^{2d}$$

$$z^{t+1} = \begin{bmatrix} w^{t+1} - w^* \\ w^t - w^* \end{bmatrix} = \begin{bmatrix} A_s(w^t - w^*) - \beta(w^{t-1} - w^*) \\ w^t - w^* \end{bmatrix}$$

$$= \begin{bmatrix} A_s & -I\beta \\ I & 0 \end{bmatrix} \begin{bmatrix} w^t - w^* \\ w^{t-1} - w^* \end{bmatrix}$$

$$= \begin{bmatrix} A_s & -I\beta \\ I & 0 \end{bmatrix} z^t$$

Proof: Convergence of Heavy Ball

$$z^{t+1} = \begin{bmatrix} w^{t+1} - w^* \\ w^t - w^* \end{bmatrix} \in \mathbb{R}^{2d}$$

$$z^{t+1} = \begin{bmatrix} w^{t+1} - w^* \\ w^t - w^* \end{bmatrix} = \begin{bmatrix} A_s(w^t - w^*) - \beta(w^{t-1} - w^*) \\ w^t - w^* \end{bmatrix}$$

$$= \begin{bmatrix} A_s & -I\beta \\ I & 0 \end{bmatrix} \begin{bmatrix} w^t - w^* \\ w^{t-1} - w^* \end{bmatrix}$$

$$= \begin{bmatrix} A_s & -I\beta \\ I & 0 \end{bmatrix} z^t$$

Simple recurrence!

Proof: Convergence of Heavy Ball

$$z^{t+1} = \begin{bmatrix} w^{t+1} - w^* \\ w^t - w^* \end{bmatrix} \in \mathbb{R}^{2d}$$

$$z^{t+1} = \begin{bmatrix} w^{t+1} - w^* \\ w^t - w^* \end{bmatrix} = \begin{bmatrix} A_s(w^t - w^*) - \beta(w^{t-1} - w^*) \\ w^t - w^* \end{bmatrix}$$

$$= \begin{bmatrix} A_s & -I\beta \\ I & 0 \end{bmatrix} \begin{bmatrix} w^t - w^* \\ w^{t-1} - w^* \end{bmatrix}$$

$$= \begin{bmatrix} A_s & -I\beta \\ I & 0 \end{bmatrix} z^t$$

Simple recurrence!

$$\|z^{t+1}\| \leq \left\| \begin{bmatrix} A_s & -I\beta \\ I & 0 \end{bmatrix} \right\| \|z^t\|$$

Proof: Convergence of Heavy Ball

$$\|z^{t+1}\| \quad \leq \quad \underbrace{\left\| \begin{bmatrix} A_s & -I\beta \\ I & 0 \end{bmatrix} \right\|} \; \|z^t\|$$

$$\|A\| \quad := \quad \max_{i=1,\dots,2n} |\lambda_i(A)|$$

Proof: Convergence of Heavy Ball

$$\|z^{t+1}\| \quad \leq \quad \underbrace{\left\| \begin{bmatrix} A_s & -I\beta \\ I & 0 \end{bmatrix} \right\|}_{\|A\| \;:=\; \max_{i=1,\ldots,2n} |\lambda_i(A)|} \|z^t\|$$

EXE on Eigenvalues:

$$\text{If } \gamma = \frac{4}{(\sqrt{L}+\sqrt{\mu})^2} \text{ and } \quad \beta = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}} \text{ then}$$

$$\left\| \begin{bmatrix} A_s & -I\beta \\ I & 0 \end{bmatrix} \right\| \quad = \quad \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$$

Proof: Convergence of Heavy Ball

$$\|z^{t+1}\| \leq \underbrace{\left\| \begin{bmatrix} A_s & -I\beta \\ I & 0 \end{bmatrix} \right\|} \|z^t\|$$

$$\|A\| := \max_{i=1,\ldots,2n} |\lambda_i(A)|$$

$$(1+\beta)I - \gamma \int_{s=0}^{1} \nabla^2 f(w^s)$$

EXE on Eigenvalues:

If $\gamma = \dfrac{4}{(\sqrt{L}+\sqrt{\mu})^2}$ and $\beta = \dfrac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$ then

$$\left\| \begin{bmatrix} A_s & -I\beta \\ I & 0 \end{bmatrix} \right\| = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$$

Adding Momentum to SGD

Rumelhart, Hinton, Geoffrey, Ronald, 1986, Nature

Stochastic Heavey Ball Method:

$$w^{t+1} = w^t - \gamma \nabla f_{j_t}(w^t) + \underbrace{\beta(w^t - w^{t-1})}_{\text{Adds “Inertia” to update}}$$

Sampled i.i.d
$j \in \{1, \ldots, n\}$
$j \sim \frac{1}{n}$

Adding Momentum to SGD

Rumelhart, Hinton, Geoffrey, Ronald, 1986, Nature

Stochastic Heavey Ball Method:

$$w^{t+1} = w^t - \gamma \nabla f_{j_t}(w^t) + \underbrace{\beta(w^t - w^{t-1})}_{\text{Adds "Inertia" to update}}$$

Sampled i.i.d
$j \in \{1, \ldots, n\}$
$j \sim \frac{1}{n}$

SGD with momentum (SGDm):

$$m^t = \beta\, m^{t-1} + \nabla f_{j_t}(w^t)$$

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

SGDm and Averaging

$$\begin{aligned} m^t &= \beta\, m^{t-1} + \nabla f_{j_t}(w^t) \\ &= \beta\, m^{t-2} + \nabla f_{j_t}(w^t) + \beta \nabla f_{j_{t-1}}(w^{t-1}) \\ &= \textstyle\sum_{i=1}^{t} \beta^i \nabla f_{j_{t-i}}(w^{t-i}) \end{aligned}$$

http://fa.bianp.net/teaching/2018/COMP-652/stochastic_gradient.html

SGDm and Averaging

$$\begin{aligned} m^t &= \beta\, m^{t-1} + \nabla f_{j_t}(w^t) \\ &= \beta\, m^{t-2} + \nabla f_{j_t}(w^t) + \beta \nabla f_{j_{t-1}}(w^{t-1}) \\ &= \textstyle\sum_{i=1}^{t} \beta^i \nabla f_{j_{t-i}}(w^{t-i}) \end{aligned}$$

$m^0 = 0$

http://fa.bianp.net/teaching/2018/COMP-652/stochastic_gradient.html

SGDm and Averaging

$$
\begin{aligned}
m^t &= \beta\, m^{t-1} + \nabla f_{j_t}(w^t) \\
&= \beta\, m^{t-2} + \nabla f_{j_t}(w^t) + \beta \nabla f_{j_{t-1}}(w^{t-1}) \\
&= \textstyle\sum_{i=1}^{t} \beta^i \nabla f_{j_{t-i}}(w^{t-i})
\end{aligned}
$$

$m^0 = 0$

SGD with momentum (SGDm):

$$
w^{t+1} = w^t - \gamma \sum_{i=1}^{t} \beta^i \nabla f_{j_{t-i}}(w^{t-i})
$$

http://fa.bianp.net/teaching/2018/COMP-652/stochastic_gradient.html

SGDm and Averaging

$$\begin{aligned} m^t &= \beta\, m^{t-1} + \nabla f_{j_t}(w^t) \\ &= \beta\, m^{t-2} + \nabla f_{j_t}(w^t) + \beta \nabla f_{j_{t-1}}(w^{t-1}) \\ &= \textstyle\sum_{i=1}^{t} \beta^i \nabla f_{j_{t-i}}(w^{t-i}) \end{aligned}$$

$m^0 = 0$

SGD with momentum (SGDm):

$$w^{t+1} = w^t - \gamma \sum_{i=1}^{t} \beta^i \nabla f_{j_{t-i}}(w^{t-i})$$

Acts like an approximate variance reduction since

http://fa.bianp.net/teaching/2018/COMP-652/stochastic_gradient.html

SGDm and Averaging

$$\begin{aligned} m^t &= \beta\, m^{t-1} + \nabla f_{j_t}(w^t) \\ &= \beta\, m^{t-2} + \nabla f_{j_t}(w^t) + \beta \nabla f_{j_{t-1}}(w^{t-1}) \\ &= \textstyle\sum_{i=1}^t \beta^i \nabla f_{j_{t-i}}(w^{t-i}) \end{aligned}$$

$m^0 = 0$

SGD with momentum (SGDm):

$$w^{t+1} = w^t - \gamma \sum_{i=1}^{t} \beta^i \nabla f_{j_{t-i}}(w^{t-i})$$

Acts like an approximate variance reduction since

$$\sum_{i=1}^{t} \beta^i \nabla f_{j_{t-i}}(w^{t-i}) \approx \sum_{i=1}^{n} \frac{1}{n} \nabla f_i(w^t) = \nabla f(w^t)$$

http://fa.bianp.net/teaching/2018/COMP-652/stochastic_gradient.html

Why Machine Learners Like SGD

Why Machine Learners like SGD

Though we solve:

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

We want to solve:

The statistical learning problem:

Minimize the expected loss over an *unknown* expectation

$$\min_{w \in \mathbf{R}^d} \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell\left(h_w(x), y\right)\right]$$

SGD can solve the statistical learning problem!

Why Machine Learners like SGD

The statistical learning problem:

Minimize the expected loss over an *unknown* expectation

$$\min_{w \in \mathbf{R}^d} \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell\left(h_w(x), y\right)\right]$$

SGD ∞.0 for learning

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

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

sample $(x, y) \sim \mathcal{D}$

calculate $v_t \in \partial \ell(h_{w^t}(x), y)$

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

Output $\overline{w}^T = \frac{1}{T}\sum_{t=1}^{T} w^t$

Bring laptops for Thursday TD !

RMG, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin and Peter Richtárik (2019), ICML
SGD: general analysis and improved rates

RMG, P. Richtarik, F. Bach (2018), preprint online
Stochastic quasi-gradient methods: Variance reduction via Jacobian sketching

N. Gazagnadou, RMG, J. Salmon (2019) , ICML 2019.
Optimal mini-batch and step sizes for SAGA

O. Sebbouh, N. Gazagnadou, S. Jelassi, F. Bach, RMG Neurips 2019, preprint online. **Towards closing the gap between the theory and practice of SVRG**