Optimization for Machine Learning

Stochastic Gradient Methods

Lecturers: Francis Bach & Robert M. Gower

Tutorials: Hadrien Hendrikx, Rui Yuan, Nidham Gazagnadou

AIMS | African Institute for Mathematical Sciences NEXT EINSTEIN INITIATIVE

African Master's in Machine Intelligence (AMMI), Kigali

Solving the Finite Sum Training Problem

Recap

Training Problem

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

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?

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 $\alpha > 0$

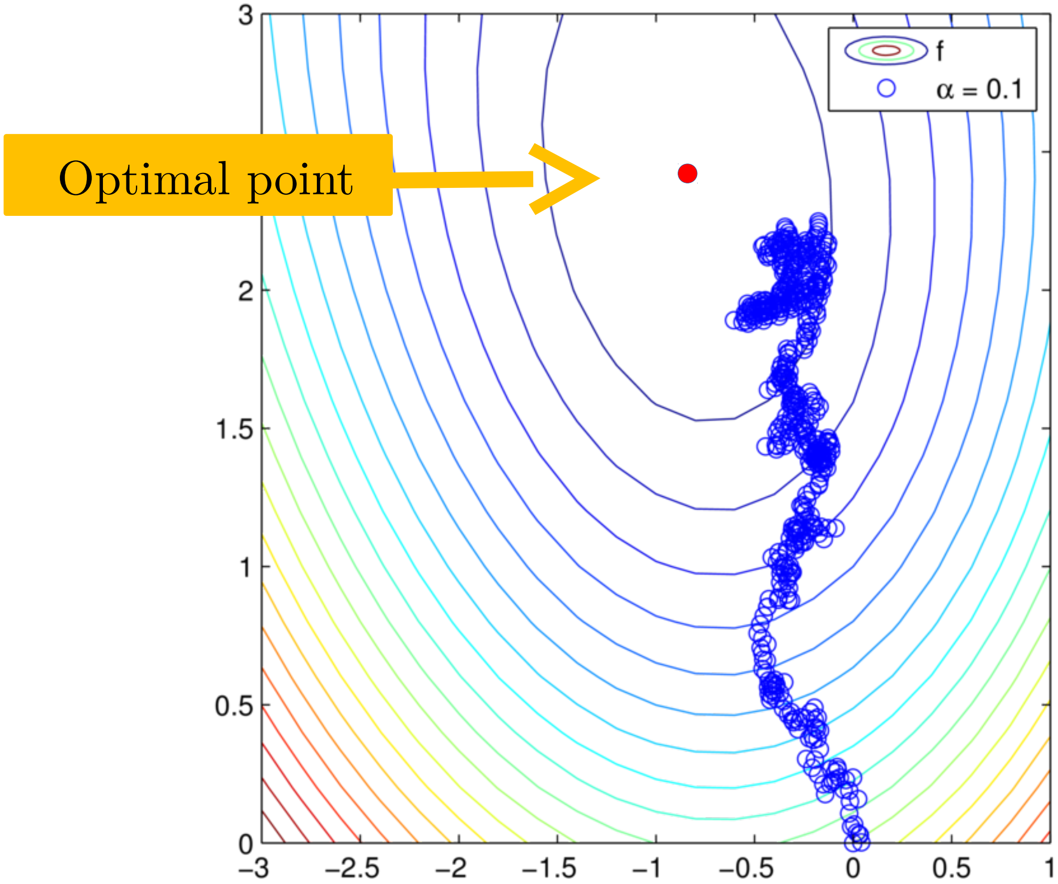
for $t = 0, 1, 2, \ldots, T-1$

  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

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

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

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

Expected Bounded Stochastic Gradients

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

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 $\alpha \approx 0$

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

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

Strong conv.

Bounded Stoch grad

Taking total expectation

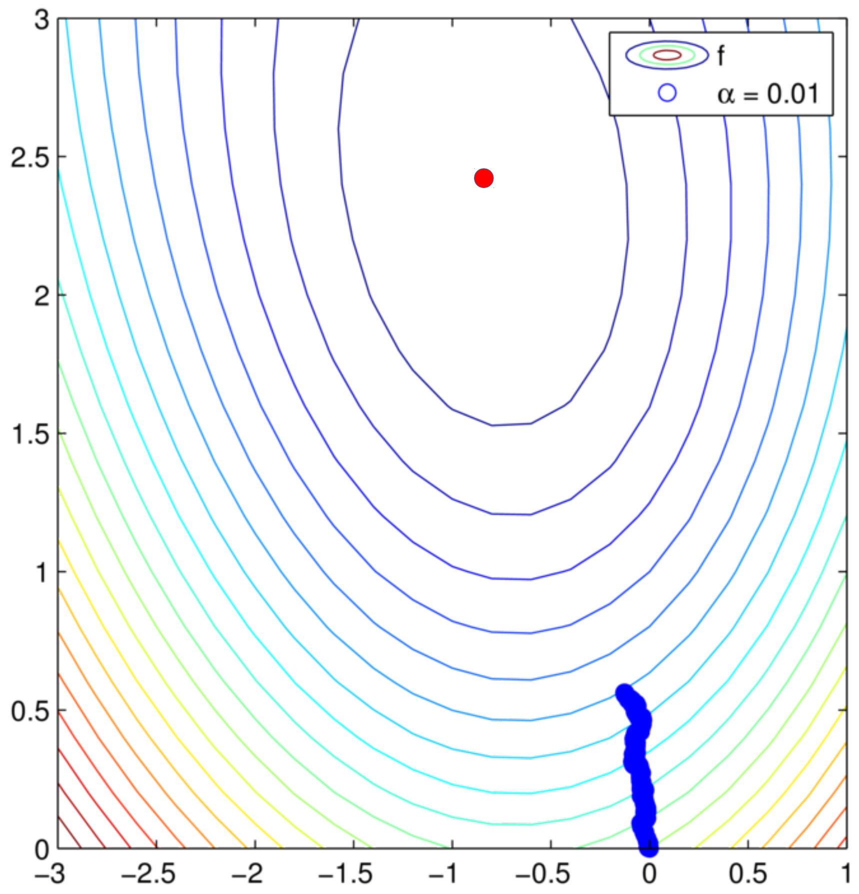
$$\begin{aligned} \mathbb{E}\left[\|w^{t+1} - w^*\|_2^2\right] &\le (1 - \alpha\lambda)\mathbb{E}\left[\|w^t - w^*\|_2^2\right] + \alpha^2 B^2 \\ &= (1 - \alpha\lambda)^{t+1}\|w^0 - w^*\|_2^2 + \textstyle\sum_{i=0}^t (1 - \alpha\lambda)^i \alpha^2 B^2 \end{aligned}$$

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

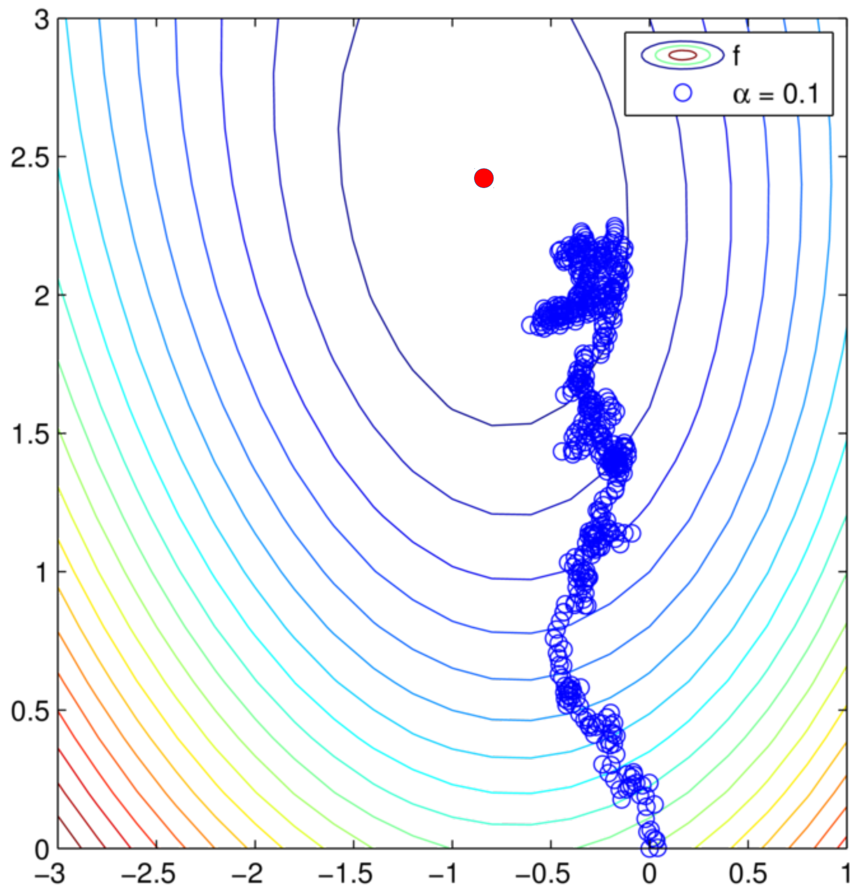
$$\mathbb{E}\left[\|w^{t+1} - w^*\|_2^2\right] \le (1 - \alpha\lambda)^{t+1}\|w^0 - w^*\|_2^2 + \tfrac{\alpha}{\lambda} B^2$$

Stochastic Gradient Descent

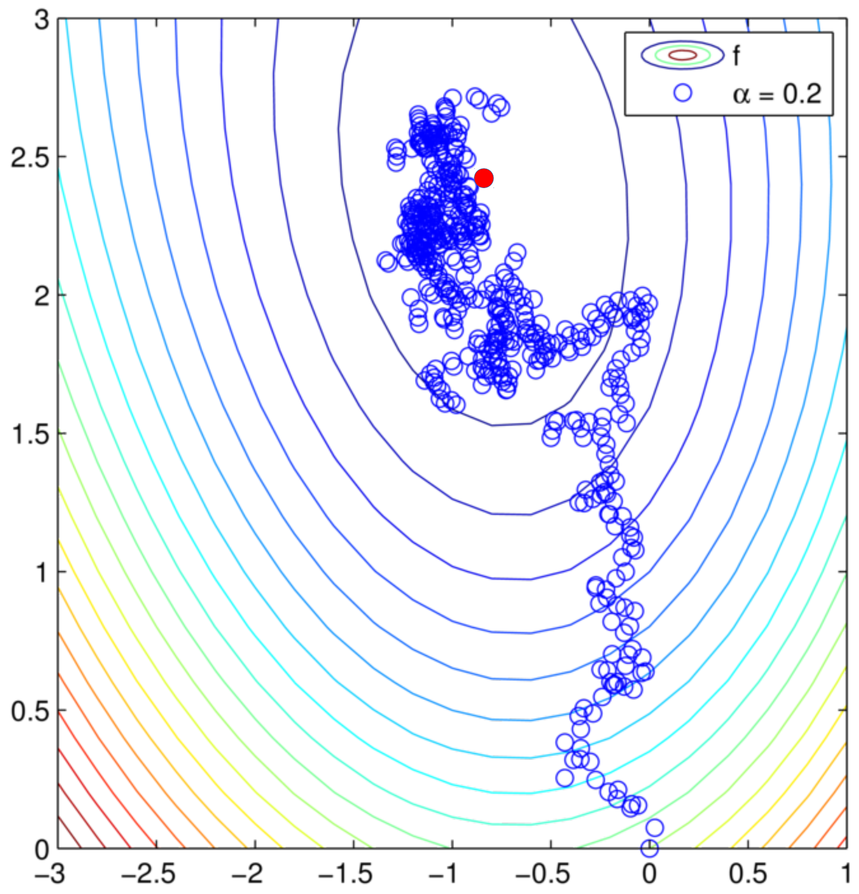
α =0.01



Stochastic Gradient Descent
α =0.1

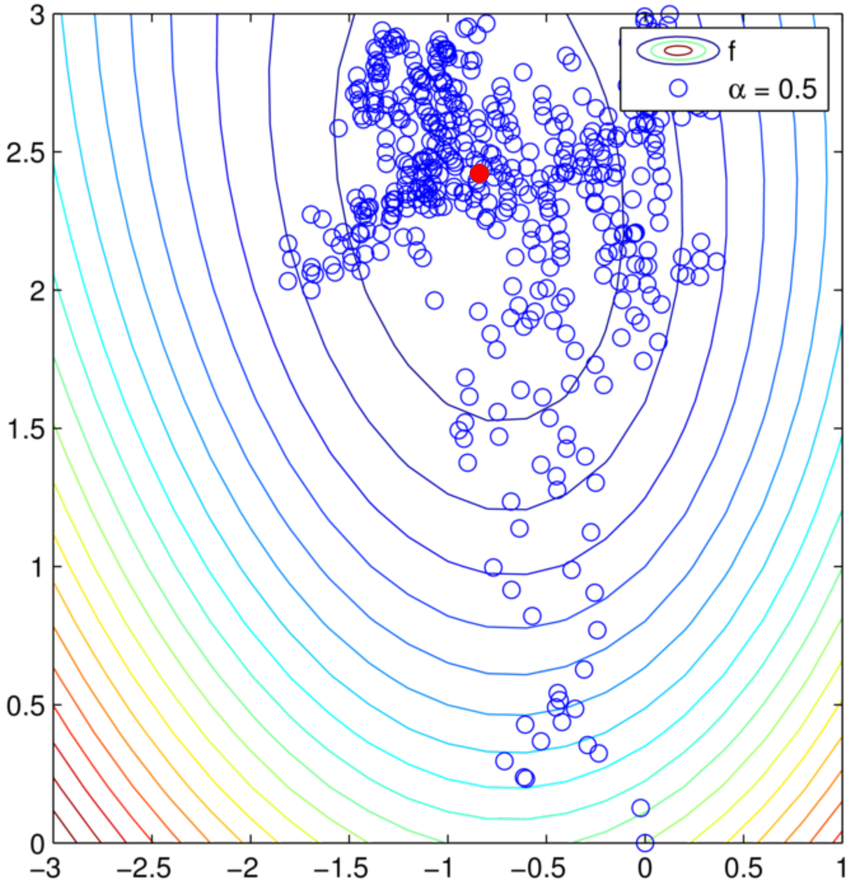


Stochastic Gradient Descent
α =0.2



Stochastic Gradient Descent

α =0.5



Assumptions for Convergence

Strong Convexity

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

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

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

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

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

Realistic assumptions for Convergence

Strongly quasi-convexity

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

Each f_i is convex and L_i smooth

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

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

Gradient Noise

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

Realistic assumptions for Convergence

Strongly quasi-convexity

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

Each f_i is convex and L_i smooth

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

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

Gradient Noise

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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 $f(w)$ be 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 $L_i = \max_{w \in {}^d} \lambda_{\max}(\nabla^2 f_i(w))$

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 $f(w)$ be 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 $L_i = \max_{w\in^d} \lambda_{\max}(\nabla^2 f_i(w))$

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

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

The final result now follows by taking the max over w, *then* max over i

Complexity / Convergence

Theorem

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

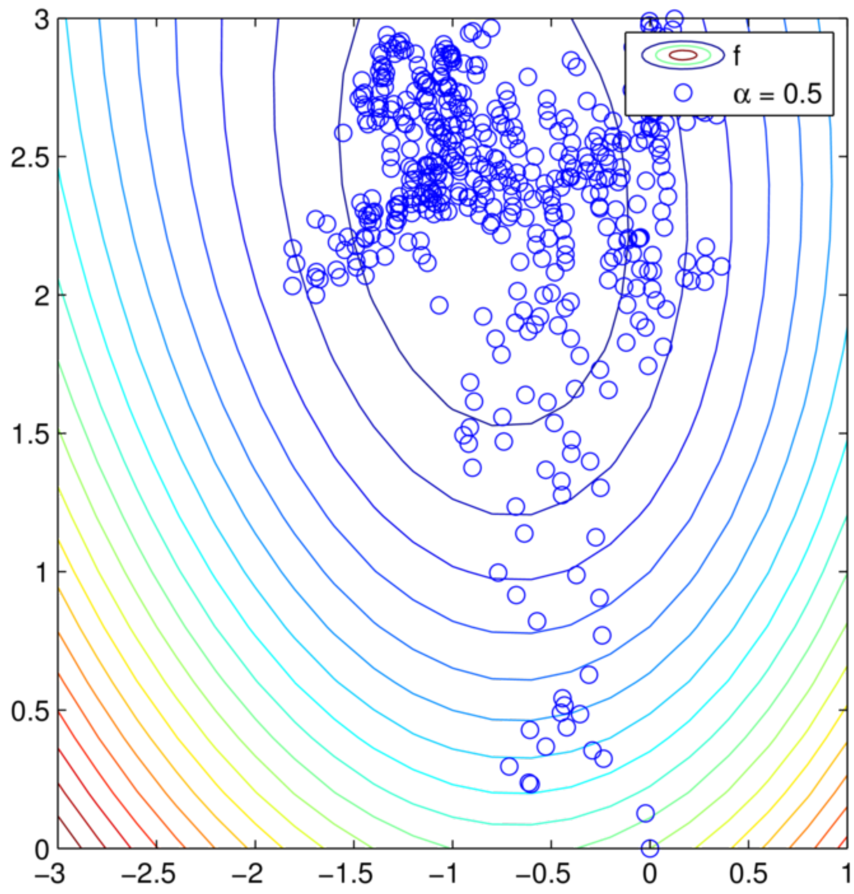
$$\mathbb{E}\left[||w^t - w^*||_2^2\right] \leq (1 - \alpha\lambda)^t ||w^0 - w^*||_2^2 + \frac{2\alpha}{\lambda}\sigma^2$$

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

RMG, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, P. Richtarik (2019) arXiv:1901.09401
SGD: General Analysis and Improved Rates.

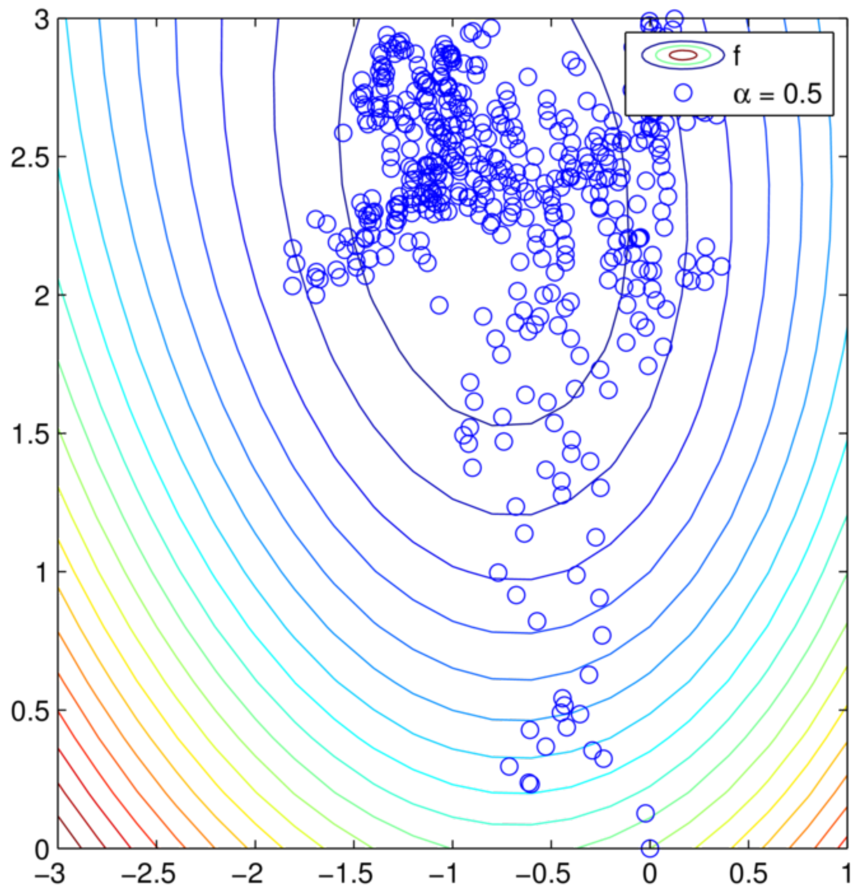
Stochastic Gradient Descent

α =0.5



Stochastic Gradient Descent

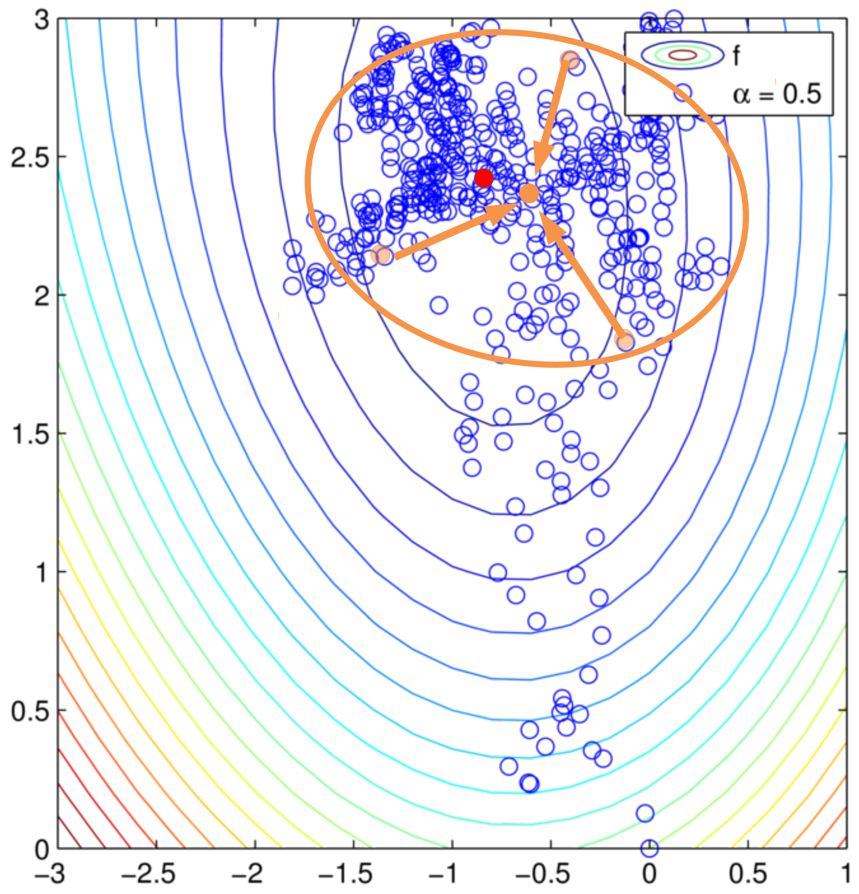
α =0.5



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

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

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

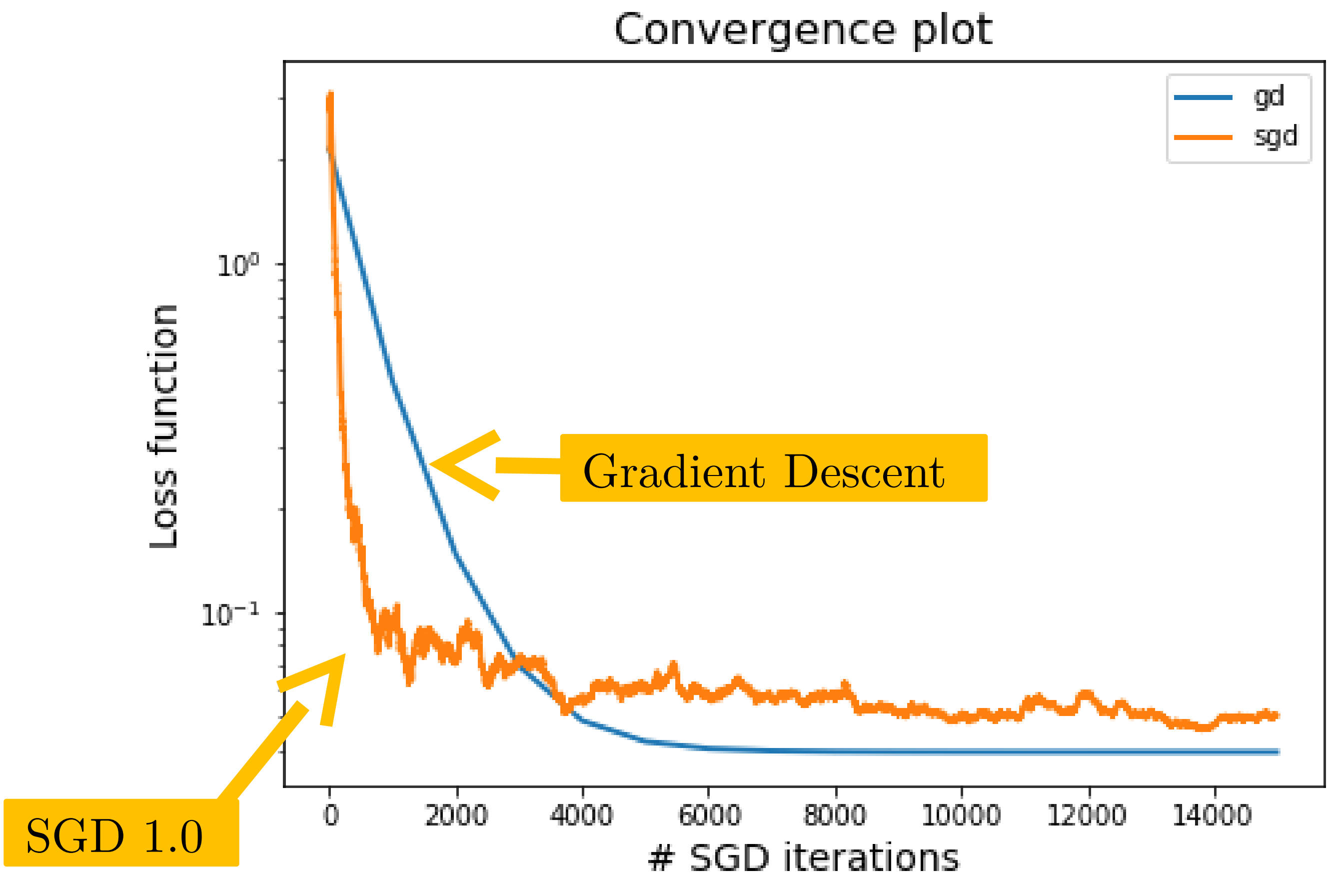
Shrinking Stepsize

How should we sample j ?

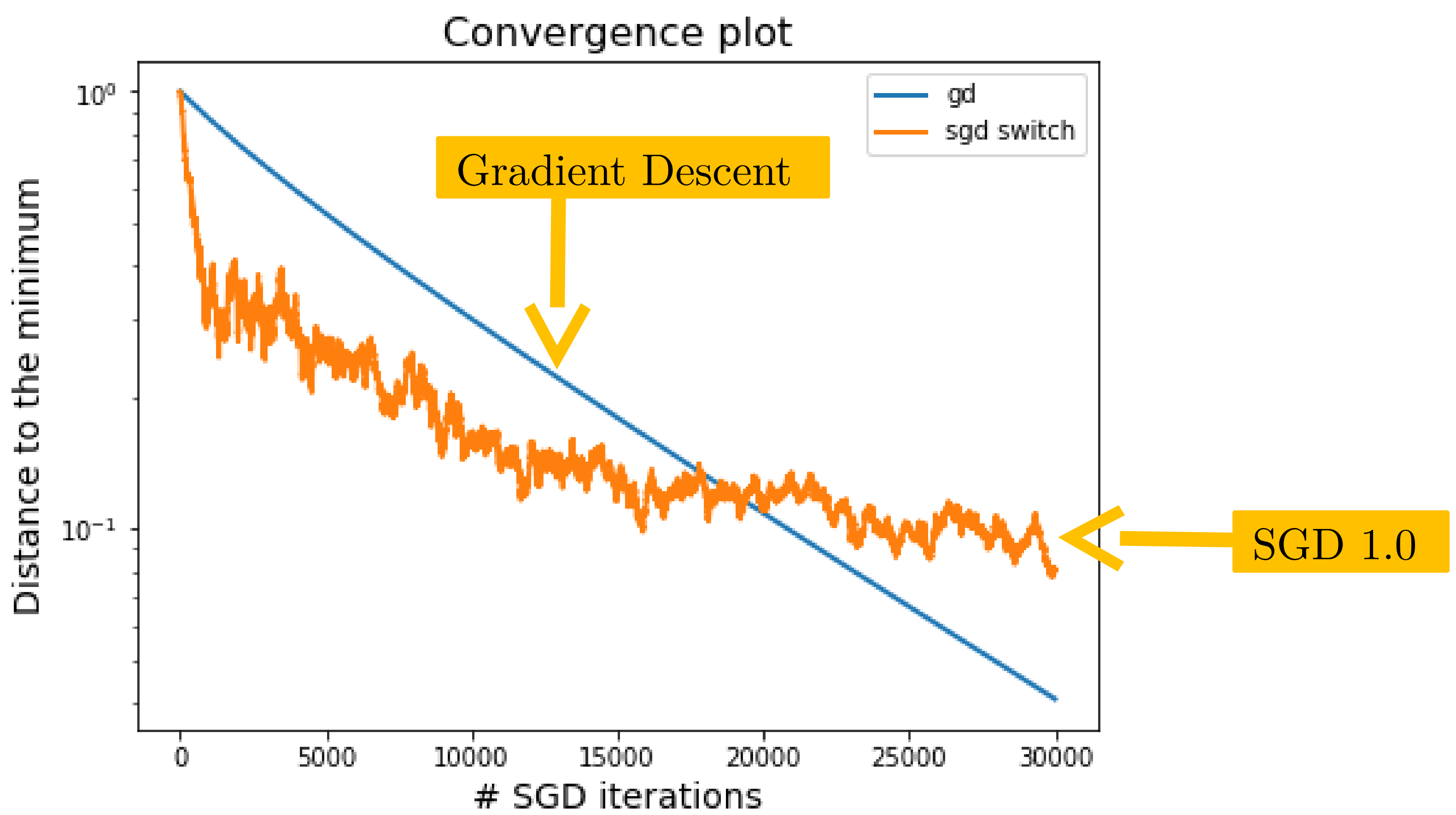
How fast $\alpha_t \rightarrow 0$?

Does this converge?

SGD with shrinking stepsize
Compared with Gradient Descent



SGD with shrinking stepsize
Compared with Gradient Descent



Complexity / Convergence

Theorem for shrinking stepsizes

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

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

Complexity / Convergence

Theorem for shrinking stepsizes

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

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

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

Complexity / Convergence

Theorem for shrinking stepsizes

Let $\mathcal{K} := L_{\max}/\mu$ and let

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

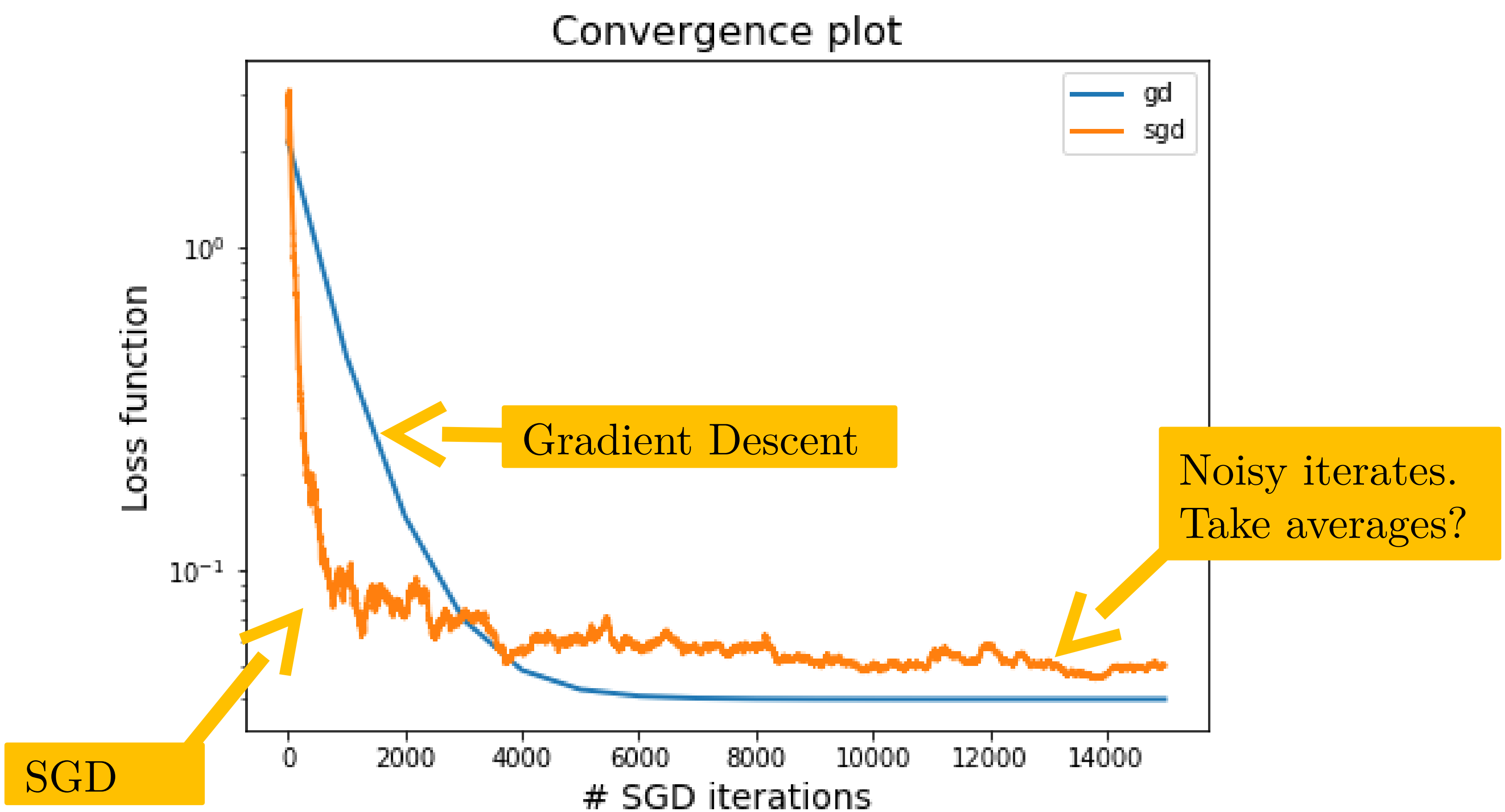
$\alpha^t = O(1/(t+1))$

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

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

In practice often $\alpha^t = C/(t+1)$ where C is tuned

Stochastic Gradient Descent
Compared with Gradient Descent



SGD with (late start) averaging

SGDA 1.1

Set $w^0 = 0$

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

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: $\overline{w} = w$

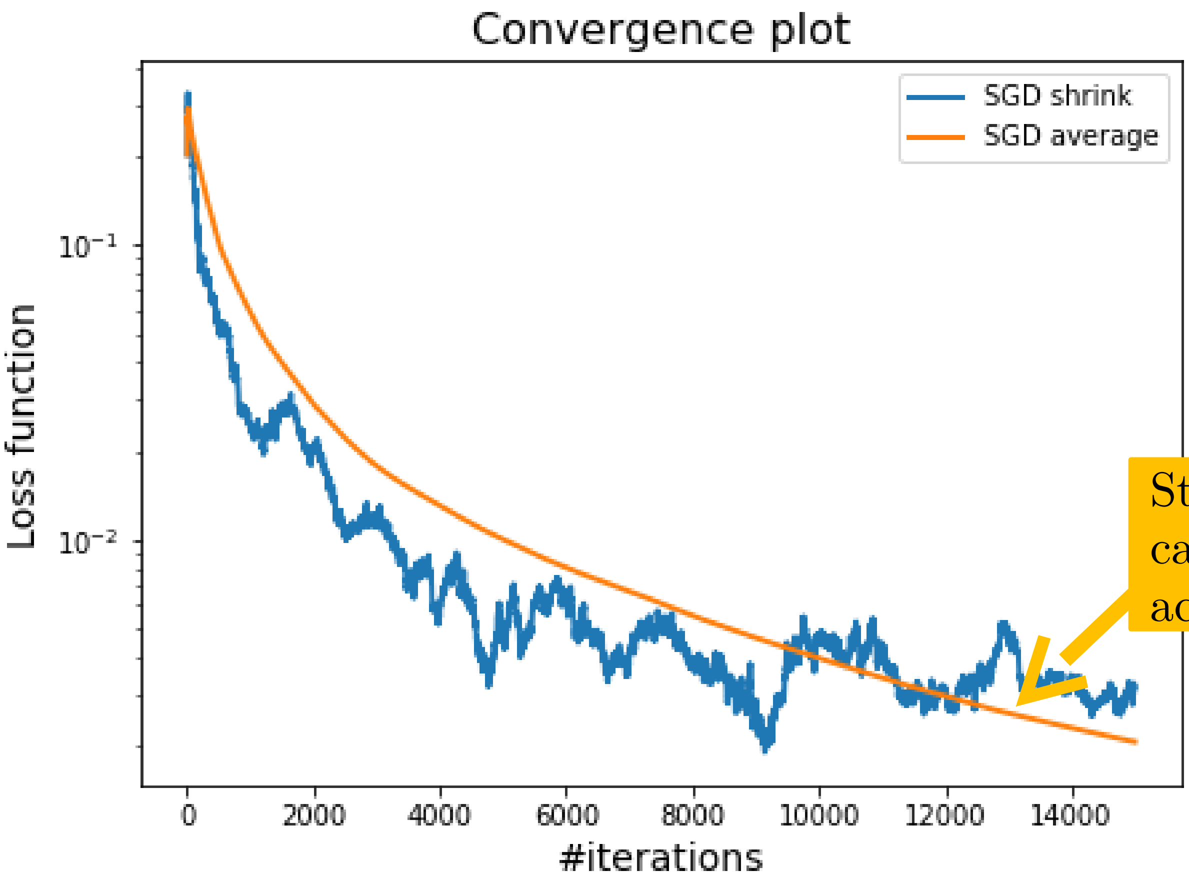
Output $\overline{w}$

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

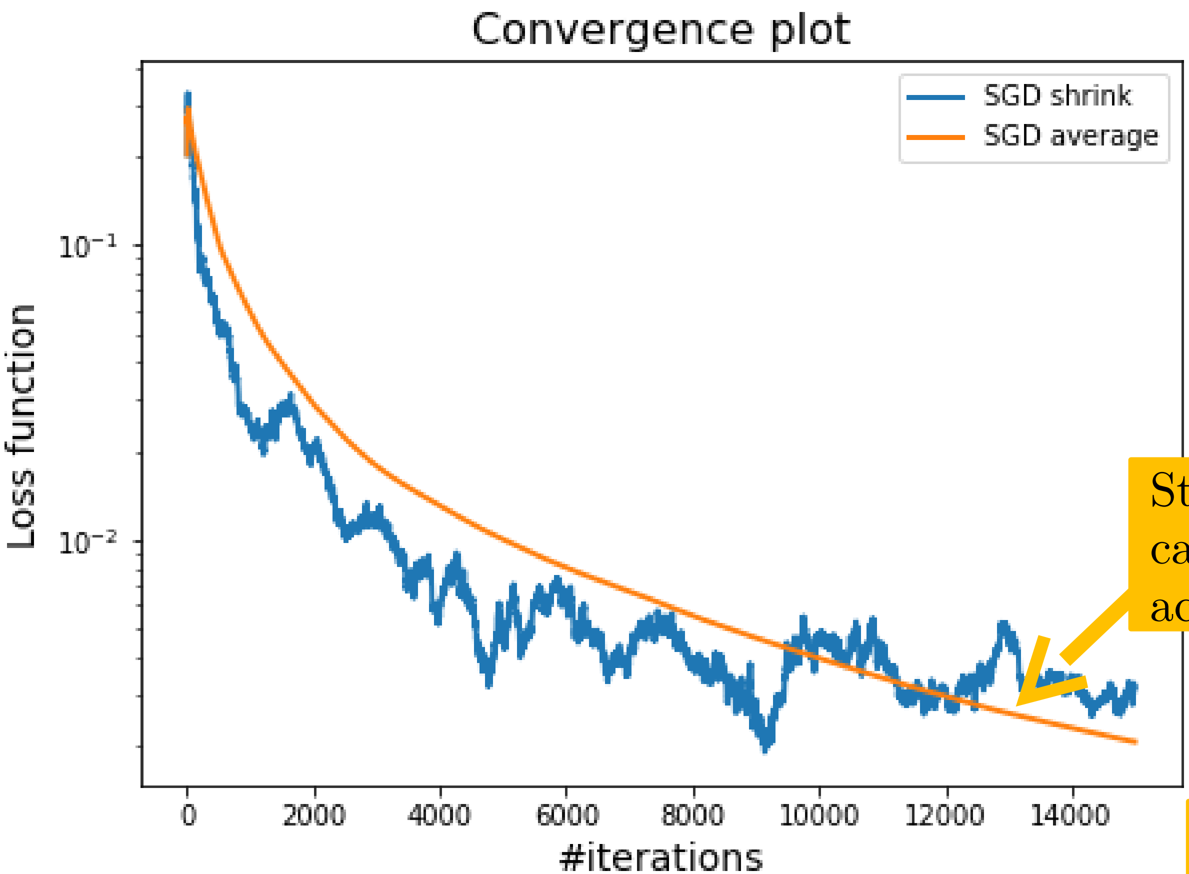
Stochastic Gradient Descent

With and without averaging



Starts slow, but can reach higher accuracy

Stochastic Gradient Descent
With and without averaging

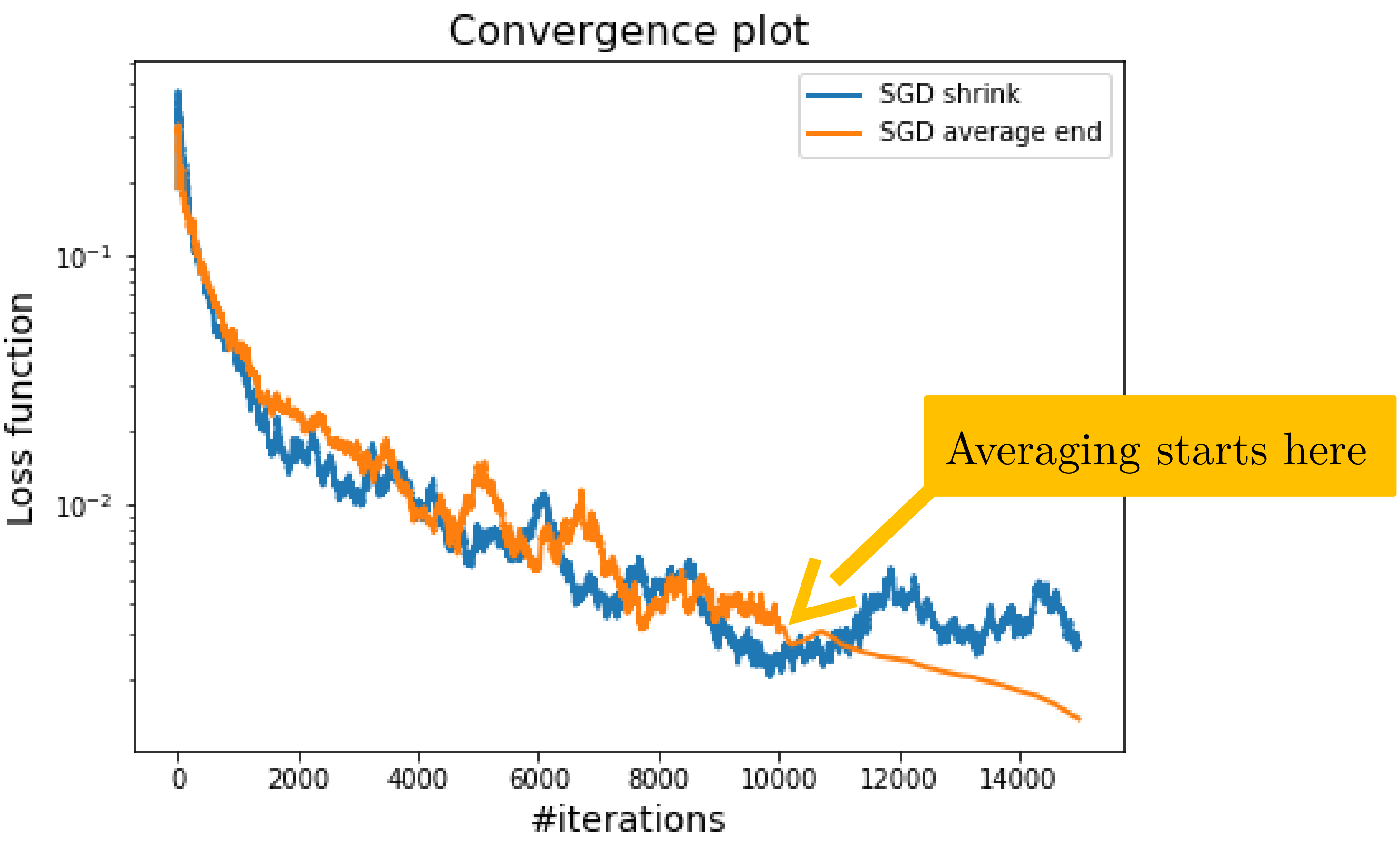


Starts slow, but can reach higher accuracy

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	$O\left(\frac{L_{\max}}{\lambda\epsilon}\right)$	$O\left(\frac{nL}{\lambda}\log\left(\frac{1}{\epsilon}\right)\right)$
Cost of an interation	$O\left(d\right)$	$O\left(nd\right)$

$$\text{Total complexity } = (\text{Iteration complexity}) \times (\text{Cost of an iteration})$$

Total complexity GD and SGD for strongly convex

Approximate solution

$$\mathbb{E}[\|w^t - w^*\|] \leq \epsilon$$

SGD

$$T \geq O\left(\frac{L_{\max}}{\lambda\epsilon}\right)$$

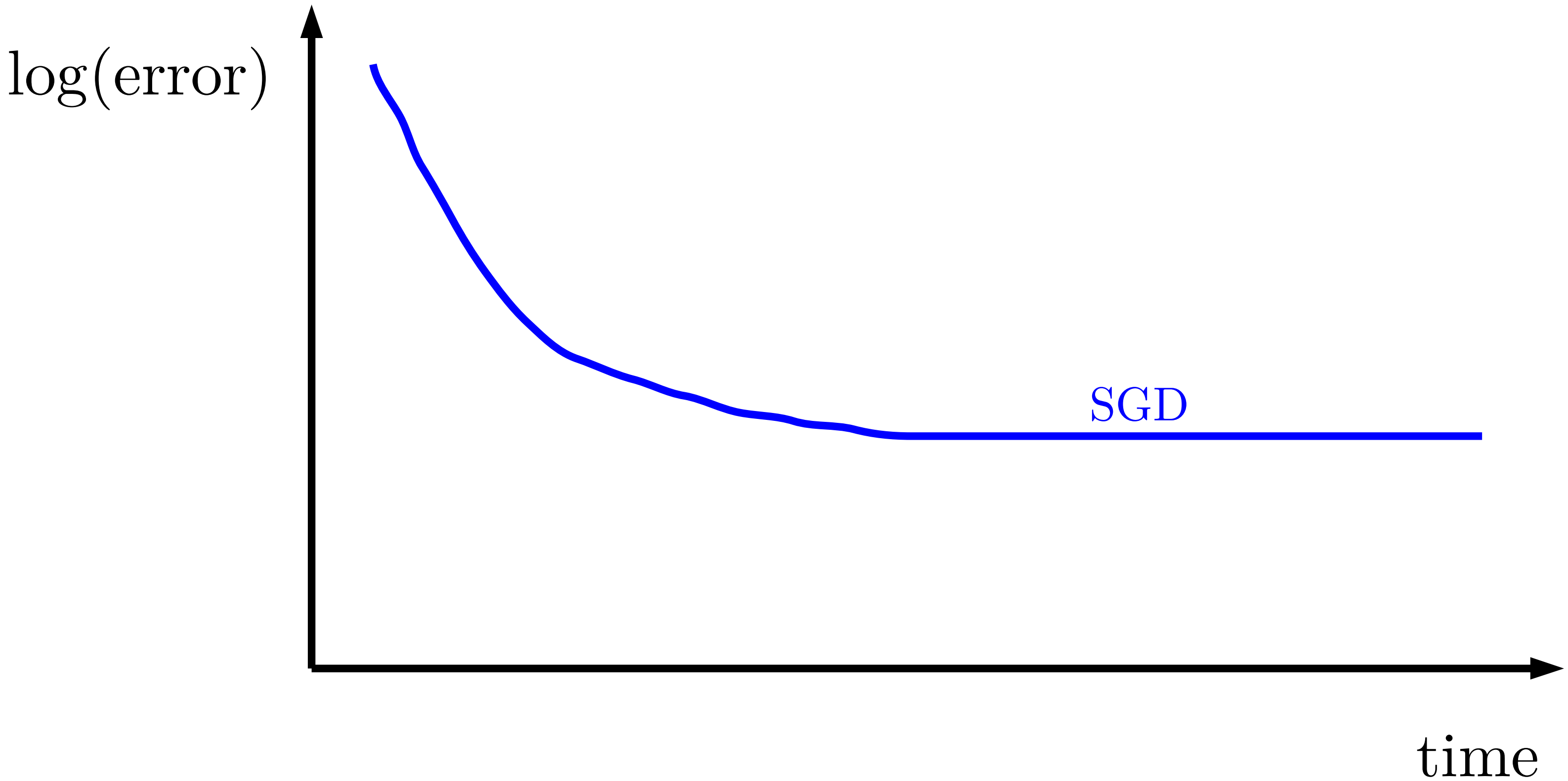
Gradient descent

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

What happens if ϵ is small?

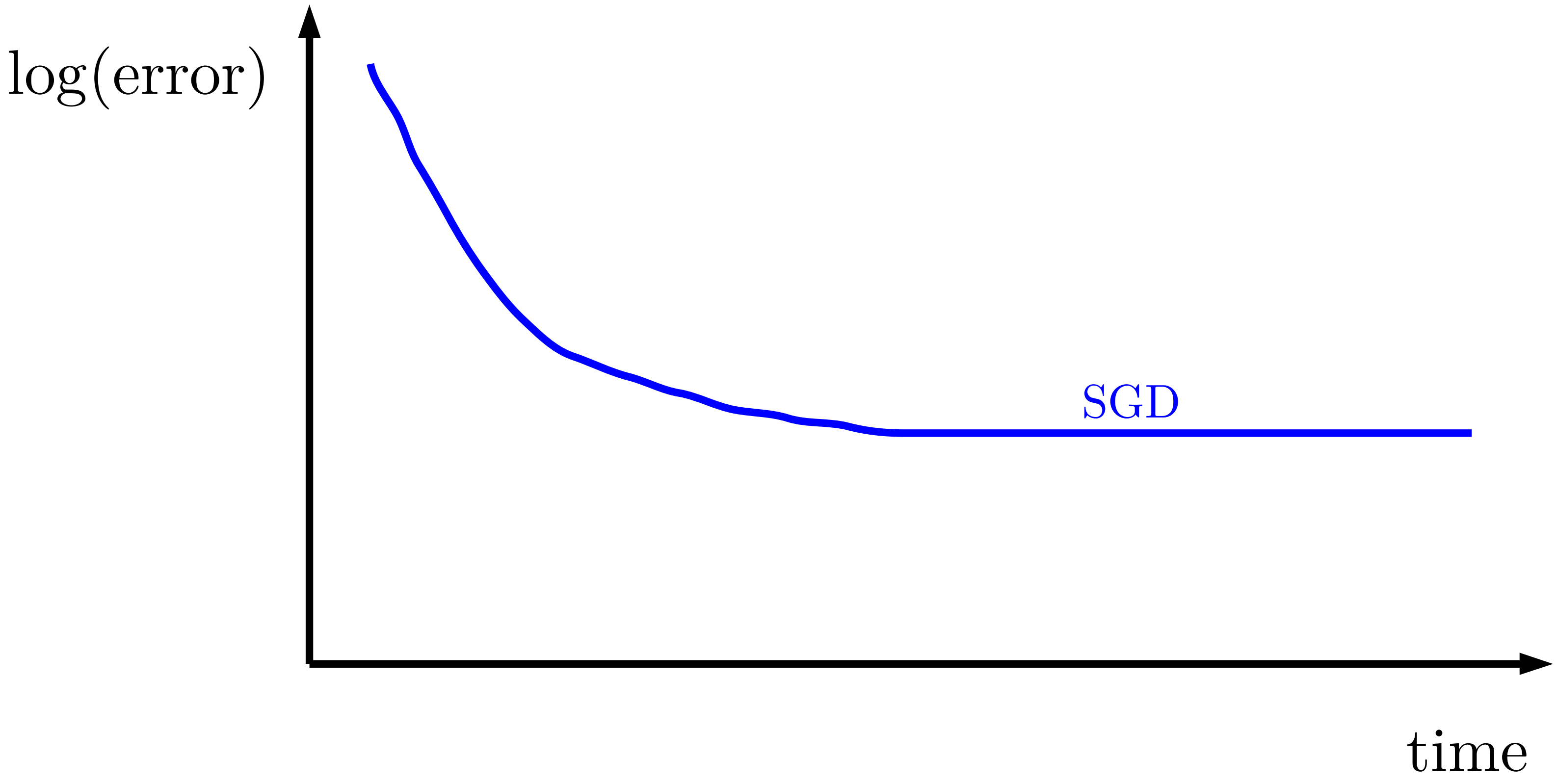
What happens if n is big?

Comparison SGD vs GD



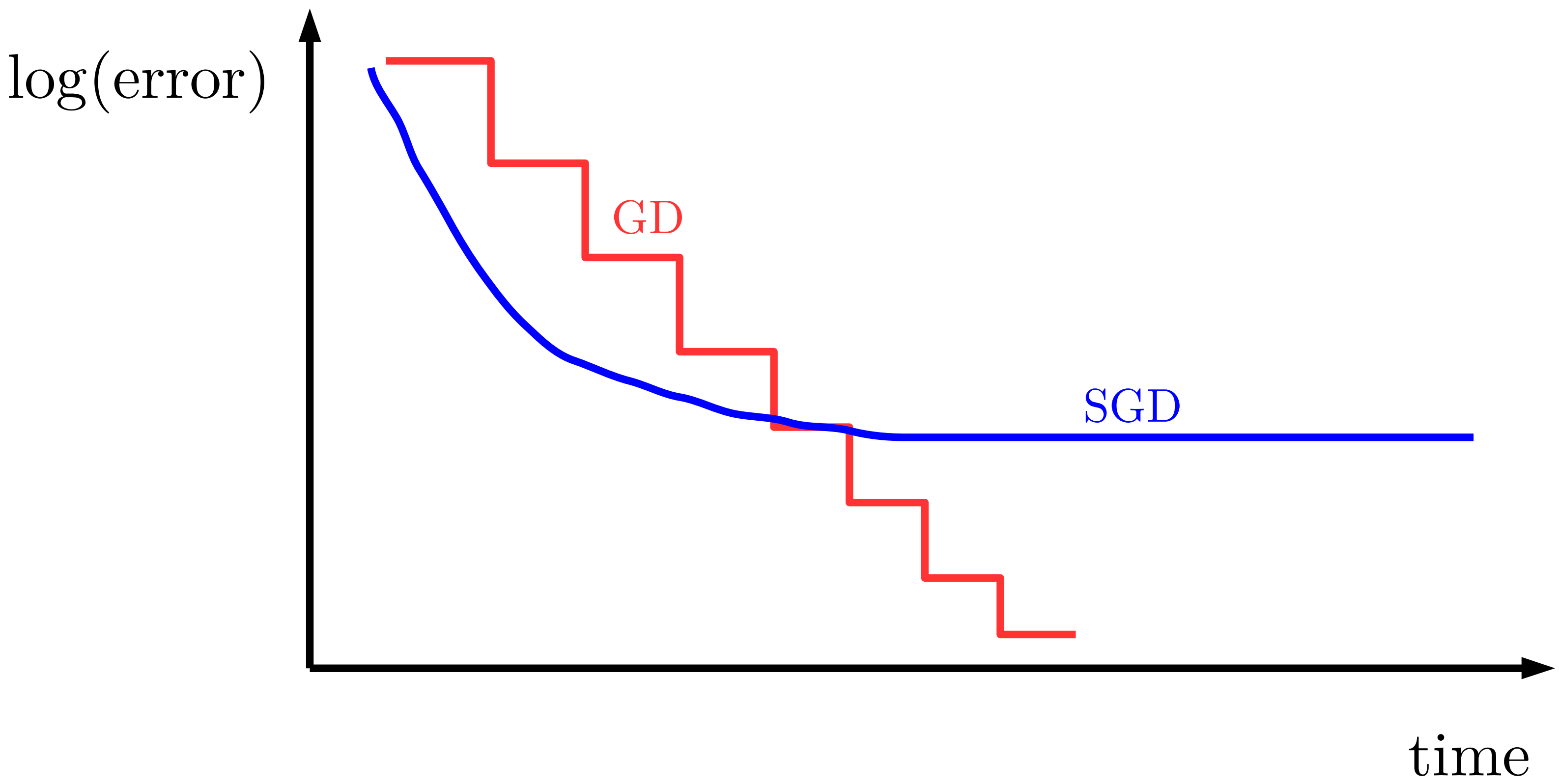
M. Schmidt, N. Le Roux, F. Bach (2016)
Mathematical Programming
Minimizing Finite Sums with the Stochastic Average Gradient.

Comparison SGD vs GD



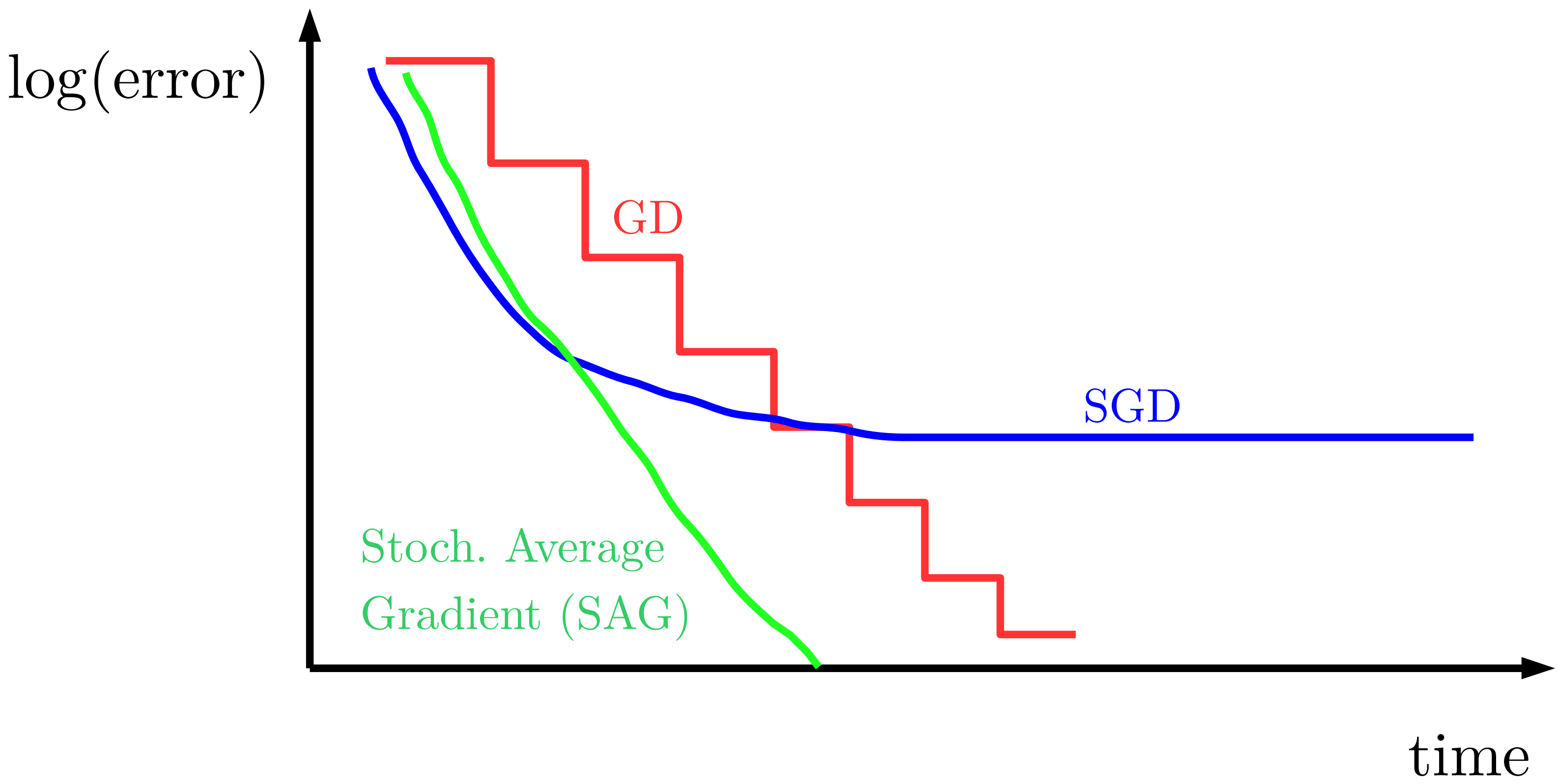
M. Schmidt, N. Le Roux, F. Bach (2016)
Mathematical Programming
Minimizing Finite Sums with the Stochastic Average Gradient.

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.

Practical SGD for Sparse Data

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?

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?

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

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 $O(s)$?

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

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

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 $O(s)$?

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

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

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

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

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

Coding time!

Appendix

Proof SGDA Part I:

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

Taking expectation with respect to j

Unbiased estimator

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

Convexity

Bounded Stoch grad

Taking total expectation and re-arranging

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

Summing up for *1* to T

$$
\begin{aligned}
\textstyle\sum_{t=1}^T (\mathbb{E}\left[f(w^t)\right] - f(w^*)) &\leq \frac{1}{2\alpha_1}||w^1 - w^*||_2^2 + \frac{1}{2}\sum_{t=1}^{T-1}\left(\frac{1}{\alpha_{t+1}} - \frac{1}{\alpha_t}\right)\mathbb{E}\left[||w^t - w^*||_2^2\right] \\
&\quad - \frac{1}{2\alpha_{T+1}}\mathbb{E}\left[||w^{T+1} - w^*||_2^2\right] + \frac{B^2}{2}\sum_{t=1}^{T}\alpha_t
\end{aligned}
$$

Proof Part II:

$$
\begin{aligned}
\sum_{t=1}^{T}\left(\mathbb{E}\left[f(w^t)\right] - f(w^*)\right) \quad \leq \quad & \frac{1}{2\alpha_1}||w^1 - w^*||_2^2 + \frac{1}{2}\sum_{t=1}^{T-1}\left(\frac{1}{\alpha_{t+1}} - \frac{1}{\alpha_t}\right)\mathbb{E}\left[||w^t - w^*||_2^2\right] \\
& - \frac{1}{2\alpha_{T+1}}\mathbb{E}\left[||w^{T+1} - w^*||_2^2\right] + \frac{B^2}{2}\sum_{t=1}^{T}\alpha_t
\end{aligned}
$$

$||w||_2^2 \ \leq \ r^2$

$\alpha_{t+1} \ \leq \ \alpha_t$

$$
\begin{aligned}
\leq \quad & \frac{2r^2}{\alpha_1} + 2r^2\sum_{t=1}^{T-1}\left(\frac{1}{\alpha_{t+1}} - \frac{1}{\alpha_t}\right) + \frac{B^2}{2}\sum_{t=1}^{T}\alpha_t \\
= \quad & \frac{2r^2}{\alpha_T} + \frac{B^2}{2}\sum_{t=1}^{T}\alpha_t
\end{aligned}
$$

Finally let $\overline{w}^T = \frac{1}{T}\sum_{t=1}^{T} w^t$ and dividing by T, using $\alpha_t = \frac{\alpha_0}{\sqrt{t}}$

$$
\begin{aligned}
\mathbb{E}[f(\bar{w}_T)] - f(w^*)) \quad \leq \quad & \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[f(w_t)] - f(w^*)) \ \leq \ \frac{r^2\sqrt{T}}{T\alpha_0} + \frac{B^2}{2T}\sum_{t=1}^{T}\frac{\alpha_0}{\sqrt{t}} \\
\leq \quad & \frac{1}{\sqrt{T}}\left(\frac{2r^2}{\alpha_0} + \alpha_0 B^2\right)
\end{aligned}
$$

Minimizing in α_0 gives $\alpha_0 = \sqrt{2}r/B$ and thus

$$
\mathbb{E}[f(\bar{w}_T)] - f(w^*)) \quad \leq \quad \frac{1}{\sqrt{T}}\left(\sqrt{2}rB + \sqrt{2}rB\right) \leq \frac{3rB}{\sqrt{T}}
$$

SGD with averaging for non-smooth and strongly convex functions

Complexity for strongly convex

Theorem (Shrinking stepsize)

If $f(w)$ is λ–strongly convex, $\overline{w}^T = \dfrac{2}{T(T+1)} \displaystyle\sum_{t=0}^{T-1} t w^t$

and $\alpha_t = \dfrac{2}{\lambda(t+1)}$ then SGD1.2 satisfies

$$\mathbb{E}[f(\overline{w}^T)] - f(w^*) \leq \frac{2B^2}{\lambda(T+1)}$$

SGD 1.2 for Strongly Convex

Set $w^0 = 0$, $\alpha_t = \frac{2}{\lambda(t+1)}$,

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

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

$w^{t+1} = \mathrm{proj}_D\left(w^t - \alpha_t \nabla f_j(w^t)\right)$

Output $\overline{w}^T$

Complexity for strongly convex

Theorem (Shrinking stepsize)

If $f(w)$ is λ–strongly convex, $\overline{w}^T = \dfrac{2}{T(T+1)} \displaystyle\sum_{t=0}^{T-1} t w^t$

and $\alpha_t = \dfrac{2}{\lambda(t+1)}$ then SGD1.2 satisfies

$$\mathbb{E}[f(\overline{w}^T)] - f(w^*) \leq \frac{2B^2}{\lambda(T+1)}$$

Faster Sublinear convergence

SGD 1.2 for Strongly Convex

Set $w^0 = 0$, $\alpha_t = \frac{2}{\lambda(t+1)}$,

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

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

$w^{t+1} = \mathrm{proj}_D\left(w^t - \alpha_t \nabla f_j(w^t)\right)$

Output $\overline{w}^T$

SGD for non-smooth functions

SGD Theory for non-smooth

Assumptions

- $f(w)$ is convex
- Subgradients bounded $\mathbb{E}_j||\nabla f_j(w^t)||_2 \leq B$
- There exists $r \in \mathbb{R}_+$ such that $w^* \in D := \{w \ : \ ||w|| \leq r\}$

SGD 1.1 theoretical

Set $w^1 = 0$, $\alpha_t \in \mathbb{R}_+$, $\alpha_t \underset{t\to\infty}{\to} 0$

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

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

$w^{t+1} = \text{proj}_D \left(w^t - \alpha_t \nabla f_j(w^t)\right)$

Output w^T

Convergence for Convex

Theorem for non-smooth (Shrinking stepsize)

If $f(w)$ is convex,

and $\alpha_t = \dfrac{\alpha_0}{\sqrt{t}}$ then SGD1.1 satisfies

$$\mathbb{E}[f(w^T)] - f(w^*) = O\left(\frac{1}{\sqrt{T}}\right)$$

Ohad Shamir and Tong Zhang (2013)
International Conference on Machine Learning
Stochastic Gradient Descent for Non-smooth Optimization: Convergence Results and Optimal Averaging Schemes.

Convergence for Convex

Theorem for non-smooth (Shrinking stepsize)

If $f(w)$ is convex,

and $\alpha_t = \dfrac{\alpha_0}{\sqrt{t}}$ then SGD1.1 satisfies

$$\mathbb{E}[f(w^T)] - f(w^*) = O\left(\frac{1}{\sqrt{T}}\right)$$

Sublinear convergence

Ohad Shamir and Tong Zhang (2013)
International Conference on Machine Learning
Stochastic Gradient Descent for Non-smooth Optimization: Convergence Results and Optimal Averaging Schemes.

Complexity for Strong. Convex

Theorem for non-smooth (Shrinking stepsize)

If $f(w)$ is λ–strongly convex,

and $\alpha_t = \dfrac{\alpha_0}{\lambda t}$ then SGD1.1 satisfies

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

Ohad Shamir and Tong Zhang (2013)
International Conference on Machine Learning
Stochastic Gradient Descent for Non-smooth Optimization: Convergence Results and Optimal Averaging Schemes.

Complexity for Strong. Convex

Theorem for non-smooth (Shrinking stepsize)

If $f(w)$ is λ–strongly convex,

and $\alpha_t = \dfrac{\alpha_0}{\lambda t}$ then SGD1.1 satisfies

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

Faster Sublinear convergence

Ohad Shamir and Tong Zhang (2013)
International Conference on Machine Learning
Stochastic Gradient Descent for Non-smooth Optimization: Convergence Results and Optimal Averaging Schemes.