Stochastic Gradient Methods

Master 2 Data Science, Univ. Paris Saclay

Robert M. Gower

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
- Quasi-Newton
- Conjugate Gradients

Two parts

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

→

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

Stochastic Gradient Descent

SGD 0.0 Constant stepsize

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

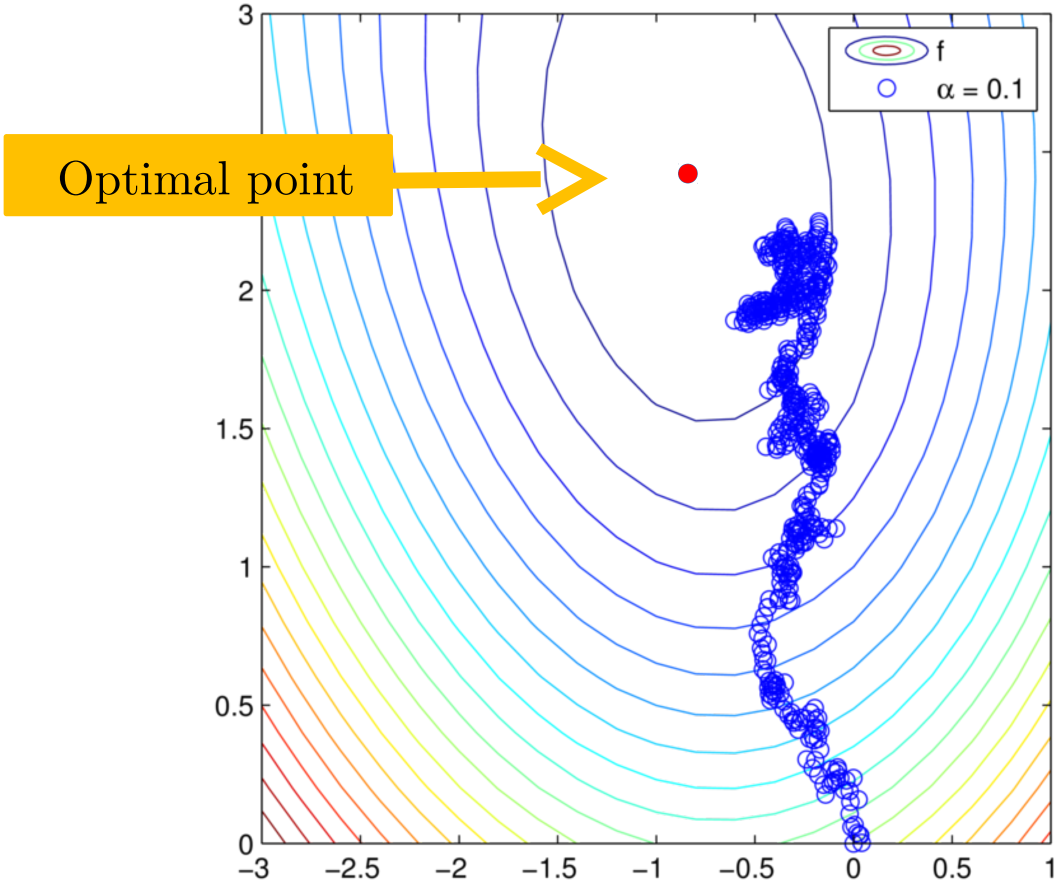
for $t = 0, 1, 2, \ldots, T-1$

  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

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

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

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

Often the same as the regularization parameter

Strong Convexity

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

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

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

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

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

EXE: **Using that**

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

Show that

Strong convexity parameter!

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

Expected Bounded Stochastic Gradients

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

Complexity / Convergence

Theorem

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

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

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

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

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

Strong conv.

Bounded
Stoch grad

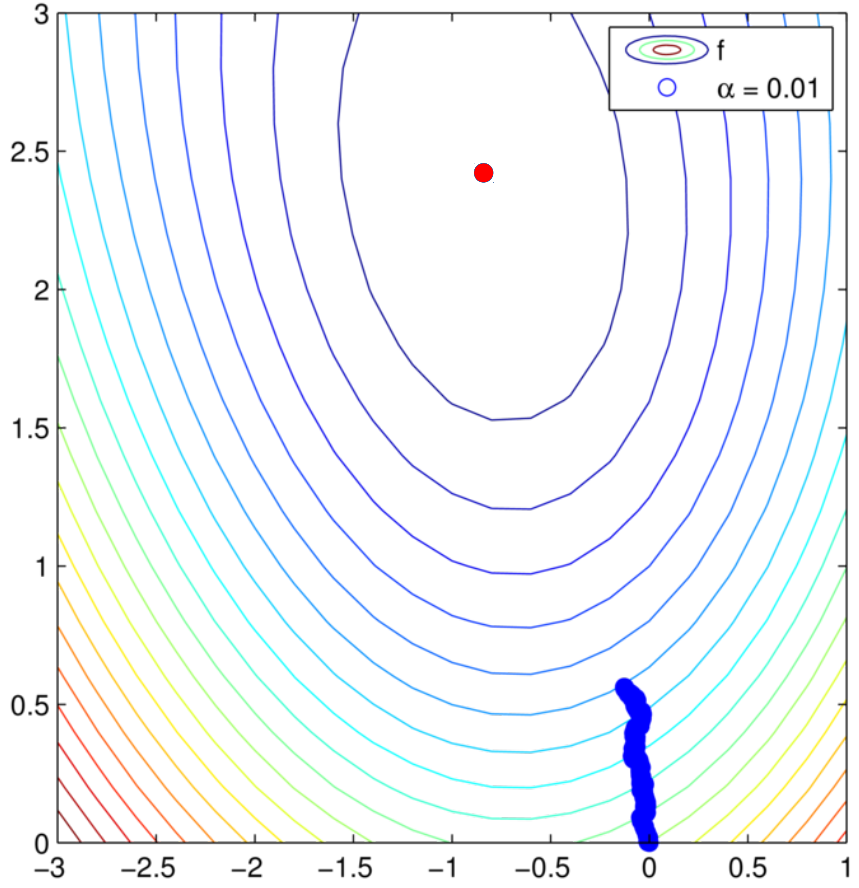
Taking total expectation

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

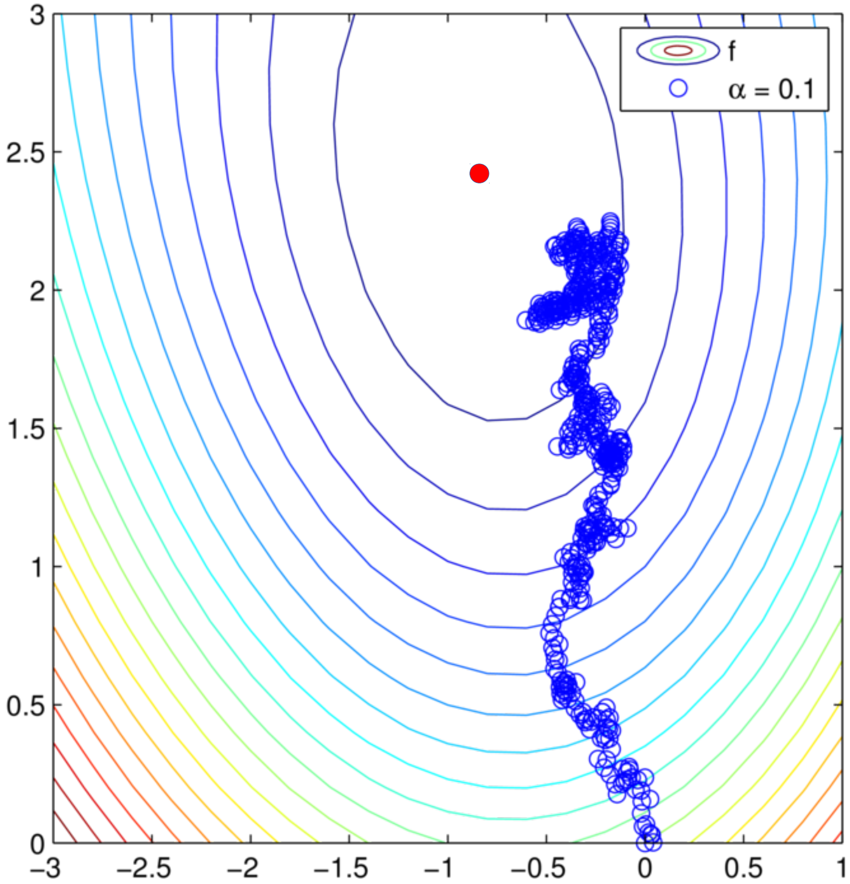
Using the geometric series sum $\displaystyle\sum_{i=0}^{t}(1-\alpha\lambda)^i = \frac{1-(1-\alpha\mu)^{t+1}}{\alpha\lambda} \leq \frac{1}{\alpha\lambda}$

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

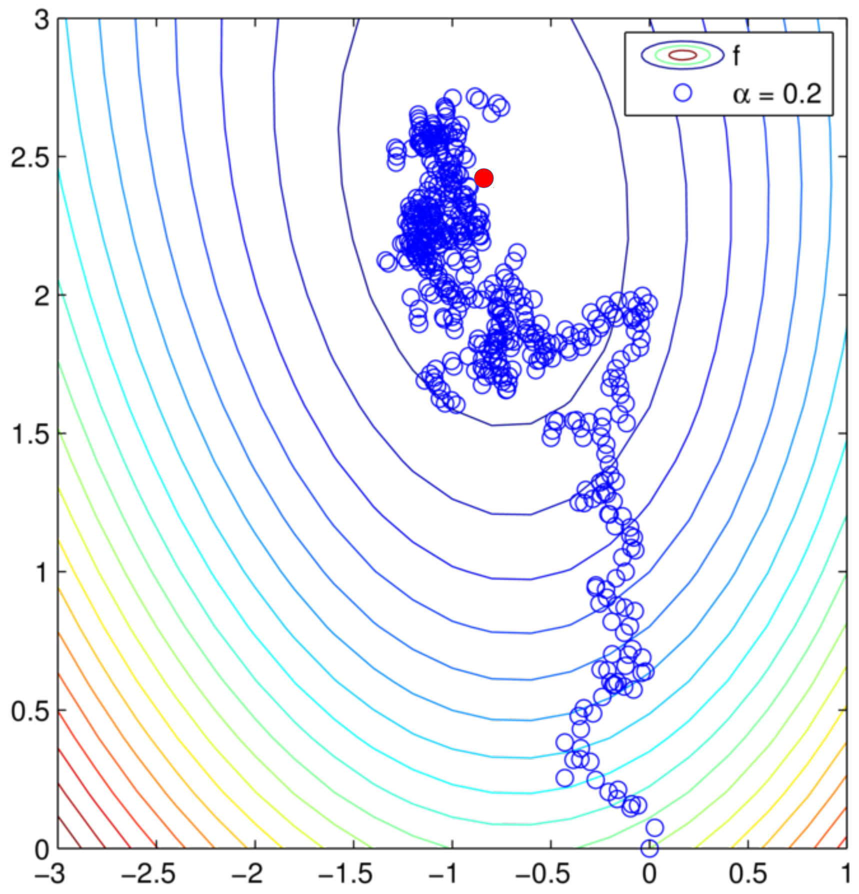
Stochastic Gradient Descent
α =0.01



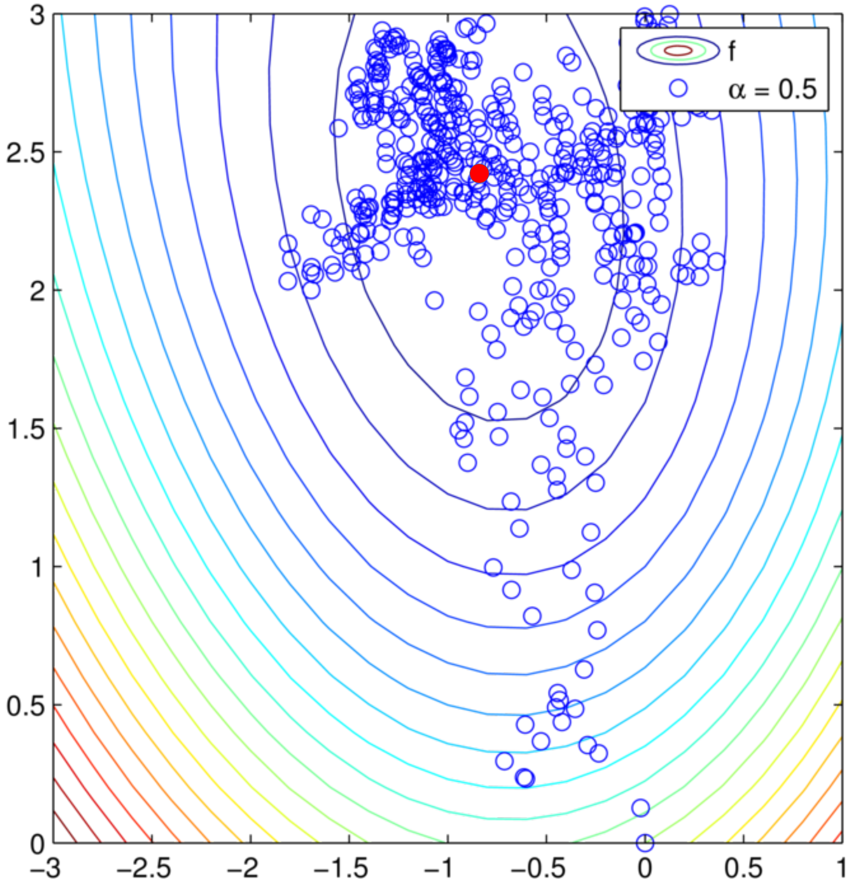
Stochastic Gradient Descent
α =0.1



Stochastic Gradient Descent
$\alpha = 0.2$



Stochastic Gradient Descent $\alpha = 0.5$



SGD shrinking stepsize

SGD 1.0: Descreasing stepsize

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

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

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

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

Output w^T

Shrinking Stepsize

why take root square

SGD shrinking stepsize

SGD 1.0: Descreasing stepsize

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

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

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

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

Output w^T

Shrinking Stepsize

Shrinking Stepsize

SGD shrinking stepsize

SGD 1.0: Descreasing stepsize

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

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

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

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

Output w^T

Shrinking Stepsize

Shrinking Stepsize

How should we sample j ?

Why is $\alpha_t \sim \dfrac{1}{\sqrt{t}}$?

Does this converge?

SGD Theoretical Properties

Convergence for Convex

- *f(w)* is convex
- Subgradients bounded

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

Convergence for Strongly Convex

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

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

Complexity for Convex

Theorem for SGD 1.1 (Shrinking stepsize)

Let $D = \{x : ||x|| \leq r\}$ and $r \in \mathbb{R}_+$

such that $||w^*||_2 \leq r$. If $\alpha_t = \dfrac{\alpha}{\sqrt{t+1}}$ for $\alpha > 0$ then

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

SGD 1.1 for Convex

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

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

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

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

Output w^T

Complexity for Convex

Theorem for SGD 1.1 (Shrinking stepsize)

Let $D = \{x : ||x|| \leq r\}$ and $r \in \mathbb{R}_+$

such that $||w^*||_2 \leq r$. If $\alpha_t = \dfrac{\alpha}{\sqrt{t+1}}$ for $\alpha > 0$ then

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

Sublinear convergence

SGD 1.1 for Convex

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

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

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

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

Output w^T

Complexity for Strong. Convex

Theorem (Shrinking stepsize)

If $f(w)$ is λ–strongly convex,

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

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

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 (Shrinking stepsize)

If $f(w)$ is λ–strongly convex,

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

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

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.

Comparison GD and SGD for strongly convex

Approximate solution

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

SGD with averaging

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

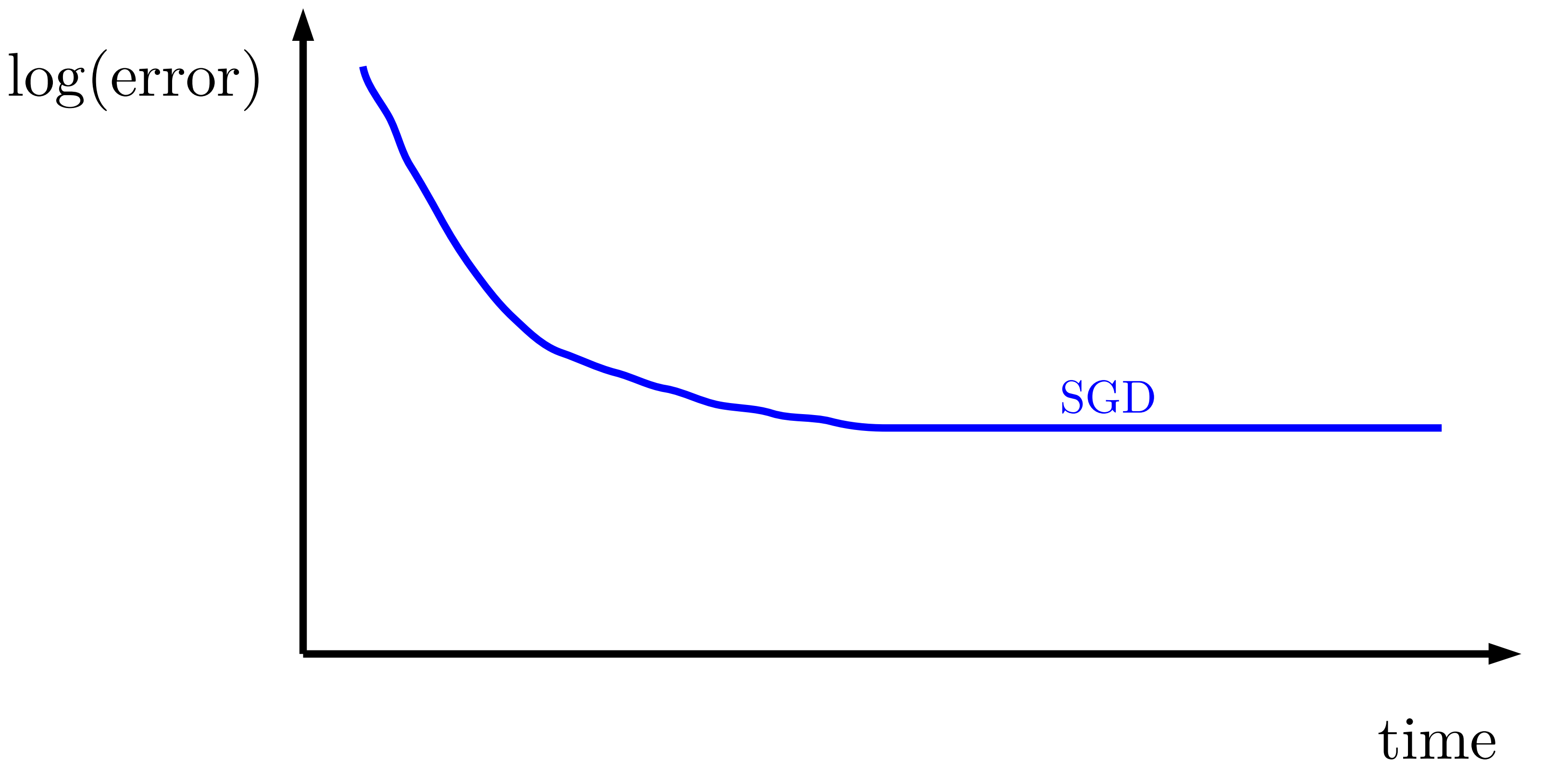
Gradient descent

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

What happens if ϵ is small?

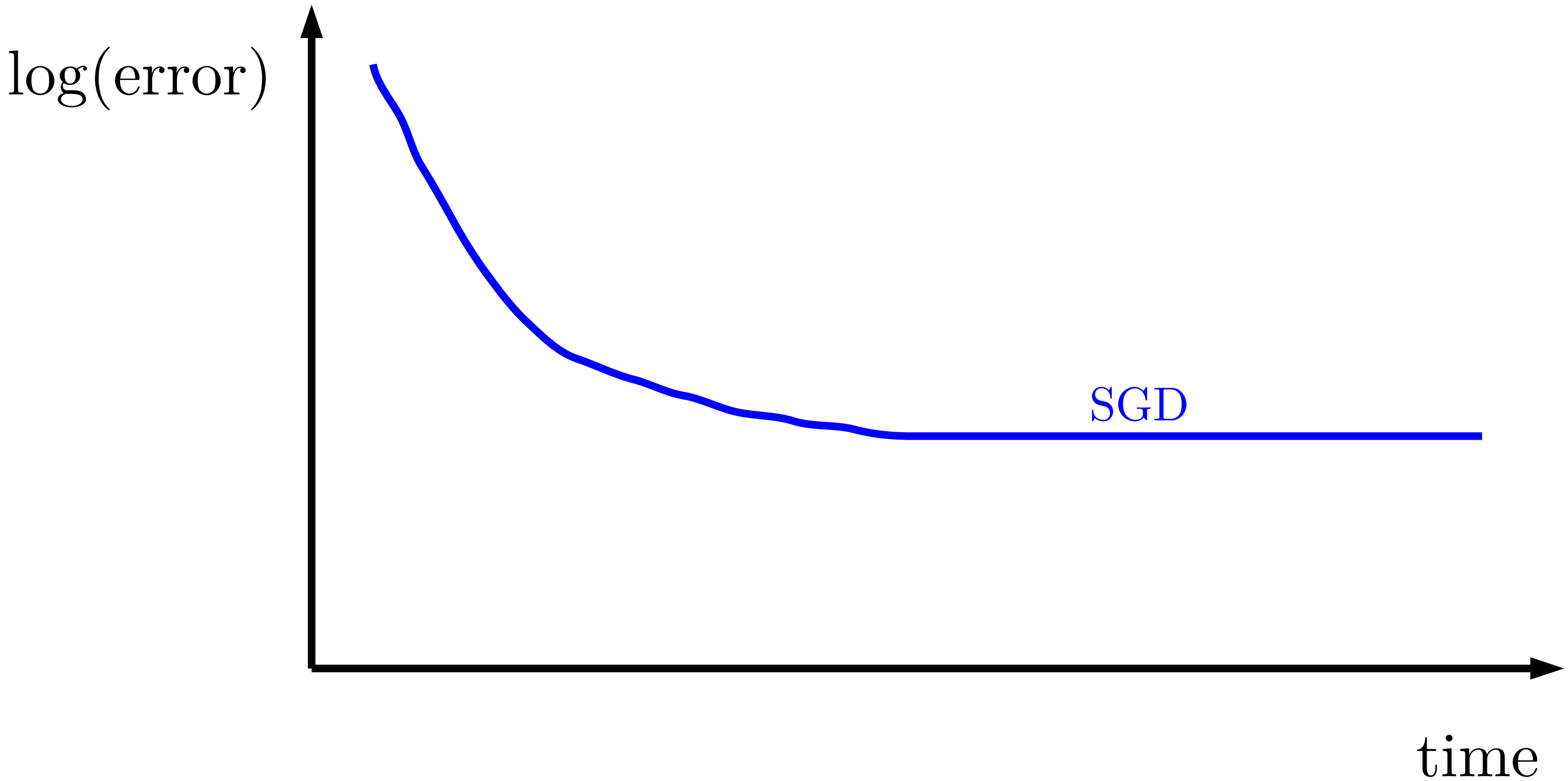
What happens if n is big?

Comparison SGD vs GD



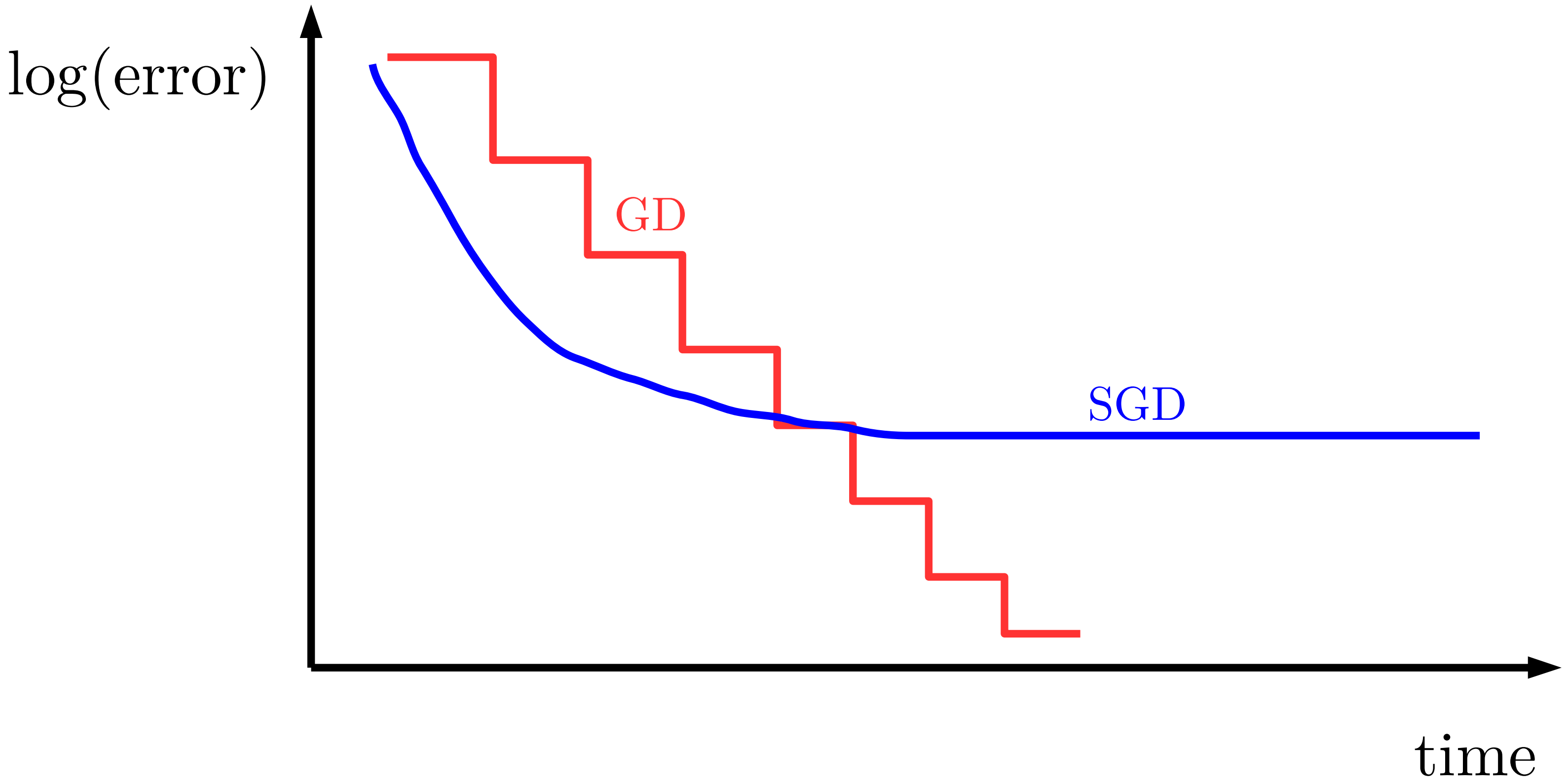
M. Schmidt, N. Le Roux, F. Bach (2016)
Mathematical Programming
Minimizing Finite Sums with the Stochastic Average Gradient.

Comparison SGD vs GD



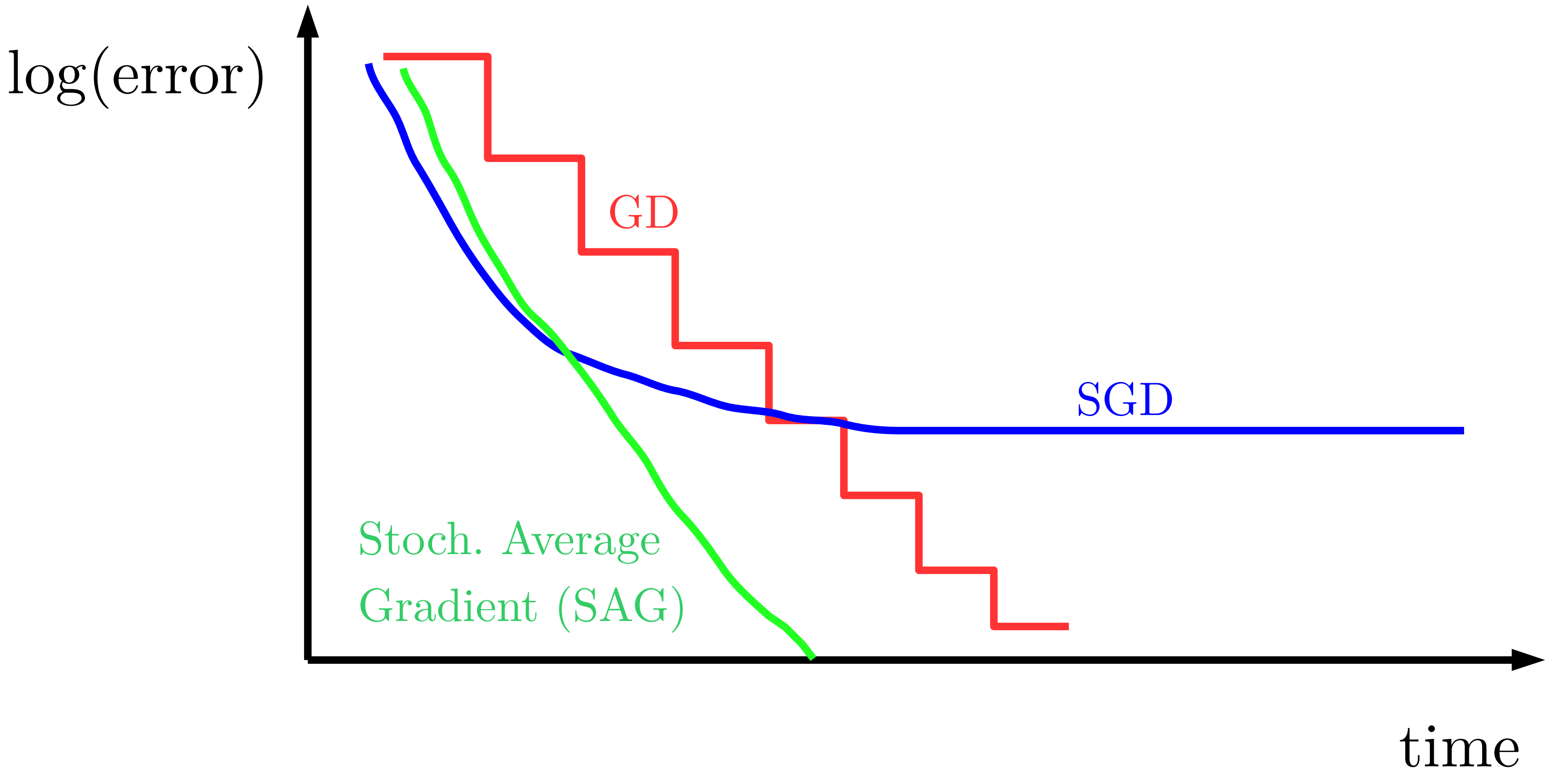
M. Schmidt, N. Le Roux, F. Bach (2016)
Mathematical Programming
Minimizing Finite Sums with the Stochastic Average Gradient.

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.

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!

Complexity for Convex SGDA

Theorem for SGD 1.1 (Shrinking stepsize)

Let $\overline{w}^T = \frac{1}{T}\sum_{t=0}^{T-1} w^t, D = \{x : ||x|| \leq r\}$ and $r \in \mathbb{R}_+$ such that $||w^*||_2 \leq r$. If $\alpha_t = \frac{2r}{B\sqrt{t+1}}$ then

$$\mathbb{E}[f(\overline{w}^T)] - f(w^*) \leq \frac{3rB}{\sqrt{T+1}}$$

SGD 1.1 for Convex

Set $w^0 = 0,\ \alpha_t = \frac{2r}{B\sqrt{t+1}}$,

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 Convex SGDA

Theorem for SGD 1.1 (Shrinking stepsize)

Let $\overline{w}^T = \frac{1}{T}\sum_{t=0}^{T-1} w^t, D = \{x : ||x|| \leq r\}$ and $r \in \mathbb{R}_+$ such that $||w^*||_2 \leq r$. If $\alpha_t = \frac{2r}{B\sqrt{t+1}}$ then

$$\mathbb{E}[f(\overline{w}^T)] - f(w^*) \leq \frac{3rB}{\sqrt{T+1}}$$

Sublinear convergence

SGD 1.1 for Convex

Set $w^0 = 0$, $\alpha_t = \frac{2r}{B\sqrt{t+1}}$,

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

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

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

Output $\overline{w}^T$

Complexity for Convex SGDA

Theorem (Shrinking stepsize)

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

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 Convex SGDA

Theorem (Shrinking stepsize)

If $f(w)$ is λ–strongly convex, $\overline{w}^T = \frac{2}{T(T+1)}\sum_{t=0}^{T-1} t w^t$ and $\alpha_t = \frac{2}{\lambda(t+1)}$ 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 $j \in \{1, \ldots, n\}$

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

Output $\overline{w}^T$