Optimization for Data Science

Stochastic Variance Reduced Gradient Methods

Pierre Ablin

Solving the Finite Sum Training Problem

Optimization Sum of Terms

A Datum Function

$$f_i(w) := \ell\left(h_w(x^i), y^i\right) + \lambda R(w)$$

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \ell\left(h_w(x^i), y^i\right) + \lambda R(w) &= \frac{1}{n}\sum_{i=1}^{n}\left(\ell\left(h_w(x^i), y^i\right) + \lambda R(w)\right)\\ &= \frac{1}{n}\sum_{i=1}^{n} f_i(w)\end{aligned}$$

Finite Sum Training Problem

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

SGD recap

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

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

Taking expectation with respect to j

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

SGD recap

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

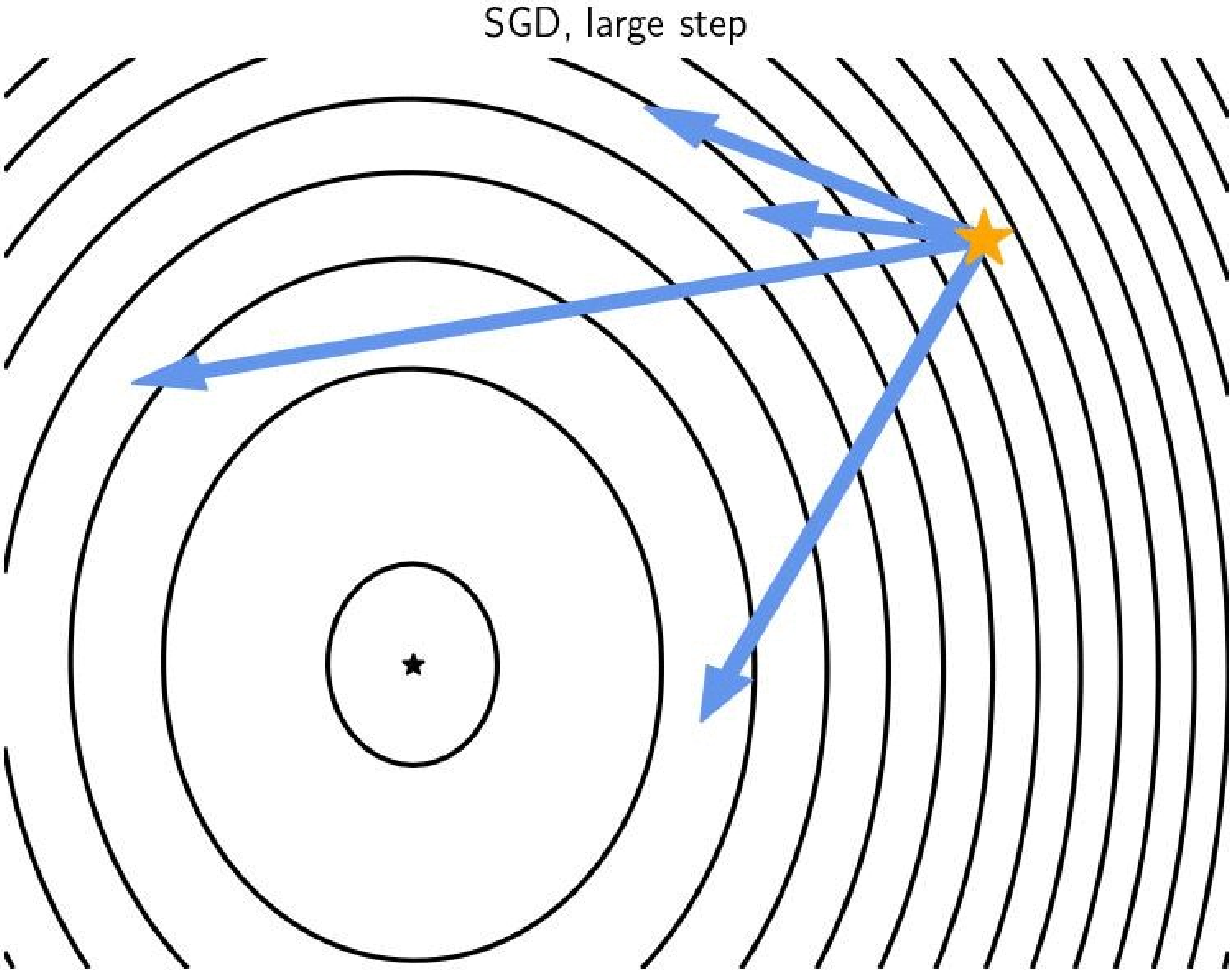
$$
\begin{aligned}
\|w^{t+1} - w^*\|_2^2 &= \|w^t - w^* - \alpha \nabla f_j(w^t)\|_2^2 \\
&= \|w^t - w^*\|_2^2 - 2\alpha \langle \nabla f_j(w^t), w^t - w^* \rangle + \alpha^2 \|\nabla f_j(w^t)\|_2^2.
\end{aligned}
$$

Taking expectation with respect to j

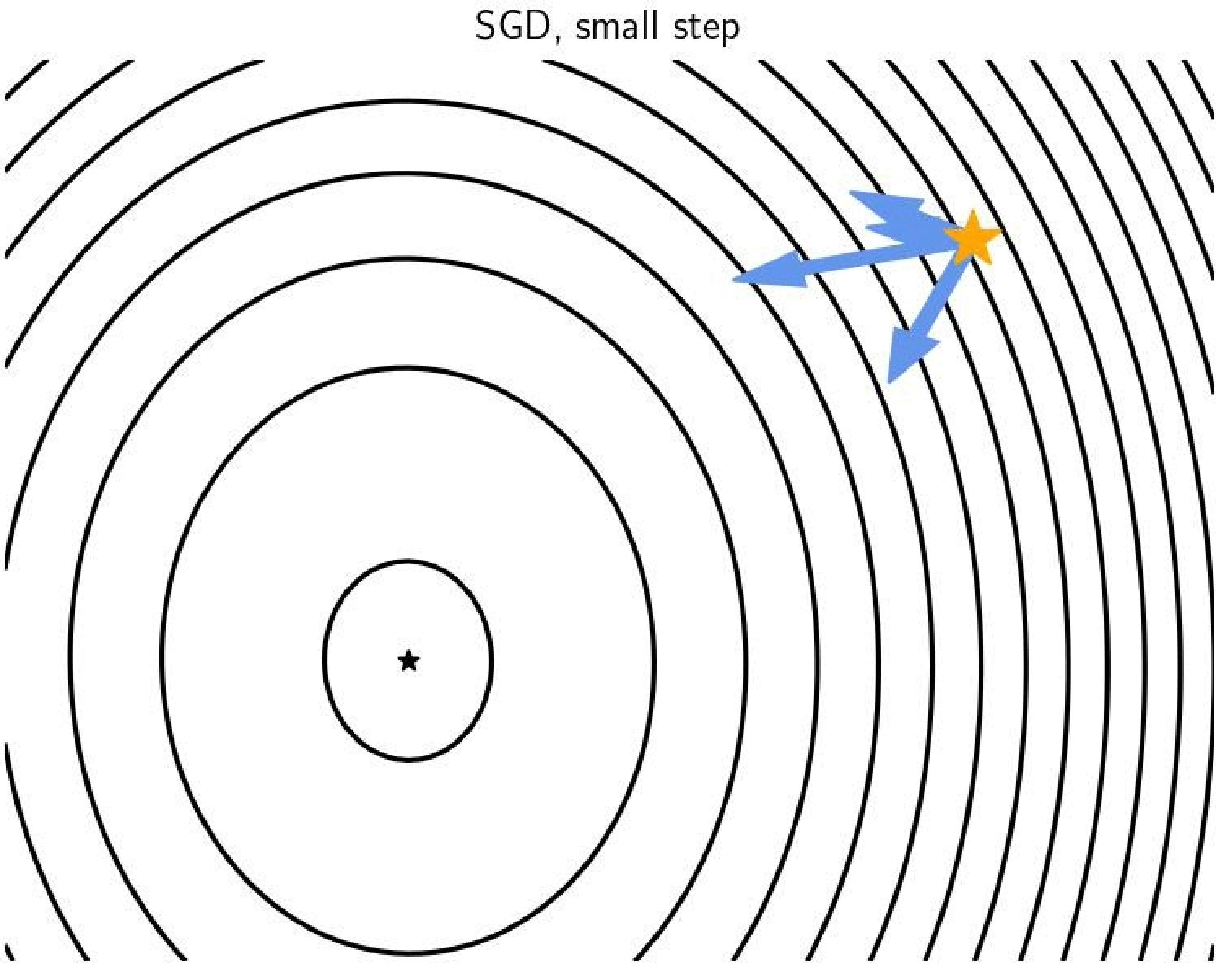
$$
\mathbb{E}_j\left[\|w^{t+1} - w^*\|_2^2\right] = \|w^t - w^*\|_2^2 - 2\alpha \langle \nabla f(w^t), w^t - w^* \rangle + \alpha^2 \mathbb{E}_j\left[\|\nabla f_j(w^t)\|_2^2\right]
$$

The Problem: This variance does not converge

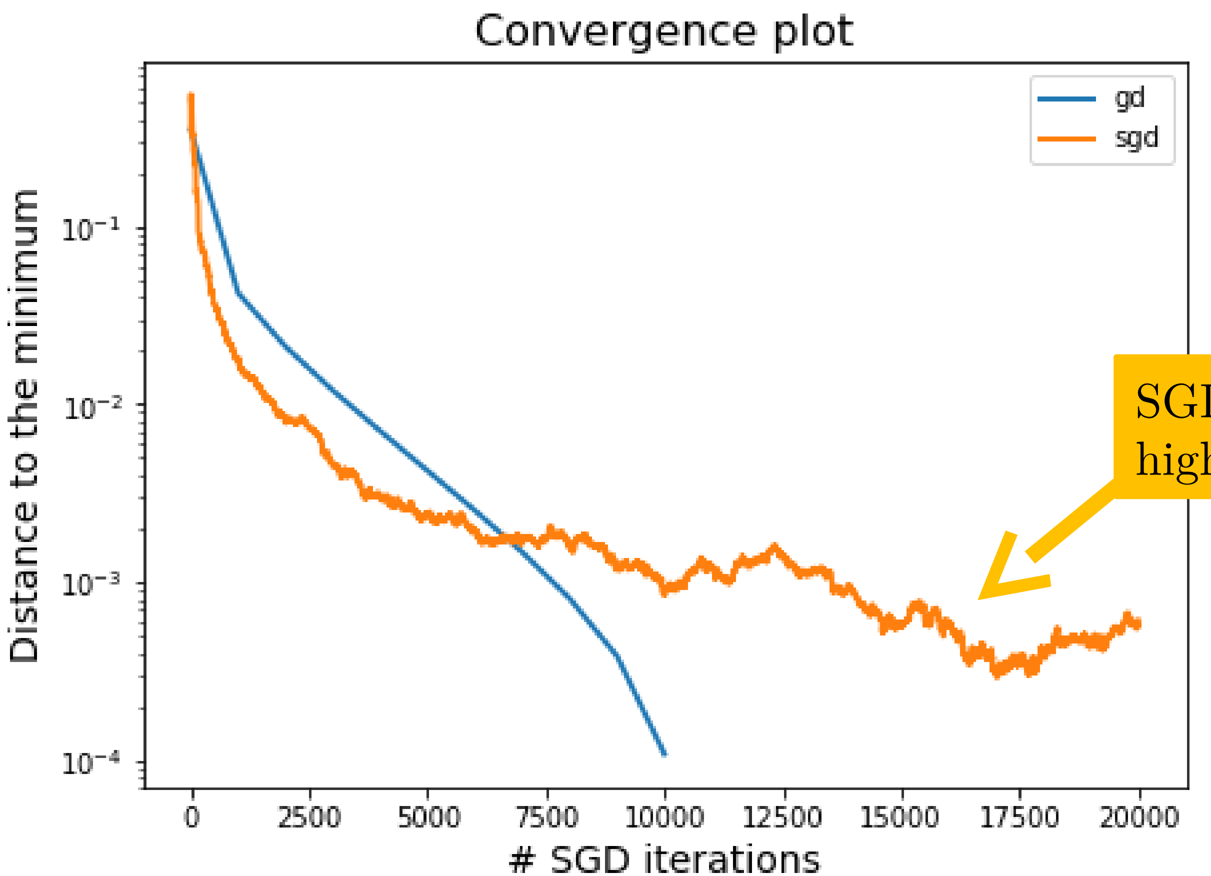
SGD trajectory



SGD trajectory



SGD initially fast, slow later



SGD suffers from high variance

SGD initially fast, slow later

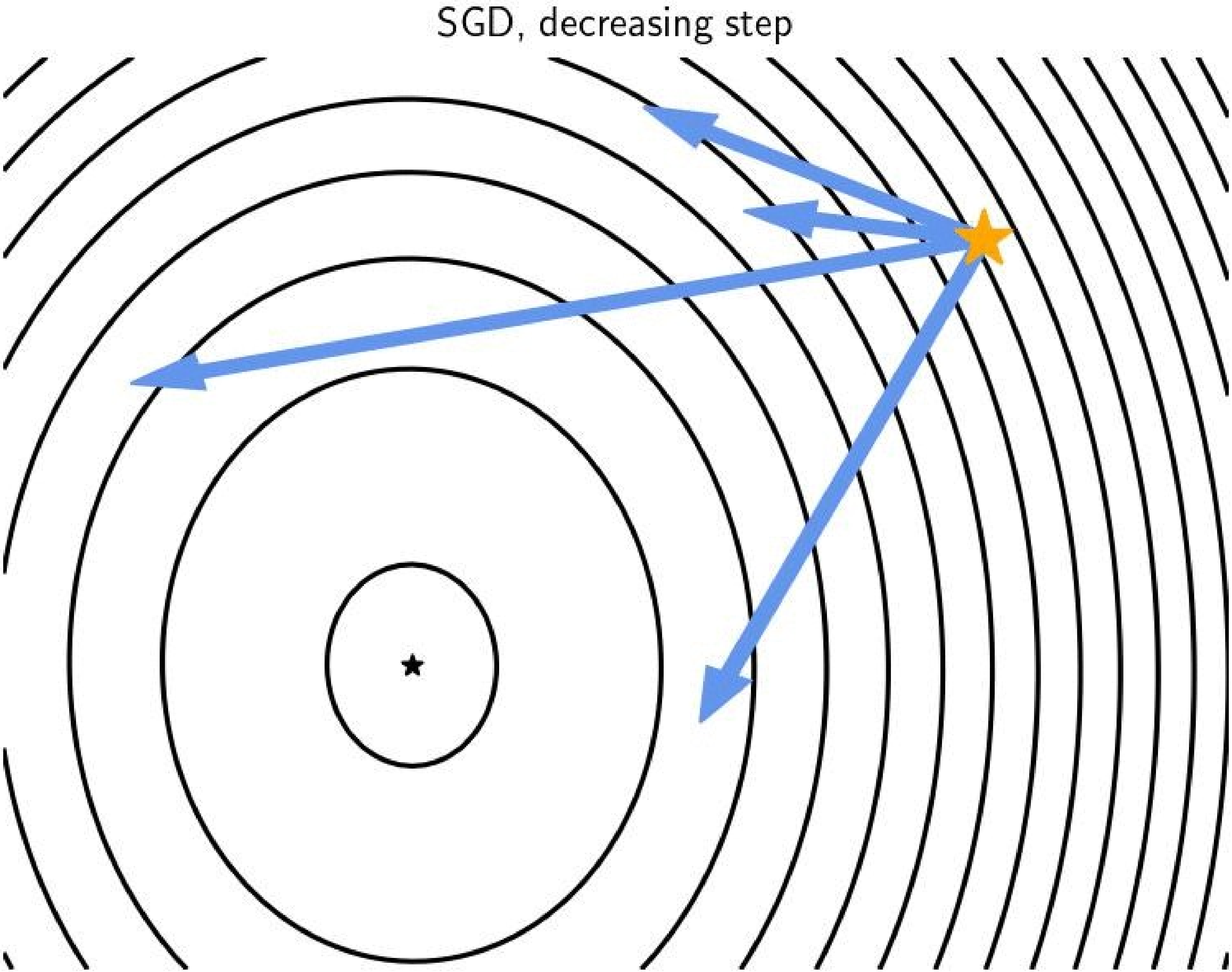
Theorem If f is $\mu-$strongly convex and $\mathbb{E}[\|\nabla f_i(w)\|^2] \leq B^2$

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

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

No convergence !

A cure: shrinking step-sizes



A cure: shrinking step-sizes

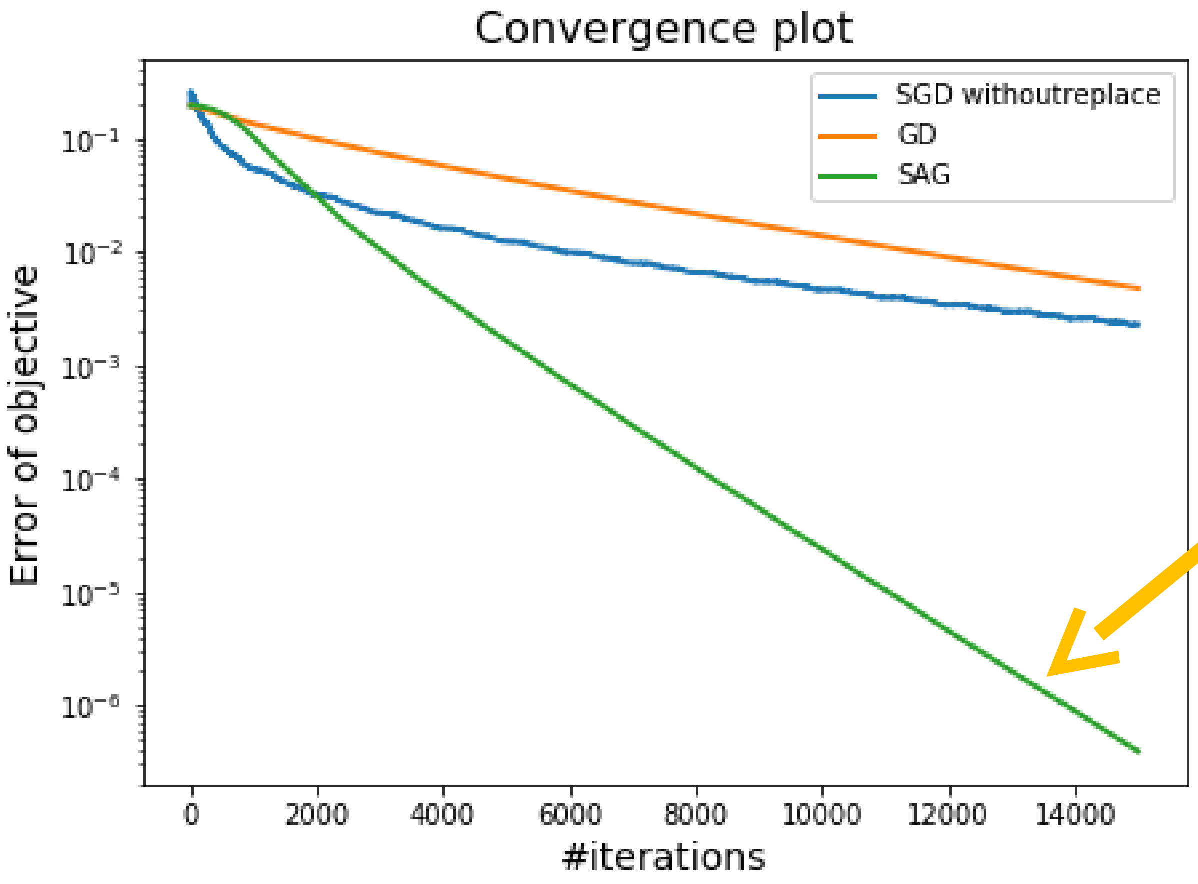
Theorem If f is $\mu-$strongly convex and $\mathbb{E}[\|\nabla f_i(w)\|^2] \leq B^2$

If α_t is such that $\sum_{t=0}^{+\infty} \alpha_t = +\infty, \; \sum_{t=0}^{+\infty} \alpha_t^2 = K \leq +\infty$, then

$$\inf_{t \leq T} \mathbb{E}\left[||w^t - w^*||_2^2\right] \leq \left(\mu \sum_{t=0}^{T} \alpha_t\right)^{-1} \times (\|w^0 - w^*\|^2 + B^2 K)$$

Convergence, but slow...

Can we get best of both?



Today we learn about methods like this one

Stochastic variance reduced methods

Build an Estimate of the Gradient

Instead of using directly $\nabla f_j(w^t) \approx \nabla f(w^t)$

Use $\nabla f_j(w^t)$ to update estimate $g_t \approx \nabla f(w^t)$

Build an Estimate of the Gradient

Instead of using directly $\nabla f_j(w^t) \approx \nabla f(w^t)$
Use $\nabla f_j(w^t)$ to update estimate $g_t \approx \nabla f(w^t)$

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

Build an Estimate of the Gradient

Instead of using directly $\nabla f_j(w^t) \approx \nabla f(w^t)$
Use $\nabla f_j(w^t)$ to update estimate $g_t \approx \nabla f(w^t)$

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

We would like gradient estimate such that:

Similar

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

Converges in *L2*

$$\mathbb{E}||g^t||_2^2 \underset{w^t \to w^*}{\longrightarrow} 0$$

Build an Estimate of the Gradient

Instead of using directly $\nabla f_j(w^t) \approx \nabla f(w^t)$
Use $\nabla f_j(w^t)$ to update estimate $g_t \approx \nabla f(w^t)$

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

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

We would like gradient estimate such that:

Similar

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

Converges in ***L2***

$$\mathbb{E}||g^t||_2^2 \underset{w^t \to w^*}{\longrightarrow} 0$$

Build an Estimate of the Gradient

Instead of using directly $\nabla f_j(w^t) \approx \nabla f(w^t)$
Use $\nabla f_j(w^t)$ to update estimate $g_t \approx \nabla f(w^t)$

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

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

We would like gradient estimate such that:

Similar

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

Solves problem of
$\alpha_t \underset{t \to \infty}{\longrightarrow} 0$

Converges in *L2*

$$\mathbb{E}||g^t||_2^2 \underset{w^t \to w^*}{\longrightarrow} 0$$

Controlled Stochastic Reformulation

Covariate functions:

$$z_i : w \mapsto z_i(w) \in \mathbb{R}, \quad \text{for } i = 1, \ldots, n$$

$$\frac{1}{n}\sum_{i=1}^{n} f_i(w) = \mathbb{E}[f_i(w)] = \mathbb{E}[f_i(w)] - \mathbb{E}[z_i(w)] + \mathbb{E}[z_i(w)]$$

$i \sim \frac{1}{n}$

Controlled Stochastic Reformulation

Covariate functions:

$$z_i : w \mapsto z_i(w) \in \mathbb{R}, \quad \text{for } i = 1, \ldots, n$$

Cancel out

$$\frac{1}{n}\sum_{i=1}^{n} f_i(w) = \mathbb{E}[f_i(w)] = \mathbb{E}[f_i(w)] - \mathbb{E}[z_i(w)] + \mathbb{E}[z_i(w)]$$

$i \sim \frac{1}{n}$

Controlled Stochastic Reformulation

Covariate functions:

$$z_i : w \mapsto z_i(w) \in \mathbb{R}, \quad \text{for } i = 1, \ldots, n$$

Cancel out

$$\frac{1}{n}\sum_{i=1}^{n} f_i(w) = \mathbb{E}[f_i(w)] = \mathbb{E}[f_i(w)] - \mathbb{E}[z_i(w)] + \mathbb{E}[z_i(w)]$$

$$i \sim \frac{1}{n} \qquad = \mathbb{E}\left[f_i(w) - z_i(w) + \mathbb{E}[z_i(w)]\right]$$

Controlled Stochastic Reformulation

Covariate functions:

$$z_i : w \mapsto z_i(w) \in \mathbb{R}, \quad \text{for } i = 1, \ldots, n$$

Cancel out

$$\frac{1}{n}\sum_{i=1}^{n} f_i(w) = \mathbb{E}[f_i(w)] = \mathbb{E}[f_i(w)] - \mathbb{E}[z_i(w)] + \mathbb{E}[z_i(w)]$$

$$= \mathbb{E}\big[f_i(w) - z_i(w) + \mathbb{E}[z_i(w)]\big]$$

$i \sim \frac{1}{n}$

Original finite sum problem

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

Controlled Stochastic Reformulation

$$\min_{w \in \mathbb{R}^d} \mathbb{E}\big[f_i(w) - z_i(w) + \mathbb{E}[z_i(w)]\big]$$

Use covariates to **control the variance**

Variance reduction as SGD

$$\min_{w \in \mathbb{R}^d} \mathbb{E}\left[f_i(w) - z_i(w) + \mathbb{E}[z_i(w)]\right]$$

Variance reduction as SGD

$$\min_{w\in\mathbb{R}^d} \mathbb{E}\left[f_i(w) - z_i(w) + \mathbb{E}[z_i(w)]\right]$$

Sample $i \sim \frac{1}{n}$

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

Variance reduction as SGD

$$\min_{w\in\mathbb{R}^d} \mathbb{E}\left[f_i(w) - z_i(w) + \mathbb{E}[z_i(w)]\right]$$

Sample $i \sim \frac{1}{n}$

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

$$g_i(w) := \nabla f_i(w) - \nabla z_i(w) + \mathbb{E}[\nabla z_i(w)]$$

Variance reduction as SGD

$$\min_{w \in \mathbb{R}^d} \mathbb{E}\left[f_i(w) - z_i(w) + \mathbb{E}[z_i(w)]\right]$$

Sample $i \sim \frac{1}{n}$

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

By design we have that $\mathbb{E}[g_i(w^t)] = \nabla f(w^t)$

$$g_i(w) := \nabla f_i(w) - \nabla z_i(w) + \mathbb{E}[\nabla z_i(w)]$$

Variance reduction as SGD

$$\min_{w\in\mathbb{R}^d} \mathbb{E}\left[f_i(w) - z_i(w) + \mathbb{E}[z_i(w)]\right]$$

By design we have that
$\mathbb{E}[g_i(w^t)] = \nabla f(w^t)$

Sample $i \sim \frac{1}{n}$

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

$$g_i(w) := \nabla f_i(w) - \nabla z_i(w) + \mathbb{E}[\nabla z_i(w)]$$

How to choose $z_i(w)$?

Variance reduction as SGD

$$\min_{w\in\mathbb{R}^d} \mathbb{E}\left[f_i(w) - z_i(w) + \mathbb{E}[z_i(w)]\right]$$

Sample $i \sim \frac{1}{n}$

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

By design we have that $\mathbb{E}[g_i(w^t)] = \nabla f(w^t)$

$$g_i(w) := \nabla f_i(w) - \nabla z_i(w) + \mathbb{E}[\nabla z_i(w)]$$

How to choose $z_i(w)$?

We could take $z_i = f_i$, but what's the problem?

Covariates

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

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

Variance Reduced Estimate:

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

Covariates

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

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

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

Variance Reduced Estimate:

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

EXE:

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

Covariates

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

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

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

Variance Reduced Estimate:

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

EXE:

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

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

Covariates

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

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

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

Variance Reduced Estimate:

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

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

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

Larger covariance between x and z is good

Covariates

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

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

Variance Reduced Estimate:

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

$$g_i(w) := \underbrace{\nabla f_i(w)}_{x} - \underbrace{\nabla z_i(w)}_{z} + \underbrace{\mathbb{E}[\nabla z_i(w)]}_{\mathbb{E}[z]}$$

$$\nabla z_i(w) \approx \nabla f_i(w) \Rightarrow \mathrm{cov}(\nabla z_i(w), \nabla f_i(w)) \gg 0$$

Choosing the covariate as a linear approximation

Sample $i \sim \frac{1}{n}$

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

$$g_i := \nabla f_i(w) - \nabla z_i(w) + \mathbb{E}[\nabla z_i(w)]$$

We would like:

Choosing the covariate as a linear approximation

$$\text{Sample } i \sim \frac{1}{n}$$

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

$$g_i := \nabla f_i(w) - \nabla z_i(w) + \mathbb{E}[\nabla z_i(w)]$$

We would like:

$$\nabla z_i(w) \approx \nabla f_i(w)$$

Choosing the covariate as a linear approximation

Sample $i \sim \frac{1}{n}$

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

$$g_i := \nabla f_i(w) - \nabla z_i(w) + \mathbb{E}[\nabla z_i(w)]$$

We would like:

$$\nabla z_i(w) \approx \nabla f_i(w)$$

Linear approximation around $\tilde{w}$

$$z_i(w) = f_i(\tilde{w}) + \langle \nabla f_i(\tilde{w}), w - \tilde{w} \rangle$$

A reference point/ snap shot

SVRG: Stochastic Variance reduced method gradient

Johnson & Zhang, 2013 NIPS

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

Reference point	$\tilde{w} \in \mathbb{R}^d$
Sample	$\nabla f_i(w^t)$, i.i.d sample with prob $\frac{1}{n}$
Grad. estimate	$g_i(w^t) = \nabla f_i(w^t) - \nabla f_i(\tilde{w}) + \nabla f(\tilde{w})$

It's unbiased because:

SVRG: Stochastic Variance reduced method gradient

Johnson & Zhang, 2013 NIPS

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

Reference point	$\tilde{w} \in \mathbb{R}^d$
Sample	$\nabla f_i(w^t)$, i.i.d sample with prob $\frac{1}{n}$
Grad. estimate	$g_i(w^t) = \nabla f_i(w^t) - \nabla f_i(\tilde{w}) + \nabla f(\tilde{w})$

It's unbiased because:

$$\mathbb{E}[g_i(w)] = \mathbb{E}[\nabla f_i(w)] - \mathbb{E}[\nabla f_i(\tilde{w})] + \nabla f(\tilde{w})$$

SVRG: Stochastic Variance reduced method gradient

Johnson & Zhang, 2013 NIPS

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

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

Sample $\nabla f_i(w^t), \quad$ i.i.d sample with prob $\frac{1}{n}$

Grad. estimate $g_i(w^t) = \nabla f_i(w^t) - \nabla f_i(\tilde{w}) + \nabla f(\tilde{w})$

It's unbiased because:

$$\begin{aligned}\mathbb{E}[g_i(w)] &= \mathbb{E}[\nabla f_i(w)] - \mathbb{E}[\nabla f_i(\tilde{w})] + \nabla f(\tilde{w}) \\ &= \nabla f(w) - \nabla f(\tilde{w}) + \nabla f(\tilde{w})\end{aligned}$$

SVRG: Stochastic Variance reduced method gradient

Johnson & Zhang, 2013 NIPS

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

Reference point	$\tilde{w} \in \mathbb{R}^d$
Sample	$\nabla f_i(w^t), \qquad$ i.i.d sample with prob $\frac{1}{n}$
Grad. estimate	$g_i(w^t) = \nabla f_i(w^t) - \nabla f_i(\tilde{w}) + \nabla f(\tilde{w})$

It's unbiased because:

$$\begin{aligned} \mathbb{E}[g_i(w)] &= \mathbb{E}[\nabla f_i(w)] - \mathbb{E}[\nabla f_i(\tilde{w})] + \nabla f(\tilde{w}) \\ &= \nabla f(w) - \cancel{\nabla f(\tilde{w})} + \cancel{\nabla f(\tilde{w})} \end{aligned}$$

SVRG: Variance

Grad. estimate

$$g_{i}(w^t) = \nabla f_{i}(w^t) - \nabla f_{i}(\tilde{w}) + \nabla f(\tilde{w})$$

Question: What is the variance of this estimate? Can you give an upper-bound?

SVRG: Variance

Grad. estimate

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

Question: What is the variance of this estimate? Can you give an upper-bound?

$$\begin{aligned}\mathbb{VAR}(g_i) &= \mathbb{E}[\|\nabla f_i(w) - \nabla f_i(\tilde{w}) - \nabla f(w) + \nabla f_i(\tilde{w})\|^2] \\ &\leq 2\mathbb{E}[\|\nabla f_i(w) - \nabla f_i(\tilde{w})\|^2] + 2\mathbb{E}[\|\nabla f(w) - \nabla f(\tilde{w})\|^2] \\ &\leq 2(L_{\max}^2 + L^2)\|w - \tilde{w}\|^2\end{aligned}$$

SVRG: Stochastic Variance Reduced Gradients

Set $\tilde{w}^0 = 0 = x_0^m$, choose $\gamma > 0, m \in \mathbb{N}$,

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

$\quad x_s^0 = x_{s-1}^m$

$\quad$ for $t = 0, 1, 2, \ldots, m-1$

$\quad\quad$ i.i.d sample $i \sim \frac{1}{n}$

$$g^t = \nabla f_i(x_s^t) - \nabla f_i(\tilde{w}^{s-1}) + \nabla f(\tilde{w}^{s-1})$$

$$x_s^{t+1} = x_s^t - \gamma g^t$$

$\quad \tilde{w}^{s+1} = x_s^m$

Output $\tilde{w}^{T+1}$

Most iterates cost *O(1)*

Tune inner loop size m

Memory methods

Another method to reduce variance

Finite dataset: let's store each gradient.

$$g_1, \ldots, g_n$$

At iteration t, sample $i \sim \frac{1}{n}$, compute $\nabla f_i(w^t)$ and update memory:

$$g_i = \nabla_i f(w^t), \text{ and } g_j \text{ stays the same for } j \neq i$$

Question: can you think of a better estimator of the gradient than g_i ?

Another method to reduce variance

Finite dataset: let's store each gradient.

$$g_1, \ldots, g_n$$

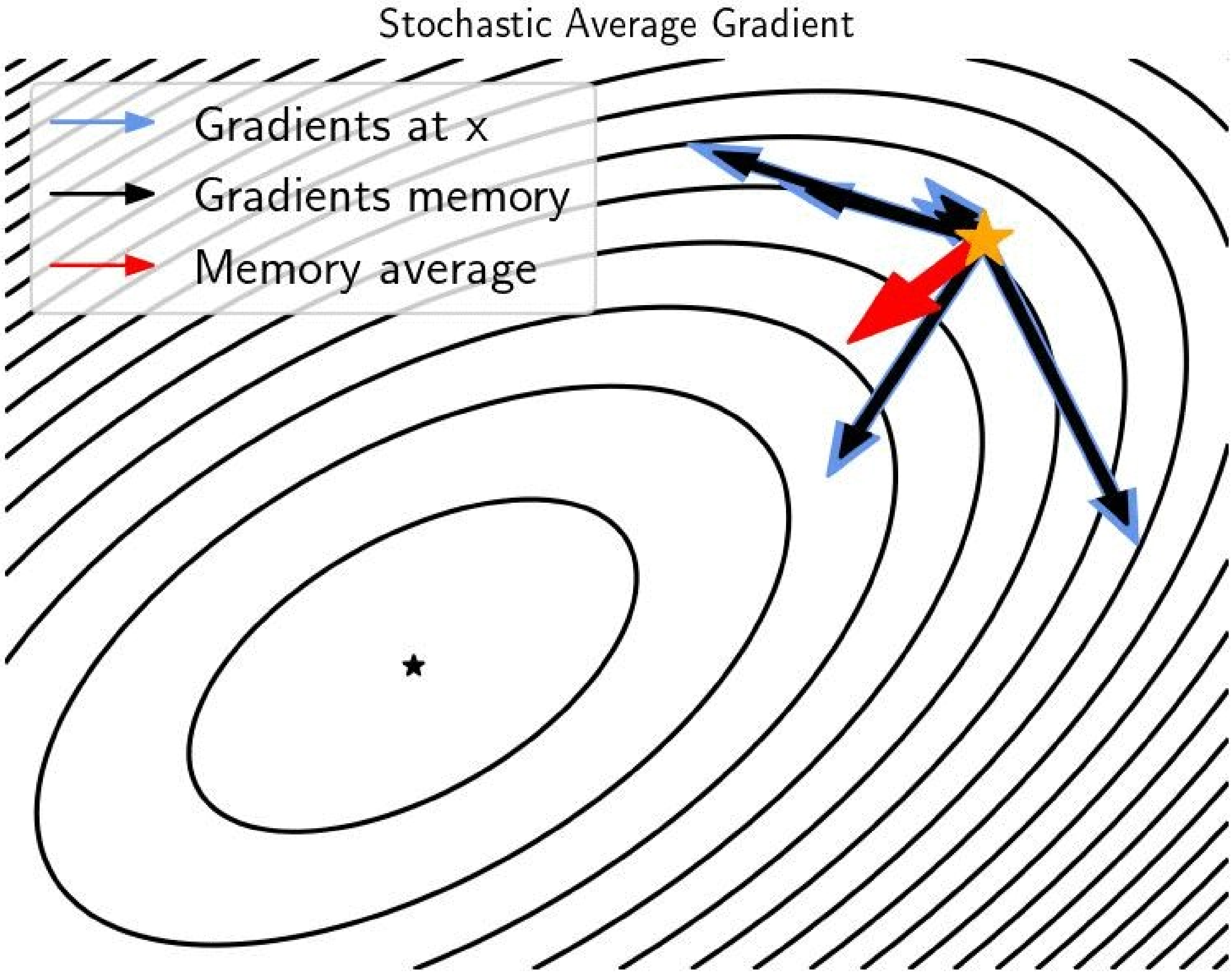
At iteration t, sample $i \sim \frac{1}{n}$, compute $\nabla f_i(w^t)$ and update memory:

$$g_i = \nabla_i f(w^t), \text{ and } g_j \text{ stays the same for } j \neq i$$

Question: can you think of a better estimator of the gradient than g_i ?

Let's take $\quad g^t = \frac{1}{n} \sum_{j=1}^{n} g_j$

SAG: Stochastic average gradient



SAG: Stochastic Average Gradient

Set $w^0 = 0, g_i = \nabla f_i(w^0),$ for $i = 1, \ldots, n$

Choose $\gamma > 0$

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

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

  $g_i = \nabla f_i(w^t)$ (update grad)

  $g^t = \frac{1}{n} \sum_{j=1}^{n} g_j$

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

Output w^T

Very easy to implement, no inner loop.

Stores a $d \times n$ matrix

SAG: Stochastic Average Gradient

Set $w^0 = 0, g_i = \nabla f_i(w^0),$ for $i = 1, \ldots, n$
Choose $\gamma > 0$
for $t = 0, 1, 2, \ldots, T-1$
    sample $i \in \{1, \ldots, n\}$
    $g_i = \nabla f_i(w^t)$ (update grad)
    $g^t = \frac{1}{n} \sum_{j=1}^n g_j$
    $w^{t+1} = w^t - \gamma g^t$
Output w^T

Very easy to implement, no inner loop.

Stores a $d \times n$ matrix

EXE: Introduce a variable $G = (1/n) \sum_{j=1} g_j$. Re-write the SAG algorithm so G is updated efficiently at each iteration.

SAG: Rationale

Let's take $$g^t = \frac{1}{n}\sum_{j=1}^{n} g_j$$

Close to convergence, $g_j \simeq \nabla f_j(w^t)$ hence

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

However, there is a problem with this method which makes analysis difficult: what is $\mathbb{E}_i[g^t]$?

SAGA: Stochastic Average Gradient

Defazio, Bach, & Lacoste-Julien, 2014 NIPs

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

Sample

$$\nabla f_i(w^t), \qquad \text{i.i.d sample with prob } \frac{1}{n}$$

Grad. estimate

$$g_i(w^t) = \nabla f_i(w^t) - \nabla f_i(w^{t_i}) + \frac{1}{n}\sum_{j=1}^{n} \nabla f_j(w^{t_j})$$

Store grad.

$$\nabla f_i(w^{t_i}) = \nabla f_i(w^t)$$

SAGA: Stochastic Average Gradient

Defazio, Bach, & Lacoste-Julien, 2014 NIPs

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

Sample: $\nabla f_i(w^t)$, i.i.d sample with prob $\frac{1}{n}$

Grad. estimate:

$$g_i(w^t) = \nabla f_i(w^t) - \nabla f_i(w^{t_i}) + \frac{1}{n}\sum_{j=1}^{n} \nabla f_j(w^{t_j})$$

$$\nabla z_i(w^t) = \nabla f_i(w^{t_i})$$

Store grad.: $\nabla f_i(w^{t_i}) = \nabla f_i(w^t)$

SAGA: Stochastic Average Gradient

Defazio, Bach, & Lacoste-Julien, 2014 NIPs

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

Sample

$\nabla f_i(w^t)$, i.i.d sample with prob $\frac{1}{n}$

Grad. estimate

$$g_i(w^t) = \nabla f_i(w^t) - \nabla f_i(w^{t_i}) + \frac{1}{n}\sum_{j=1}^{n} \nabla f_j(w^{t_j})$$

$$z_i(w) = f_i(w^{t_i}) + \langle \nabla f_i(w^{t_i}), w - w^{t_i} \rangle$$

$$\nabla z_i(w^t) = \nabla f_i(w^{t_i})$$

Store grad.

$$\nabla f_i(w^{t_i}) = \nabla f_i(w^t)$$

SAGA: Stochastic Average Gradient

Defazio, Bach, & Lacoste-Julien, 2014 NIPs

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

Sample

$\nabla f_i(w^t)$, i.i.d sample with prob $\frac{1}{n}$

Grad. estimate

$$g_i(w^t) = \nabla f_i(w^t) - \nabla f_i(w^{t_i}) + \frac{1}{n}\sum_{j=1}^{n} \nabla f_j(w^{t_j})$$

$$z_i(w) = f_i(w^{t_i}) + \langle \nabla f_i(w^{t_i}), w - w^{t_i} \rangle$$

$\nabla z_i(w^t) = \nabla f_i(w^{t_i})$

$\mathbb{E}[\nabla z_i(w^t)]$

Store grad.

$$\nabla f_i(w^{t_i}) = \nabla f_i(w^t)$$

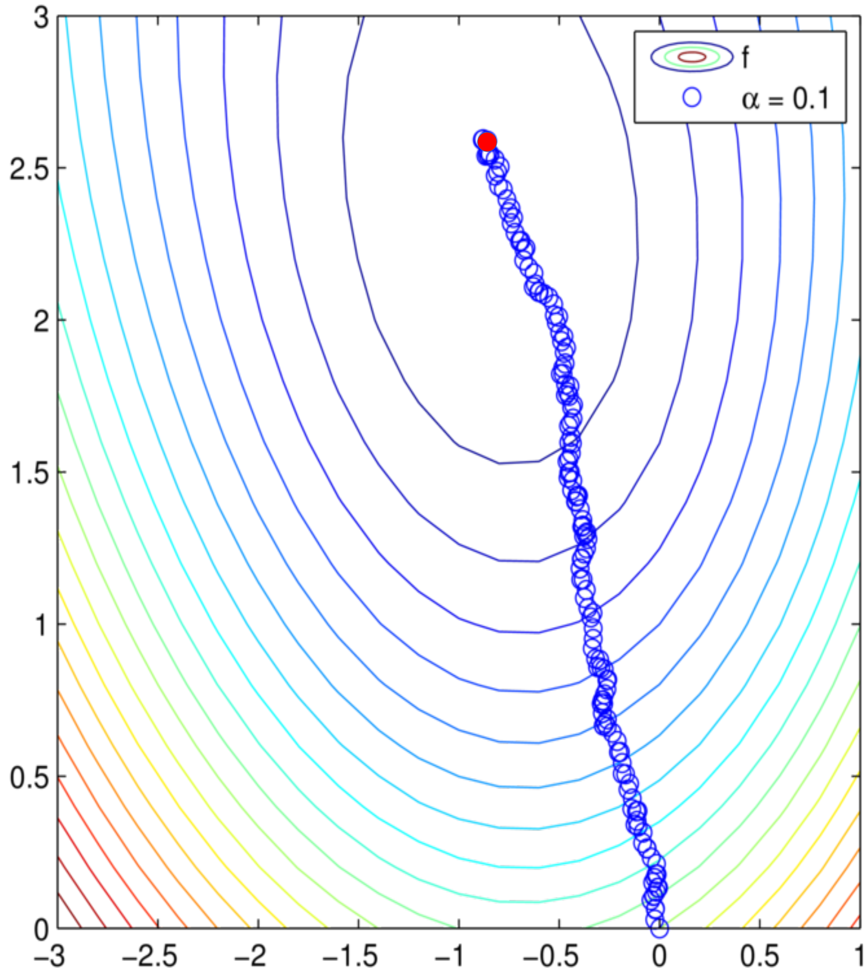
SAGA: Stochastic Average Gradient

Set $w^0 = 0, g_i = \nabla f_i(w^0)$, for $i = 1 \ldots, n$
Choose $\gamma > 0$
for $t = 0, 1, 2, \ldots, T-1$
    sample $i \in \{1, \ldots, n\}$
    $g^t = \nabla f_i(w^t) - g_i + \frac{1}{n}\sum_{j=1}^n g_j$
    $w^{t+1} = w^t - \gamma g^t$
    $g_i = \nabla f_i(w^t)$
Output w^T

No inner loop, rolling update

Stores a $d \times n$ matrix

The Stochastic Average Gradient



Convergence Theorems: exercise

Convergence Theorems

Assumptions for Convergence

Strong Convexity

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

Smoothness + convexity

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

$$f_i(w) \geq f_i(y) + \langle \nabla f_i(y), w - y \rangle \qquad \text{for } i = 1, \ldots, n$$

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

Convergence SAGA

Theorem SAGA

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

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

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

A practical convergence result !

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

Convergence SAG

Theorem SAG

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

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

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

Less practical convergence result compared to SAGA

Because of biased gradients, very hard proof that relies on computer assisted steps

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

From Convergence to Complexity

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

$$\frac{||w^T - w^*||_2^2}{||w^1 - w^*||_2^2} \leq \epsilon \quad \implies \quad T \geq \frac{1}{1-\rho}\log\left(\frac{1}{\epsilon}\right)$$

$$\frac{||w^T - w^*||_2^2}{||w^1 - w^*||_2^2} \leq \epsilon \quad \implies \quad T \geq \max\left\{4n, \frac{3L_{\max}}{\mu}\right\}\log\left(\frac{1}{\epsilon}\right)$$

Comparisons in total complexity for strongly convex

Approximate solution

$$\mathbb{E}[f(w^T)] - f(w^*) \leq \epsilon \quad \text{or} \quad \mathbb{E}\|w^t - w^*\|^2 \leq \epsilon$$

SGD

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

Gradient descent

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

SVRG/SAGA/SAG

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

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

How did I get these complexity results from the convergence results?

Practicals implementation of SAG for Linear Classifiers

Finite Sum Training Problem

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

L2 regularizer + linear hypothesis

Practicals implementation of SAG for Linear Classifiers

L2 regularizer + linear hypothesis

Finite Sum Training Problem

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

$$\nabla f_i(w) = \ell'(\langle w, x^i\rangle, y^i)x^i + \lambda w$$

Practicals implementation of SAG for Linear Classifiers

Finite Sum Training Problem

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

L2 regularizer + linear hypothesis

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

Practicals implementation of SAG for Linear Classifiers

L2 regularizer + linear hypothesis

Finite Sum Training Problem

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

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

Reduce Storage to $O(n)$

Only store real number $\qquad \beta_i = \ell'(\langle w^{t_i}, x^i \rangle, y^i)$

Stoch. gradient estimate $\qquad \nabla f_i(w^{t_i}) = \beta_i x^i + \lambda w^t$

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

Take for home Variance Reduction

- Variance reduced methods use only **one stochastic gradient per iteration** and converge linearly on strongly convex functions

- Choice of **fixed stepsize** possible

- **SAGA** only needs to know the smoothness parameter to work, but requires storing n past stochastic gradients

- **SVRG** only has $O(d)$ storage, but requires full gradient computations every so often. Has an extra "number of inner iterations" parameter to tune

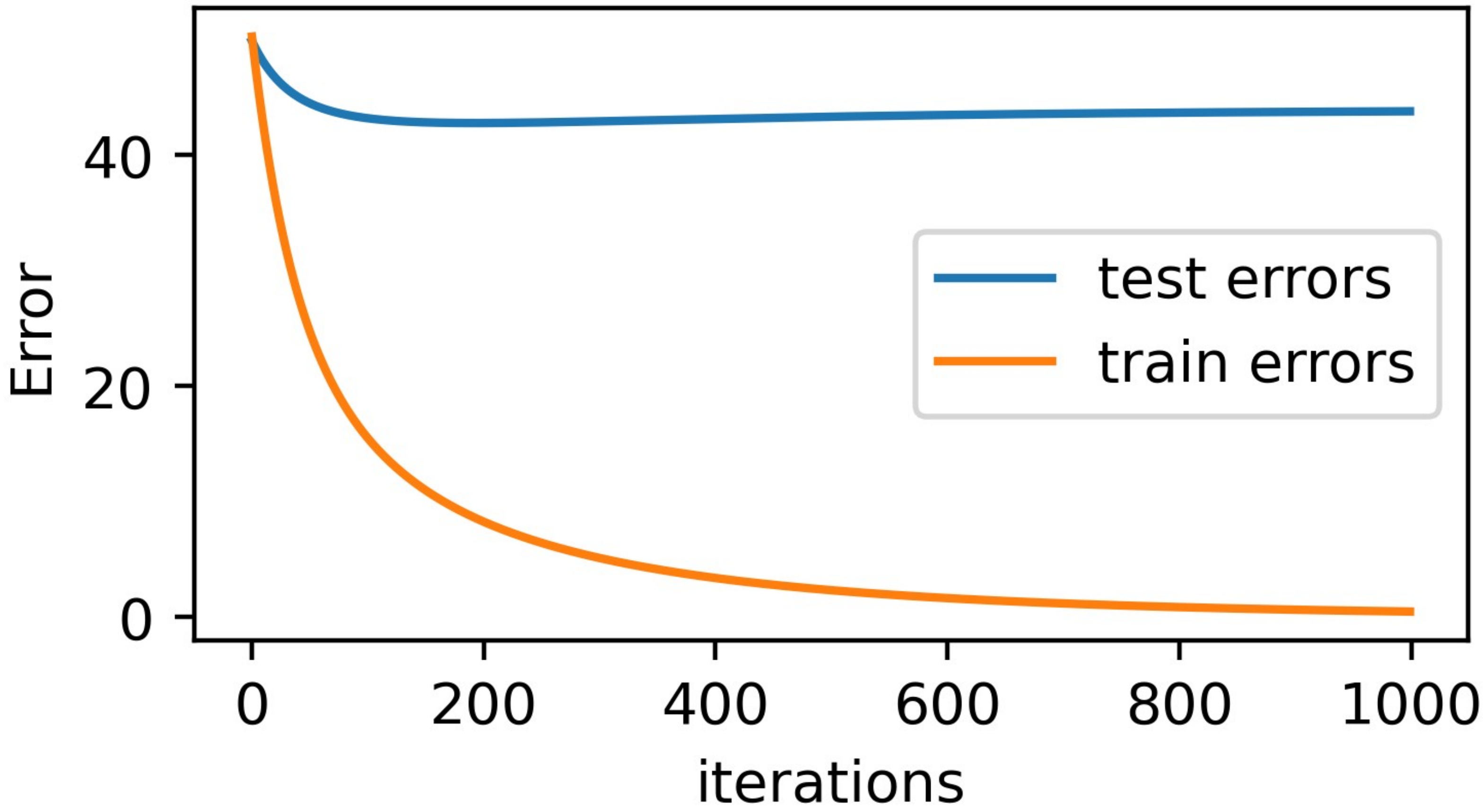
Implicit bias of gradient descent

Underdetermined regression problem

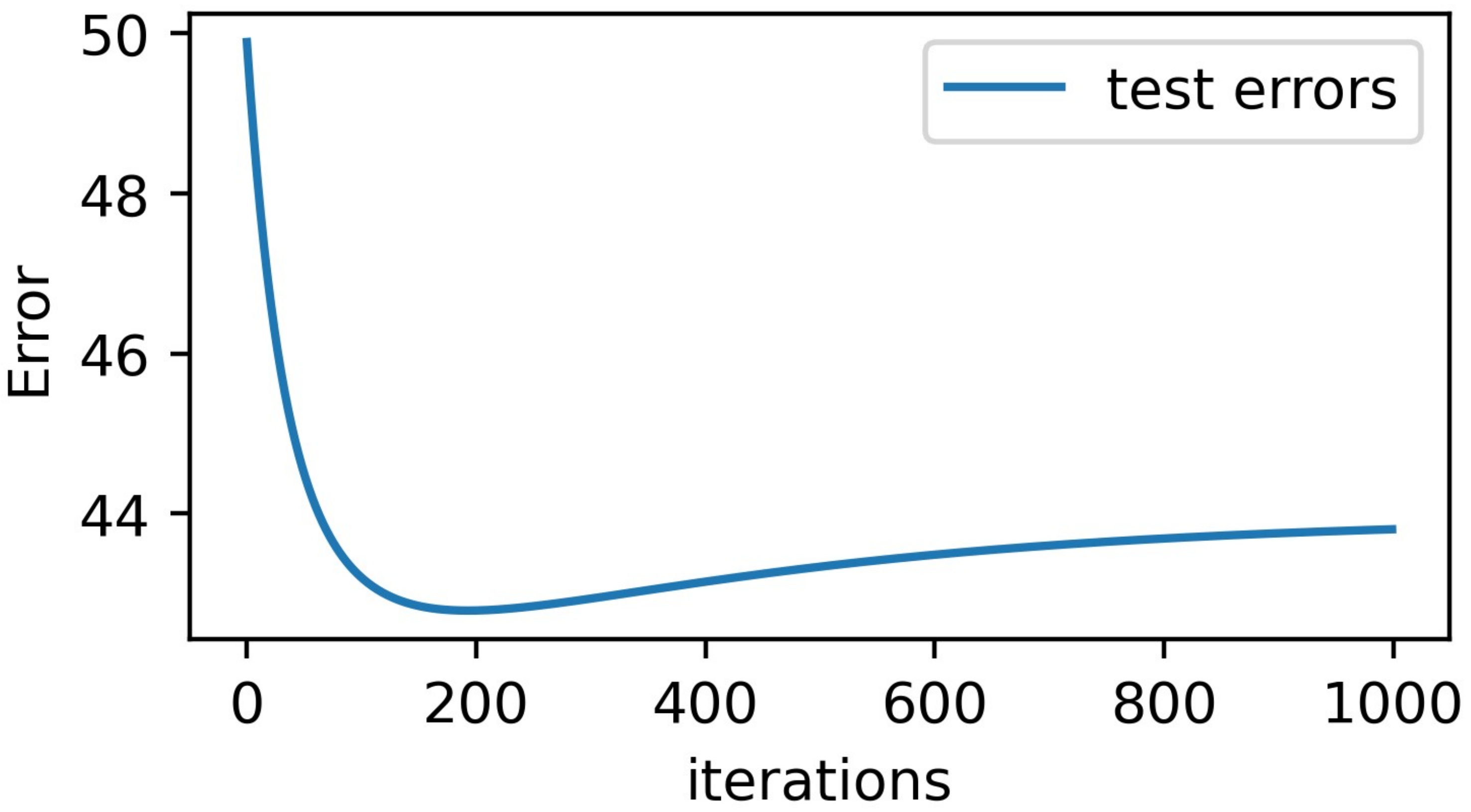
- Take a dataset X of size n, p with n < p, and n scalar y. Split between train and test. Solve
$$\min_w \frac{1}{2}\|X_{train}w - y_{train}\|^2$$
with gradient descent to find training parameters w

- Compute the train and test error during training. What do you expect to see?

Underdetermined regression problem



Underdetermined regression problem



What is going on?

Exercises 2.pdf