ML X Science Summer School

Optimization for ML

Robert M. Gower

Outline of my classes

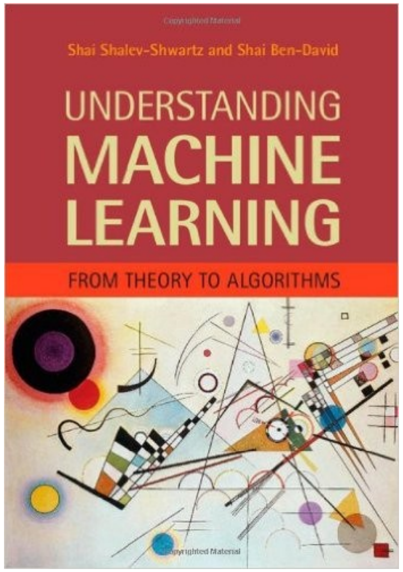
- Intro to empirical risk problem and gradient descent (GD)
- (Stochastic Gradient) SGD for convex optimization. Theory and variants
- SGD with momentum and some tricks
- Lecture slides, exercises, & jupyter notebook: gowerrobert.github.io/

Part I: An Introduction to Supervised Learning

References for my lectures

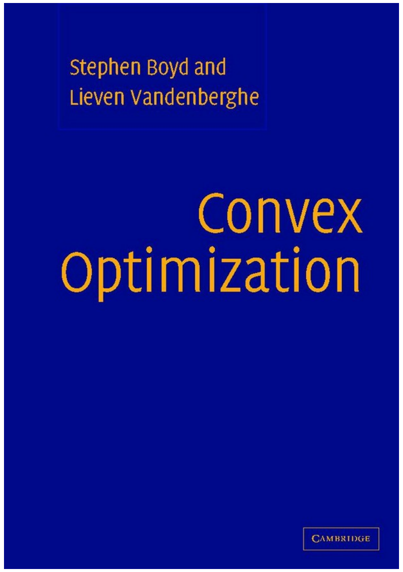
Chapter 2

Understanding Machine Learning: From Theory to Algorithms

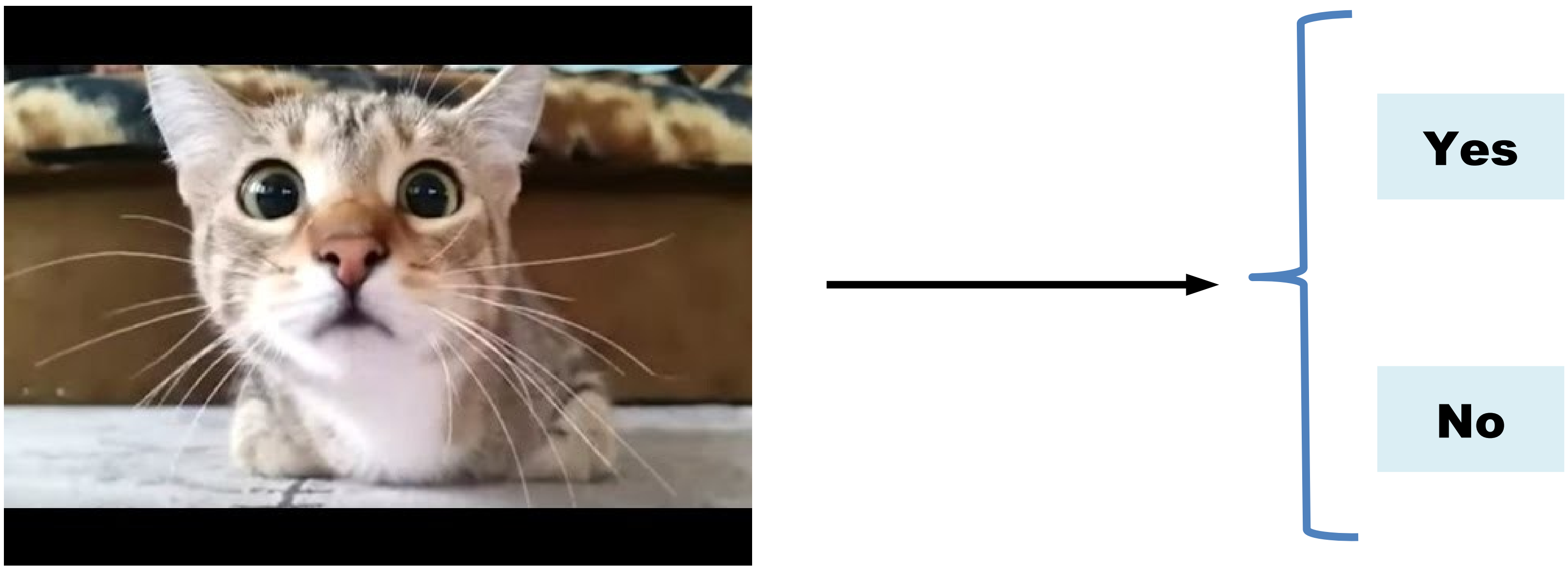


Pages 67 to 79

Convex Optimization, Stephen Boyd



Is There a Cat in the Photo?



Is There a Cat in the Photo?

Yes

Is There a Cat in the Photo?

Yes

Is There a Cat in the Photo?

No

Is There a Cat in the Photo?

Yes

Is There a Cat in the Photo?

$\xrightarrow{h}$

Yes

No

x: Input/Feature

y: Output/Target

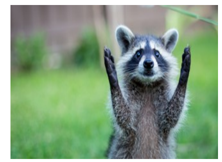
Find mapping *h* that assigns the “correct” target to each input

$$h : x \in \mathbf{R}^d \longrightarrow y \in \mathbf{R}$$

Labeled Data: The training set

x^1 { 
$y^1 = 1$

x^2 { 
$y^2 = 1$

x^3 { 
$y^3 = -1$

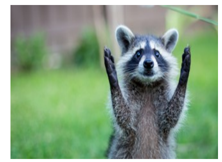
$\cdots$ x^n { 
$y^n = 1$

y= *-1* means no/false

Labeled Data: The training set

$x^1 \{$  $y^1 = 1$

$x^2 \{$  $y^2 = 1$

$x^3 \{$  $y^3 = -1$

$\cdots\ x^n \{$  $y^n = 1$

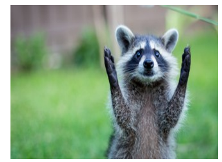
y= -1 means no/false

Training Algorithm

Labeled Data: The training set

x^1 { 
$y^1 = 1$

x^2 { 
$y^2 = 1$

x^3 { 
$y^3 = -1$

$\cdots$ x^n { 
$y^n = 1$

y= -1 means no/false

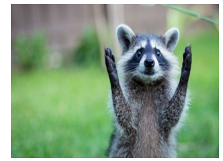
Training Algorithm

$$h : x \in \mathbf{R}^d \rightarrow y \in \mathbf{R}$$

Labeled Data: The training set

x^1 {  $y^1 = 1$

x^2 {  $y^2 = 1$

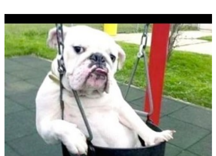
x^3 {  $y^3 = -1$

$\cdots$ x^n {  $y^n = 1$

y= -1 means no/false

Training Algorithm

$$h : x \in \mathbf{R}^d \rightarrow y \in \mathbf{R}$$

$h\Big($  $\Big)$ $\longrightarrow$ **-1**

Example: Linear Regression for Height

Male = 0
Female = 1

Labelled data $x \in \mathbf{R}^2, y \in \mathbf{R}_+$

x_1^1	Sex	0
x_2^1	Age	30
y^1	Height	1,72 cm

$\cdots$

x_1^n	Sex	1
x_2^n	Age	70
y^n	Height	1,52 cm

Example: Linear Regression for Height

Male = 0
Female = 1

Labelled data $x \in \mathbf{R}^2, y \in \mathbf{R}_+$

x_1^1	Sex	0
x_2^1	Age	30
y^1	Height	1,72 cm

$\cdots$

x_1^n	Sex	1
x_2^n	Age	70
y^n	Height	1,52 cm

Example Hypothesis: Linear Model

$$h_w(x_1, x_2) = w_0 + x_1 w_1 + x_2 w_2 \overset{x_0=1}{=} \langle w, x \rangle$$

Example: Linear Regression for Height

Male = 0
Female = 1

Labelled data $x \in \mathbf{R}^2, y \in \mathbf{R}_+$

x_1^1	Sex	0
x_2^1	Age	30
y^1	Height	1,72 cm

$\cdots$

x_1^n	Sex	1
x_2^n	Age	70
y^n	Height	1,52 cm

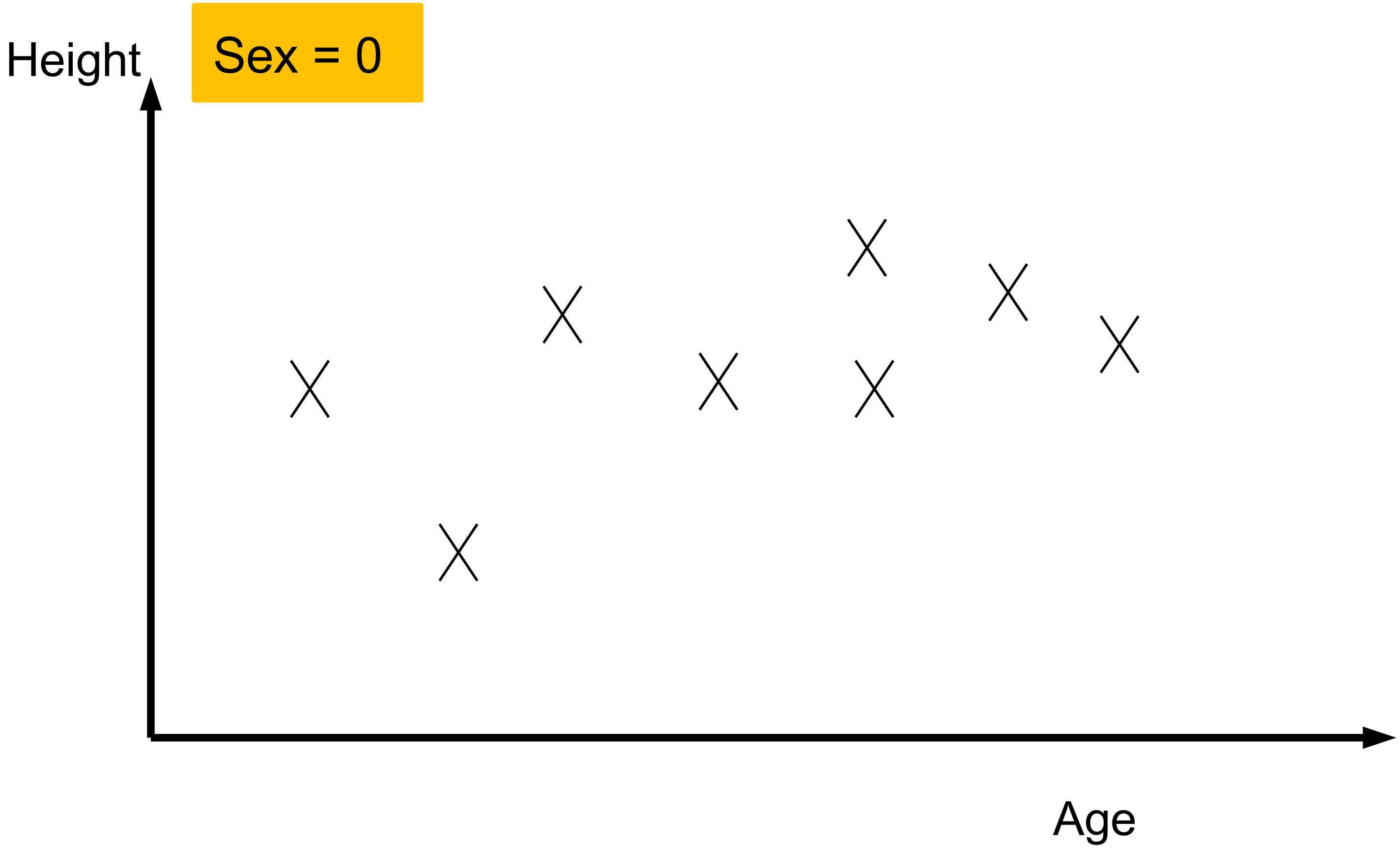
Example Hypothesis: Linear Model

$$h_w(x_1, x_2) = w_0 + x_1 w_1 + x_2 w_2 \overset{x_0=1}{=} \langle w, x \rangle$$

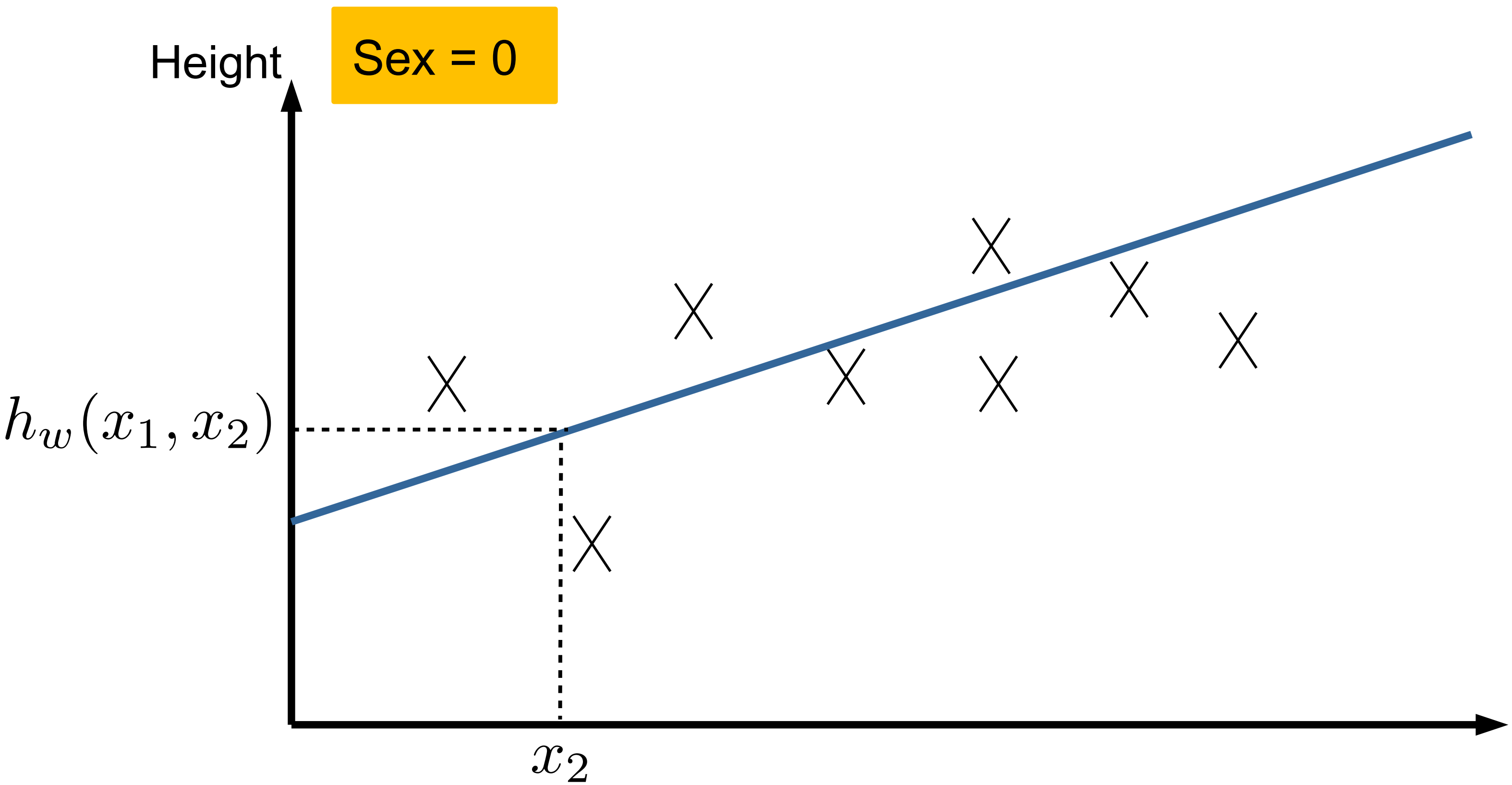
Example Training Problem:

$$\min_{w \in \mathbf{R}^3} \frac{1}{n} \sum_{i=1}^{n} \left(h_w(x_1^i, x_2^i) - y^i \right)^2$$

Linear Regression for Height



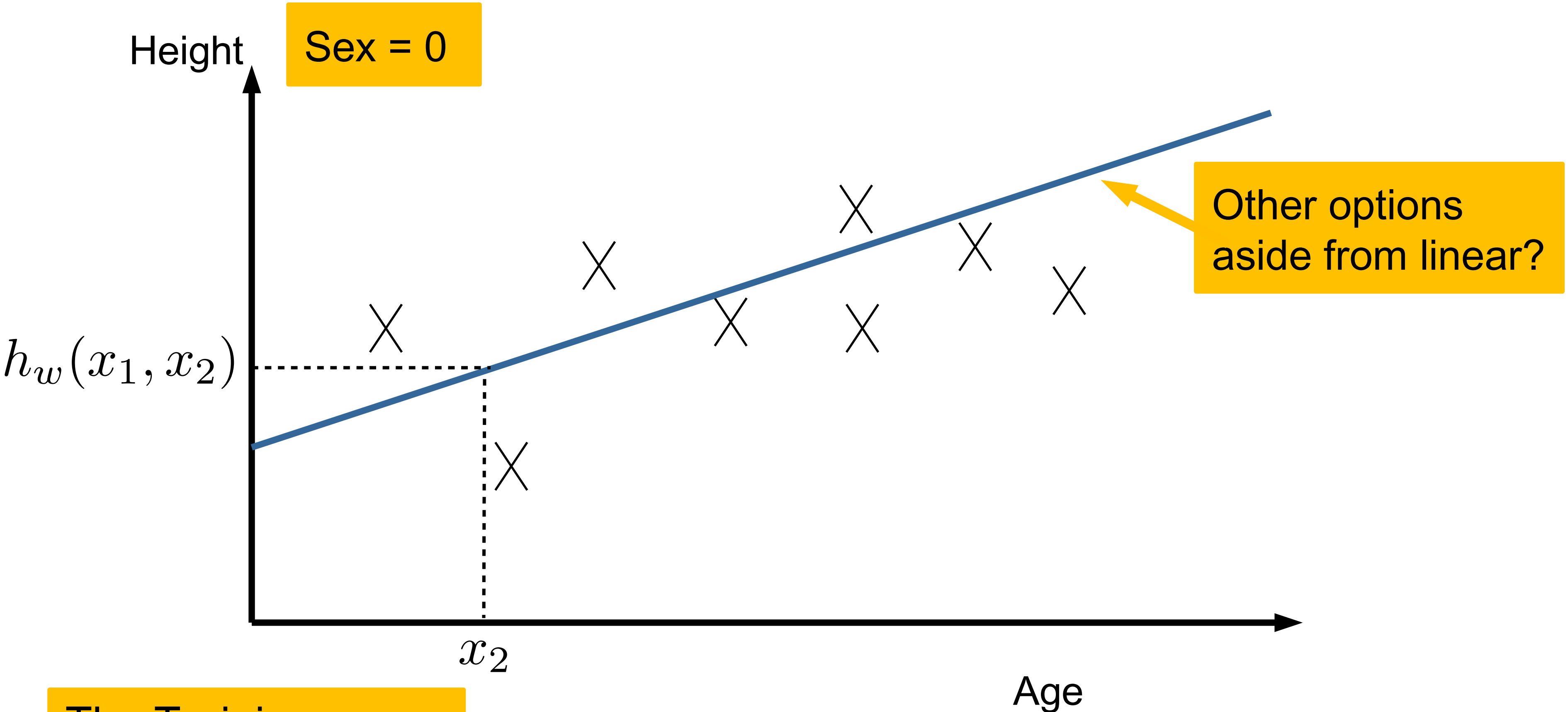
Linear Regression for Height



The Training Algorithm

$$\min_{w \in \mathbf{R}^3} \frac{1}{n} \sum_{i=1}^{n} \left(h_w(x_1^i, x_2^i) - y^i\right)^2$$

Linear Regression for Height



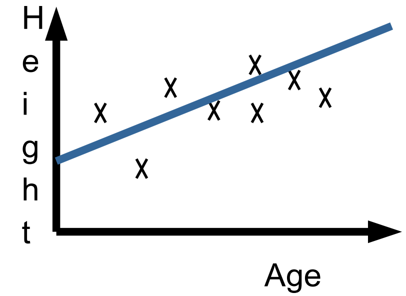
The Training Algorithm

$$\min_{w \in \mathbf{R}^3} \frac{1}{n} \sum_{i=1}^{n} \left(h_w(x_1^i, x_2^i) - y^i\right)^2$$

Parametrizing the Hypothesis

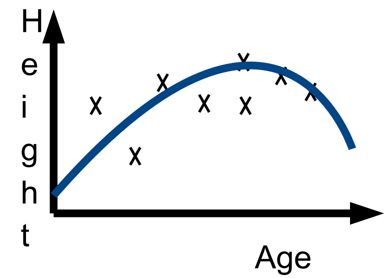
Linear:

$$h_w(x) = \sum_{i=0}^{d} w_i x_i$$

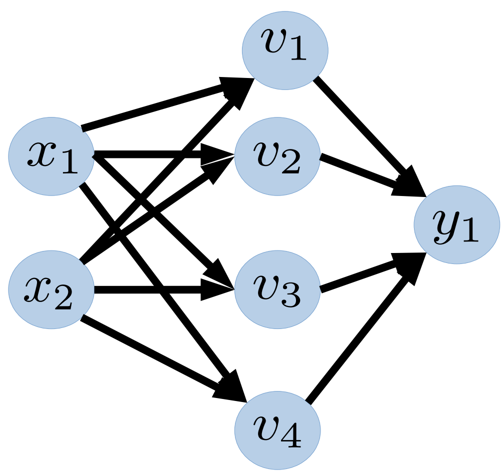


Polinomial:

$$h_w(x) = \sum_{i,j=0}^{d} w_{ij} x_i x_j$$



Neural Net:



$exe:$

$$v_1 = \mathrm{sign}(w_{11}x_1 + w_{12}x_2)$$

$$v_4 = 1/\left(1 + \exp(w_{41}x_1 + w_{42}x_2)\right)$$

Loss Functions

$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left(h_w(x^i) - y^i \right)^2$$

Why a Squared Loss?

Loss Functions

$$\min_{w\in\mathbf{R}^d} \frac{1}{n}\sum_{i=1}^{n}\left(h_w(x^i)-y^i\right)^2$$

Why a Squared Loss?

$$\text{Let } y_h := h_w(x)$$

Loss Functions

$$\begin{array}{rccc} \ell: & \mathbf{R}\times\mathbf{R} & \rightarrow & \mathbf{R}_+ \\ & (y_h, y) & \rightarrow & \ell(y_h, y) \end{array}$$

The Training Problem

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

Loss Functions

$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left(h_w(x^i) - y^i\right)^2$$

Why a Squared Loss?

$$\text{Let } y_h := h_w(x)$$

Loss Functions

$$\begin{aligned} \ell : \quad \mathbf{R} \times \mathbf{R} &\rightarrow \mathbf{R}_+ \\ (y_h, y) &\rightarrow \ell(y_h, y) \end{aligned}$$

Typically a convex function

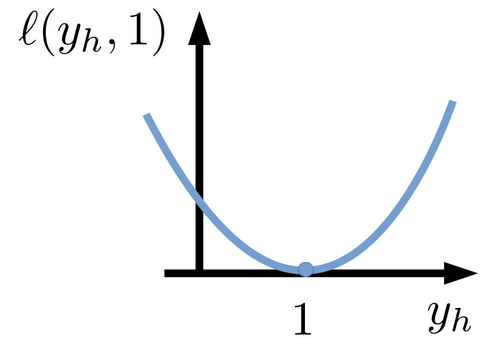
The Training Problem

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

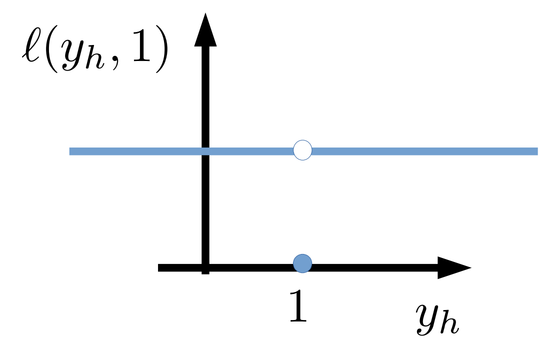
Different the Loss Functions

Let $y_h := h_w(x)$

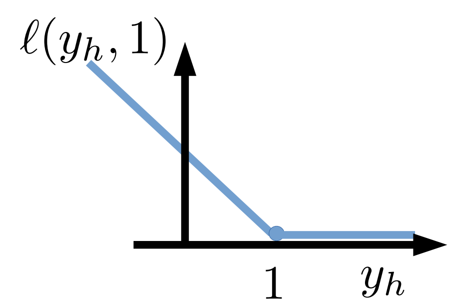
Square Loss $\quad \ell(y_h, y) = (y_h - y)^2$



Binary Loss $\quad \ell(y_h, y) = \begin{cases} 0 & \text{if } y_h = y \\ 1 & \text{if } y_h \neq y \end{cases}$



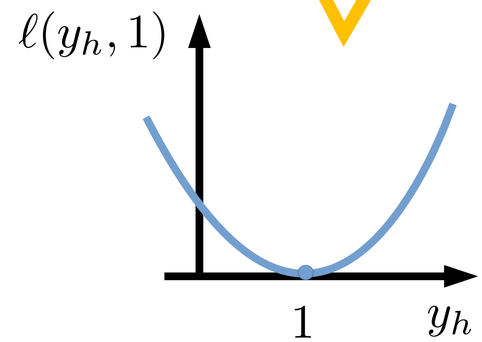
Hinge Loss $\quad \ell(y_h, y) = \max\{0, 1 - y_h y\}$



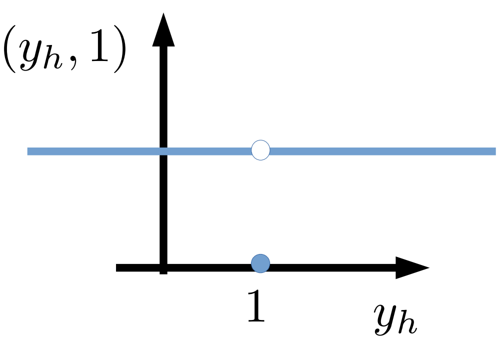
Different the Loss Functions

y=1 in all figures

Let $y_h := h_w(x)$

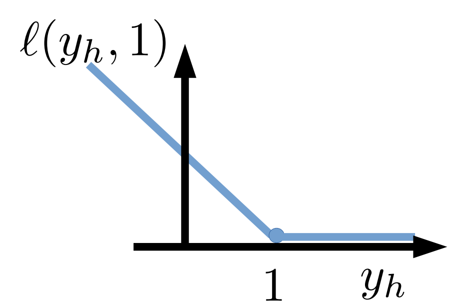


Square Loss $\quad \ell(y_h, y) = (y_h - y)^2$



Binary Loss

$$\ell(y_h, y) = \begin{cases} 0 & \text{if } y_h = y \\ 1 & \text{if } y_h \neq y \end{cases}$$



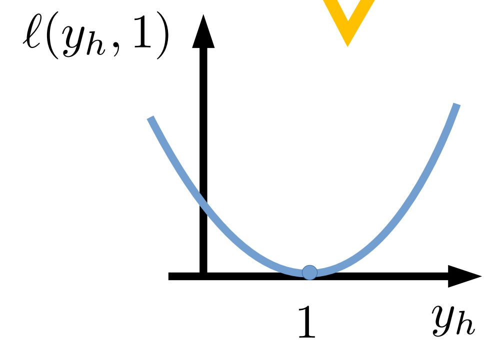
Hinge Loss $\quad \ell(y_h, y) = \max\{0, 1 - y_h y\}$

Different the Loss Functions

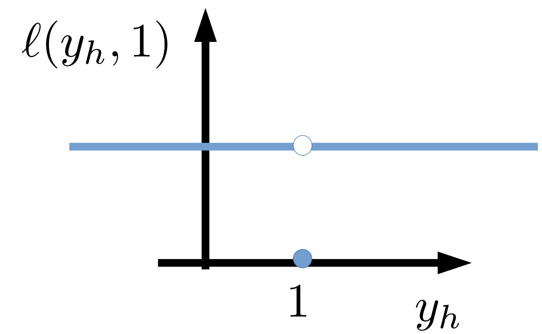
y=1 in all figures

Let $y_h := h_w(x)$

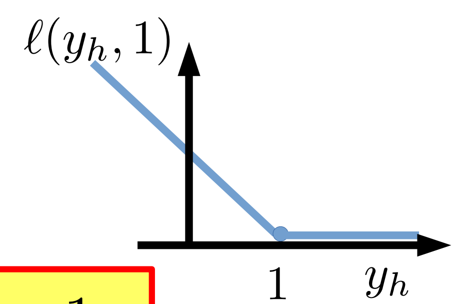
Square Loss $\ell(y_h, y) = (y_h - y)^2$



Binary Loss $\ell(y_h, y) = \begin{cases} 0 & \text{if } y_h = y \\ 1 & \text{if } y_h \neq y \end{cases}$



Hinge Loss $\ell(y_h, y) = \max\{0, 1 - y_h y\}$



EXE: Plot the binary and hinge loss function in when $y = -1$

Are we done?

Is a notion of Loss enough?

What happens when we do not have enough data?

Are we done?

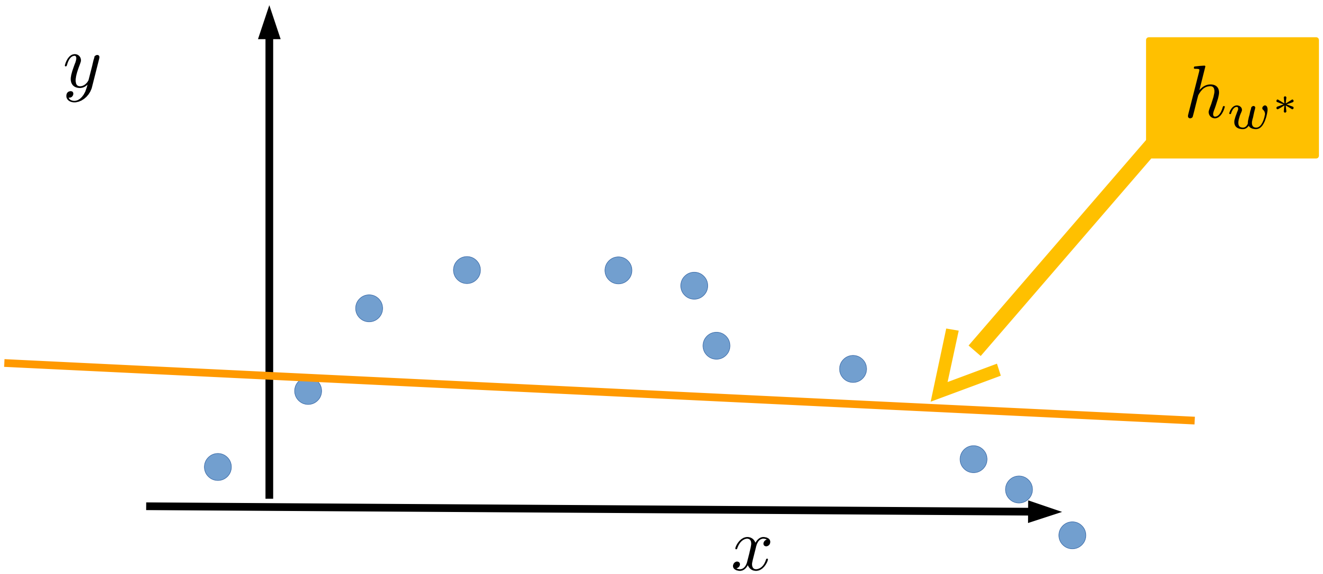
The Training Problem

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

Is a notion of Loss enough?

What happens when we do not have enough data?

Overfitting and Model Complexity

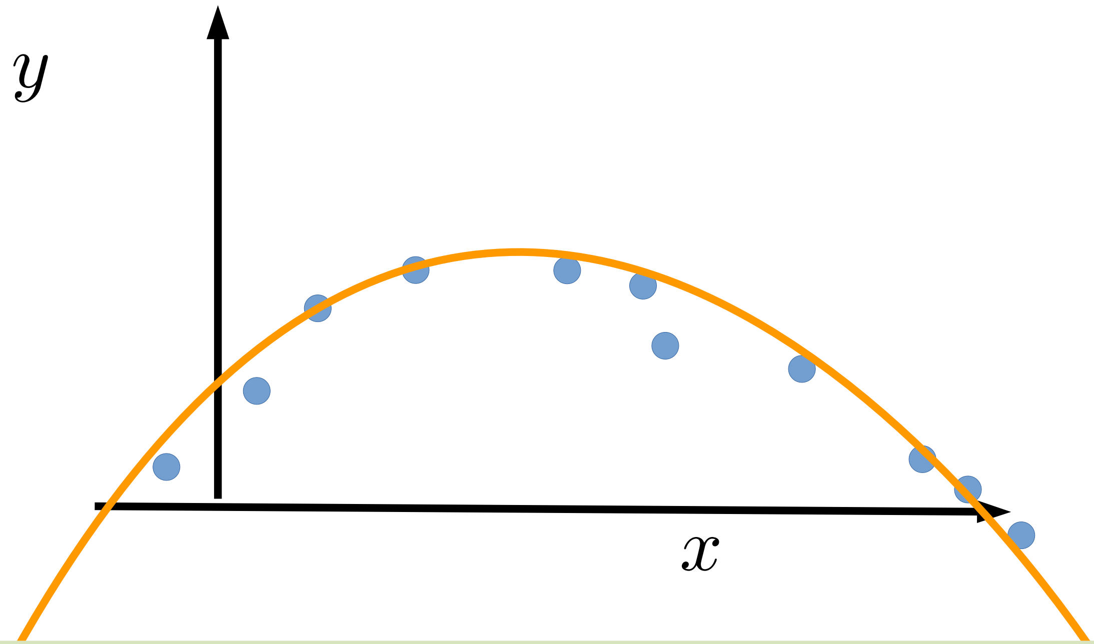


Fitting 1st order polynomial

$$h_w = \langle w, x \rangle$$

$$w^* = \arg \min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left(h_w(x^i) - y^i \right)^2$$

Overfitting and Model Complexity

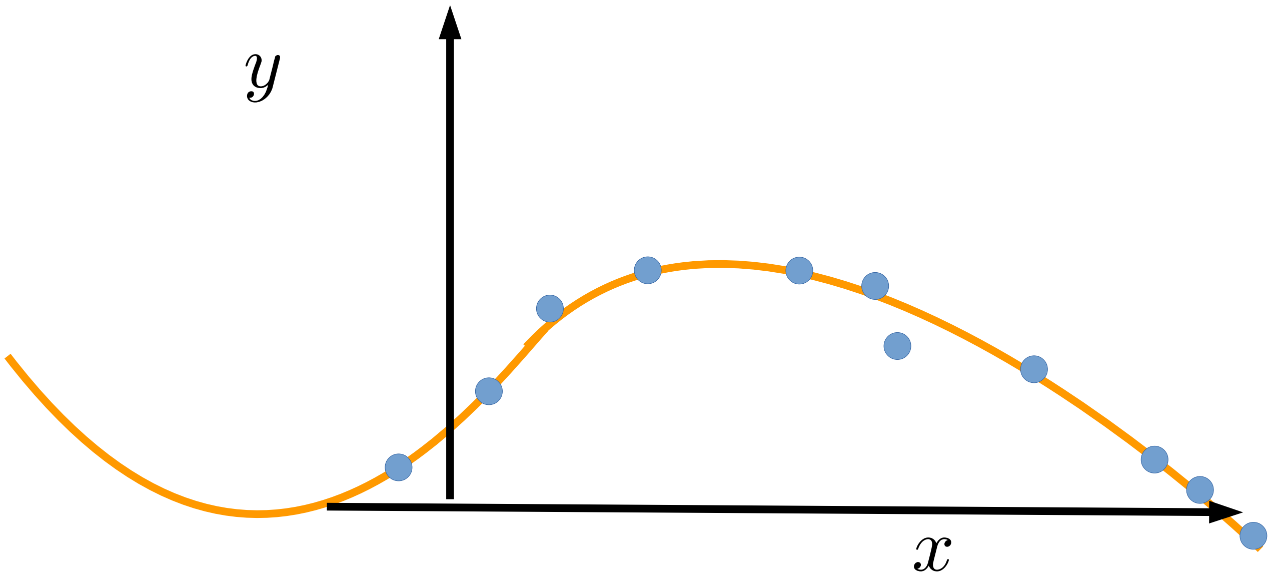


Fitting 2nd order polynomial

$$h_w = w_0 + w_1 x + w_2 x^2$$

$$w^* = \arg \min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left(h_w(x^i) - y^i \right)^2$$

Overfitting and Model Complexity

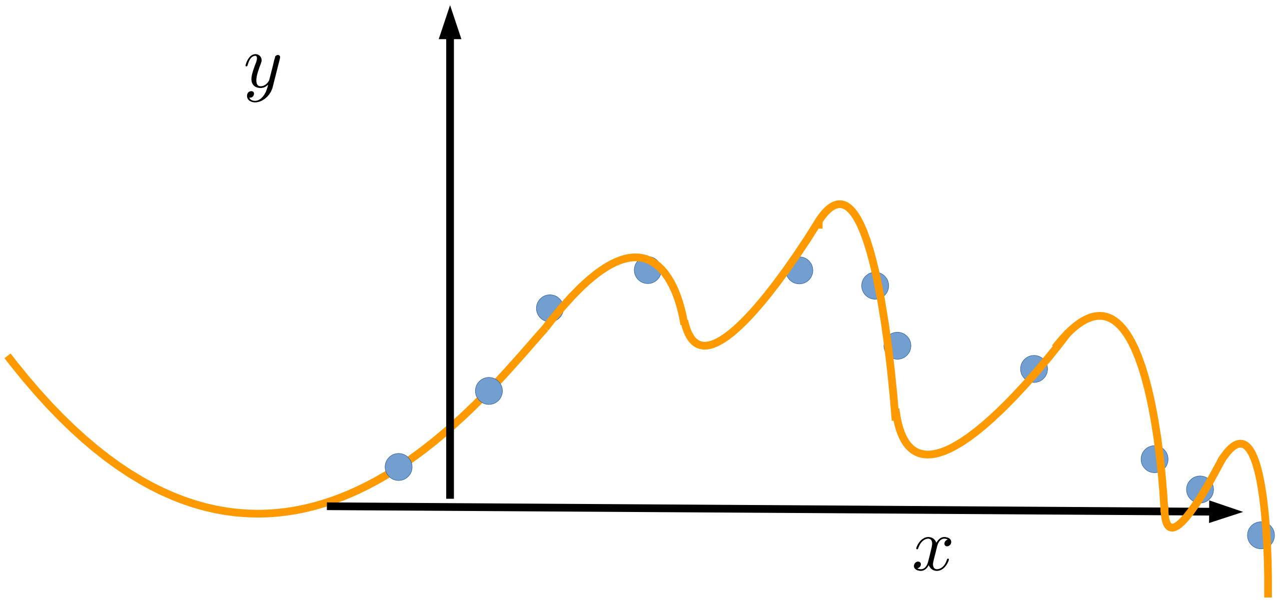


Fitting 3rd order polynomial

$$h_w = \sum_{i=0}^{3} w_i x^i$$

$$w^* = \arg \min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left(h_w(x^i) - y^i\right)^2$$

Overfitting and Model Complexity



Fitting 9th order polynomial

$$h_w = \sum_{i=0}^{9} w_i x^i$$

$$w^* = \arg\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left(h_w(x^i) - y^i \right)^2$$

Regularization/Prior

Regularizor Functions

$$\begin{aligned} R: \quad \mathbf{R}^d &\to \mathbf{R}_+ \\ w &\to R(w) \end{aligned}$$

General Training Problem

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

Regularization/Prior

Regularizor Functions

$$\begin{aligned} R: \quad \mathbf{R}^d &\rightarrow \mathbf{R}_+ \\ w &\rightarrow R(w) \end{aligned}$$

General Training Problem

$$\min_{w \in \mathbf{R}^d} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \ell\left(h_w(x^i), y^i\right)}_{\text{Goodness of fit, fidelity term ...etc}} + \lambda R(w)$$

Regularization/Prior

Regularizor Functions

$$\begin{aligned} R: \quad \mathbf{R}^d &\to \mathbf{R}_+ \\ w &\to R(w) \end{aligned}$$

General Training Problem

$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \underbrace{\sum_{i=1}^{n} \ell\left(h_w(x^i), y^i\right)}_{\text{Goodness of fit, fidelity term ...etc}} + \lambda \underbrace{R(w)}_{\text{Penalizes complexity}}$$

Regularization/Prior

Regularizor Functions

$$\begin{aligned} R: \quad \mathbf{R}^d &\to \mathbf{R}_+ \\ w &\to R(w) \end{aligned}$$

Controls tradeoff between fit and complexity

General Training Problem

$$\min_{w \in \mathbf{R}^d} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \ell\left(h_w(x^i), y^i\right)}_{\text{Goodness of fit, fidelity term ...etc}} + \lambda \underbrace{R(w)}_{\text{Penalizes complexity}}$$

Regularization/Prior

Regularizor Functions

$$R: \quad \mathbf{R}^d \rightarrow \mathbf{R}_+$$
$$\quad\quad w \rightarrow R(w)$$

Controls tradeoff between fit and complexity

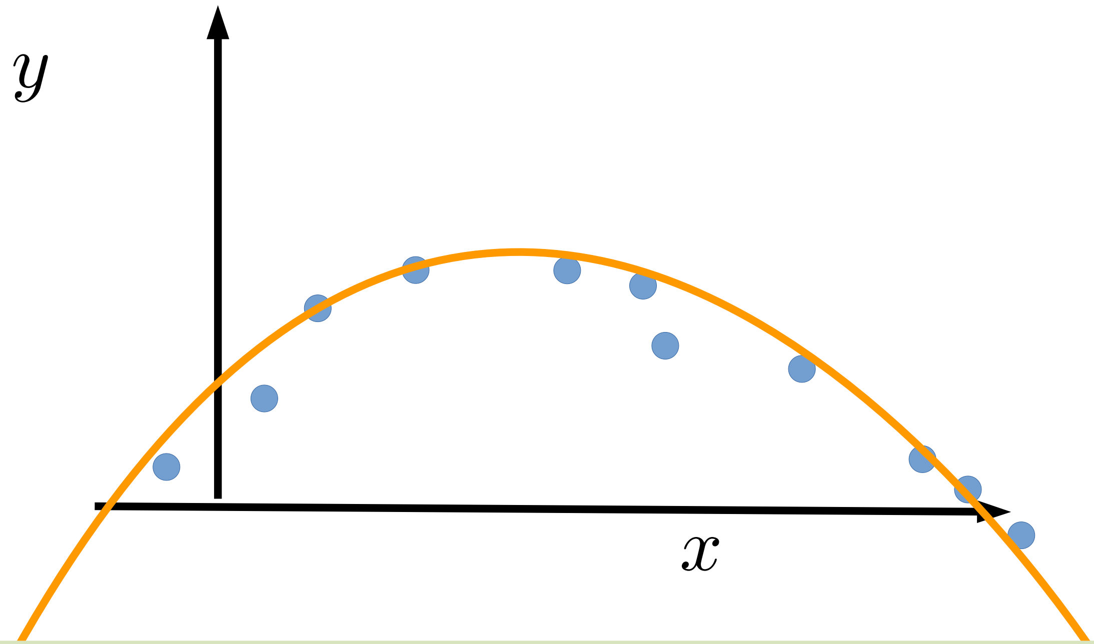
General Training Problem

$$\min_{w \in \mathbf{R}^d} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell\left(h_w(x^i), y^i\right)}_{\text{Goodness of fit, fidelity term ...etc}} + \lambda \underbrace{R(w)}_{\text{Penalizes complexity}}$$

Exe:

$$R(w) = ||w||_2^2, \quad ||w||_1, \quad ||w||_p, \quad \text{other norms} \quad \ldots$$

Overfitting and Model Complexity

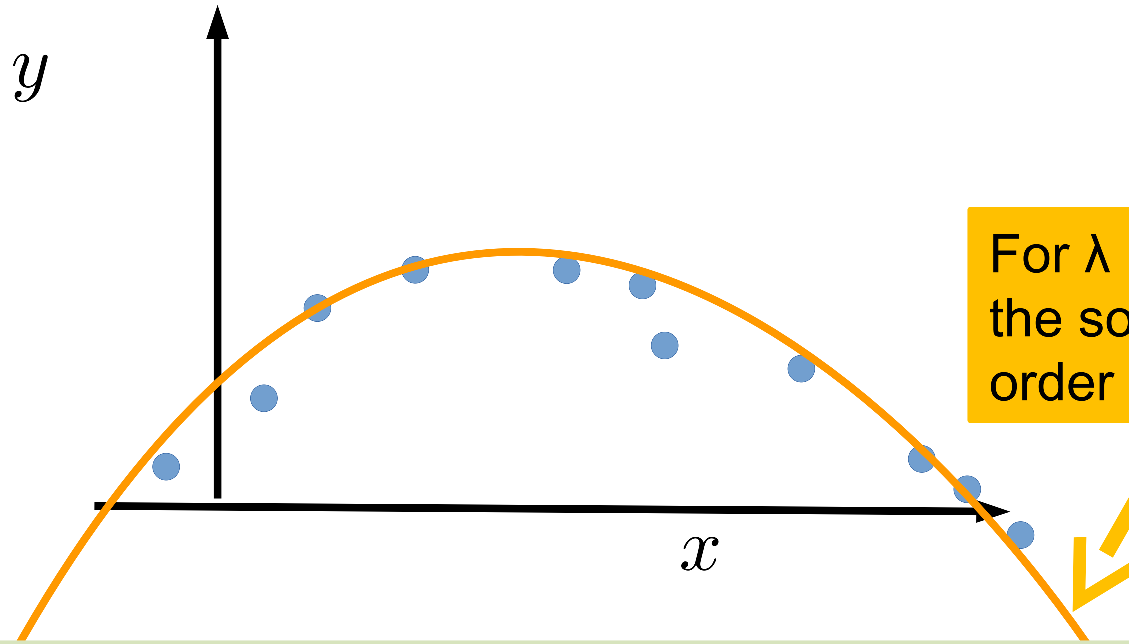


Fitting kth order polynomial

$$h_w = \sum_{i=0}^{k} w_i x^i$$

$$w^* = \arg\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left(h_w(x^i) - y^i\right)^2 + \lambda ||w||_1$$

Overfitting and Model Complexity



For λ big enough, the solution is a 2nd order polynomial

Fitting k^th^ order polynomial

$$h_w = \sum_{i=0}^{k} w_i x^i$$

$$w^* = \arg \min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left(h_w(x^i) - y^i \right)^2 + \lambda ||w||_1$$

Exe: Ridge Regression

Linear hypothesis

$$h_w(x) = \langle w, x \rangle$$

L2 regularizor

$$R(w) = ||w||_2^2$$

L2 loss

$$\ell(y_h, y) = (y_h - y)^2$$

Ridge Regression

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

Exe: Support Vector Machines

Linear hypothesis

$$h_w(x) = \langle w, x \rangle$$

L2 regularizor

$$R(w) = ||w||_2^2$$

Hinge loss

$$\ell(y_h, y) = \max\{0, 1 - y_h y\}$$

SVM with soft margin

$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \max\{0, 1 - y^i \langle w, x^i \rangle\} + \lambda ||w||_2^2$$

Exe: Logistic Regression

Linear hypothesis

$$h_w(x) = \langle w, x \rangle$$

+

L2 regularizor

$$R(w) = ||w||_2^2$$

Logistic loss

$$\ell(y_h, y) = \ln(1 + e^{-y y_h})$$

Logistic Regression

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

ML as seen by Optimizer

$$(1) \quad \text{Get the labeled data: } (x^1, y^1), \ldots, (x^n, y^n)$$

ML as seen by Optimizer

(1) Get the labeled data: $(x^1, y^1), \ldots, (x^n, y^n)$

(2) Choose a parametrization for hypothesis: $h_w(x)$

ML as seen by Optimizer

(1) Get the labeled data: $(x^1, y^1), \ldots, (x^n, y^n)$

(2) Choose a parametrization for hypothesis: $h_w(x)$

(3) Choose a loss function: $\ell(h_w(x), y) \geq 0$

ML as seen by Optimizer

(1) Get the labeled data: $(x^1, y^1), \ldots, (x^n, y^n)$

(2) Choose a parametrization for hypothesis: $h_w(x)$

(3) Choose a loss function: $\ell(h_w(x), y) \geq 0$

(4) Solve the *training problem*:

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

ML as seen by Optimizer

(1) Get the labeled data: $(x^1, y^1), \ldots, (x^n, y^n)$

(2) Choose a parametrization for hypothesis: $h_w(x)$

(3) Choose a loss function: $\ell(h_w(x), y) \geq 0$

(4) Solve the *training problem*:

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

(5) Test and cross-validate. If fail, go back a few steps

ML as seen by Optimizer

(1) Get the labeled data: $(x^1, y^1), \ldots, (x^n, y^n)$

(2) Choose a parametrization for hypothesis: $h_w(x)$

(3) Choose a loss function: $\ell(h_w(x), y) \geq 0$

(4) Solve the *training problem*:

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

(5) Test and cross-validate. If fail, go back a few steps

Part II: Solving the Training Problem

Re-writing as 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

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

Ignore all structure for now

The Training Problem

Solving the *training problem*:

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

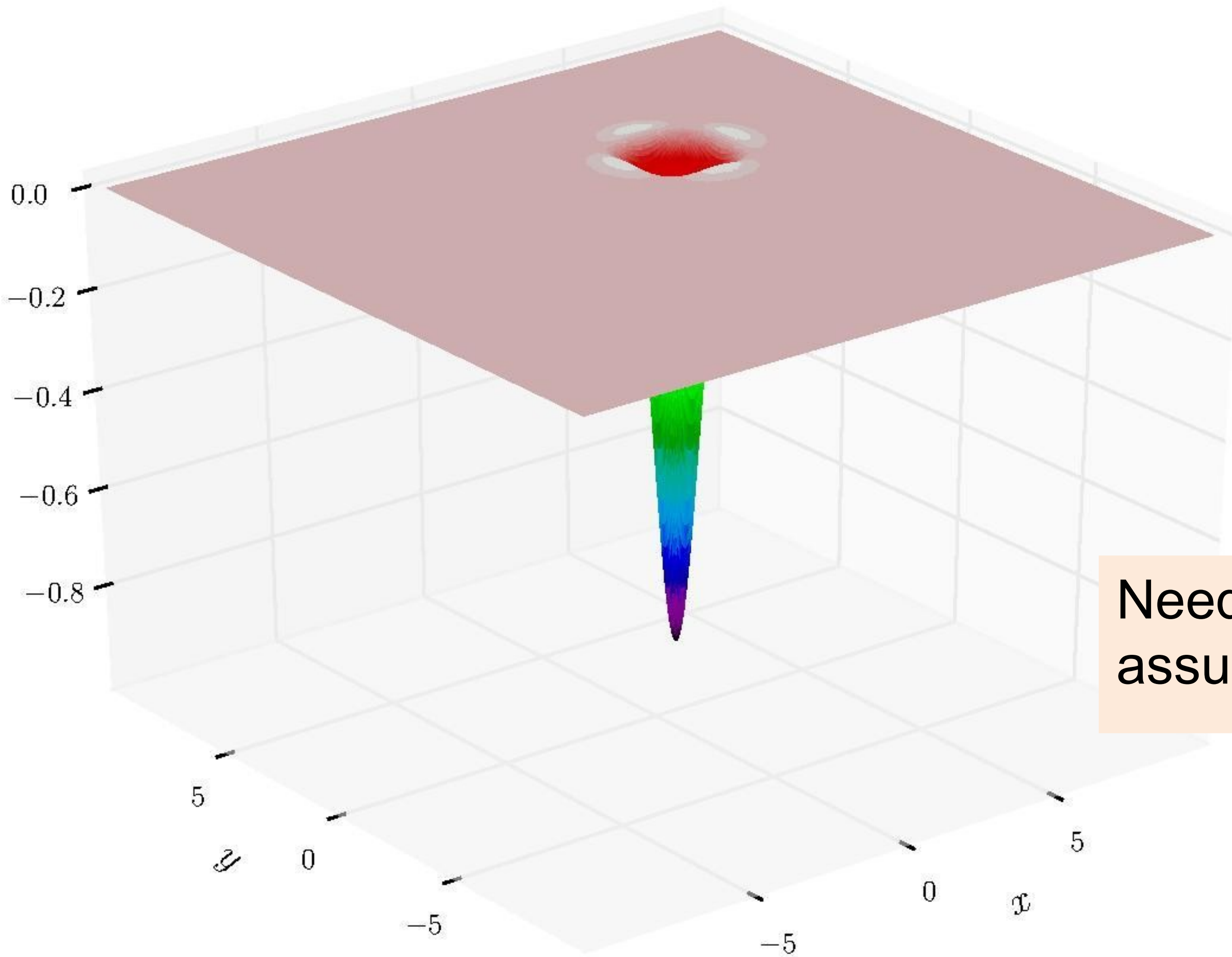
Reference method: Gradient descent

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

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

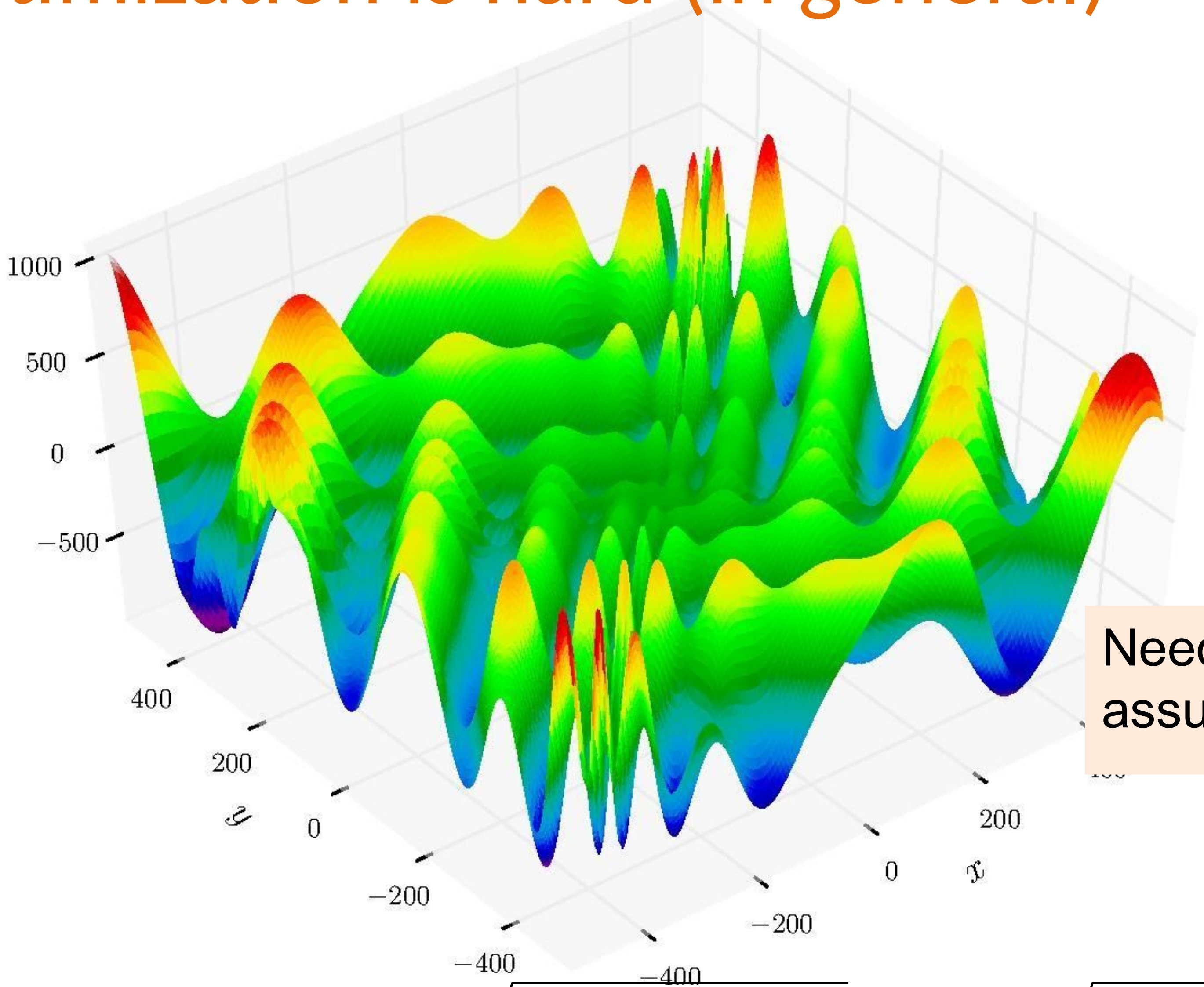
Optimization is hard (in general)



Need assumptions!

$$f(x, y) = -\cos(x)\cos(y)\exp\left(-(x-\pi)^2 - (y-\pi)^2\right)$$

Optimization is hard (in general)



Need assumptions!

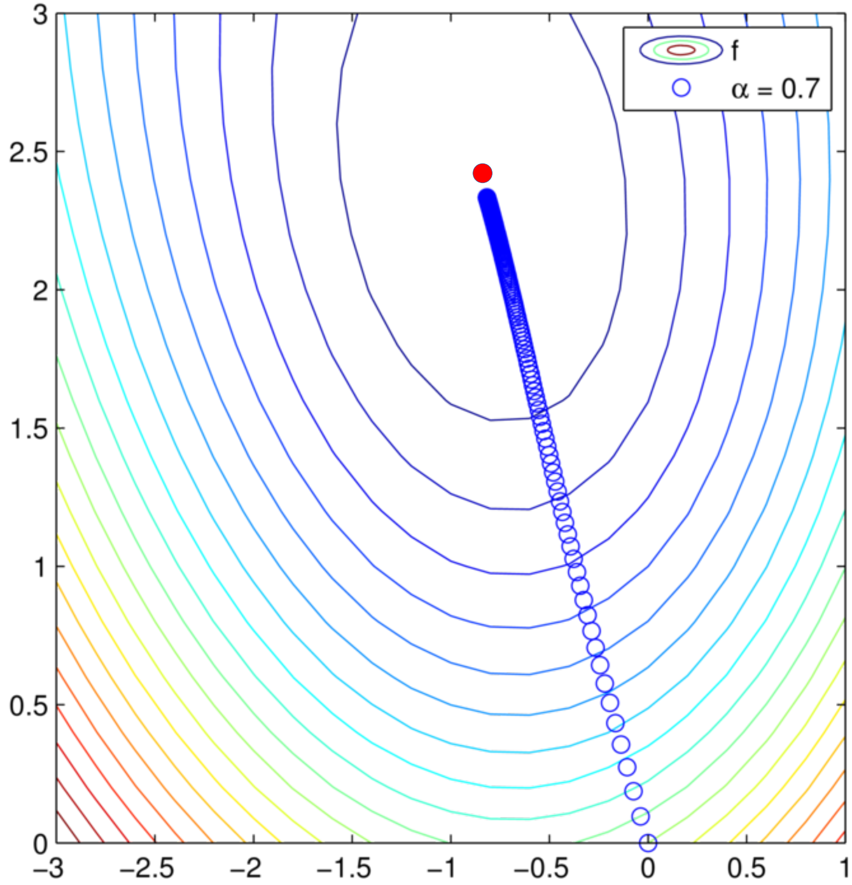
$$f(x,y) = -(y+47)\sin\sqrt{\left|\frac{x}{2}+(y+47)\right|} - x\sin\sqrt{\left|\frac{x}{2}-(y+47)\right|}$$

Gradient Descent Example

A Logistic Regression problem using the fourclass labelled data from LIBSVM
$(n, d) = (862,2)$

Logistic Regression

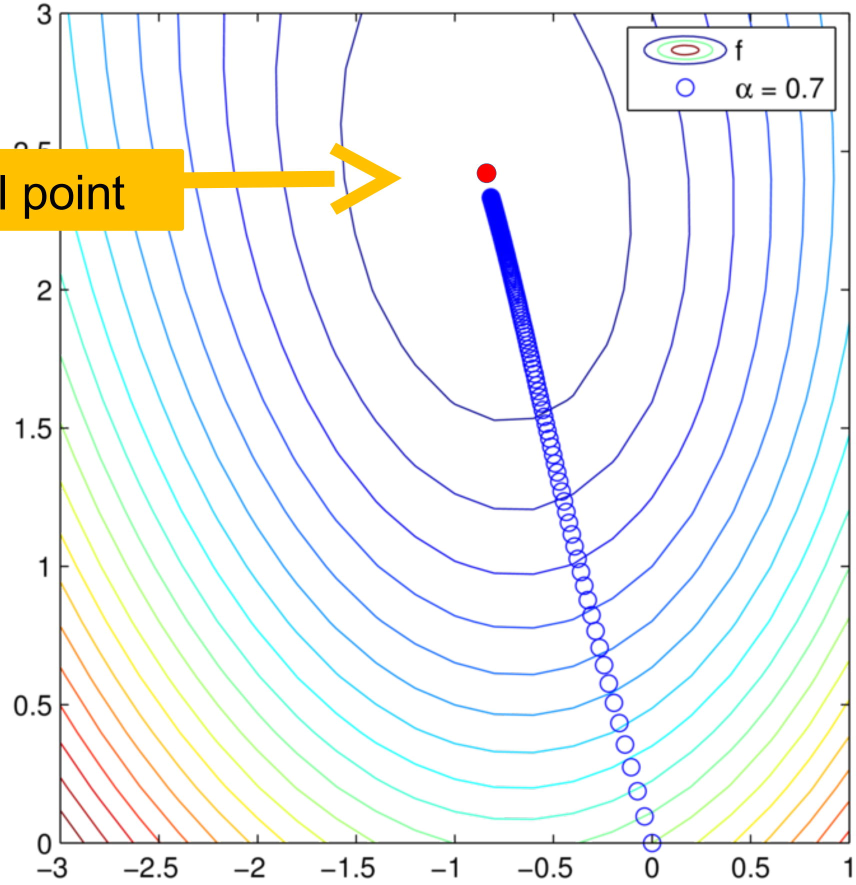
$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ln(1 + e^{-y^i \langle w, x^i \rangle}) + \lambda ||w||_2^2$$



Can we prove that this always works?

Gradient Descent Example

Optimal point



A Logistic Regression problem using the fourclass labelled data from LIBSVM

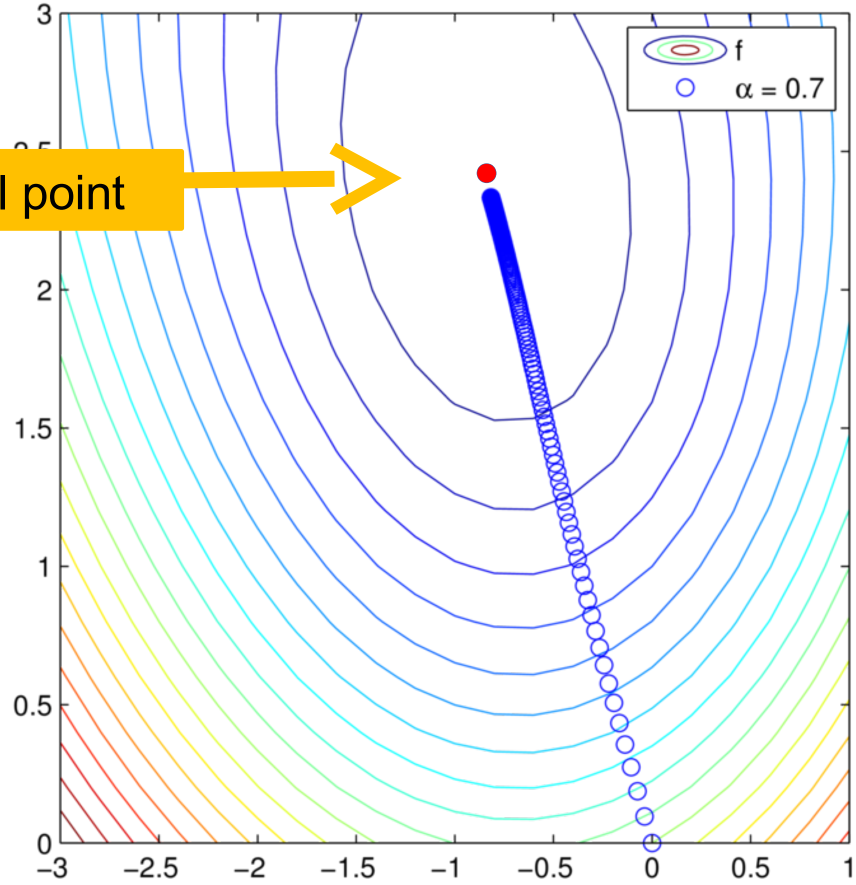
(n, d)= (862,2)

Logistic Regression

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

Can we prove that this always works?

Gradient Descent Example



Optimal point

A Logistic Regression problem using the fourclass labelled data from LIBSVM
(n, d)= (862,2)

Logistic Regression

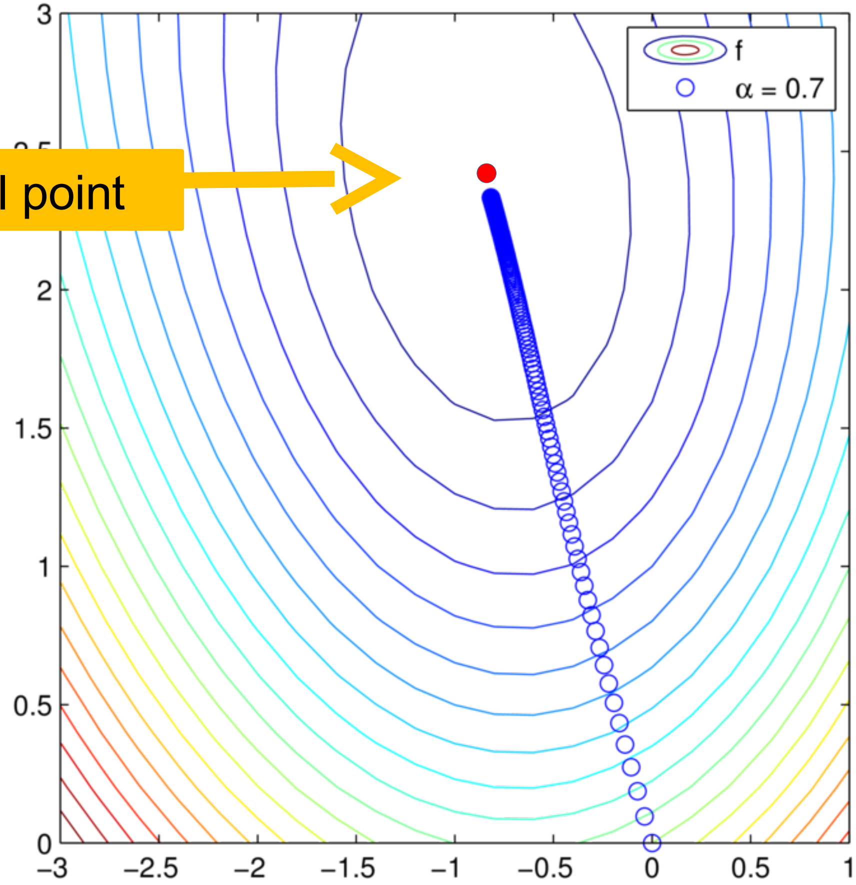
$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ln(1 + e^{-y^i \langle w, x^i \rangle}) + \lambda ||w||_2^2$$

Can we prove that this always works?

No! There is no universal optimization method. The "no free lunch" of Optimization

Gradient Descent Example

Optimal point



A Logistic Regression problem using the fourclass labelled data from LIBSVM
$(n, d) = (862,2)$

Logistic Regression

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

Can we prove that this always works?

No! There is no universal optimization method. The "no free lunch" of Optimization

Specialize

Convex and smooth training problems

Main assumption

Nice property

$$\text{If } \nabla f(w^*) = 0 \quad \text{then} \quad f(w^*) \leq f(w), \quad \forall w \in \mathbb{R}^d$$

All stationary points are global minima

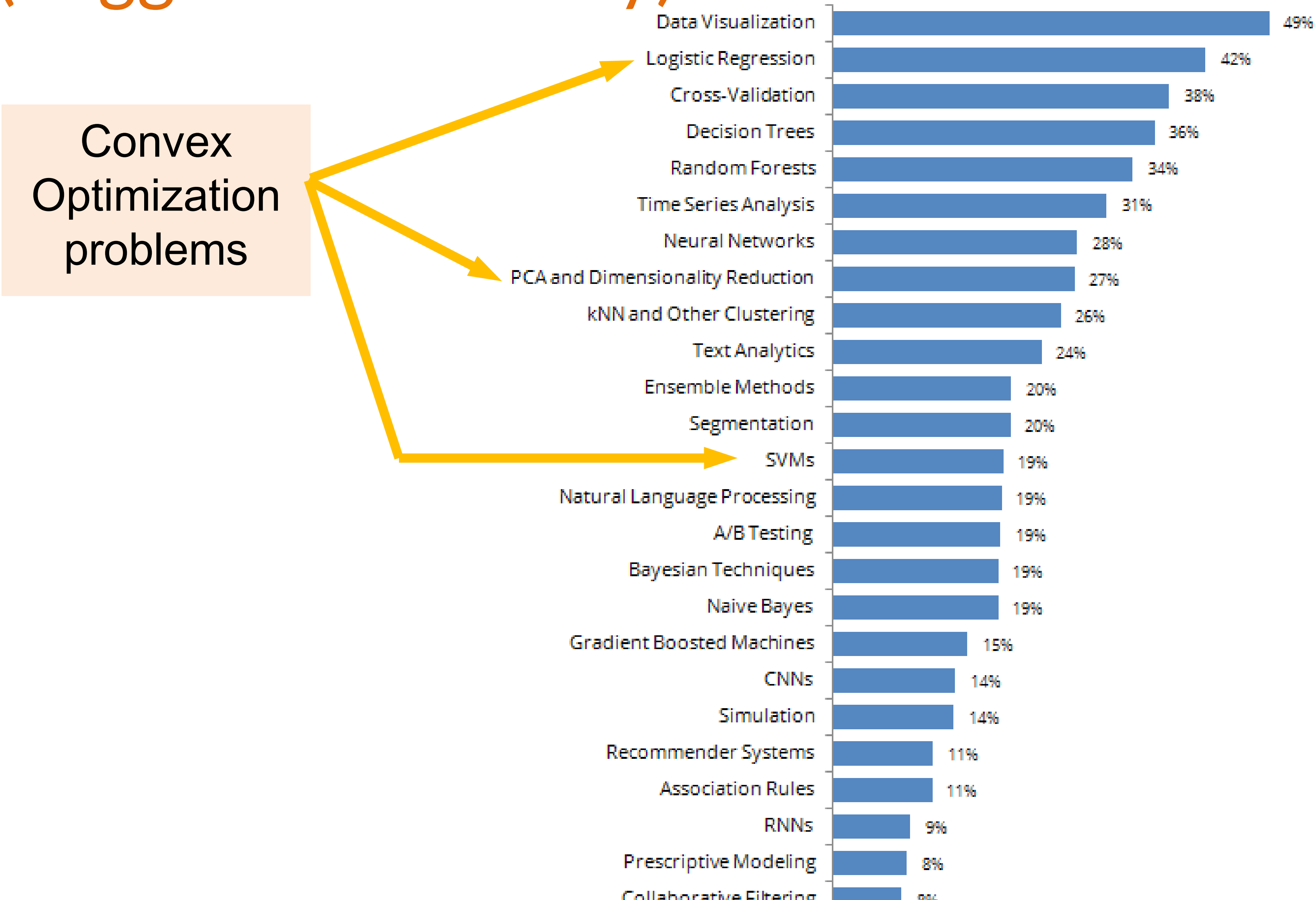
Lemma: Convexity => Nice property

$$\text{If } \ f(w) \geq f(y) + \langle \nabla f(y), w - y \rangle, \quad \forall w, y \in \mathbb{R}^d$$

$$\text{then nice property holds}$$

PROOF: $\text{Choose } y = w^*$

Data science methods most used (Kaggle 2017 survey)

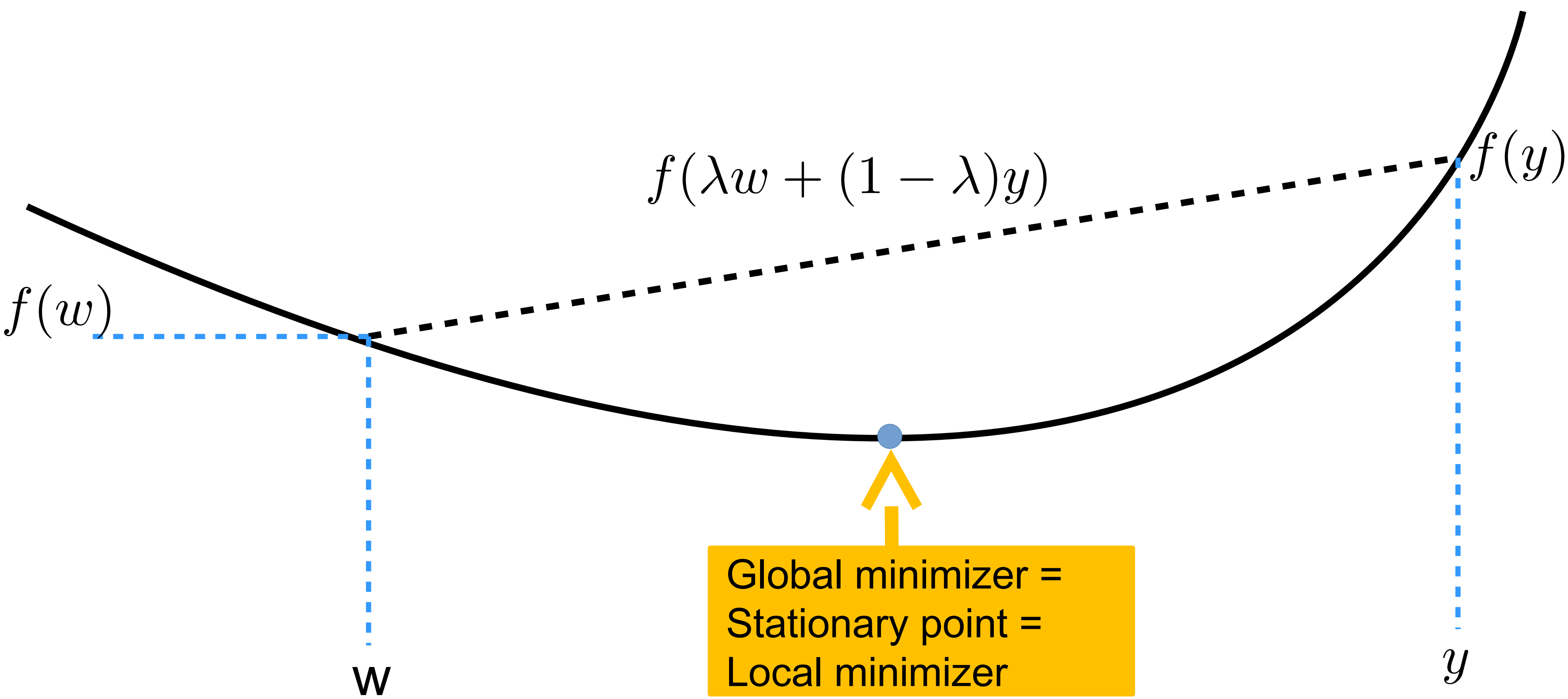


Part II: Convexity, Smoothness, Gradient Descent

Convexity

We say $f : \mathrm{dom}(f) \subset \mathbb{R}^n \to \mathbb{R}$ is convex if $\mathrm{dom}(f)$ is convex and

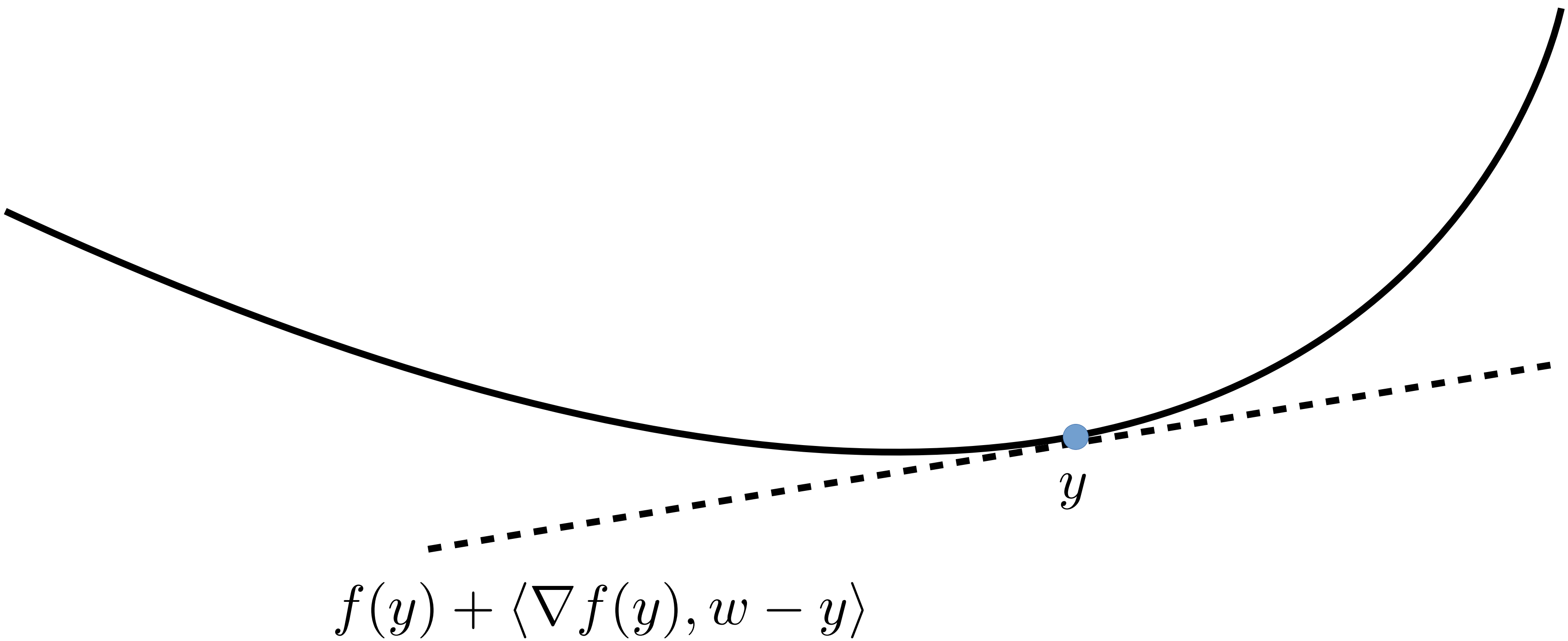
$$f(\lambda w + (1-\lambda)y) \leq \lambda f(w) + (1-\lambda) f(y), \quad \forall w, y \in C, \lambda \in [0, 1]$$



Convexity: First derivative

A differentiable function $f : \mathrm{dom}(f) \subset \mathbb{R}^n \to \mathbb{R}$ is convex iff

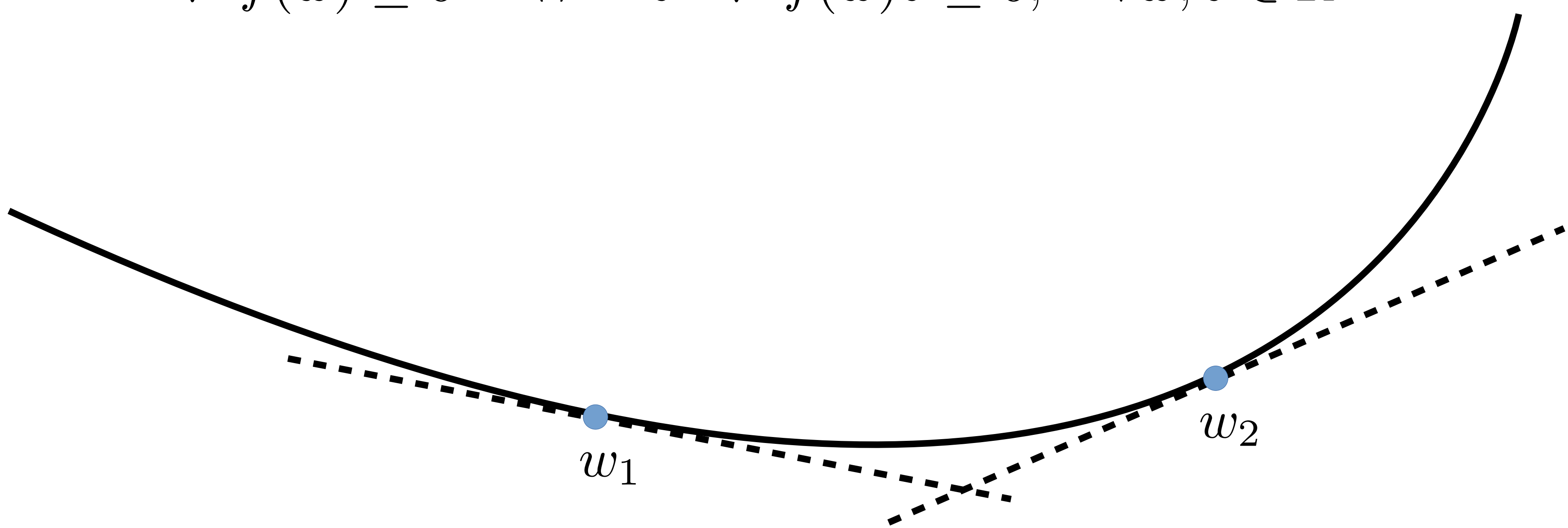
$$f(w) \geq f(y) + \langle \nabla f(y), w - y \rangle$$



Convexity: Second derivative

A twice differentiable function $f : \mathrm{dom}(f) \subset \mathbb{R}^n \to \mathbb{R}$ is convex iff

$$\nabla^2 f(w) \succeq 0 \quad \Leftrightarrow \quad v^\top \nabla^2 f(w) v \geq 0, \quad \forall w, v \in \mathbb{R}^n$$



$$w_1 \leq w_2 \quad \Rightarrow f'(w_1) \leq f'(w_2)$$

Convexity: Examples

Extended-value extension: $$f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$$

$$f(x) = \infty, \quad \forall x \notin \mathrm{dom}(f)$$

Norms and squared norms: $$x \mapsto ||x||$$

$$x \mapsto ||x||^2$$

Proof is in the "Convexity & smoothness" exercise list

Negative log and logistic: $$x \mapsto -\log(x)$$

$$x \mapsto \log\left(1 + e^{-y\langle a, x\rangle}\right)$$

Hinge loss $$x \mapsto \max\{0, 1 - yx\}$$

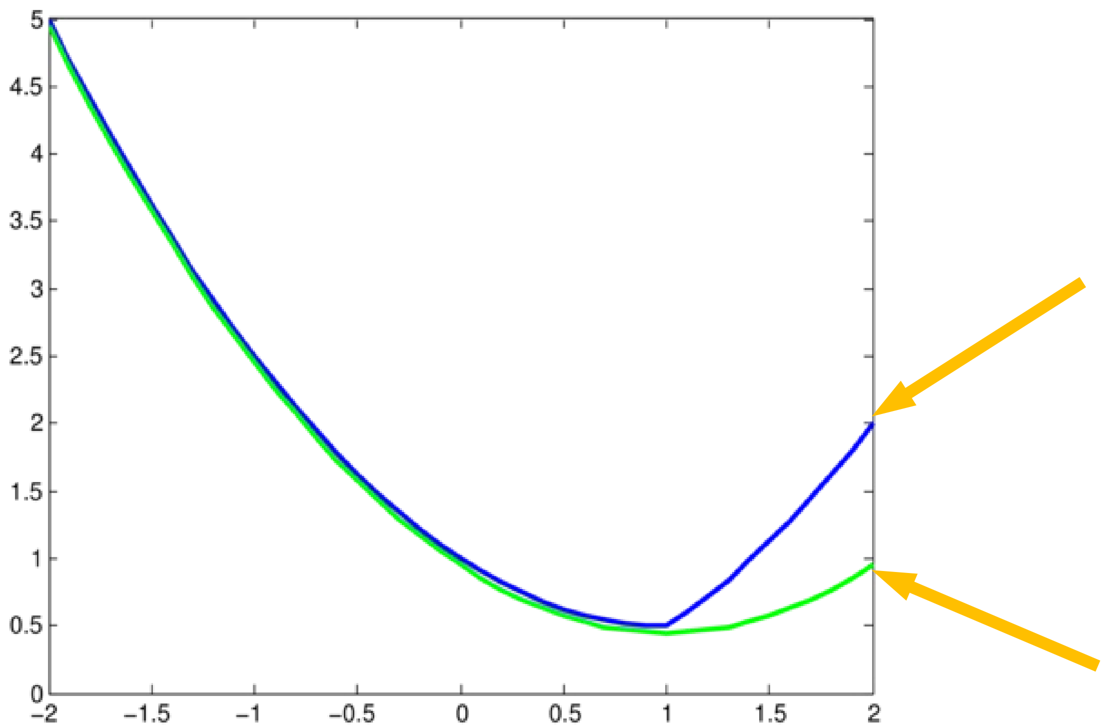
Negatives log determinant, exponentiation … etc

Assumption: Strong convexity

We say $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is μ–strongly convex if

$$f(w) \geq f(y) + \langle \nabla f(y), w - y \rangle + \frac{\mu}{2}||w - y||^2, \quad \forall w, y \in \mathbb{R}^n$$

EXE:



Hinge loss + L2

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

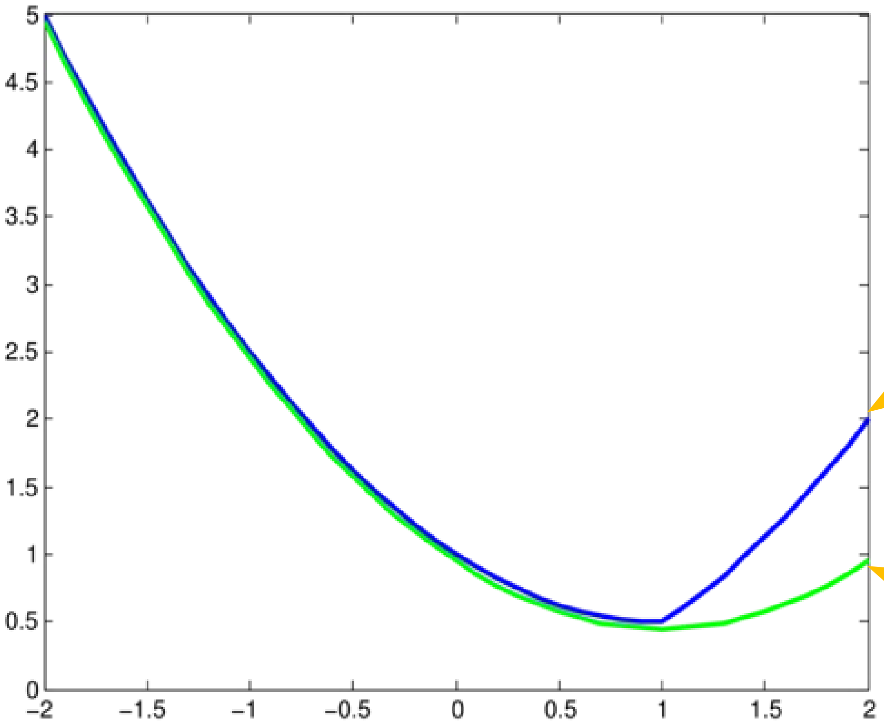
Quadratic lower bound

Assumption: Strong convexity

We say $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is μ–strongly convex if

$$\boxed{f(w)} \geq f(y) + \langle \nabla f(y), w - y \rangle + \frac{\mu}{2}||w-y||^2, \quad \forall w, y \in \mathbb{R}^n$$

EXE:



Hinge loss + L2

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

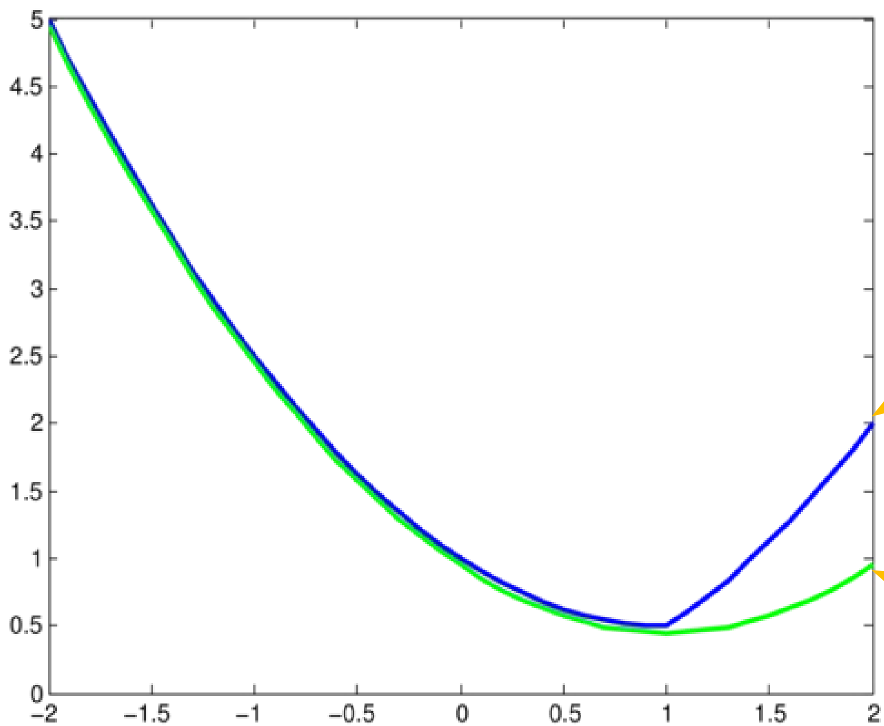
Quadratic lower bound

Assumption: Strong convexity

We say $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is μ–strongly convex if

$$\boxed{f(w)} \geq \boxed{f(y) + \langle \nabla f(y), w - y \rangle + \frac{\mu}{2}||w - y||^2}, \quad \forall w, y \in \mathbb{R}^n$$

EXE:



Hinge loss + L2

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

Quadratic lower bound

Assumption: Strong convexity

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

$$\text{strongly convex} \quad = \quad \text{convex} \quad + \quad \frac{1}{2}\|w\|^2$$

Example: SVM with soft margin

$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \max\{0, 1 - y^i \langle w, x^i \rangle\} + \frac{\lambda}{2} \|w\|_2^2$$

Not an Example: Neural networks, dictionary learning, Matrix completion, and more

Assumption: Smoothness

We say $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is smooth if

$$||\nabla f(x) - \nabla f(y)|| \leq L||x - y||, \quad \forall x, y \in \mathbb{R}^n$$

Assumption: Smoothness

We say $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is smooth if

$$||\nabla f(x) - \nabla f(y)|| \leq L||x - y||, \quad \forall x, y \in \mathbb{R}^n$$

If a twice differentiable $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is L–smooth then

1) $$d^\top \nabla^2 f(x) d \leq L \cdot ||d||_2^2, \quad \forall x, d \in \mathbb{R}^n$$

2) $$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2}||x - y||^2, \quad \forall x, y \in \mathbb{R}^n$$

Assumption: Smoothness

We say $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is smooth if

$$||\nabla f(x) - \nabla f(y)|| \leq L||x - y||, \quad \forall x, y \in \mathbb{R}^n$$

If a twice differentiable $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is L–smooth then

1) $$d^\top \nabla^2 f(x) d \leq L \cdot ||d||_2^2, \quad \forall x, d \in \mathbb{R}^n$$

2) $$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2}||x - y||^2, \quad \forall x, y \in \mathbb{R}^n$$

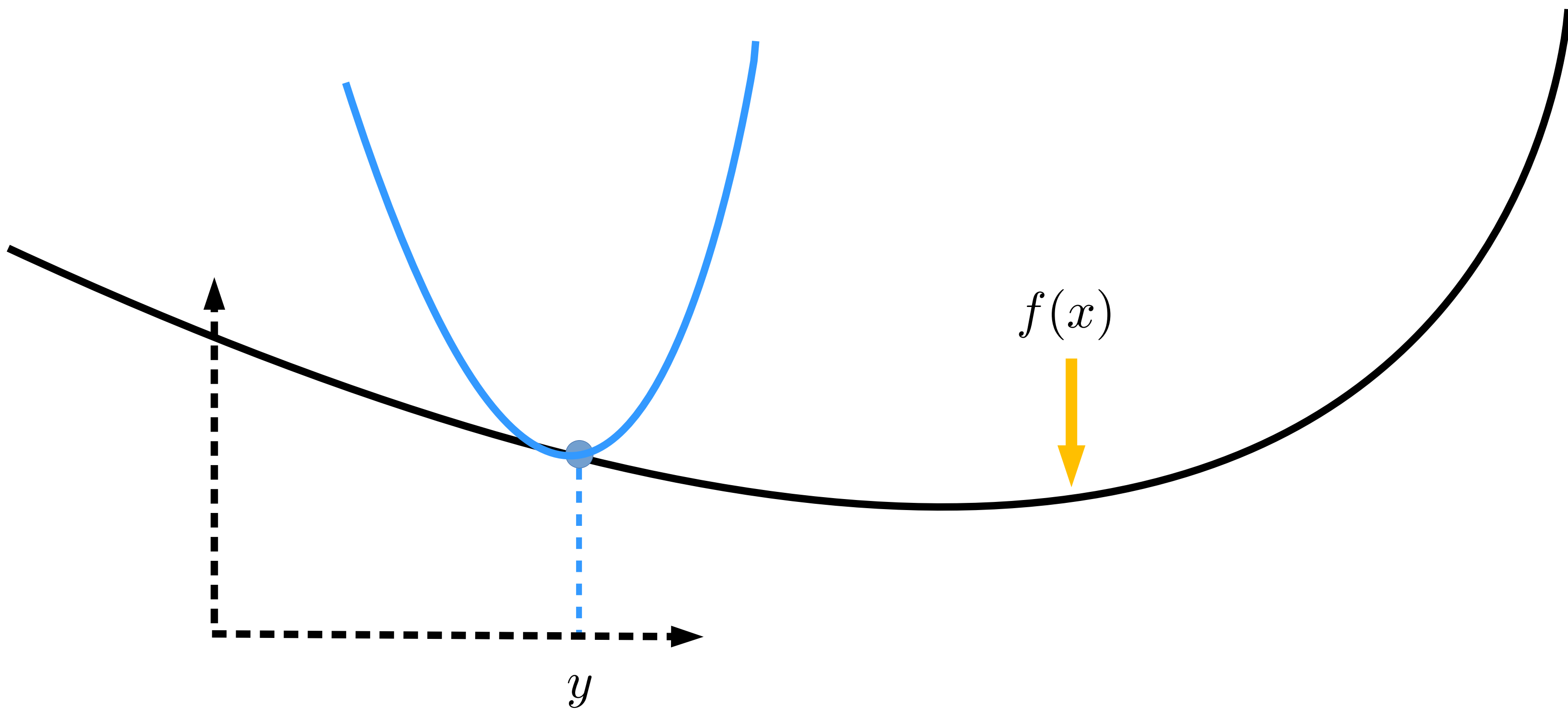
EXE: Using that $$\sigma_{\max}(X)^2 ||d||_2^2 \geq ||X^\top d||_2^2$$

Show that $\frac{1}{2}||X^\top w - b||_2^2$ is $\sigma_{\max}(X)^2$–smooth

Important consequences of Smoothness

If $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is L–smooth then

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2}||x - y||^2, \quad \forall x, y \in \mathbb{R}^n$$



Smoothness: Examples

Convex quadratics: $x \mapsto x^\top A x + b^\top x + c$

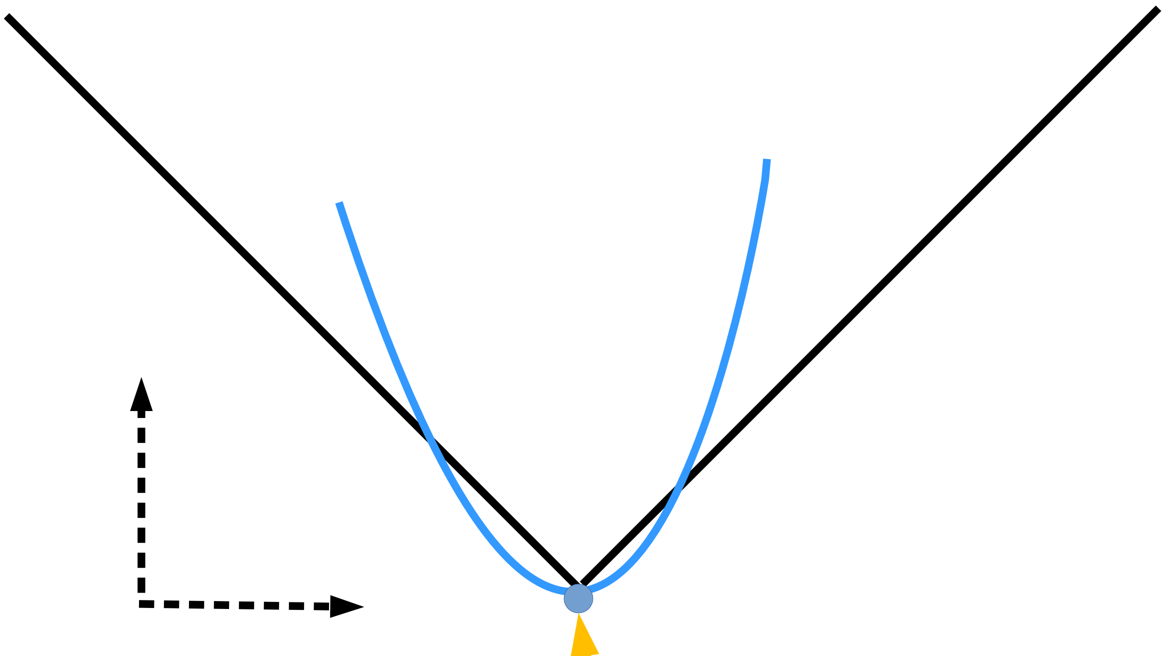
Logistic: $x \mapsto \log\left(1 + e^{-y\langle a,x\rangle}\right)$

Trigonometric: $x \mapsto \cos(x), \sin(x)$

Proof is an exercise!

Smoothness: Convex counter-example

$$f(w) = ||w||_1 = \sum_{i=1}^{n} |w_i|$$



Non-smooth can be solved with proximal SGD

Does not fit.
Not smooth

Gradient Descent via Smoothness

$$\boxed{f(w) \leq f(y) + \langle \nabla f(y), w - y \rangle + \frac{L}{2}||w - y||^2, \quad \forall w, y \in \mathbb{R}^n}$$

Minimizing the upper bound in w we get:

$$\nabla_w \left(f(y) + \langle \nabla f(y), w - y \rangle + \frac{L}{2}||w - y||^2 \right) = \nabla f(y) + L(w - y) = 0$$

Gradient Descent via Smoothness

$$\boxed{f(w) \leq f(y) + \langle \nabla f(y), w - y \rangle + \frac{L}{2}||w - y||^2, \quad \forall w, y \in \mathbb{R}^n}$$

Minimizing the upper bound in w we get:

$$\nabla_w \left(f(y) + \langle \nabla f(y), w - y \rangle + \frac{L}{2}||w - y||^2 \right) = \nabla f(y) + L(w - y) = 0$$

$$\Downarrow$$

$$w = y - \frac{1}{L} \nabla f(y)$$

Gradient Descent via Smoothness

$$\boxed{f(w) \leq f(y) + \langle \nabla f(y), w - y \rangle + \frac{L}{2}||w - y||^2, \quad \forall w, y \in \mathbb{R}^n}$$

Minimizing the upper bound in w we get:

$$\nabla_w \left(f(y) + \langle \nabla f(y), w - y \rangle + \frac{L}{2}||w - y||^2 \right) = \nabla f(y) + L(w - y) = 0$$

A gradient descent step !

$$w = y - \frac{1}{L}\nabla f(y)$$

Gradient Descent via Smoothness

$$\boxed{f(w) \leq f(y) + \langle \nabla f(y), w - y \rangle + \frac{L}{2}||w - y||^2, \quad \forall w, y \in \mathbb{R}^n}$$

Minimizing the upper bound in w we get:

$$\nabla_w \left(f(y) + \langle \nabla f(y), w - y \rangle + \frac{L}{2}||w - y||^2 \right) = \nabla f(y) + L(w - y) = 0$$

Smoothness Lemma (EXE):

If *f* is *L*-smooth, show that

$$f(y - \tfrac{1}{L}\nabla f(y)) - f(y) \leq -\frac{1}{2L}||\nabla f(y)||_2^2, \; \forall y$$

$$f(w^*) - f(w) \leq -\frac{1}{2L}||\nabla f(w)||_2^2, \quad \forall w \in \mathbb{R}^n$$

$$\text{where } f(w^*) \leq f(w), \quad \forall w \in \mathbb{R}^n$$

$$w = y - \frac{1}{L}\nabla f(y)$$

A gradient descent step !

Convergence GD strongly convex

Theorem

Let *f* be μ-strongly convex and *L*-smooth.

$$||w^t - w^*||_2^2 \leq \left(1 - \frac{\mu}{L}\right)^t ||w^1 - w^*||_2^2$$

Where

$$w^{t+1} = w^t - \frac{1}{L}\nabla f(w^t), \quad \text{for } t = 1, \dots, T$$

Convergence GD strongly convex

Theorem

Let f be μ-strongly convex and L-smooth.

$$||w^t - w^*||_2^2 \leq \left(1 - \frac{\mu}{L}\right)^t ||w^1 - w^*||_2^2$$

Where

$$w^{t+1} = w^t - \frac{1}{L}\nabla f(w^t), \quad \text{for } t = 1, \ldots, T$$

$$\Rightarrow \text{for } \frac{||w^T - w^*||_2^2}{||w^1 - w^*||_2^2} \leq \epsilon \text{ we need } T \geq \frac{L}{\mu}\log\left(\frac{1}{\epsilon}\right) = O\left(\log\left(\frac{1}{\epsilon}\right)\right)$$

Convergence GD strongly convex

Theorem

Let f be μ-strongly convex and L-smooth.

$$||w^t - w^*||_2^2 \leq \left(1 - \frac{\mu}{L}\right)^t ||w^1 - w^*||_2^2$$

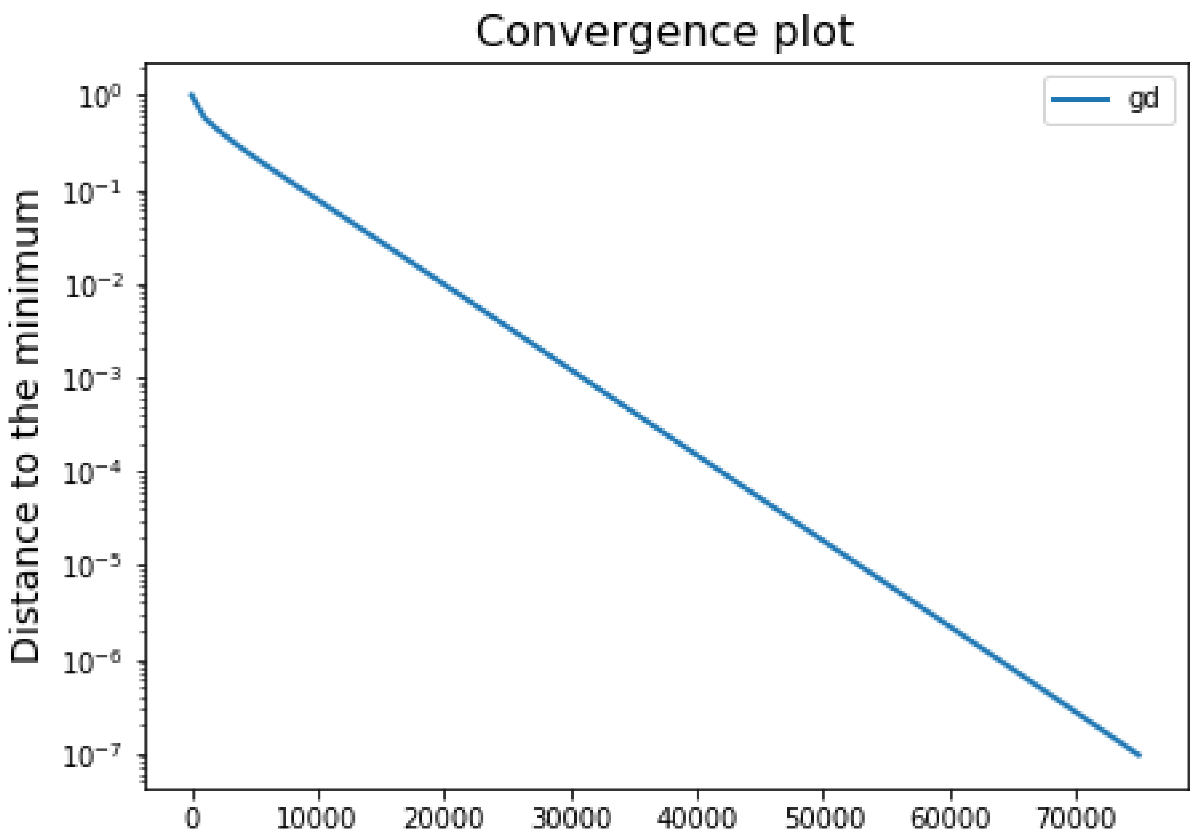
Where

$$w^{t+1} = w^t - \frac{1}{L}\nabla f(w^t), \quad \text{for } t = 1, \ldots, T$$

$$\Rightarrow \text{for } \frac{||w^T - w^*||_2^2}{||w^1 - w^*||_2^2} \leq \epsilon \text{ we need } T \geq \frac{L}{\mu}\log\left(\frac{1}{\epsilon}\right) = O\left(\log\left(\frac{1}{\epsilon}\right)\right)$$

EXE: Solve the questions in "Complexity rates.pdf"

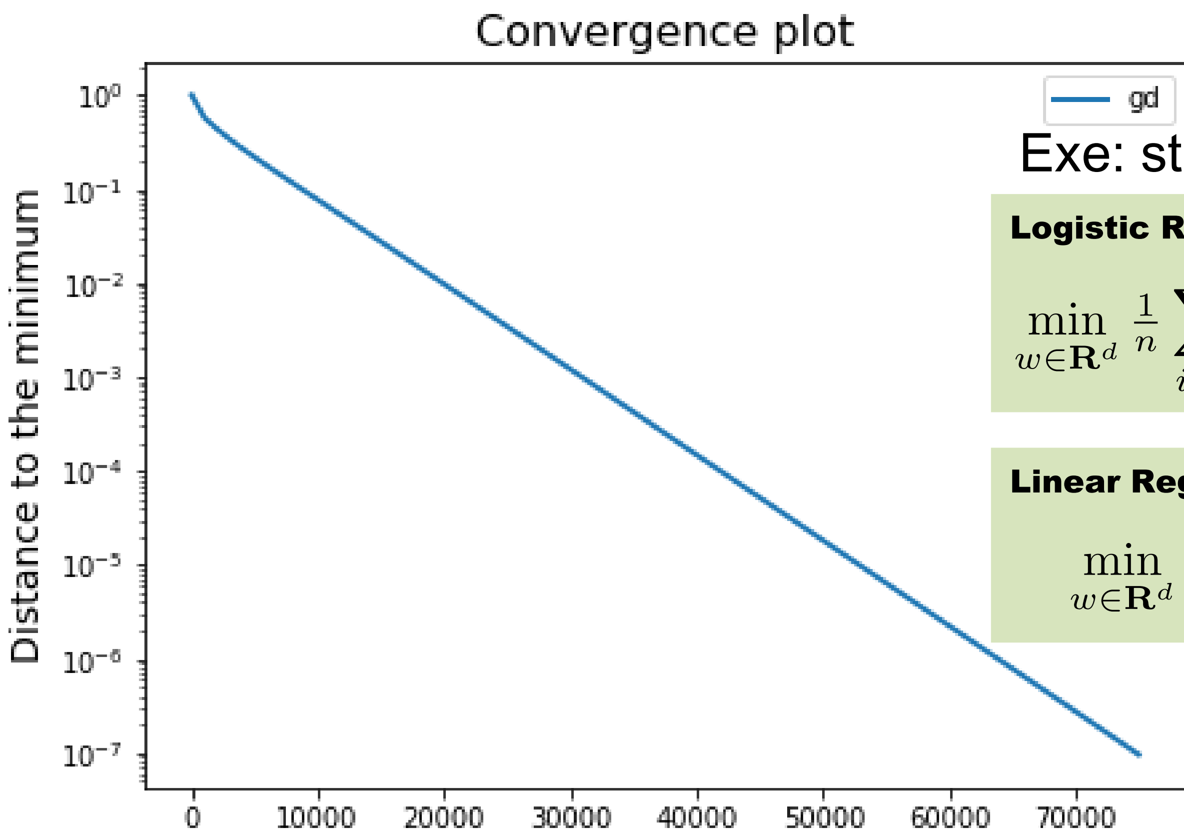
Gradient Descent Example: logistic



$$y\text{–axis} = \frac{||w^t - w^*||_2^2}{||w^1 - w^*||_2^2}$$

$$\log\left(\frac{||w^t - w^*||_2^2}{||w^1 - w^*||_2^2}\right) \leq t \log\left(1 - \frac{\mu}{L}\right)$$

Gradient Descent Example: logistic



Exe: strongly convex + smooth

Logistic Regression

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

Linear Regression

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

$$y\text{–axis} = \frac{||w^t - w^*||_2^2}{||w^1 - w^*||_2^2} \quad \Rightarrow \quad \log\left(\frac{||w^t - w^*||_2^2}{||w^1 - w^*||_2^2}\right) \leq t \log\left(1 - \frac{\mu}{L}\right)$$

Proof Convergence GD strongly convex + smooth

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

Proof:

$$||w^{t+1} - w^*||_2^2 = ||w^t - w^* - \frac{1}{L}\nabla f(w^t)||_2^2$$

$$= ||w^t - w^*||_2^2 + \frac{2}{L}\langle \nabla f(w^t), w^* - w^t\rangle + \frac{1}{L^2}||\nabla f(w^t)||_2^2$$

Proof Convergence GD strongly convex + smooth

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

Proof:

$$||w^{t+1} - w^*||_2^2 = ||w^t - w^* - \frac{1}{L}\nabla f(w^t)||_2^2$$

$$= ||w^t - w^*||_2^2 + \frac{2}{L}\boxed{\langle \nabla f(w^t), w^* - w^t \rangle} + \frac{1}{L^2}||\nabla f(w^t)||_2^2$$

Proof Convergence GD strongly convex + smooth

Proof:

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

$$||w^{t+1} - w^*||_2^2 = ||w^t - w^* - \frac{1}{L}\nabla f(w^t)||_2^2$$

$$= ||w^t - w^*||_2^2 + \frac{2}{L}\boxed{\langle \nabla f(w^t), w^* - w^t \rangle} + \frac{1}{L^2}||\nabla f(w^t)||_2^2$$

Strong convexity:

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

$$\langle \nabla f(w), w^* - w \rangle \leq -\frac{\mu}{2}||w - w^*||^2 - (f(w) - f(w^*))$$

Proof Convergence GD strongly convex + smooth

Proof:

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

$$||w^{t+1} - w^*||_2^2 = ||w^t - w^* - \frac{1}{L}\nabla f(w^t)||_2^2$$

$$= ||w^t - w^*||_2^2 + \frac{2}{L}\boxed{\langle \nabla f(w^t), w^* - w^t \rangle} + \frac{1}{L^2}||\nabla f(w^t)||_2^2$$

Strong convexity:

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

$$\langle \nabla f(w), w^* - w \rangle \leq -\frac{\mu}{2}||w - w^*||^2 - (f(w) - f(w^*))$$

$$||w^{t+1} - w^*||_2^2 \leq \left(1 - \frac{\mu}{L}\right)\|w^t - w^*\|^2 - \frac{2}{L}(f(w^t) - f(w^*)) + \frac{1}{L^2}\|\nabla f(w^t)\|^2$$

Proof Convergence GD strongly convex + smooth

$$||w^{t+1} - w^*||_2^2 \leq \left(1 - \frac{\mu}{L}\right) \|w^t - w^*\|^2 - \frac{2}{L}(f(w^t) - f(w^*)) + \frac{1}{L^2}\boxed{\|\nabla f(w^t)\|^2}$$

Proof Convergence GD strongly convex + smooth

$$||w^{t+1} - w^*||_2^2 \leq \left(1 - \frac{\mu}{L}\right) \|w^t - w^*\|^2 - \frac{2}{L}(f(w^t) - f(w^*)) + \frac{1}{L^2}\boxed{\|\nabla f(w^t)\|^2}$$

Smoothness Lemma (EXE):

$$f(w^*) - f(w) \leq -\frac{1}{2L}||\nabla f(w)||_2^2$$

$$\Rightarrow \quad ||\nabla f(w)||_2^2 \leq 2L(f(w) - f(w^*))$$

Proof Convergence GD strongly convex + smooth

$$||w^{t+1} - w^*||_2^2 \leq \left(1 - \frac{\mu}{L}\right) \|w^t - w^*\|^2 - \frac{2}{L}(f(w^t) - f(w^*)) + \frac{1}{L^2}\boxed{\|\nabla f(w^t)\|^2}$$

Smoothness Lemma (EXE):

$$f(w^*) - f(w) \leq -\frac{1}{2L}||\nabla f(w)||_2^2$$

$$||\nabla f(w)||_2^2 \leq 2L(f(w) - f(w^*))$$

$$||w^{t+1} - w^*||_2^2 \leq \left(1 - \frac{\mu}{L}\right) \|w^t - w^*\|^2 - \frac{2}{L}(f(w^t) - f(w^*)) + \frac{2}{L}(f(w^t) - f(w^*))$$

$$= \left(1 - \frac{\mu}{L}\right) \|w^t - w^*\|^2 \quad \blacksquare$$

Proof Convergence GD strongly convex + smooth

$$||w^{t+1} - w^*||_2^2 \leq \left(1 - \frac{\mu}{L}\right) \|w^t - w^*\|^2 - \frac{2}{L}(f(w^t) - f(w^*)) + \frac{1}{L^2}\boxed{\|\nabla f(w^t)\|^2}$$

Smoothness Lemma (EXE):

$$f(w^*) - f(w) \leq -\frac{1}{2L}||\nabla f(w)||_2^2$$

$$||\nabla f(w)||_2^2 \leq 2L(f(w) - f(w^*))$$

$$||w^{t+1} - w^*||_2^2 \leq \left(1 - \frac{\mu}{L}\right) \|w^t - w^*\|^2 - \frac{2}{L}(f(w^t) - f(w^*)) + \frac{2}{L}(f(w^t) - f(w^*))$$

$$= \left(1 - \frac{\mu}{L}\right) \|w^t - w^*\|^2 \quad \blacksquare$$

(EXE): Repeat proof for $w^{t+1} = w^t - \alpha \nabla f(w^t)$ where $\alpha > 0$.

For what values of α does $w^t \to w^*$ converge?

Convergence GD for smooth + convex

Theorem

Let f be convex and L-smooth.

$$f(w^t) - f(w^*) \leq \frac{2L||w^1 - w^*||_2^2}{t-1} = O\left(\frac{1}{t}\right).$$

Where

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

$$\Rightarrow \text{for } \frac{f(w^T) - f(w^*)}{||w^1 - w^*||_2^2} \leq \epsilon \text{ we need } T \geq \frac{2L}{\epsilon} = O\left(\frac{1}{\epsilon}\right)$$

Convex and Smooth Properties

Co-coercivity Lemma

If $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ convex and L–smooth then

$$f(y) - f(x) \leq \langle \nabla f(y), y - x \rangle - \frac{1}{2L} ||\nabla f(y) - \nabla f(x)||_2^2$$

$$\text{and} \quad \langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \frac{1}{L} ||\nabla f(x) - \nabla f(y)||_2$$

Proof:

Adding together the last two inequalities gives the result.

Convex and Smooth Properties

Co-coercivity Lemma

If $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ convex and L–smooth then

$$f(y) - f(x) \leq \langle \nabla f(y), y - x \rangle - \frac{1}{2L} ||\nabla f(y) - \nabla f(x)||_2^2$$

$$\text{and} \quad \langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \frac{1}{L} ||\nabla f(x) - \nabla f(y)||_2$$

Proof: $$f(y) - f(x) = \overbrace{f(y) - f(z)}^{\text{Use convexity}} + \overbrace{f(z) - f(x)}^{\text{Use smoothness}}$$

Adding together the last two inequalities gives the result.

Convex and Smooth Properties

Co-coercivity Lemma

If $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ convex and L–smooth then

$$f(y) - f(x) \leq \langle \nabla f(y), y - x \rangle - \frac{1}{2L} ||\nabla f(y) - \nabla f(x)||_2^2$$

$$\text{and} \qquad \langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \frac{1}{L} ||\nabla f(x) - \nabla f(y)||_2$$

Proof:

$$f(y) - f(x) = \overbrace{f(y) - f(z)}^{\text{Use convexity}} + \overbrace{f(z) - f(x)}^{\text{Use smoothness}}$$

$$\leq \langle \nabla f(y), y - z \rangle + \langle \nabla f(x), z - x \rangle + \frac{L}{2} ||z - x||^2, \quad \forall z$$

Adding together the last two inequalities gives the result.

Convex and Smooth Properties

Co-coercivity Lemma

If $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ convex and L–smooth then

$$f(y) - f(x) \leq \langle \nabla f(y), y - x \rangle - \frac{1}{2L} ||\nabla f(y) - \nabla f(x)||_2^2$$

$$\text{and} \quad \langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \frac{1}{L} ||\nabla f(x) - \nabla f(y)||_2$$

Proof:

$$f(y) - f(x) = \overbrace{f(y) - f(z)}^{\text{Use convexity}} + \overbrace{f(z) - f(x)}^{\text{Use smoothness}}$$

$$\leq \langle \nabla f(y), y - z \rangle + \langle \nabla f(x), z - x \rangle + \frac{L}{2} ||z - x||^2, \quad \forall z$$

Minimizing in z gives: $$z = x - \frac{1}{L}(\nabla f(x) - \nabla f(y)).$$

Adding together the last two inequalities gives the result.

Convex and Smooth Properties

Co-coercivity Lemma

If $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ convex and L–smooth then

$$f(y) - f(x) \leq \langle \nabla f(y), y - x \rangle - \frac{1}{2L} ||\nabla f(y) - \nabla f(x)||_2^2$$

$$\text{and} \qquad \langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \frac{1}{L} ||\nabla f(x) - \nabla f(y)||_2$$

Proof:
$$f(y) - f(x) = \overbrace{f(y) - f(z)}^{\text{Use convexity}} + \overbrace{f(z) - f(x)}^{\text{Use smoothness}}$$

$$\leq \langle \nabla f(y), y - z \rangle + \langle \nabla f(x), z - x \rangle + \frac{L}{2} ||z - x||^2, \quad \forall z$$

Minimizing in *z* gives: $$z = x - \frac{1}{L}(\nabla f(x) - \nabla f(y)).$$

Inserting this *z* in bound (and after some computations) gives:

$$f(y) - f(x) \leq \langle \nabla f(y), y - x \rangle - \frac{1}{2L} ||\nabla f(y) - \nabla f(x)||_2^2$$

Adding together the last two inequalities gives the result.

Convex and Smooth Properties

Co-coercivity Lemma

If $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ convex and L–smooth then

$$f(y) - f(x) \leq \langle \nabla f(y), y - x\rangle - \frac{1}{2L}||\nabla f(y) - \nabla f(x)||_2^2$$

$$\text{and} \qquad \langle \nabla f(y) - \nabla f(x), y - x\rangle \geq \frac{1}{L}||\nabla f(x) - \nabla f(y)||_2$$

Proof:

$$f(y) - f(x) = \overbrace{f(y) - f(z)}^{\text{Use convexity}} + \overbrace{f(z) - f(x)}^{\text{Use smoothness}}$$

$$\leq \langle \nabla f(y), y - z\rangle + \langle \nabla f(x), z - x\rangle + \frac{L}{2}||z - x||^2, \quad \forall z$$

Minimizing in *z* gives:
$$z = x - \frac{1}{L}(\nabla f(x) - \nabla f(y)).$$

Inserting this *z* in bound (and after some computations) gives:

$$f(y) - f(x) \leq \langle \nabla f(y), y - x\rangle - \frac{1}{2L}||\nabla f(y) - \nabla f(x)||_2^2$$

Switching *x* for *y* gives:

$$f(x) - f(y) \leq \langle \nabla f(x), x - y\rangle - \frac{1}{2L}||\nabla f(y) - \nabla f(x)||_2^2$$

Adding together the last two inequalities gives the result.

Proof *Sketch* of GD smooth + convex

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

Proof *Sketch* of GD smooth + convex

$$
\begin{aligned}
||w^{t+1} - w^*||_2^2 &= ||w^t - w^* - \frac{1}{L}\nabla f(w^t)||_2^2 \\
&= ||w^t - w^*||_2^2 + \frac{2}{L}\boxed{\langle \nabla f(w^t), w^* - w^t \rangle} + \frac{1}{L^2}||\nabla f(w^t)||_2^2
\end{aligned}
$$

Proof *Sketch* of GD smooth + convex

$$||w^{t+1} - w^*||_2^2 = ||w^t - w^* - \frac{1}{L}\nabla f(w^t)||_2^2$$

$$= ||w^t - w^*||_2^2 + \frac{2}{L}\boxed{\langle \nabla f(w^t), w^* - w^t\rangle} + \frac{1}{L^2}||\nabla f(w^t)||_2^2$$

Use co-coercivity

Proof *Sketch* of GD smooth + convex

$$||w^{t+1} - w^*||_2^2 = ||w^t - w^* - \frac{1}{L}\nabla f(w^t)||_2^2$$

$$= ||w^t - w^*||_2^2 + \frac{2}{L}\boxed{\langle \nabla f(w^t), w^* - w^t \rangle} + \frac{1}{L^2}||\nabla f(w^t)||_2^2$$

Use co-coercivity

Co-coercivity: $\langle \nabla f(y) - \nabla f(w), y - w \rangle \geq \frac{1}{L}||\nabla f(w) - \nabla f(y)||_2$

$$\text{With } y = w^* \text{ gives } \langle \nabla f(w), w^* - w \rangle \leq -\frac{1}{L}||\nabla f(w)||_2$$

Proof *Sketch* of GD smooth + convex

$$||w^{t+1} - w^*||_2^2 = ||w^t - w^* - \frac{1}{L}\nabla f(w^t)||_2^2$$

$$= ||w^t - w^*||_2^2 + \frac{2}{L}\boxed{\langle \nabla f(w^t), w^* - w^t\rangle} + \frac{1}{L^2}||\nabla f(w^t)||_2^2$$

Use co-coercivity

Co-coercivity: $\langle \nabla f(y) - \nabla f(w), y - w\rangle \geq \frac{1}{L}||\nabla f(w) - \nabla f(y)||_2$

$$\text{With } y = w^* \text{ gives } \langle \nabla f(w), w^* - w\rangle \leq -\frac{1}{L}||\nabla f(w)||_2$$

Inserting above shows decreasing

$$||w^{t+1} - w^*||_2^2 \leq ||w^t - w^*||_2^2 - \frac{1}{L^2}||\nabla f(w^t)||_2^2$$

Thus $\|w^t - w^*\|$ is a decreasing sequence :

$$\|w^{t+1} - w^*\| \leq \|w^t - w^*\| \leq \cdots \leq \|w^1 - w^*\|$$

Proof *Sketch* of GD smooth + convex

Decreasing: $\|w^{t+1} - w^*\| \leq \|w^t - w^*\| \leq \cdots \leq \|w^1 - w^*\|$

Proof *Sketch* of GD smooth + convex

Decreasing: $\|w^{t+1} - w^*\| \leq \|w^t - w^*\| \leq \cdots \leq \|w^1 - w^*\|$

Subtracting *f(w*) =f** from the Smoothness Lemma bound gives

$$f(w^{t+1}) - f^* \leq f(w^t) - f^* - \frac{1}{2L}||\nabla f(w^t)||_2^2$$

Proof *Sketch* of GD smooth + convex

Decreasing: $\|w^{t+1} - w^*\| \leq \|w^t - w^*\| \leq \cdots \leq \|w^1 - w^*\|$

Subtracting *f(w*) =f** from the Smoothness Lemma bound gives

$$f(w^{t+1}) - f^* \leq f(w^t) - f^* \boxed{- \frac{1}{2L}||\nabla f(w^t)||_2^2}$$

Proof *Sketch* of GD smooth + convex

Decreasing: $\|w^{t+1} - w^*\| \leq \|w^t - w^*\| \leq \cdots \leq \|w^1 - w^*\|$

Subtracting *f(w*) =f** from the Smoothness Lemma bound gives

$$f(w^{t+1}) - f^* \leq f(w^t) - f^* \boxed{- \frac{1}{2L} ||\nabla f(w^t)||_2^2}$$

Using convexity:

$$\begin{aligned} f(w^t) - f^* &\leq \langle \nabla f(w^t), w^t - w^* \rangle \\ &\leq ||\nabla f(w^t)||_2 ||w^t - w^*||_2 \\ &\leq ||\nabla f(w^t)||_2 ||w^1 - w^*||_2 \end{aligned}$$

$$\Rightarrow \quad -||\nabla f(w^t)||_2 \leq -\frac{f(w^t) - f^*}{||w^t - w^*||_2}$$

Proof *Sketch* of GD smooth + convex

Decreasing: $\|w^{t+1} - w^*\| \le \|w^t - w^*\| \le \cdots \le \|w^1 - w^*\|$

Subtracting *f(w*) =f** from the Smoothness Lemma bound gives

$$f(w^{t+1}) - f^* \le f(w^t) - f^* \boxed{- \frac{1}{2L}||\nabla f(w^t)||_2^2}$$

Using convexity:

$$\begin{aligned} f(w^t) - f^* &\le \langle \nabla f(w^t), w^t - w^* \rangle \\ &\le ||\nabla f(w^t)||_2 ||w^t - w^*||_2 \\ &\le ||\nabla f(w^t)||_2 ||w^1 - w^*||_2 \end{aligned} \Rightarrow -||\nabla f(w^t)||_2 \le -\frac{f(w^t) - f^*}{||w^t - w^*||_2}$$

Returning to smoothness bound

$$f(w^{t+1}) - f^* \le f(w^t) - f^* - \frac{1}{2L} \frac{(f(w^t) - f^*)^2}{\|w^t - w^1\|^2}$$

See “Gradient convergence notes.pdf” for a solution to this nonlinear recurrence relation of the form $\delta_{t+1} \le \delta_t - C\delta_t^2$

Acceleration and lower bounds

The Accelerated gradient method

$$\min_{w \in \mathbb{R}^d} f(w)$$

Let *f* be μ-strongly convex and *L*-smooth.

Accelerated gradient for strong convex

Set $w^1 = 0 = y^1$

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

$$y^{t+1} = w^t - \frac{1}{L}\nabla f(w^t)$$

$$w^{t+1} = y^{t+1} + \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}\right)(y^{t+1} - w^t)$$

Output w^{T+1}

The Accelerated gradient method

$$\min_{w \in \mathbb{R}^d} f(w)$$

Let *f* be μ-strongly convex and *L*-smooth.

Accelerated gradient for strong convex

Set $w^1 = 0 = y^1$

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

$$y^{t+1} = w^t - \frac{1}{L}\nabla f(w^t)$$

$$w^{t+1} = y^{t+1} + \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}\right)(y^{t+1} - w^t)$$

Output w^{T+1}

Weird extrapolation, but it works

Convergence lower bounds strongly convex

Yuri Nesterov (1998), Springer Publishing, **Introductory Lectures on Convex Optimization: A Basic Course**

Theorem (Nesterov)

For any optimization algorithm where

$$w^{t+1} \in w^t + \text{span}\left(\nabla f(w^1), \nabla f(w^2), \ldots, \nabla f(w^t)\right)$$

There exists a function $f(w)$ that is L–smooth and μ–strongly convex such that

$$f(w^T) - f(w^*) \geq \frac{\mu}{2}\left(1 - \frac{2}{\sqrt{\kappa + 1}}\right)^{2(T-1)} ||w^1 - w^*||_2^2$$

$$= O\left(\left(1 - \frac{1}{\sqrt{\kappa}}\right)^{2T}\right)$$

$\kappa := \frac{L}{\mu}$

Accelerated gradient has this rate!

Convergence lower bounds strongly convex

Yuri Nesterov (1998), Springer Publishing, **Introductory Lectures on Convex Optimization: A Basic Course**

Theorem (Nesterov)

For any optimization algorithm where

$$w^{t+1} \in w^t + \mathrm{span}\left(\nabla f(w^1), \nabla f(w^2), \ldots, \nabla f(w^t)\right)$$

There exists a function $f(w)$ that is L–smooth and μ–strongly convex such that

$$f(w^T) - f(w^*) \geq \frac{\mu}{2}\left(1 - \frac{2}{\sqrt{\kappa+1}}\right)^{2(T-1)} ||w^1 - w^*||_2^2$$

$$= O\left(\left(1 - \frac{1}{\sqrt{\kappa}}\right)^{2T}\right)$$

$\kappa := \frac{L}{\mu}$

Accelerated gradient has this rate!

Convergence lower bounds strongly convex

Yuri Nesterov (1998), Springer Publishing, **Introductory Lectures on Convex Optimization: A Basic Course**

Theorem (Nesterov)

For any optimization algorithm where

$$w^{t+1} \in w^t + \operatorname{span}\left(\nabla f(w^1), \nabla f(w^2), \ldots, \nabla f(w^t)\right)$$

There exists a function $f(w)$ that is L–smooth and μ–strongly convex such that

$$f(w^T) - f(w^*) \geq \frac{\mu}{2}\left(1 - \frac{2}{\sqrt{\kappa+1}}\right)^{2(T-1)} ||w^1 - w^*||_2^2$$

$$\kappa := \frac{L}{\mu}$$

$$= O\left(\left(1 - \frac{1}{\sqrt{\kappa}}\right)^{2T}\right)$$

Accelerated gradient has this rate!

$$\Rightarrow \text{for } \frac{||w^T - w^*||_2^2}{||w^1 - w^*||_2^2} \leq \epsilon \text{ we need } T \geq \sqrt{\frac{L}{\mu}} \log\left(\frac{1}{\epsilon}\right) = O\left(\log\left(\frac{1}{\epsilon}\right)\right)$$

The Accelerated gradient method

$$\min_{w \in \mathbb{R}^d} f(w)$$

Let *f* be convex and *L*-smooth.

Accelerated gradient for convex

Set $w^1 = 0 = y^1, \alpha^1 = 1$

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

$$y^{t+1} = w^t - \frac{1}{L}\nabla f(w^t)$$

$$\alpha^{t+1} = \frac{1+\sqrt{1+\alpha^t}}{2}$$

$$w^{t+1} = y^{t+1} + \left(\frac{\alpha^t - 1}{\alpha^{t+1}}\right)(y^{t+1} - w^t)$$

Output w^{T+1}

The Accelerated gradient method

$$\min_{w \in \mathbb{R}^d} f(w)$$

Let *f* be convex and *L*-smooth.

Accelerated gradient for convex

Set $w^1 = 0 = y^1, \alpha^1 = 1$

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

$$y^{t+1} = w^t - \frac{1}{L}\nabla f(w^t)$$

$$\alpha^{t+1} = \frac{1+\sqrt{1+\alpha^t}}{2}$$

$$w^{t+1} = y^{t+1} + \left(\frac{\alpha^t - 1}{\alpha^{t+1}}\right)(y^{t+1} - w^t)$$

Output w^{T+1}

Weird extrapolation, but it works

Convergence lower bounds convex

Theorem (Nesterov)

For any optimization algorithm where

$$w^{t+1} \in w^t + \mathrm{span}\left(\nabla f(w^1), \nabla f(w^2), \ldots, \nabla f(w^t)\right)$$

There exists a function $f(w)$ that is L–smooth and convex such that

Accelerated gradient has this rate!

$$\min_{i=1,\ldots,T} f(w^i) - f(w^*) \geq \frac{3L\|w^1 - w^*\|_2^2}{32(T+1)^2} = O\left(\frac{1}{T^2}\right).$$

Yuri Nesterov (1998), Springer Publishing, **Introductory Lectures on Convex Optimization: A Basic Course**

Convergence lower bounds convex

Theorem (Nesterov)

For any optimization algorithm where

$$w^{t+1} \in w^t + \text{span}\left(\nabla f(w^1), \nabla f(w^2), \ldots, \nabla f(w^t)\right)$$

There exists a function $f(w)$ that is L–smooth and convex

Accelerated gradient has this rate!

such that

$$\min_{i=1,\ldots,T} f(w^i) - f(w^*) \geq \frac{3L\|w^1 - w^*\|_2^2}{32(T+1)^2} = O\left(\frac{1}{T^2}\right).$$

Yuri Nesterov (1998), Springer Publishing, **Introductory Lectures on Convex Optimization: A Basic Course**

Exercises !

Solve **Exercises lists:**

- Complexity rates
- Convexity & smoothness
- Ridge Regression

> gowerrobert.github.io <

Exercises !

Solve **Exercises lists:**

- Complexity rates
- Convexity & smoothness
- Ridge Regression

> gowerrobert.github.io <

Part III: Stochastic Gradient Descent

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*, …, *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*, …, *n*} selected uniformly at random. Then

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

$$\text{Use } \nabla f_j(w) \approx \nabla f(w)$$

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, …, 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

More reason why ML likes SGD

The training problem

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

More reason why ML likes SGD

The training problem

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

But we already know these labels

More reason why ML likes SGD

The training problem

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

Test problem

But we already know these labels

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 be applied to the statistical learning problem!

Why Machine Learners like SGD

The statistical learning problem:

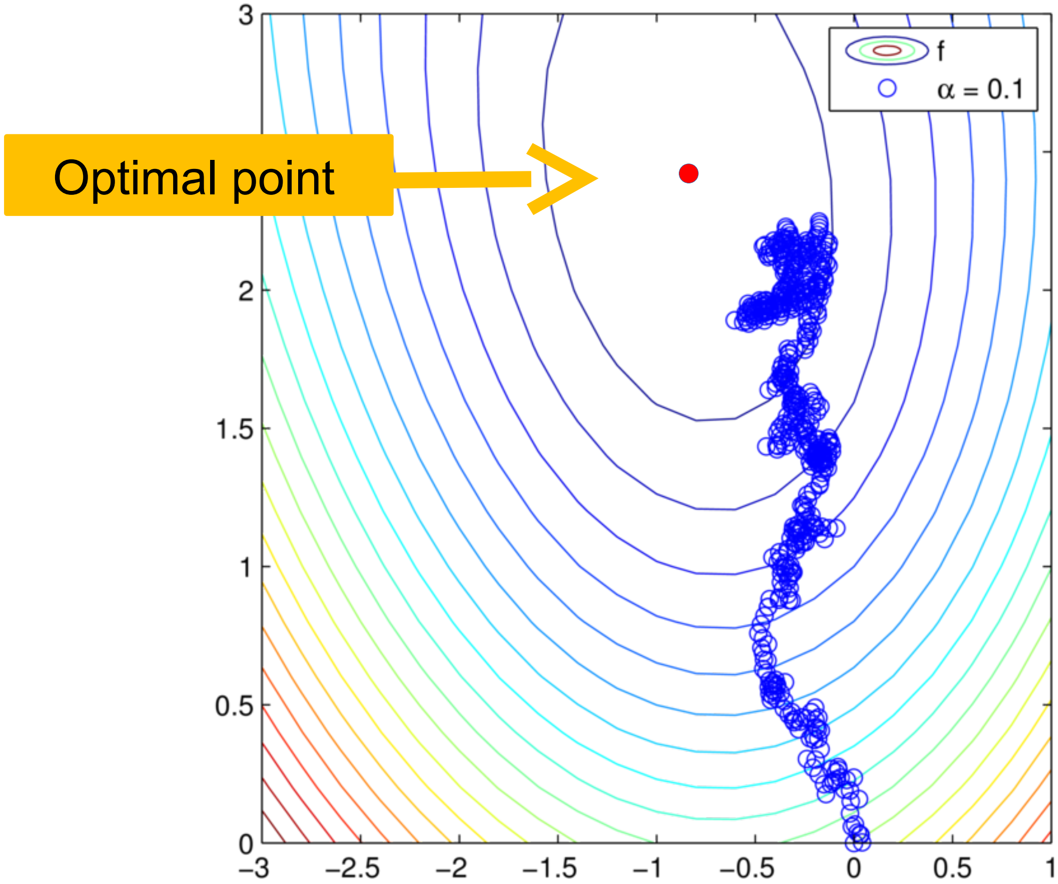
Minimize the expected loss over an *unknown* expectation

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

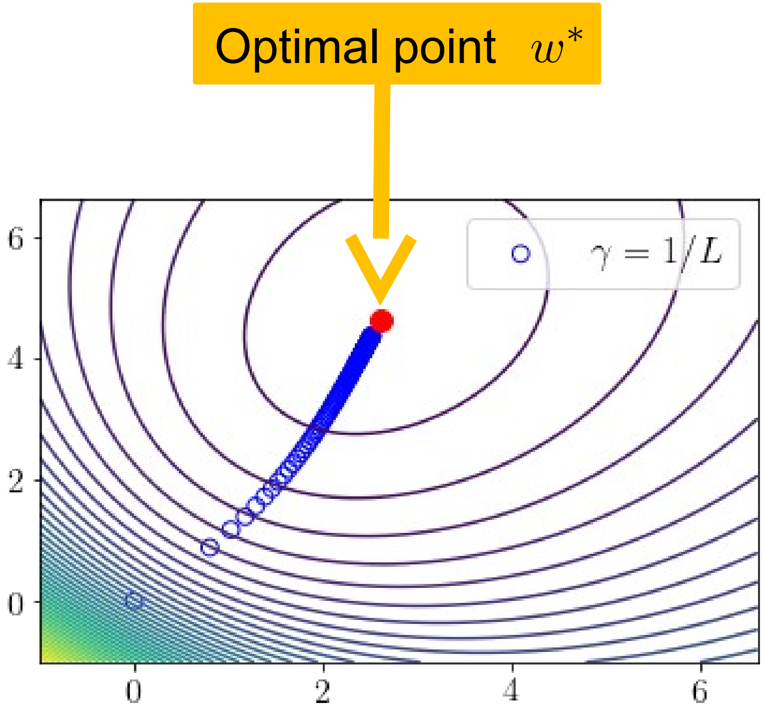
SGD for learning

$$
\begin{aligned}
&\text{Set } w^0 = 0,\ \alpha_t > 0\\
&\text{for } t = 0, 1, 2, \ldots, T-1\\
&\qquad \text{sample } (x, y) \sim \mathcal{D}\\
&\qquad w^{t+1} = w^t - \alpha_t \nabla \ell(h_{w^t}(x), y)\\
&\text{Output } \overline{w}^T = \tfrac{1}{T}\textstyle\sum_{t=1}^T w^t
\end{aligned}
$$

Stochastic Gradient Descent



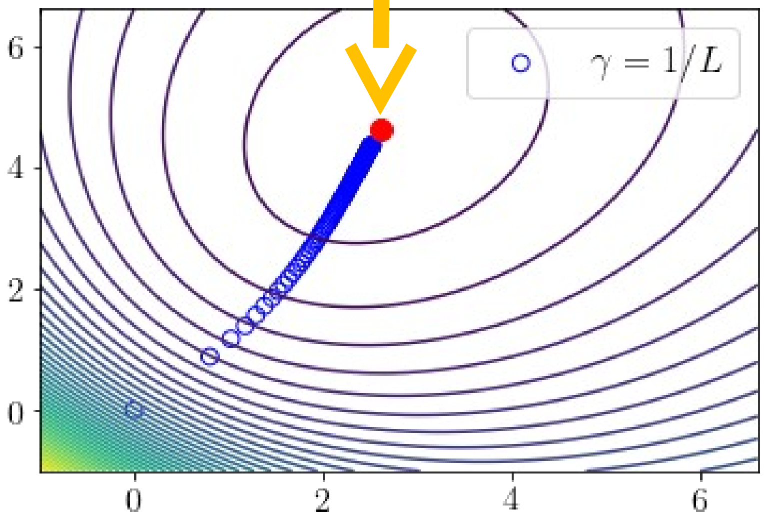
GD vs Stochastic Gradient Descent



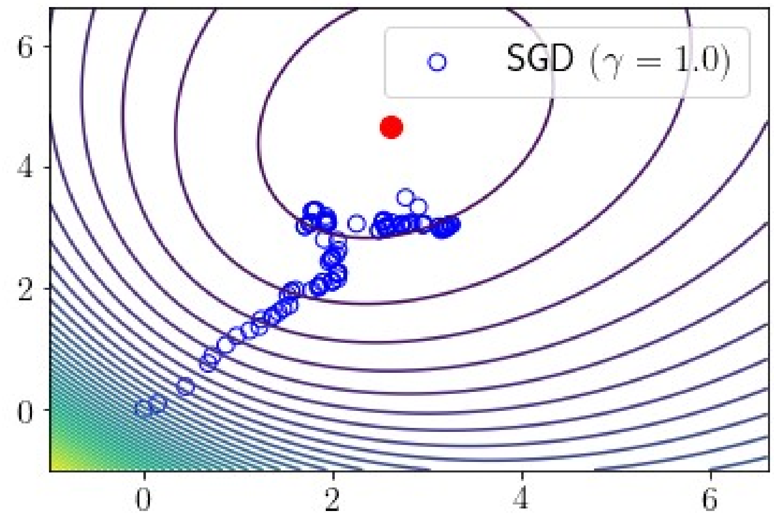
Gradient Descent

GD vs Stochastic Gradient Descent

Optimal point w^*



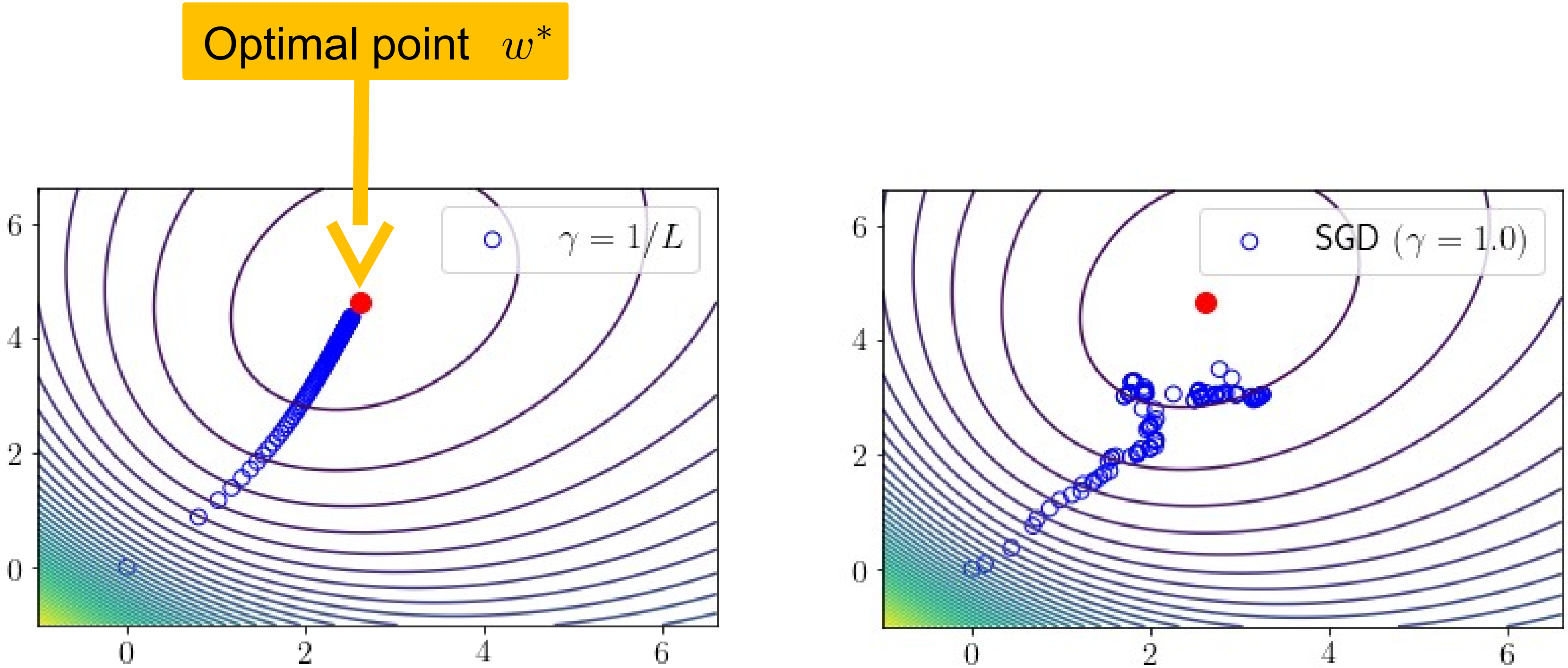
Gradient Descent



Stochastic Gradient Descent

Why does this happen?

GD vs Stochastic Gradient Descent



Gradient Descent

Stochastic Gradient Descent

Why does this happen? → Need Assumptions

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

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

Assumptions for Convergence

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

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

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

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

Assumptions for Convergence

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

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

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

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

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

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

Assumptions for Convergence

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

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

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

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

$$\nabla^2 f_i(w) = x_i x_i^\top + \lambda \quad \preceq \quad (||x_i||_2^2 + \lambda) I \quad = \quad L_i\, I$$

Assumptions for Convergence

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

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

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

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

$$1. \quad f(w) = \frac{1}{2n}||X^\top w - y||_2^2 + \frac{\lambda}{2}||w||_2^2 = \frac{1}{n}\sum_{i=1}^n \left(\frac{1}{2}(x_i^\top w - y_i)^2 + \frac{\lambda}{2}||w||_2^2\right)$$

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

$$\nabla^2 f_i(w) = x_i x_i^\top + \lambda \quad \preceq \quad (||x_i||_2^2 + \lambda)I \quad = \quad L_i\, I$$

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

Assumptions for Convergence

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

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

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

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

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

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

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

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

Complexity / Convergence

Theorem

If f is μ–str. convex, f_i is convex, L_i–smooth, $\alpha \in [0, \frac{1}{2L_{\max}}]$ then the iterates of the SGD satisfy

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

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

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

Shows that $\alpha \approx 0$

PDF Adobe

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

Lemma If $f_i : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$ convex and $L_{\max}$–smooth then

$$\mathbb{E}[\|\nabla f_{j}(w)\|^2] \;\leq\; 4L_{\max}(f(w) - f(w^*)) + 2\sigma^2$$

Proof:

Lemma If $f_i : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$ convex and $L_{\max}$–smooth then

$$\mathbb{E}[\|\nabla f_j(w)\|^2] \; \leq \; 4L_{\max}(f(w) - f(w^*)) + 2\sigma^2$$

Proof:

Co-coercivity Lemma (recall slide 55)

$$f_i(y) - f_i(x) \leq \langle \nabla f_i(y), y - x \rangle - \frac{1}{2L_{\max}} ||\nabla f_i(y) - \nabla f_i(x)||_2^2$$

Lemma If $f_i : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ convex and $L_{\max}$–smooth then

$$\mathbb{E}[\|\nabla f_j(w)\|^2] \;\leq\; 4L_{\max}(f(w) - f(w^*)) + 2\sigma^2$$

Proof:

Co-coercivity Lemma (recall slide 55)

$$f_i(y) - f_i(x) \leq \langle \nabla f_i(y), y - x\rangle - \frac{1}{2L_{\max}}||\nabla f_i(y) - \nabla f_i(x)||_2^2$$

$$\frac{1}{n}\sum_{i=1}^n ||\nabla f_i(y) - \nabla f_i(x)||_2^2 \leq 2L_{\max}\frac{1}{n}\sum_{i=1}^n (f_i(x) - f_i(y) + \langle \nabla f_i(y), y - x\rangle)$$

Lemma If $f_i : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$ convex and $L_{\max}$–smooth then

$$\mathbb{E}[\|\nabla f_{j}(w)\|^2] \;\leq\; 4L_{\max}(f(w) - f(w^*)) + 2\sigma^2$$

Proof:

Co-coercivity Lemma (recall slide 55)

$$f_i(y) - f_i(x) \leq \langle \nabla f_i(y), y - x\rangle - \frac{1}{2L_{\max}}||\nabla f_i(y) - \nabla f_i(x)||_2^2$$

$$\frac{1}{n}\sum_{i=1}^{n}||\nabla f_i(y) - \nabla f_i(x)||_2^2 \leq 2L_{\max}\frac{1}{n}\sum_{i=1}^{n}\left(f_i(x) - f_i(y) + \langle \nabla f_i(y), y - x\rangle\right)$$

$$= 2L_{\max}\left(f(x) - f(y) + \langle \nabla f(y), y - x\rangle\right)$$

Lemma If $f_i : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$ convex and $L_{\max}$–smooth then

$$\mathbb{E}[||\nabla f_j(w)||^2] \;\leq\; 4L_{\max}(f(w) - f(w^*)) + 2\sigma^2$$

Proof:

Co-coercivity Lemma (recall slide 55)

$$f_i(y) - f_i(x) \leq \langle \nabla f_i(y), y - x \rangle - \frac{1}{2L_{\max}} ||\nabla f_i(y) - \nabla f_i(x)||_2^2$$

$$\frac{1}{n}\sum_{i=1}^n ||\nabla f_i(y) - \nabla f_i(x)||_2^2 \leq 2L_{\max}\frac{1}{n}\sum_{i=1}^n (f_i(x) - f_i(y) + \langle \nabla f_i(y), y - x \rangle)$$

$$= 2L_{\max}\,(f(x) - f(y) + \langle \nabla f(y), y - x \rangle)$$

Take $y = x^* \in \arg\min f(x)$, thus $\nabla f(x^*) = 0$ and

$$(*) \qquad \frac{1}{n}\sum_{i=1}^n ||\nabla f_i(x^*) - \nabla f_i(x)||_2^2 \leq 2L_{\max}\,(f(x) - f(x^*))$$

Lemma If $f_i : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ convex and $L_{\max}$–smooth then

$$\mathbb{E}[\|\nabla f_j(w)\|^2] \;\leq\; 4L_{\max}(f(w) - f(w^*)) + 2\sigma^2$$

Proof:

Co-coercivity Lemma (recall slide 55)

$$f_i(y) - f_i(x) \leq \langle \nabla f_i(y), y - x\rangle - \frac{1}{2L_{\max}}||\nabla f_i(y) - \nabla f_i(x)||_2^2$$

$$\frac{1}{n}\sum_{i=1}^n ||\nabla f_i(y) - \nabla f_i(x)||_2^2 \leq 2L_{\max}\frac{1}{n}\sum_{i=1}^n (f_i(x) - f_i(y) + \langle \nabla f_i(y), y - x\rangle)$$

$$= 2L_{\max}\left(f(x) - f(y) + \langle \nabla f(y), y - x\rangle\right)$$

Take $y = x^* \in \arg\min f(x)$, thus $\nabla f(x^*) = 0$ and

$$(*) \qquad \frac{1}{n}\sum_{i=1}^n ||\nabla f_i(x^*) - \nabla f_i(x)||_2^2 \leq 2L_{\max}\left(f(x) - f(x^*)\right)$$

Using $\quad ||\nabla f_i(x)||_2^2 \leq 2||\nabla f_i(x^*) - \nabla f_i(x)||_2^2 + 2||\nabla f_i(x^*)||_2^2$

$$\mathbb{E}_j||\nabla f_j(x)||_2^2 = \frac{1}{n}\sum_{i=1}^n ||\nabla f_i(x)||_2^2 \leq \frac{1}{n}\sum_{i=1}^n ||\nabla f_i(x^*) - \nabla f_i(x)||_2^2 + 2\sigma^2$$

Lemma If $f_i : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$ convex and $L_{\max}$–smooth then

$$\mathbb{E}[\|\nabla f_j(w)\|^2] \;\leq\; 4L_{\max}(f(w) - f(w^*)) + 2\sigma^2$$

Co-coercivity Lemma (recall slide 55)

$$f_i(y) - f_i(x) \leq \langle \nabla f_i(y), y - x\rangle - \frac{1}{2L_{\max}}||\nabla f_i(y) - \nabla f_i(x)||_2^2$$

Proof:

$$\frac{1}{n}\sum_{i=1}^n ||\nabla f_i(y) - \nabla f_i(x)||_2^2 \leq 2L_{\max}\frac{1}{n}\sum_{i=1}^n (f_i(x) - f_i(y) + \langle \nabla f_i(y), y - x\rangle)$$

$$= 2L_{\max}\left(f(x) - f(y) + \langle \nabla f(y), y - x\rangle\right)$$

Take $y = x^* \in \arg\min f(x)$, thus $\nabla f(x^*) = 0$ and

$$(*) \qquad \frac{1}{n}\sum_{i=1}^n ||\nabla f_i(x^*) - \nabla f_i(x)||_2^2 \leq 2L_{\max}\left(f(x) - f(x^*)\right)$$

Using $\quad ||\nabla f_i(x)||_2^2 \leq 2||\nabla f_i(x^*) - \nabla f_i(x)||_2^2 + 2||\nabla f_i(x^*)||_2^2$

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

$$\mathbb{E}_j||\nabla f_j(x)||_2^2 = \frac{1}{n}\sum_{i=1}^n ||\nabla f_i(x)||_2^2 \leq \frac{1}{n}\sum_{i=1}^n ||\nabla f_i(x^*) - \nabla f_i(x)||_2^2 + 2\sigma^2$$

Lemma If $f_i : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$ convex and $L_{\max}$–smooth then

$$\mathbb{E}[\|\nabla f_j(w)\|^2] \;\leq\; 4L_{\max}(f(w) - f(w^*)) + 2\sigma^2$$

Proof:

Co-coercivity Lemma (recall slide 55)

$$f_i(y) - f_i(x) \leq \langle \nabla f_i(y), y - x\rangle - \frac{1}{2L_{\max}}||\nabla f_i(y) - \nabla f_i(x)||_2^2$$

$$\frac{1}{n}\sum_{i=1}^{n}||\nabla f_i(y) - \nabla f_i(x)||_2^2 \leq 2L_{\max}\frac{1}{n}\sum_{i=1}^{n}\left(f_i(x) - f_i(y) + \langle \nabla f_i(y), y - x\rangle\right)$$

$$= 2L_{\max}\left(f(x) - f(y) + \langle \nabla f(y), y - x\rangle\right)$$

Take $y = x^* \in \arg\min f(x)$, thus $\nabla f(x^*) = 0$ and

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

$$(*) \qquad \frac{1}{n}\sum_{i=1}^{n}||\nabla f_i(x^*) - \nabla f_i(x)||_2^2 \leq 2L_{\max}\left(f(x) - f(x^*)\right)$$

Using $\quad ||\nabla f_i(x)||_2^2 \leq 2||\nabla f_i(x^*) - \nabla f_i(x)||_2^2 + 2||\nabla f_i(x^*)||_2^2$

$$\mathbb{E}_j||\nabla f_j(x)||_2^2 = \frac{1}{n}\sum_{i=1}^{n}||\nabla f_i(x)||_2^2 \leq \frac{1}{n}\sum_{i=1}^{n}||\nabla f_i(x^*) - \nabla f_i(x)||_2^2 + 2\sigma^2$$

$$\overset{(*)}{\leq} 4L_{\max}\left(f(x) - f(x^*)\right) + 2\sigma^2$$

Proof is SUPER EASY:

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

Taking expectation with respect to $j \sim \frac{1}{n}$

Proof is SUPER EASY:

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

Taking expectation with respect to $j \sim \frac{1}{n}$

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

Proof is SUPER EASY:

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

Taking expectation with respect to $j \sim \frac{1}{n}$

$\mathbb{E}[\nabla f_j(w)] = \nabla f(w)$

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

Proof is SUPER EASY:

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

Taking expectation with respect to $j \sim \frac{1}{n}$

$\mathbb{E}[\nabla f_j(w)] = \nabla f(w)$

quasi strong conv

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

Proof is SUPER EASY:

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

Taking expectation with respect to $j \sim \frac{1}{n}$

$\mathbb{E}[\nabla f_j(w)] = \nabla f(w)$

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

quasi strong conv

Lemma

$\mathbb{E}[\|\nabla f_j(w)\|^2] \leq 4L_{\max}(f(w) - f(w^*)) + 2\sigma^2$

Proof is SUPER EASY:

$$||w^{t+1} - w^*||_2^2 = ||w^t - w^* - \gamma \nabla f_j(w^t)||_2^2$$

$$= ||w^t - w^*||_2^2 - 2\gamma\langle \nabla f_j(w^t), w^t - w^*\rangle + \gamma^2 ||\nabla f_j(w^t)||_2^2.$$

Taking expectation with respect to $j \sim \frac{1}{n}$

$\mathbb{E}[\nabla f_j(w)] = \nabla f(w)$

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

quasi strong conv

$$\leq (1-\gamma\mu)||w^t - w^*||_2^2 - 2\gamma(f(w^t) - f(w^*)) + \gamma^2 \mathbb{E}_j \left[||\nabla f_j(w^t)||_2^2\right]$$

$$\leq (1-\gamma\mu)||w^t - w^*||_2^2 + 2\gamma(2\gamma L_{\max} - 1)(f(w) - f(w^*)) + 2\gamma^2\sigma^2$$

$\gamma \leq \frac{1}{2L_{\max}}$

$$\leq (1-\gamma\mu)||w^t - w^*||_2^2 + 2\gamma^2\sigma^2$$

Lemma

$$\mathbb{E}[\|\nabla f_j(w)\|^2] \leq 4L_{\max}(f(w) - f(w^*)) + 2\sigma^2$$

Proof is SUPER EASY:

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

Taking expectation with respect to $j \sim \frac{1}{n}$

$\mathbb{E}[\nabla f_j(w)] = \nabla f(w)$

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

quasi strong conv

$\gamma \leq \frac{1}{2L_{\max}}$

Lemma

$\mathbb{E}[\|\nabla f_j(w)\|^2] \leq 4L_{\max}(f(w)-f(w^*)) + 2\sigma^2$

Taking total expectation

$$\begin{aligned}\mathbb{E}\left[||w^{t+1}-w^*||_2^2\right] &\leq (1-\gamma\mu)\mathbb{E}\left[||w^t-w^*||_2^2\right] + 2\gamma^2\sigma^2 \\ &= (1-\gamma\mu)^{t+1}||w^0-w^*||_2^2 + 2\sum_{i=0}^t (1-\gamma\mu)^i\gamma^2\sigma^2 \\ &\leq (1-\gamma\mu)^{t+1}||w^0-w^*||_2^2 + \frac{2\gamma\sigma^2}{\mu}\end{aligned}$$

Proof is SUPER EASY:

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

Taking expectation with respect to $j \sim \frac{1}{n}$

$\mathbb{E}[\nabla f_j(w)] = \nabla f(w)$

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

quasi strong conv

$$\leq (1-\gamma\mu)||w^t - w^*||_2^2 - 2\gamma(f(w^t) - f(w^*)) + \gamma^2 \mathbb{E}_j \left[||\nabla f_j(w^t)||_2^2\right]$$

$$\leq (1-\gamma\mu)||w^t - w^*||_2^2 + 2\gamma(2\gamma L_{\max} - 1)(f(w) - f(w^*)) + 2\gamma^2\sigma^2$$

$\gamma \leq \frac{1}{2L_{\max}}$

$$\leq (1-\gamma\mu)||w^t - w^*||_2^2 + 2\gamma^2\sigma^2$$

Lemma

$$\mathbb{E}[\|\nabla f_j(w)\|^2] \leq 4L_{\max}(f(w) - f(w^*)) + 2\sigma^2$$

Taking total expectation

$$\begin{aligned} \mathbb{E}\left[||w^{t+1} - w^*||_2^2\right] &\leq (1-\gamma\mu)\mathbb{E}\left[||w^t - w^*||_2^2\right] + 2\gamma^2\sigma^2 \\ &= (1-\gamma\mu)^{t+1}||w^0 - w^*||_2^2 + 2\sum_{i=0}^t (1-\gamma\mu)^i \gamma^2\sigma^2 \\ &\leq (1-\gamma\mu)^{t+1}||w^0 - w^*||_2^2 + \frac{2\gamma\sigma^2}{\mu} \end{aligned}$$

$$\sum_{i=0}^t (1-\gamma\mu)^i = \frac{1-(1-\gamma\mu)^{t+1}}{\gamma\mu} \leq \frac{1}{\gamma\mu}$$

77 / 120

Complexity / Convergence

Theorem

If f is μ–str. convex, f_i is convex, L_i–smooth, $\alpha \in [0, \frac{1}{2L_{\max}}]$ then the iterates of the SGD satisfy

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

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

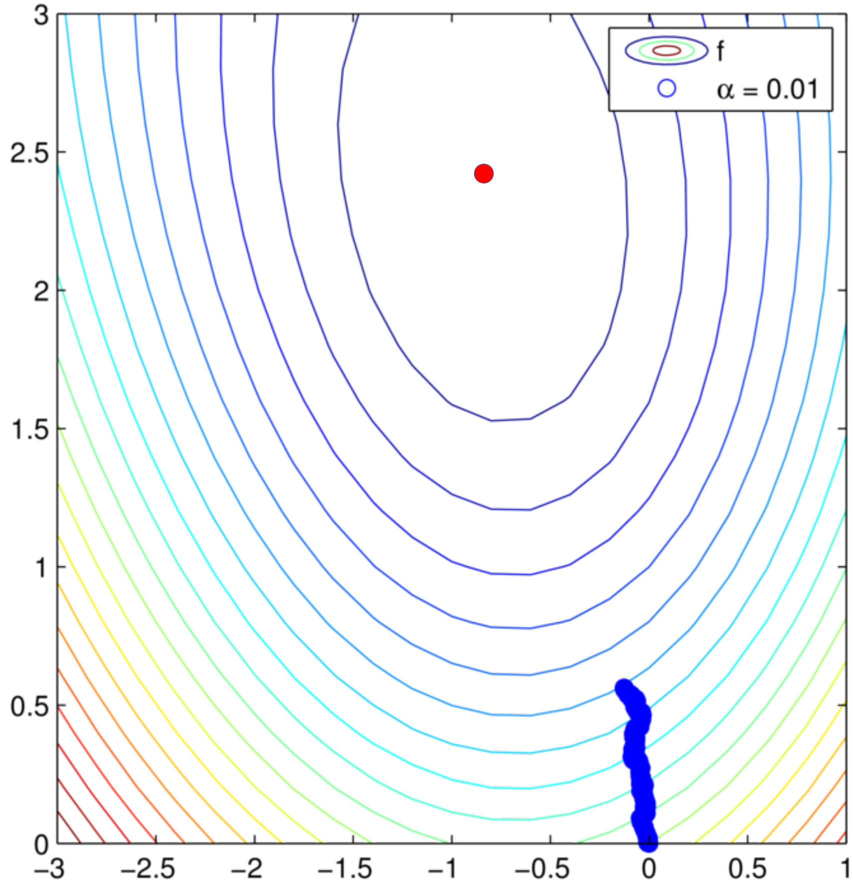
Shows that $\alpha \approx \frac{1}{\mu}$

Shows that $\alpha \approx 0$

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

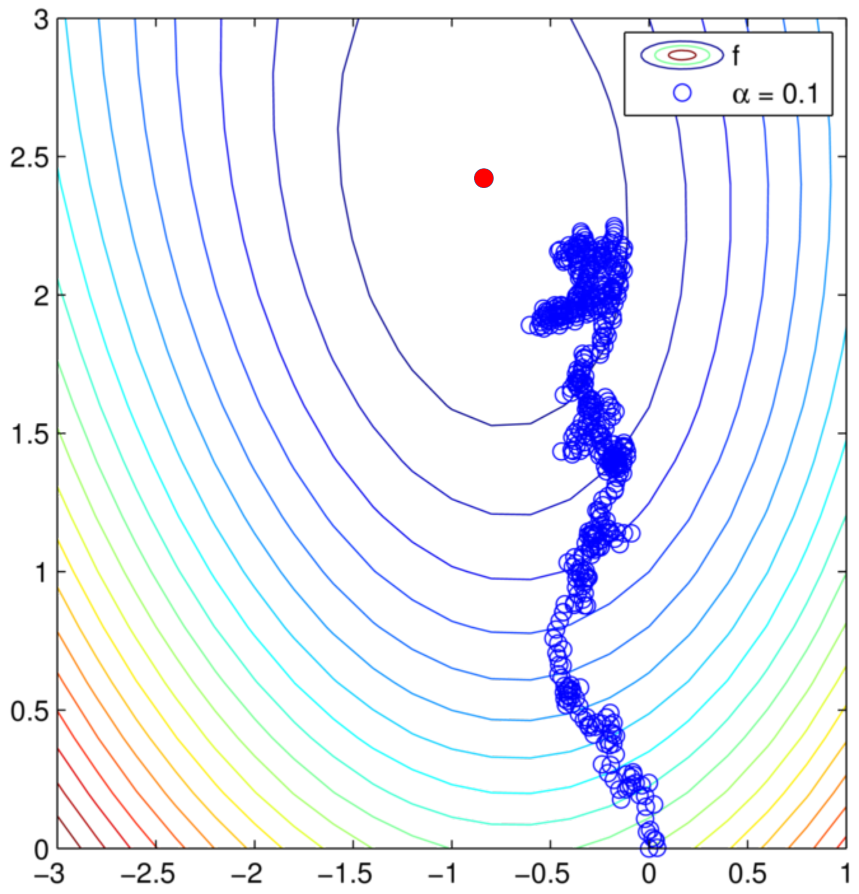
Stochastic Gradient Descent

$\alpha = 0.01$



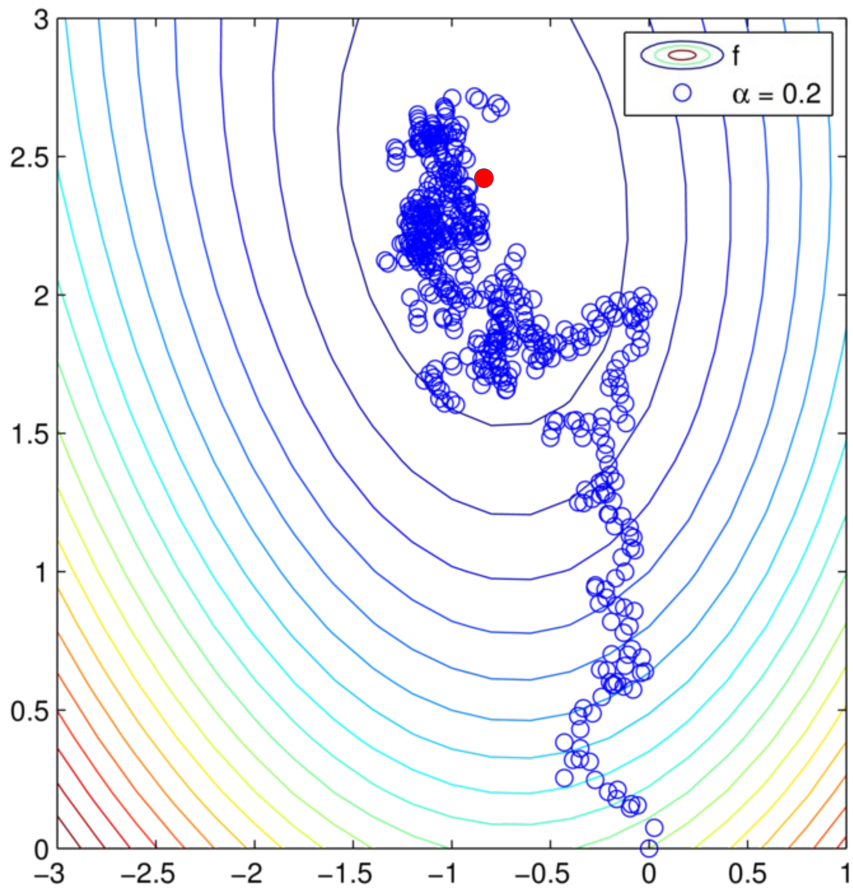
Stochastic Gradient Descent

α =0.1



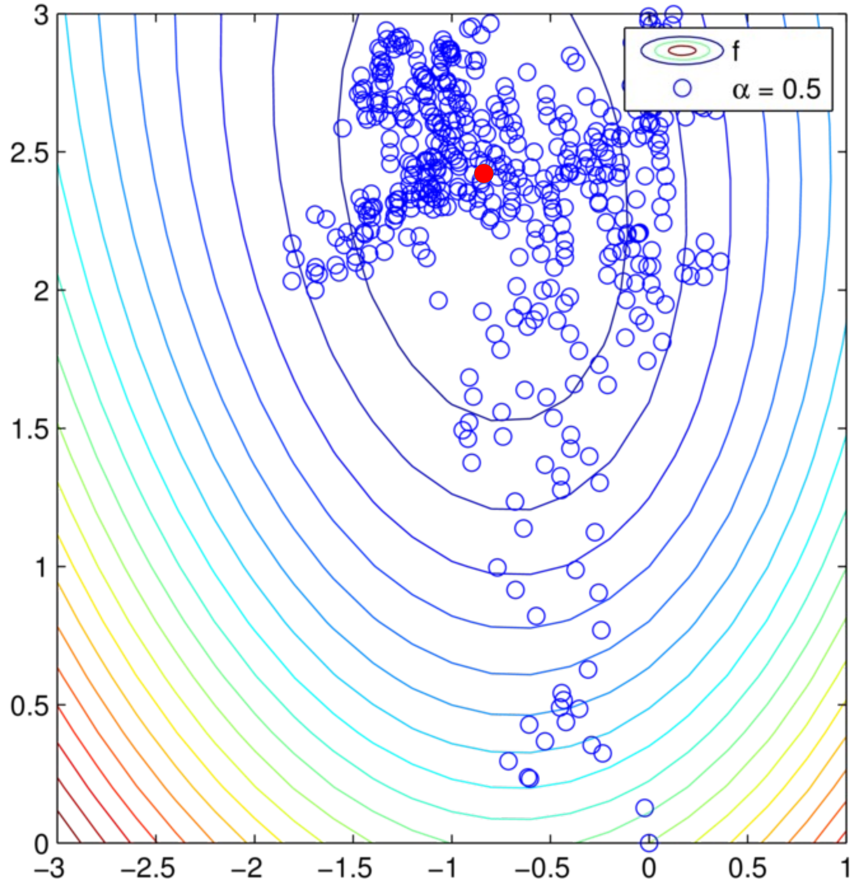
Stochastic Gradient Descent

α =0.2



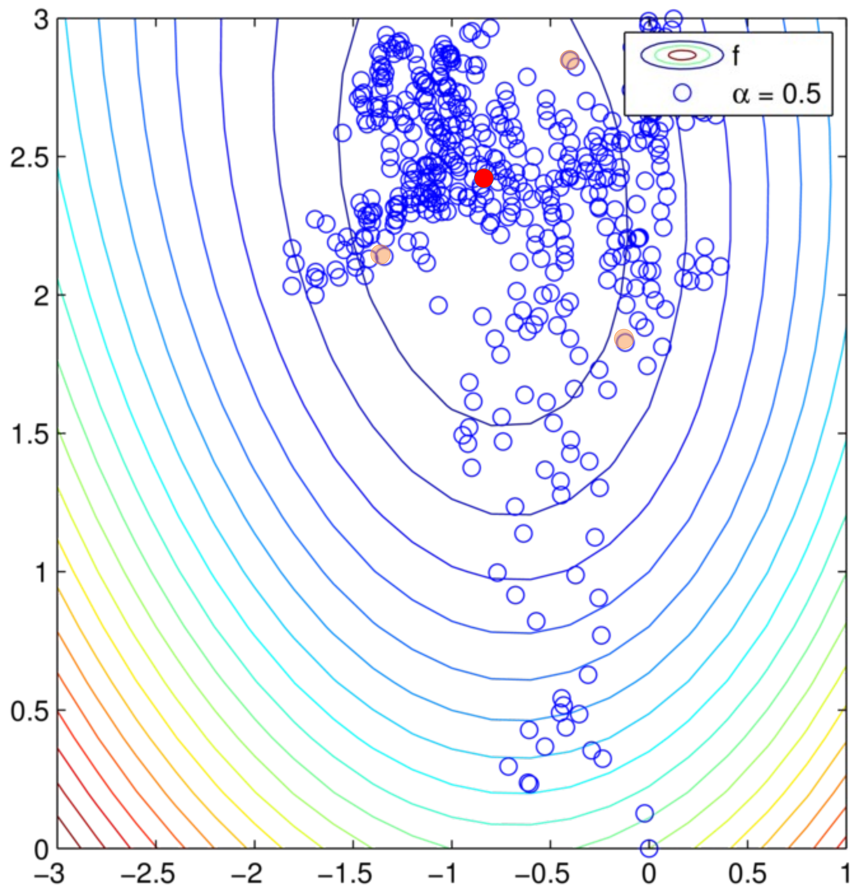
Stochastic Gradient Descent

α =0.5



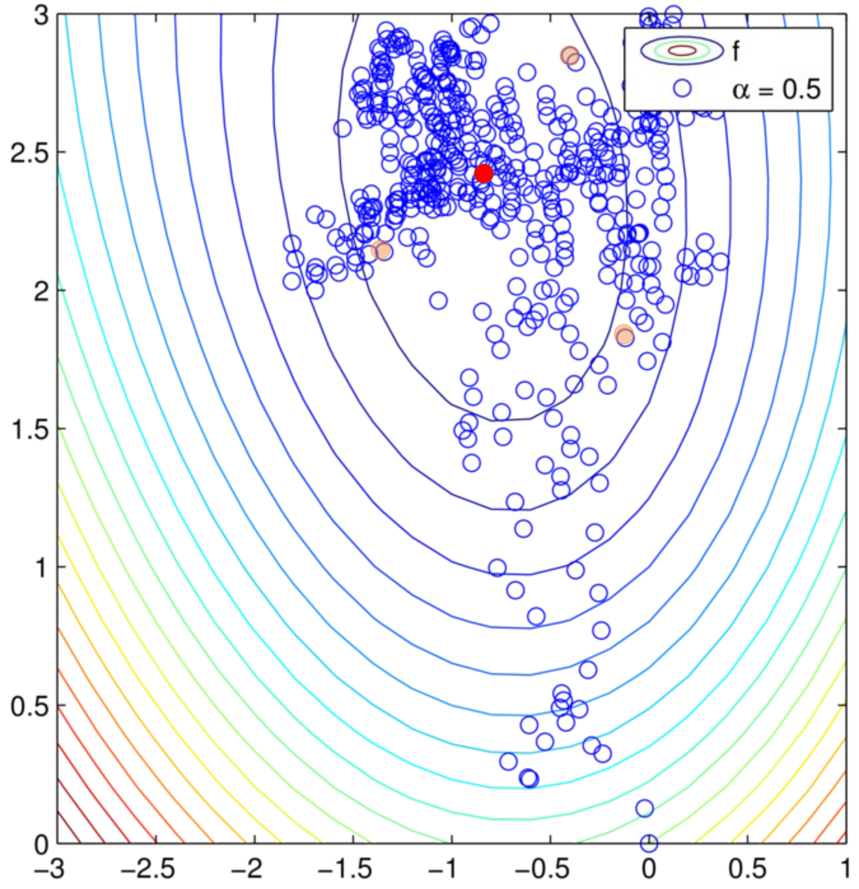
Stochastic Gradient Descent

α =0.5



Stochastic Gradient Descent

$\alpha = 0.5$

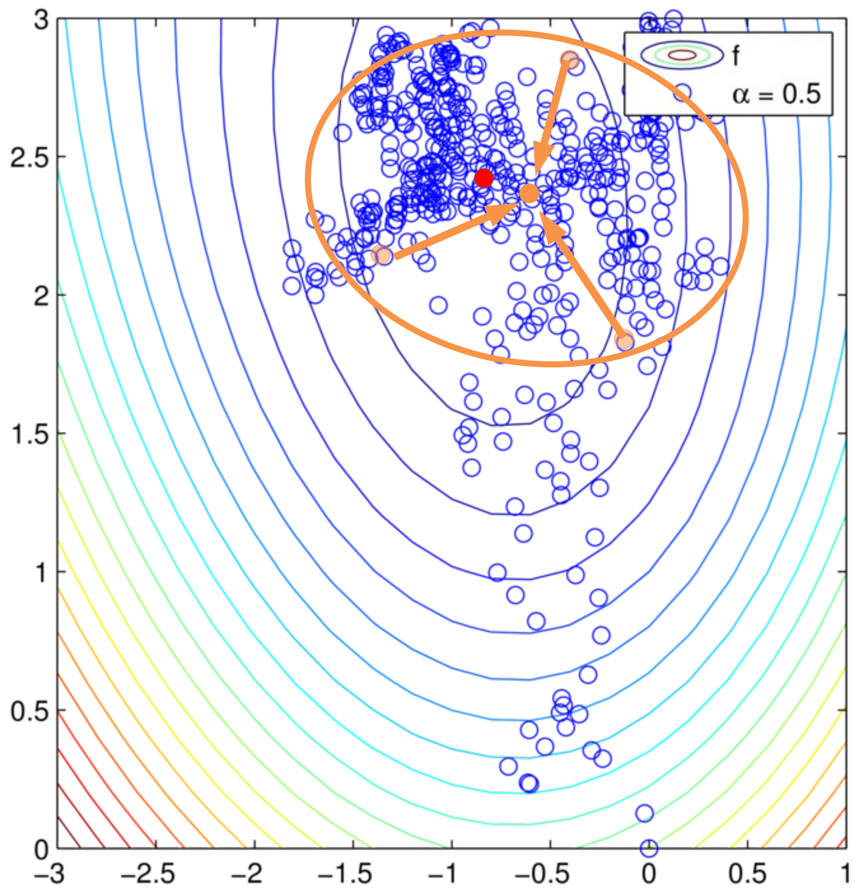


1) Start with big steps and end with smaller steps

2) Try averaging the points

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 Shrinking 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 Shrinking 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?

Complexity / Convergence

Theorem for switching to shrinking stepsizes

If f is μ–str. convex, f_i is convex and 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 the SGD iterates converge

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

Complexity / Convergence

Theorem for switching to shrinking stepsizes

If f is μ–str. convex, f_i is convex and 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 the SGD iterates converge

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

Complexity / Convergence

Theorem for switching to shrinking stepsizes

If f is μ–str. convex, f_i is convex and L_i–smooth.

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

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

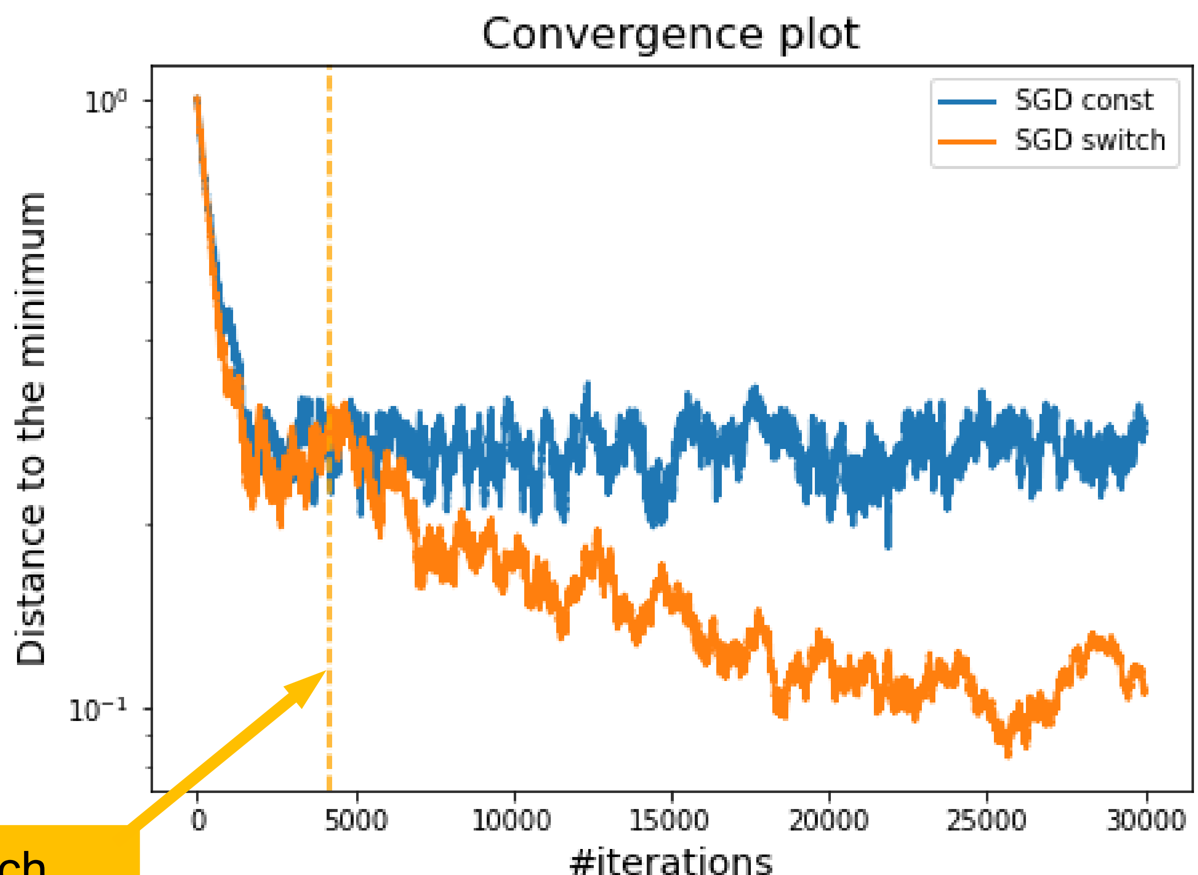
$$\alpha^t = O(1/(t+1))$$

If $t \geq 4\lceil \mathcal{K} \rceil$, then the SGD iterates converge

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

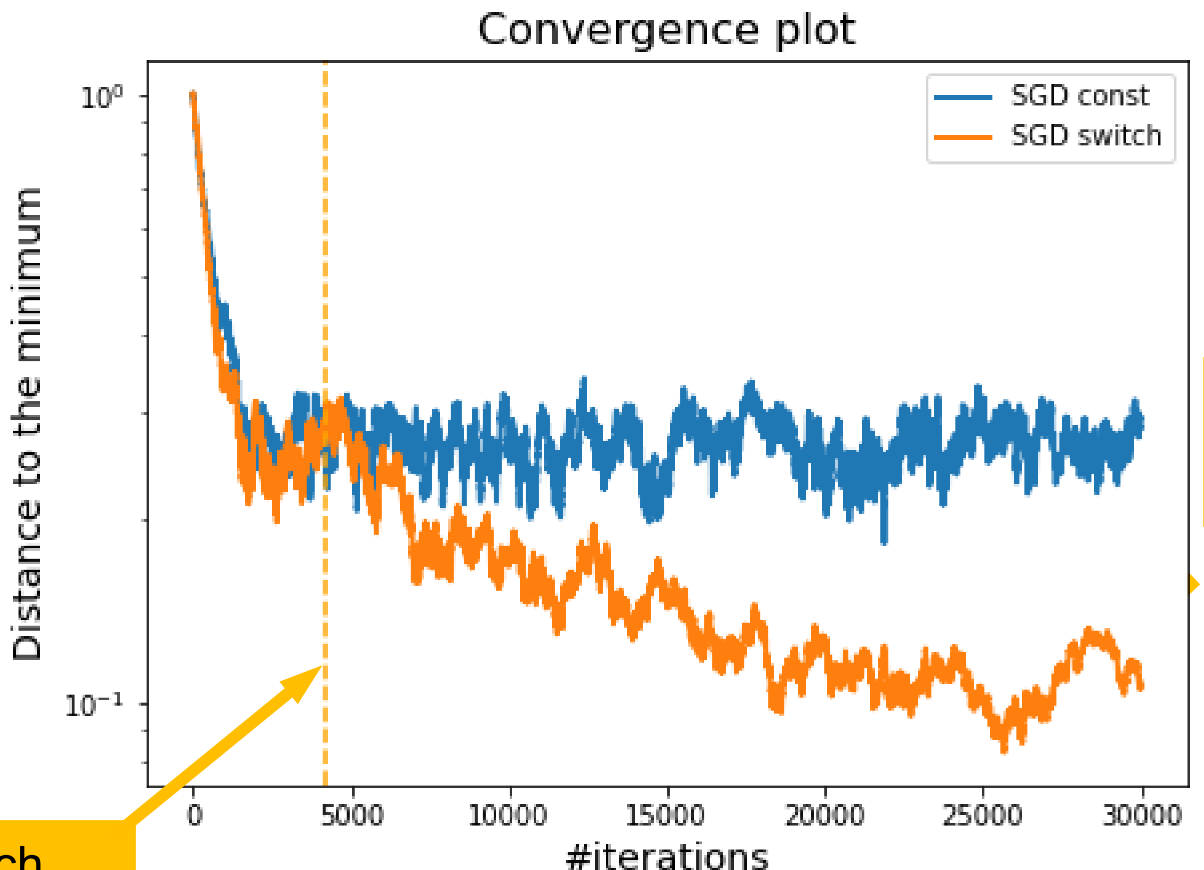
In practice often $\alpha^t = C/\sqrt{t+1}$ where C is tuned

Stochastic Gradient Descent with switch to decreasing stepsizes



Switch point
$t = 4\lceil \mathcal{K} \rceil$

Stochastic Gradient Descent with switch to decreasing stepsizes



Noisy iterates. Take averages?

Switch point $t = 4\lceil\mathcal{K}\rceil$

SGD with (late start) averaging

SGD with late averaging

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

SGD with late averaging

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 $\overline{w}$

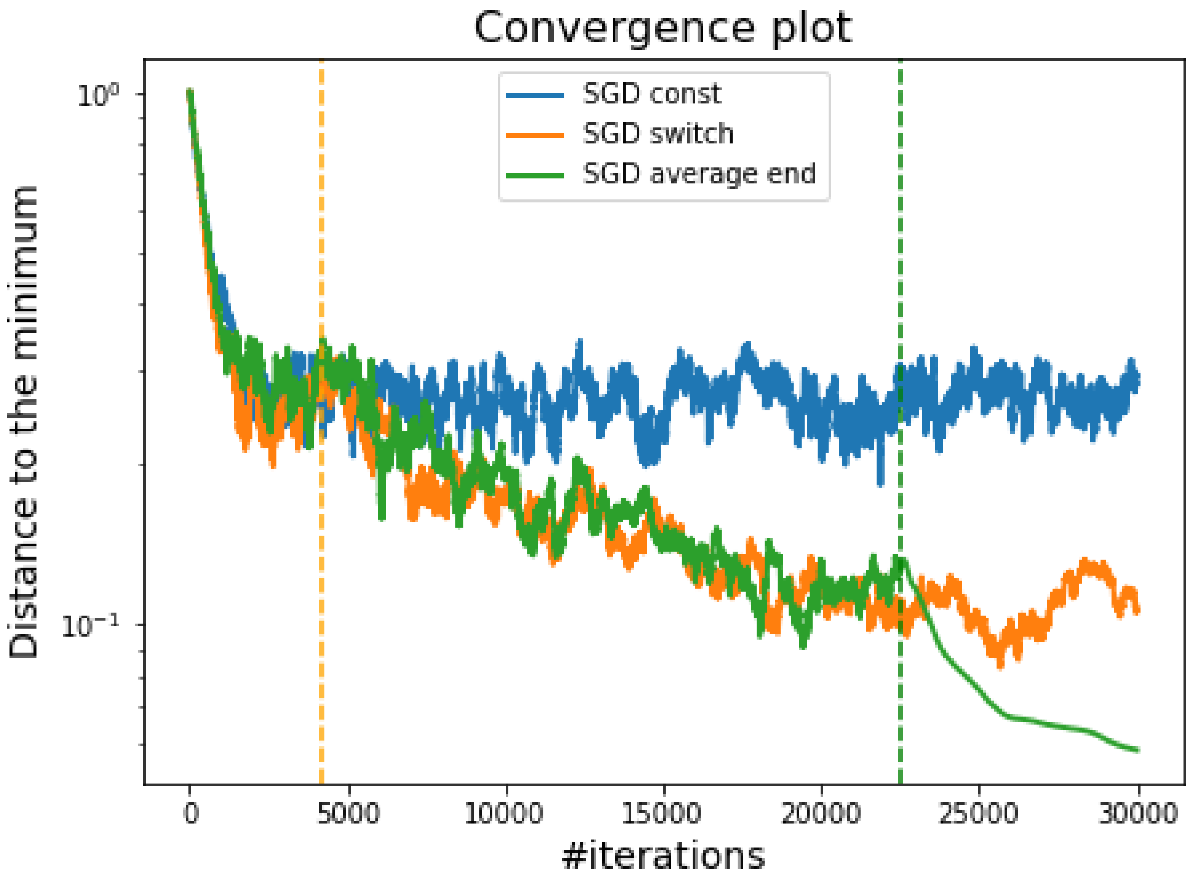
This is not efficient. How to make this efficient?

B. T. Polyak and A. B. Juditsky, SIAM Journal on Control and Optimization (1992)

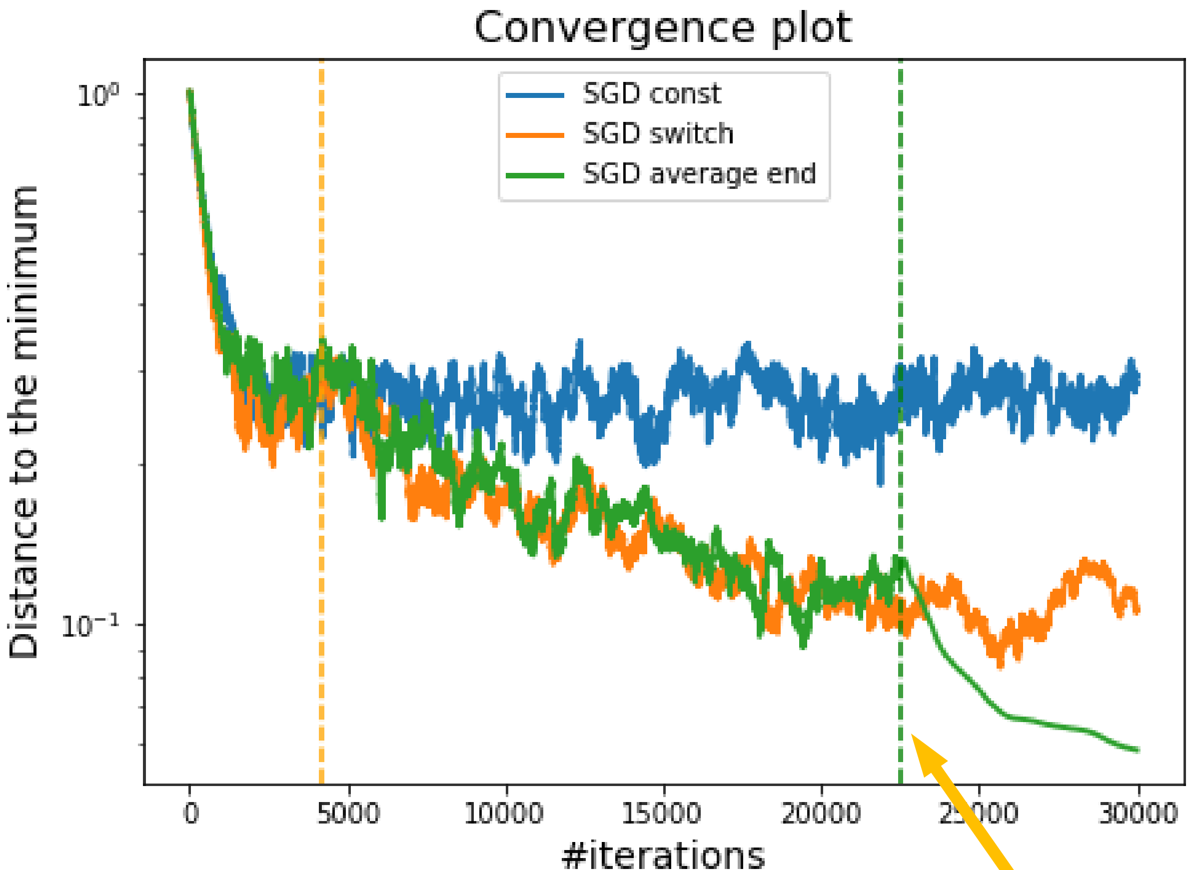
Acceleration of stochastic approximation by averaging

Stochastic Gradient Descent

Averaging the last few iterates



Stochastic Gradient Descent
Averaging the last few iterates



Averaging starts here

Part III.2: Stochastic Gradient Descent 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

Let x^i have at most $s \in \mathbb{N}$ nonzero elements for all i.
How many operations does each SGD step cost?

Sparse Examples:
encoding of categorical variables (hot one encoding), word2vec, recommendation systems ...etc

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

Let x^i have at most $s \in \mathbb{N}$ nonzero elements for all i.
How many operations does each SGD step cost?

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

Sparse Examples:
encoding of categorical variables (hot one encoding), word2vec, recommendation systems ...etc

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

Let x^i have at most $s \in \mathbb{N}$ nonzero elements for all i.
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 $O(s)$ $=$ $O(d)$

Sparse Examples:
encoding of categorical variables (hot one encoding), word2vec, recommendation systems ...etc

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 $O(s)$?

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

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 $O(s)$?

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

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 $O(s)$?

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

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

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

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

Part IV: Momentum and gradient descent

Back to 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

GD motivated through local rate of change

Local rate of change

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

GD motivated through local rate of change

Local rate of change

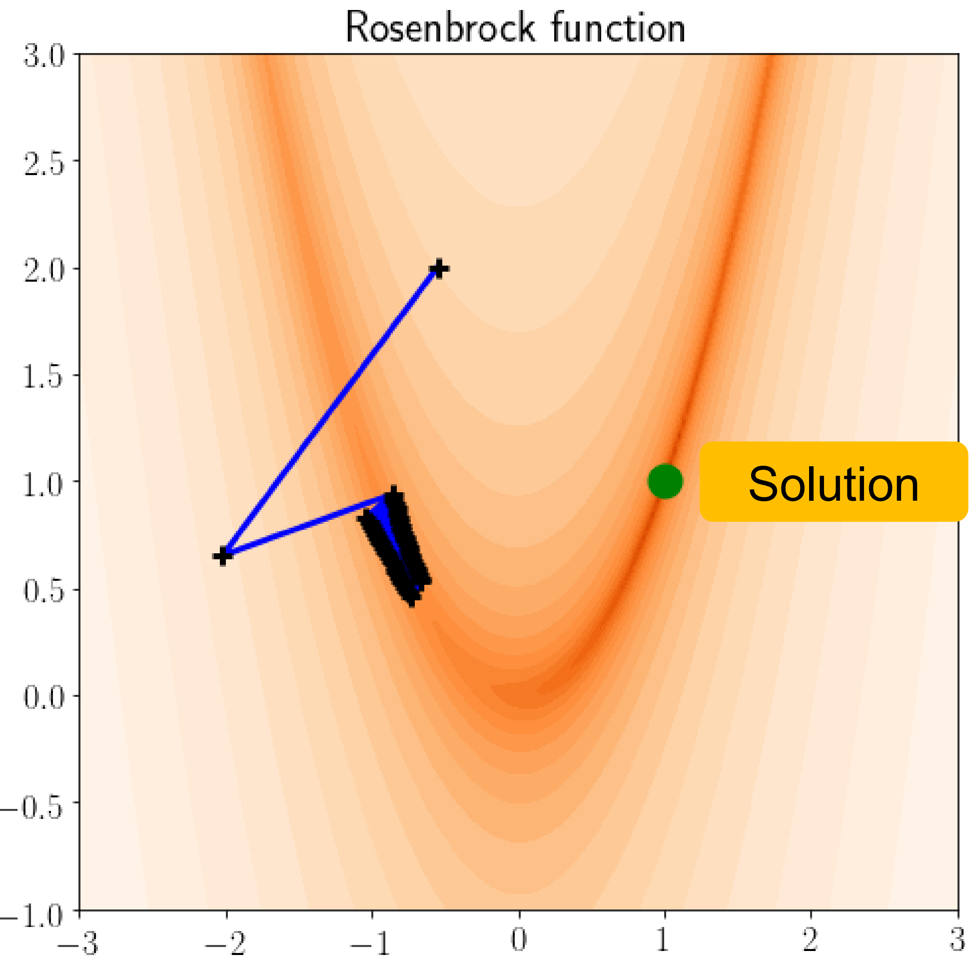
$$\Delta(d) \;:=\; \lim_{s \to 0^+} \frac{f(x+ds) - f(x)}{s}$$

Max local rate

$$\frac{\nabla f(w^t)}{\|\nabla f(w^t)\|} \;:=\; \max_{w \in \mathbb{R}^d} \Delta(d)$$
$$\text{subject to} \quad \|d\| \;=\; 1$$

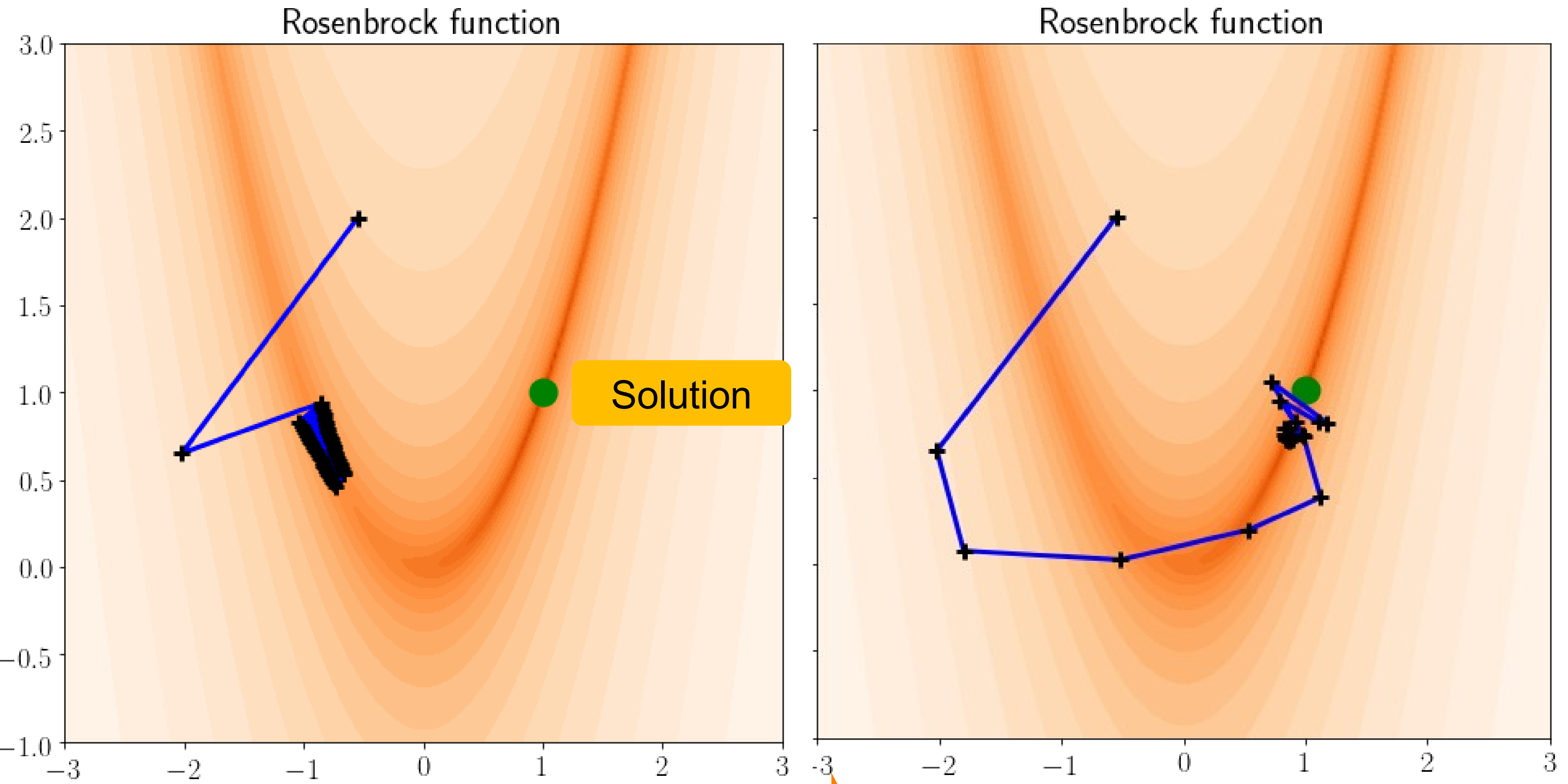
GD is the “steepest descent”

Local motivation not good for global



Get's stuck in "flat" valleys

Local motivation not good for global



Get's stuck in "flat" valleys

Give momentum to keep going

Adding Momentum to GD

Additional momentum parameter ≈ 0.99

Heavey Ball Method:

$$w^{t+1} = w^{t} - \gamma \nabla f(w^{t}) + \beta \underbrace{(w^{t} - w^{t-1})}_{\text{Adds “Inertia” to update, like friction for a heavy ball}}$$

Equivalent Momentum formulation

Heavey Ball Method:

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

Equivalent Momentum formulation

Heavey Ball Method:

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

GD with momentum (GDm):

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

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

Equivalent Momentum formulation

Heavey Ball Method:

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

Adds “Momentum” to update

Adds “Inertia” to update

GD with momentum (GDm):

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

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

Equivalent Momentum formulation

GD with momentum:

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

Equivalent Momentum formulation

GD with momentum:

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

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

Equivalent Momentum formulation

GD with momentum:

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

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

Equivalent Momentum formulation

GD with momentum:

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

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

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

Equivalent Momentum formulation

GD with momentum:

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

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

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

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

Equivalent Momentum formulation

GD with momentum:

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

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

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

Heavey Ball Method:

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

Equivalent Iterate Averaging formulation

Heavey Ball Method:

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

Equivalent Iterate Averaging formulation

Heavey Ball Method:

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

Iterate Averaging: Let $\eta > 0, \alpha \in [0,\ 1]$

$$z^t = z^{t-1} - \eta \nabla f(w^t)$$

$$w^{t+1} = \frac{\alpha}{\alpha + 1} w^t + \frac{1}{\alpha + 1} z^t$$

Equivalent Iterate Averaging formulation

Heavey Ball Method:

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

Additional sequence of variables

Iterate Averaging: Let $\eta > 0, \alpha \in [0,\ 1]$

New parameters

$$z^t = z^{t-1} - \eta \nabla f(w^t)$$

$$w^{t+1} = \frac{\alpha}{\alpha+1} w^t + \frac{1}{\alpha+1} z^t$$

Averaging of variables

Equivalent Iterate Averaging formulation

Iterate Averaging: Let $\eta > 0, \alpha \in [0,\ 1]$

$$z^t = z^{t-1} - \eta \nabla f(x^t)$$

$$w^{t+1} = \frac{\alpha}{\alpha+1} w^t + \frac{1}{\alpha+1} z^t$$

Define: $\gamma = \dfrac{\eta}{\alpha+1}$ and $\beta = \dfrac{\alpha}{\alpha+1}$

Equivalent Iterate Averaging formulation

Iterate Averaging: Let $\eta > 0, \alpha \in [0,\ 1]$

$$z^t = z^{t-1} - \eta \nabla f(x^t)$$
$$w^{t+1} = \frac{\alpha}{\alpha+1} w^t + \frac{1}{\alpha+1} z^t$$

Define: $\gamma = \dfrac{\eta}{\alpha+1}$ and $\beta = \dfrac{\alpha}{\alpha+1}$

$$w^{t+1} = \beta w^t + \frac{1}{\alpha+1} z^t$$

Equivalent Iterate Averaging formulation

Iterate Averaging: Let $\eta > 0, \alpha \in [0,\ 1]$

$$z^t = z^{t-1} - \eta \nabla f(x^t)$$

$$w^{t+1} = \frac{\alpha}{\alpha + 1} w^t + \frac{1}{\alpha + 1} z^t$$

Define: $\gamma = \dfrac{\eta}{\alpha + 1}$ and $\beta = \dfrac{\alpha}{\alpha + 1}$

$$\begin{aligned} w^{t+1} &= \beta w^t + \frac{1}{\alpha + 1} z^t \\ &= \beta w^t + \frac{1}{\alpha + 1} \left(z^{t-1} - \eta \nabla f(w^t) \right) \end{aligned}$$

Equivalent Iterate Averaging formulation

Iterate Averaging: Let $\eta > 0, \alpha \in [0,\ 1]$

$$z^t = z^{t-1} - \eta \nabla f(x^t)$$

$$w^{t+1} = \frac{\alpha}{\alpha+1} w^t + \frac{1}{\alpha+1} z^t$$

Define: $\gamma = \frac{\eta}{\alpha+1}$ and $\beta = \frac{\alpha}{\alpha+1}$

$t \leftarrow t-1$

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

$$w^{t+1} = \beta w^t + \frac{1}{\alpha+1} z^t$$

$$= \beta w^t + \frac{1}{\alpha+1}\left(z^{t-1} - \eta \nabla f(w^t)\right)$$

Equivalent Iterate Averaging formulation

Iterate Averaging: Let $\eta > 0, \alpha \in [0,\ 1]$

$$z^t = z^{t-1} - \eta \nabla f(x^t)$$

$$w^{t+1} = \frac{\alpha}{\alpha+1} w^t + \frac{1}{\alpha+1} z^t$$

Define: $\gamma = \dfrac{\eta}{\alpha+1}$ and $\beta = \dfrac{\alpha}{\alpha+1}$

$t \leftarrow t-1$

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

$$w^{t+1} = \beta w^t + \frac{1}{\alpha+1} z^t$$

$$= \beta w^t + \frac{1}{\alpha+1}\left(z^{t-1} - \eta \nabla f(w^t)\right)$$

Equivalent Iterate Averaging formulation

Iterate Averaging: Let $\eta > 0, \alpha \in [0,\ 1]$

$$z^t = z^{t-1} - \eta \nabla f(x^t)$$
$$w^{t+1} = \frac{\alpha}{\alpha+1} w^t + \frac{1}{\alpha+1} z^t$$

Define: $\gamma = \frac{\eta}{\alpha+1}$ and $\beta = \frac{\alpha}{\alpha+1}$

$t \leftarrow t-1$

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

$$\begin{aligned} w^{t+1} &= \beta w^t + \frac{1}{\alpha+1} z^t \\ &= \beta w^t + \frac{1}{\alpha+1}\left(z^{t-1} - \eta \nabla f(w^t)\right) \\ &= \beta w^t + \frac{1}{\alpha+1}\left((\alpha+1)w^t - \alpha w^{t-1} - \eta \nabla f(w^t)\right) \end{aligned}$$

Equivalent Iterate Averaging formulation

Iterate Averaging: Let $\eta > 0, \alpha \in [0,\ 1]$

$$z^t = z^{t-1} - \eta \nabla f(x^t)$$
$$w^{t+1} = \frac{\alpha}{\alpha+1} w^t + \frac{1}{\alpha+1} z^t$$

Define: $\gamma = \dfrac{\eta}{\alpha+1}$ and $\beta = \dfrac{\alpha}{\alpha+1}$

$t \leftarrow t-1$

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

$$\begin{aligned}
w^{t+1} &= \beta w^t + \frac{1}{\alpha+1} z^t \\
&= \beta w^t + \frac{1}{\alpha+1}\left(z^{t-1} - \eta \nabla f(w^t)\right) \\
&= \beta w^t + \frac{1}{\alpha+1}\left((\alpha+1)w^t - \alpha w^{t-1} - \eta \nabla f(w^t)\right) \\
&= w^t - \gamma \nabla f(w^t) + \beta(w^t - w^{t-1})
\end{aligned}$$

Equivalent Iterate Averaging formulation

Iterate Averaging: Let $\eta > 0, \alpha \in [0,\ 1]$

$$z^t = z^{t-1} - \eta \nabla f(x^t)$$

$$w^{t+1} = \frac{\alpha}{\alpha+1} w^t + \frac{1}{\alpha+1} z^t$$

Define: $\gamma = \frac{\eta}{\alpha+1}$ and $\beta = \frac{\alpha}{\alpha+1}$

$t \leftarrow t-1$

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

$$\begin{aligned} w^{t+1} &= \beta w^t + \frac{1}{\alpha+1} z^t \\ &= \beta w^t + \frac{1}{\alpha+1}\left(z^{t-1} - \eta \nabla f(w^t)\right) \\ &= \beta w^t + \frac{1}{\alpha+1}\left((\alpha+1)w^t - \alpha w^{t-1} - \eta \nabla f(w^t)\right) \end{aligned}$$

Heavey Ball Method:

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

Part IV.2: Convergence of Momentum with gradient descent

Convergence of Gradient Descent

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 $\gamma = \dfrac{2}{L+\mu}$ then Gradient Descent converges

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

$\kappa := L/\mu \geq 1$

Convergence of Gradient Descent

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 $\gamma = \dfrac{2}{L+\mu}$ then Gradient Descent converges

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

$\kappa := L/\mu \geq 1$

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

Convergence of Gradient Descent with Momentum

Polyak 1964

Theorem Let $f \in C^2$ be μ–strongly convex and L–smooth, that is

stepsize

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

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

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

$\kappa := L/\mu \geq 1$

Convergence of Gradient Descent with Momentum

Polyak 1964

Theorem Let $f \in C^2$ 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 $\gamma = \frac{4}{(\sqrt{L}+\sqrt{\mu})^2}$ and $\beta = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$ then SGDm converges

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

$\kappa := L/\mu \geq 1$

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

Convergence of Gradient Descent with Momentum

Polyak 1964

Theorem Let $f \in C^2$ be μ–strongly convex and L–smooth, that is

stepsize

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

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

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

$\kappa := L/\mu \geq 1$

Optimal iteration complexity for this function class

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

Proof: Convergence of Heavy Ball. Two time steps

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: Convergence of Heavy Ball. Two time steps

Fundamental Theorem of Calculus

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

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

Proof: Convergence of Heavy Ball. Two time steps

Fundamental Theorem of Calculus

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

$+w^* - w^*$

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

Proof: Convergence of Heavy Ball. Two time steps

Fundamental Theorem of Calculus

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

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

Proof: Convergence of Heavy Ball. Two time steps

Fundamental Theorem of Calculus

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

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

Proof: Convergence of Heavy Ball. Two time steps

Fundamental Theorem of Calculus

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

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

Depends on two times steps

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

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

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

Proof: Convergence of Heavy Ball

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

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

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

Proof: Convergence of Heavy Ball

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

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

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

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

Proof: Convergence of Heavy Ball

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

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

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

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

Simple recurrence!

Proof: Convergence of Heavy Ball

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

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

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

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

Simple recurrence!

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

Proof: Convergence of Heavy Ball

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

Proof: Convergence of Heavy Ball

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

EXE on Eigenvalues:

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

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

Proof: Convergence of Heavy Ball

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

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

EXE on Eigenvalues:

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

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

Part V: Momentum with SGD

Adding Momentum to SGD

Rumelhart, Hinton, Geoffrey, Ronald, 1986, Nature

Stochastic Heavey Ball Method:

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

SGD with momentum:

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

Sampled i.i.d

$j_t \in \{1, \ldots, n\}$

$\mathbb{P}[j = j_t] = 1/n$

Iterate Averaging:

$$z^t = z^{t-1} - \eta \nabla f_{j_t}(x^t)$$
$$w^{t+1} = \frac{\alpha}{\alpha+1} w^t + \frac{1}{\alpha+1} z^t$$

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

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

$m^0 = 0$

Momentum as exponentiated average:

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

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

SGDm and Averaging

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

$m^0 = 0$

Momentum as exponentiated average:

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

Acts like an approximate variance reduction since

SGDm and Averaging

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

$m^0 = 0$

Momentum as exponentiated average:

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

Acts like an approximate variance reduction since

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

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

SGDm and Averaging

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

$m^0 = 0$

Momentum as exponentiated average:

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

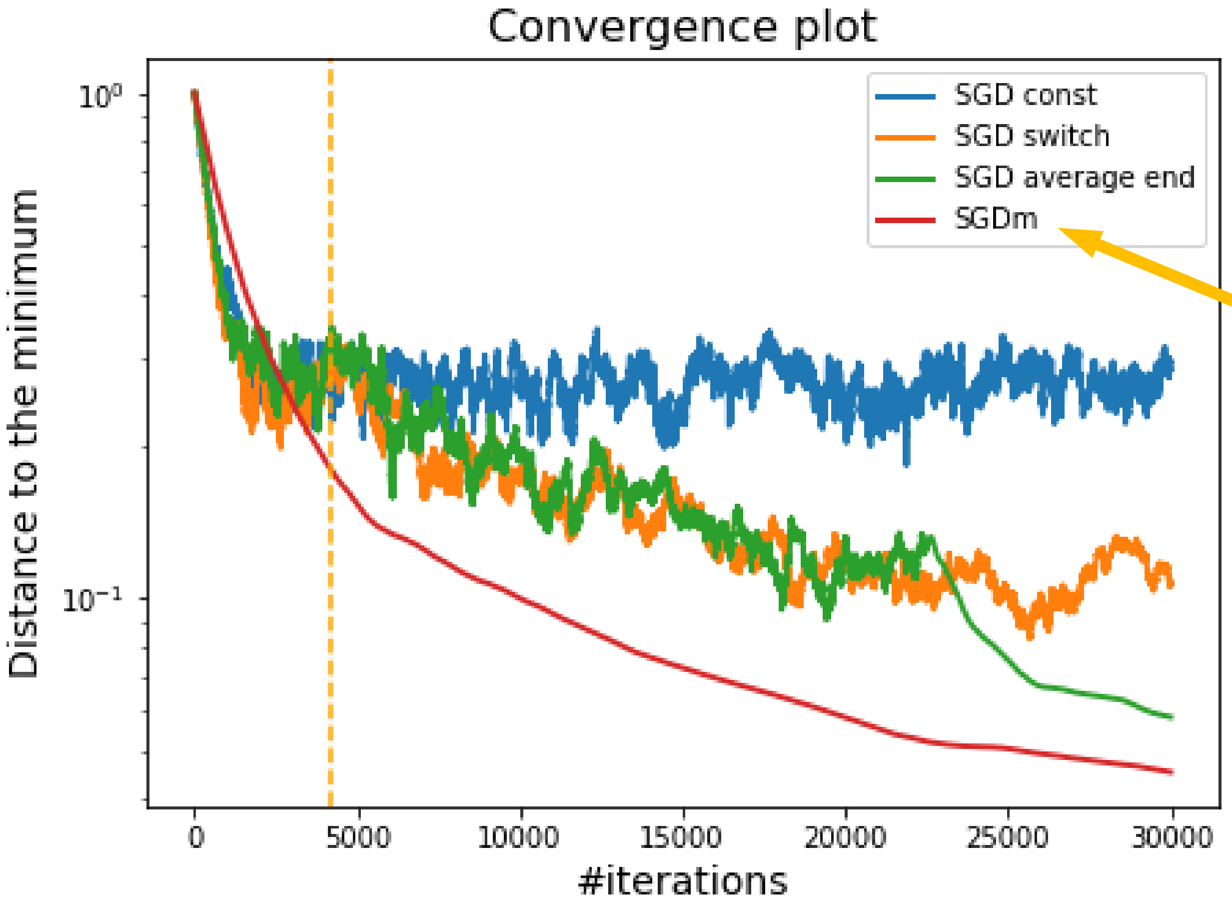
Acts like an approximate variance reduction since

This is why momentum works well with SGD

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

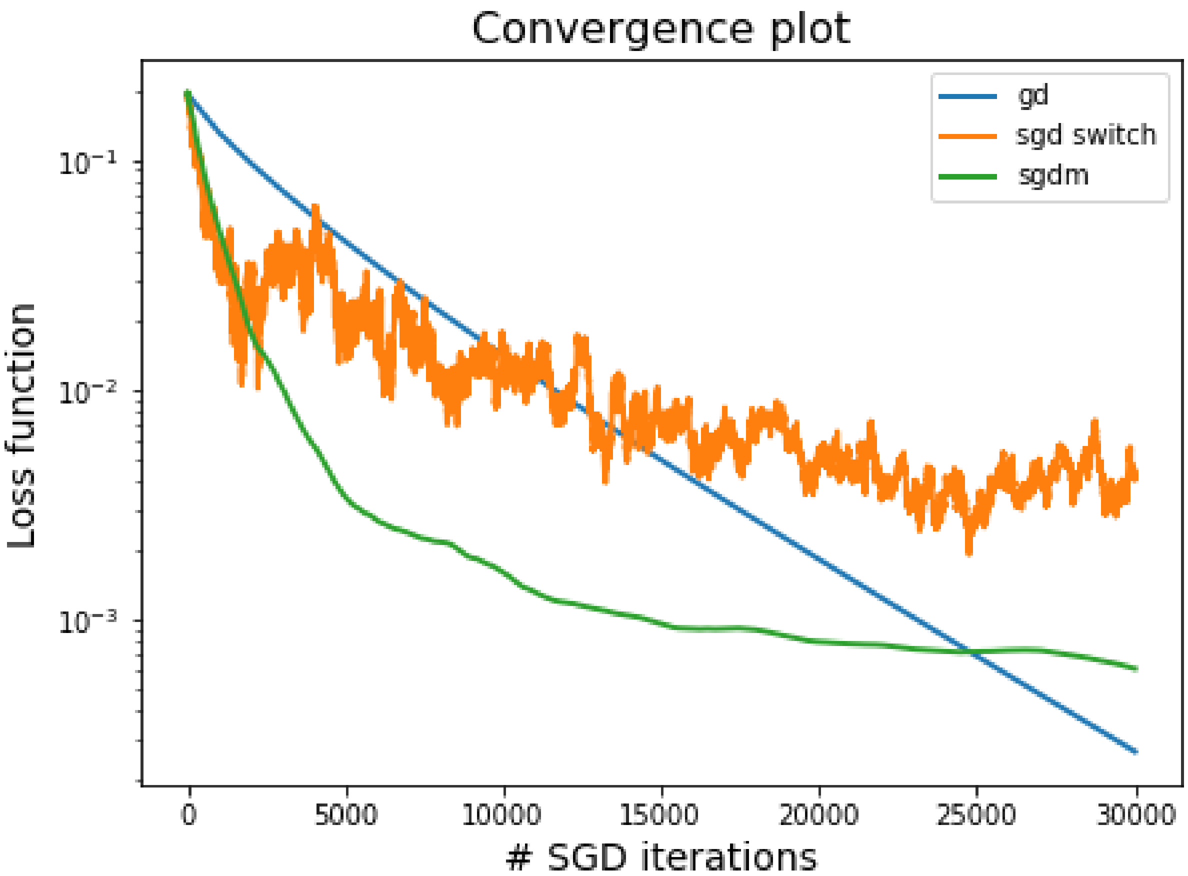
http://fa.bianp.net/teaching/2018/COMP-652/stochastic_gradient.html

Stochastic Gradient Descent with momentum



SGDm = SGD with momentum

Stochastic Gradient Descent with momentum vs GD



Can we prove momentum always works?

Difficult: Recent 2019 results only

Convergence of Gradient Descent with Momentum

Does momentum make SGD converge faster?

Not clear, recently same rate as SGD + averaging

Convergence of Gradient Descent with Momentum

Does momentum make SGD converge faster?

→

Not clear, recently same rate as SGD + averaging

Convergence of Gradient Descent with Momentum

Does momentum make SGD converge faster?

Not clear, recently same rate as SGD + averaging

f is μ–strongly convex,

f_i is convex and L_i–smooth

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

Convergence of Gradient Descent with Momentum

Does momentum make SGD converge faster?

Not clear, recently same rate as SGD + averaging

f is μ–strongly convex,
f_i is convex and L_i–smooth

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

f_i is convex and L_i–smooth

$$t \geq O\left(\frac{1}{\epsilon^2}\right)$$

Convergence of Gradient Descent with Momentum

Does momentum make SGD converge faster?

Not clear, recently same rate as SGD + averaging

f is μ–strongly convex,

f_i is convex and L_i–smooth

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

f_i is convex and L_i–smooth

$$t \geq O\left(\frac{1}{\epsilon^2}\right)$$

Sebbouth, Defazio, RMG, online soon, 2020

Convergence of Gradient Descent with Momentum

Does momentum make SGD converge faster?

Not clear, recently same rate as SGD + averaging

f is μ–strongly convex,
f_i is convex and L_i–smooth

f_i is convex and L_i–smooth

Results use iterate averaging to crack the proof!

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

$$t \geq O\left(\frac{1}{\epsilon^2}\right)$$

Sebbouth, Defazio, RMG, online soon, 2020

Part V: Test error and Validation

Validation Error

$$X := \begin{bmatrix} x_1 & x_2 & \cdots & x_T & x_{T+1} & \cdots & x_n \end{bmatrix} \in \mathbb{R}^{d \times n}$$

$$y := \begin{bmatrix} y_1 & y_2 & \cdots & y_T & y_{T+1} & \cdots & y_n \end{bmatrix} \in \mathbb{R}^{n}$$

Validation Error

$$X := \left[\begin{array}{cccc|ccc} x_1 & x_2 & \cdots & x_T & x_{T+1} & \cdots & x_n \end{array}\right] \in \mathbb{R}^{d \times n}$$

$$y := \left[\begin{array}{cccc|ccc} y_1 & y_2 & \cdots & y_T & y_{T+1} & \cdots & y_n \end{array}\right] \in \mathbb{R}^{n}$$

Validation Error

	Train set	Validation set	
$X :=$	$\big[x_1 \quad x_2 \quad \cdots \quad x_T$	$x_{T+1} \quad \cdots \quad x_n\big]$	$\in \mathbb{R}^{d\times n}$
$y :=$	$\big[y_1 \quad y_2 \quad \cdots \quad y_T$	$y_{T+1} \quad \cdots \quad y_n\big]$	$\in \mathbb{R}^{n}$

Validation Error

Train set | Validation set

$$X := \begin{bmatrix} x_1 & x_2 & \cdots & x_T & x_{T+1} & \cdots & x_n \end{bmatrix} \in \mathbb{R}^{d \times n}$$

$$y := \begin{bmatrix} y_1 & y_2 & \cdots & y_T & y_{T+1} & \cdots & y_n \end{bmatrix} \in \mathbb{R}^{n}$$

Use to train

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

Validation Error

Train set | Validation set

$$X := \begin{bmatrix} x_1 & x_2 & \cdots & x_T & x_{T+1} & \cdots & x_n \end{bmatrix} \in \mathbb{R}^{d\times n}$$

$$y := \begin{bmatrix} y_1 & y_2 & \cdots & y_T & y_{T+1} & \cdots & y_n \end{bmatrix} \in \mathbb{R}^{n}$$

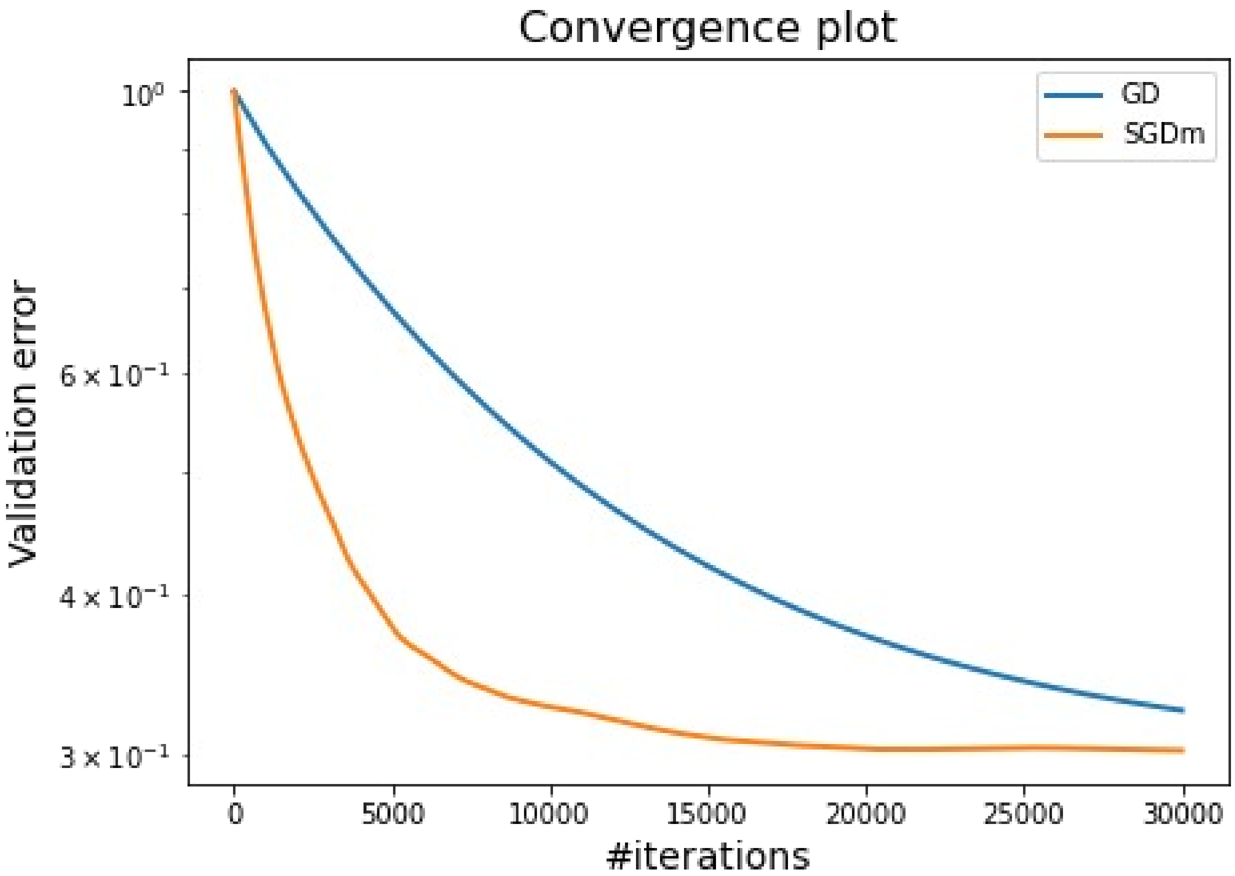
Use to train

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

Use to validate

$$\text{loss}(w^t) = \frac{1}{n-T}\sum_{i=T+1}^{n} \ell\left(h_{w^t}(x^i), y^i\right) + \lambda R(w^t)$$

Stochastic Gradient Descent with momentum vs GD on validation set



This is why SGD is popular in ML