Optimization for Machine Learning

Introduction into supervised learning

Lecturers: Francis Bach & Robert M. Gower

Tutorials: Hadrien Hendrikx, Rui Yuan, Nidham Gazagnadou

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

Core Info

- **Where** : AMMI Kigali
- **Room** : ?
- **Volume** : 24 hours
- **When** : 04/02 – 15/02
- **Online:** https://perso.telecom-paristech.fr/rgower/teaching.html
- **Lab projects:** Students send their lab python projects to: gowerrobert@gmail.com

Course Outline

Week 1: Francis Bach and Hadrien Hendrikx

- Mon. 4/2 (4-6pm): Introduction to ML – introduction to Optimization
- Tue. 5/2 (9-11pm): Exercises
- Wed. 6/2 (9-11pm): Proximal methods
- Wed. 6/2 (2-4pm): Exercises
- Fri. 8/2 (9-10pm): Quiz
- Fri. 8/2 (10.15pm – 12.15pm + 2-4pm): Lab with week 1 and week 2 teachers!

Course Outline

Week 2: Robert Gower and Rui Yuan, Nidham Gazagnadou

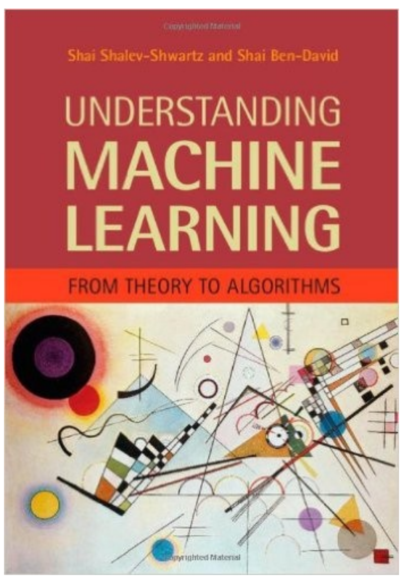
- Mon 11/2 (9-11pm): Stochastic gradient
- Mon 11/2 (2-4pm): Exercises
- Wed 13/2 (9-11pm): Variance reduction
- Wed 13/12 (2-4pm): Exercises
- Fri. 15/2 (9-10pm): Quiz
- Fri. 15/2 (10.15pm – 12.15pm + 2-4pm): Lab

An Introduction to Supervised Learning

References classes today

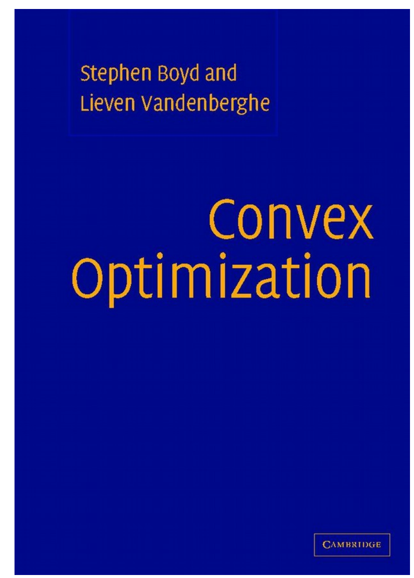
Chapter 2

Understanding Machine Learning: From Theory to Algorithms



Pages 67 to 79

Convex Optimization, Stephen Boyd



Is There a Cat in the Photo?

Yes

No

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

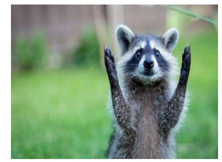
Find mapping *h* that assigns the "correct" target to each input

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

Labeled Data: The training set

x^1 { 
$y^1 = 1$

x^2 { 
$y^2 = 1$

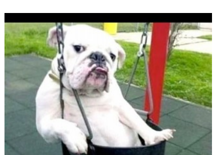
x^3 { 
$y^3 = -1$

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

y= -1 means no/false

Learning Algorithm

$h : x \in X \to y \in \mathbf{R}$

h () $\longrightarrow$ **-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

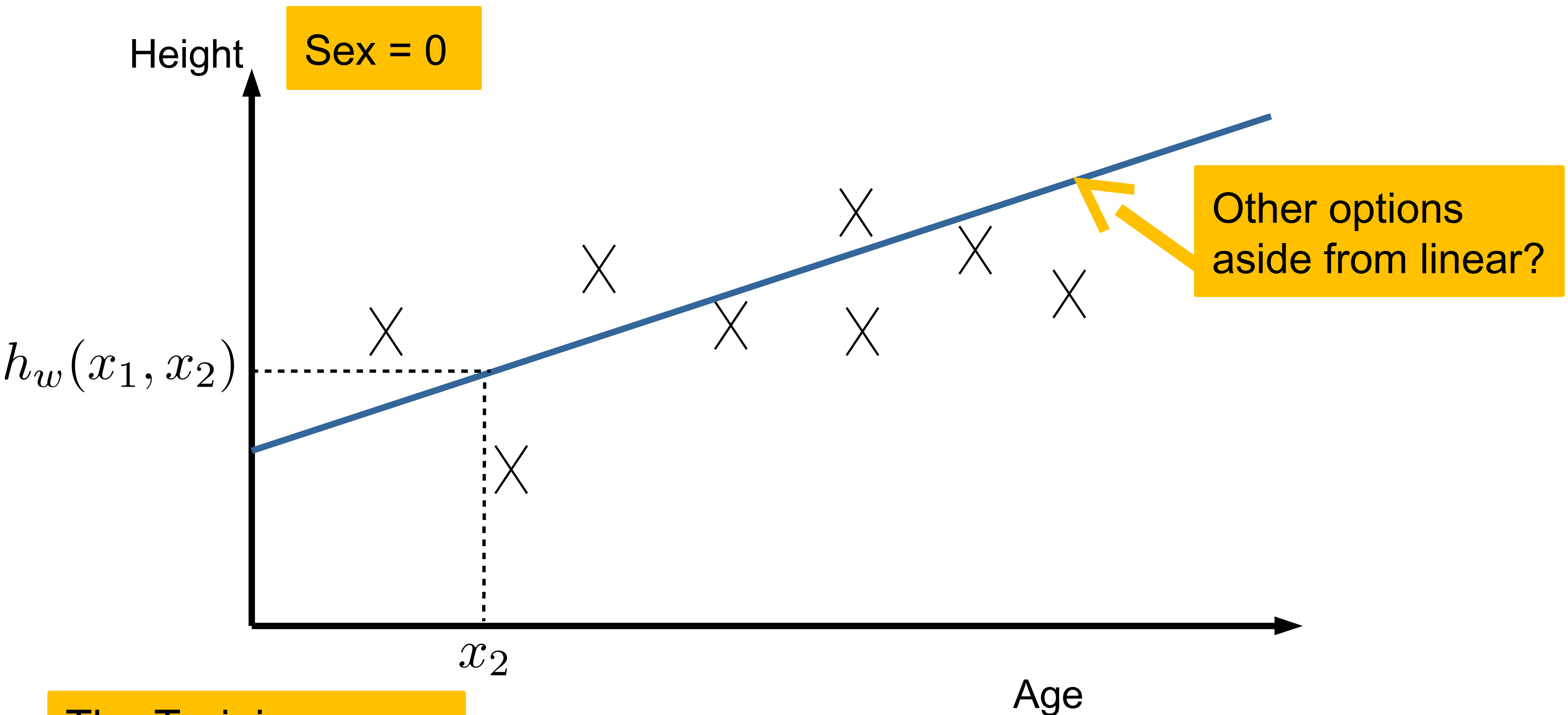
Example Hypothesis: Linear Model

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

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



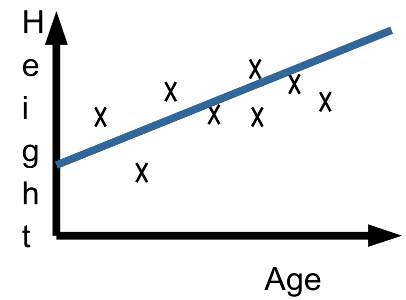
The Training Algorithm

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

Parametrizing the Hypothesis

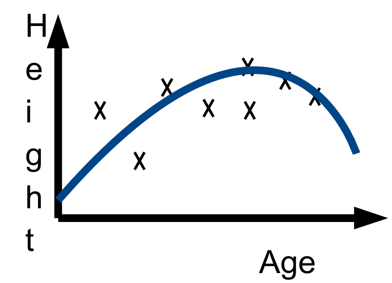
Linear:

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

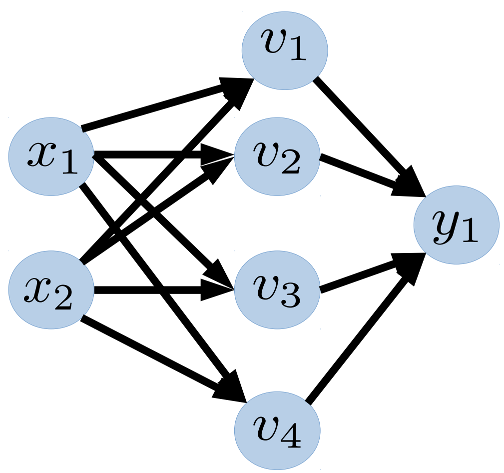


Polinomial:

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



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?

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

Loss Functions

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

Typically a convex function

The Training Problem

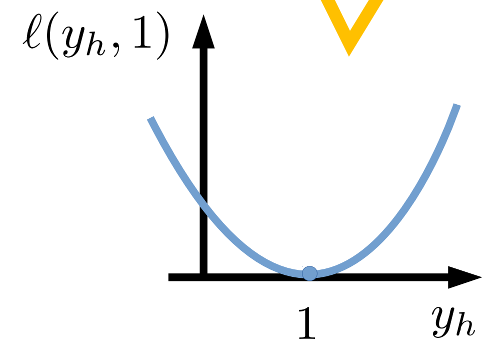
$$\min_{w\in\mathbf{R}^d} \frac{1}{n}\sum_{i=1}^{n}\ell\left(h_w(x^i), y^i\right)$$

Choosing the Loss Function

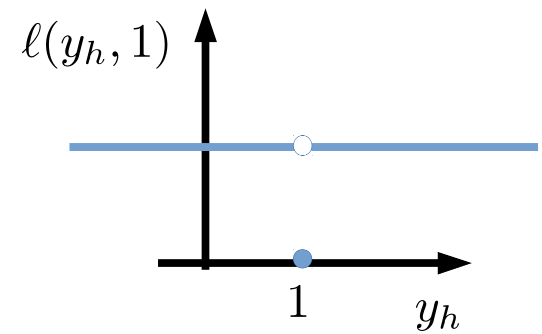
y=1 in all figures

Let $y_h := h_w(x)$

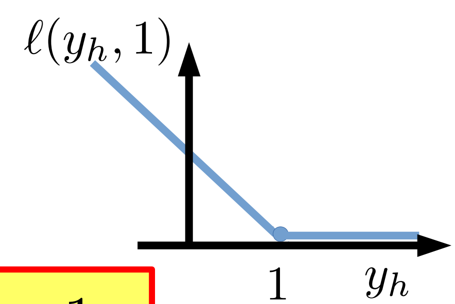
Quadratic Loss $\ell(y_h, y) = (y_h - y)^2$



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 $y = -1$

Loss Functions

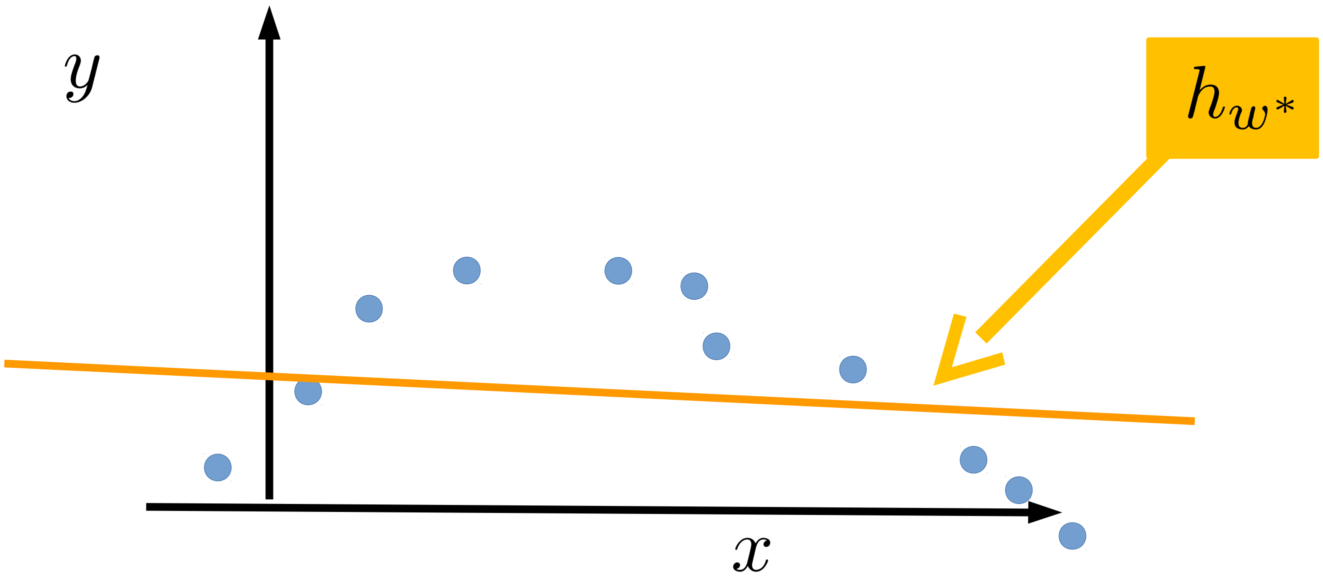
The Training Problem

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

Is a notion of Loss enough?

What happens when we do not have enough data?

Overfitting and Model Complexity

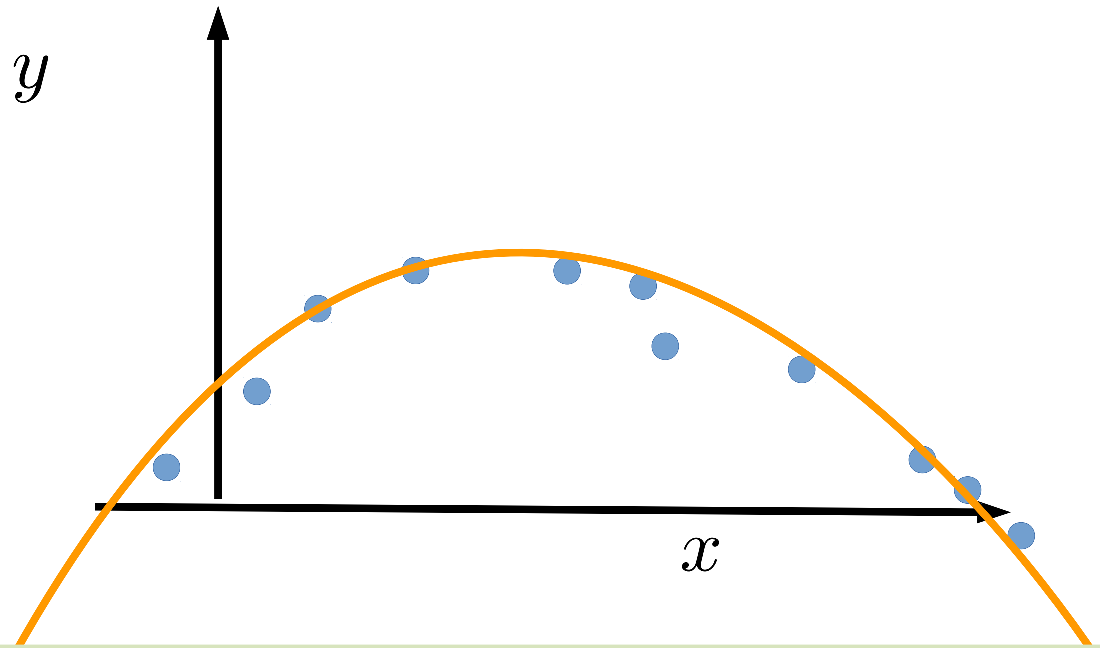


Fitting 1st order polynomial

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

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

Overfitting and Model Complexity

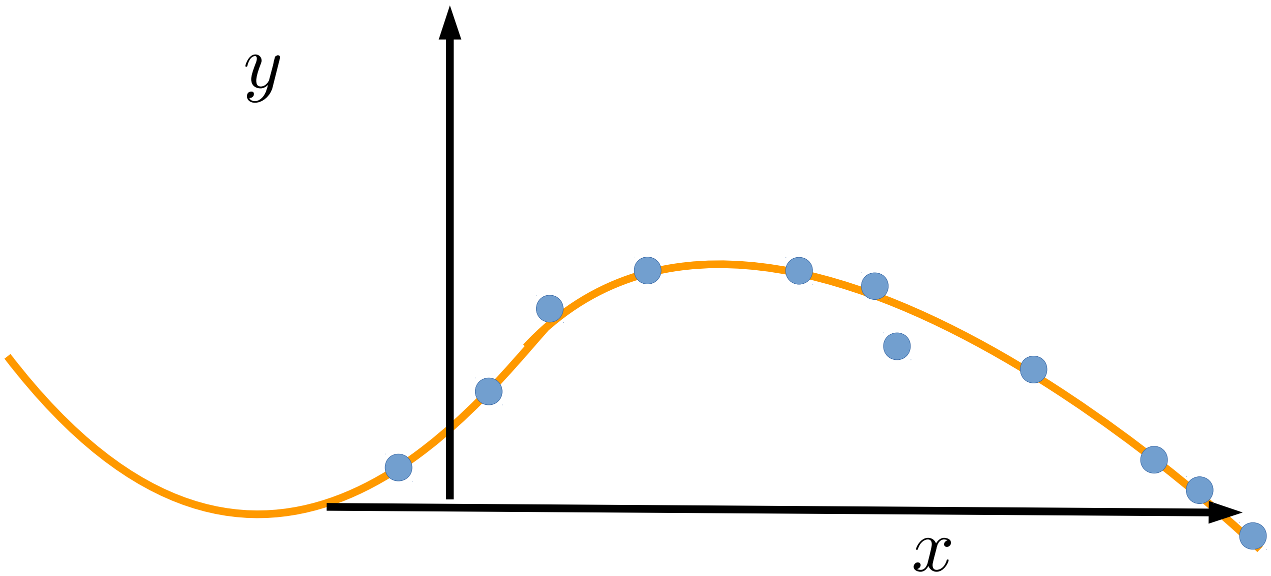


Fitting 2nd order polynomial

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

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

Overfitting and Model Complexity

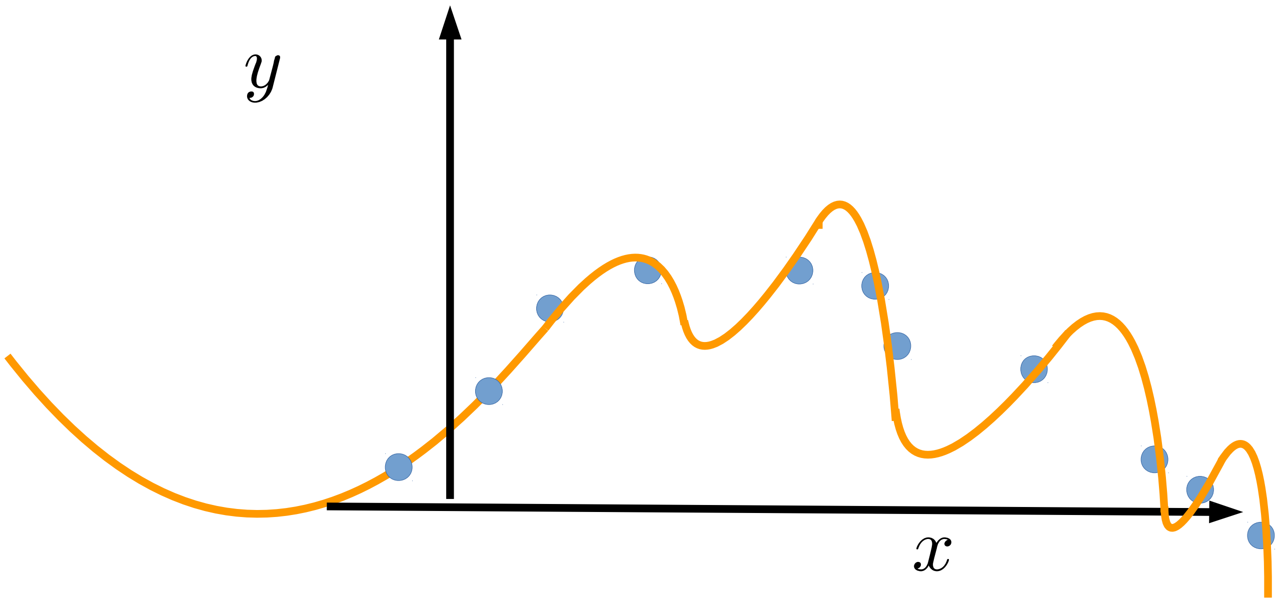


Fitting 3rd order polynomial

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

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

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

Regularizor Functions

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

Controls tradeoff between fit and complexity

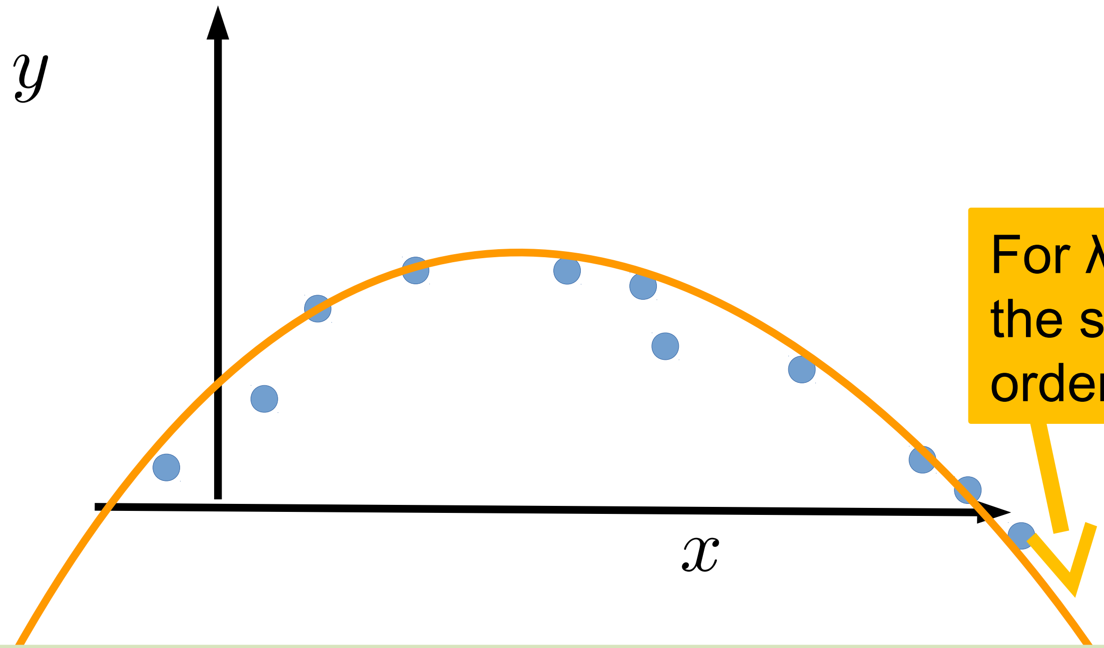
General Training Problem

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

Exe:

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

Overfitting and Model Complexity



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

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

Logistic Regression

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

The Machine Learners Job

(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: $\ell(h_w(x), y) \geq 0$

$$\boxed{\begin{array}{l}\text{(4) Solve the \textit{training problem}:} \\ \displaystyle \min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell\left(h_w(x^i), y^i\right) + \lambda R(w)\end{array}}$$

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

The Statistical Learning Problem: The hard truth

Do we really care if the loss $\ell\left(h_w(x^i), y^i\right)$ is small on the ***known*** labelled data paris *(x^i,y^i) ?* **Nope**

We really want to have a small loss on new unlabelled Observations!

Assume data sampled $(x, y) \sim \mathcal{D}$ where $\mathcal{D}$ is an unknown distribution

The Statistical Learning Problem: The hard truth

The statistical learning problem:

Minimize the expected loss over an *unknown* expectation

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

Variance of sample mean:

$$\left|\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell\left(h_w(x), y\right)\right] - \frac{1}{n}\sum_{i=1}^{n} \ell\left(h_w(x_i), y_i\right)\right|^2 = O\left(\frac{1}{n}\right)$$

Optimization for Machine Learning

Convexity, Smoothness and the Gradient Method

Lecturers: Francis Bach & Robert M. Gower

Tutorials: Hadrien Hendrikx, Rui Yuan, Nidham Gazagnadou

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

Solving the Finite Sum Training Problem

Optimization Sum of Terms

A Datum Function

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

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

Finite Sum Training Problem

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

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 $\alpha > 0$.

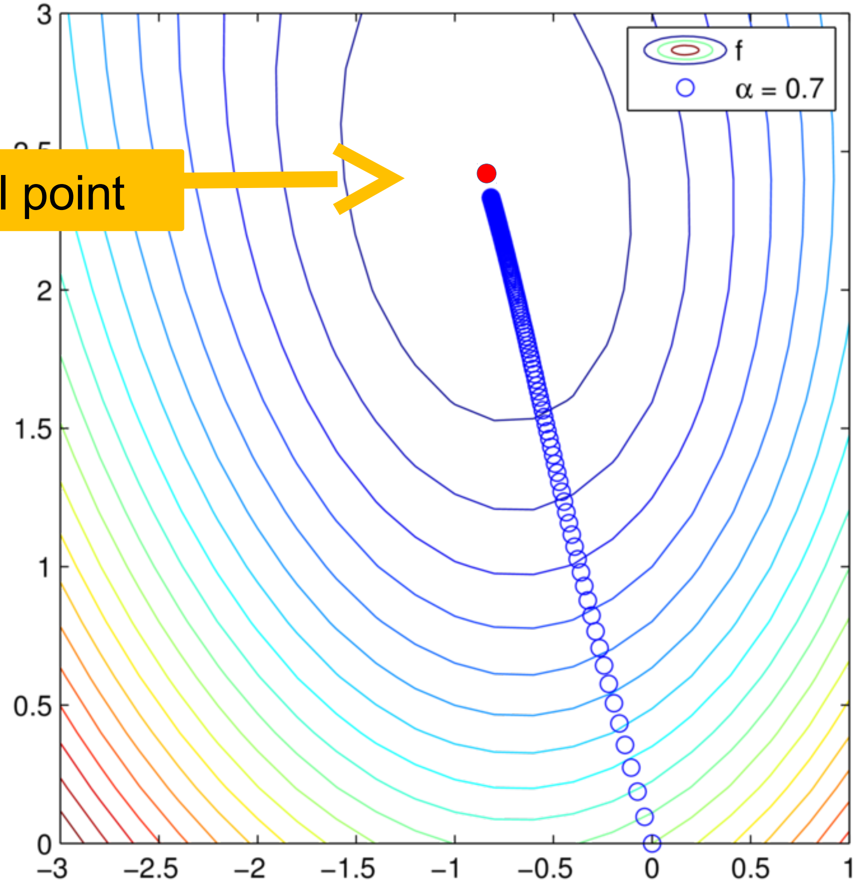
for $t = 1, 2, 3, \ldots, T$

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

Output w^{T+1}

Gradient Descent Example

Optimal point



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

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

Convergence GD

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

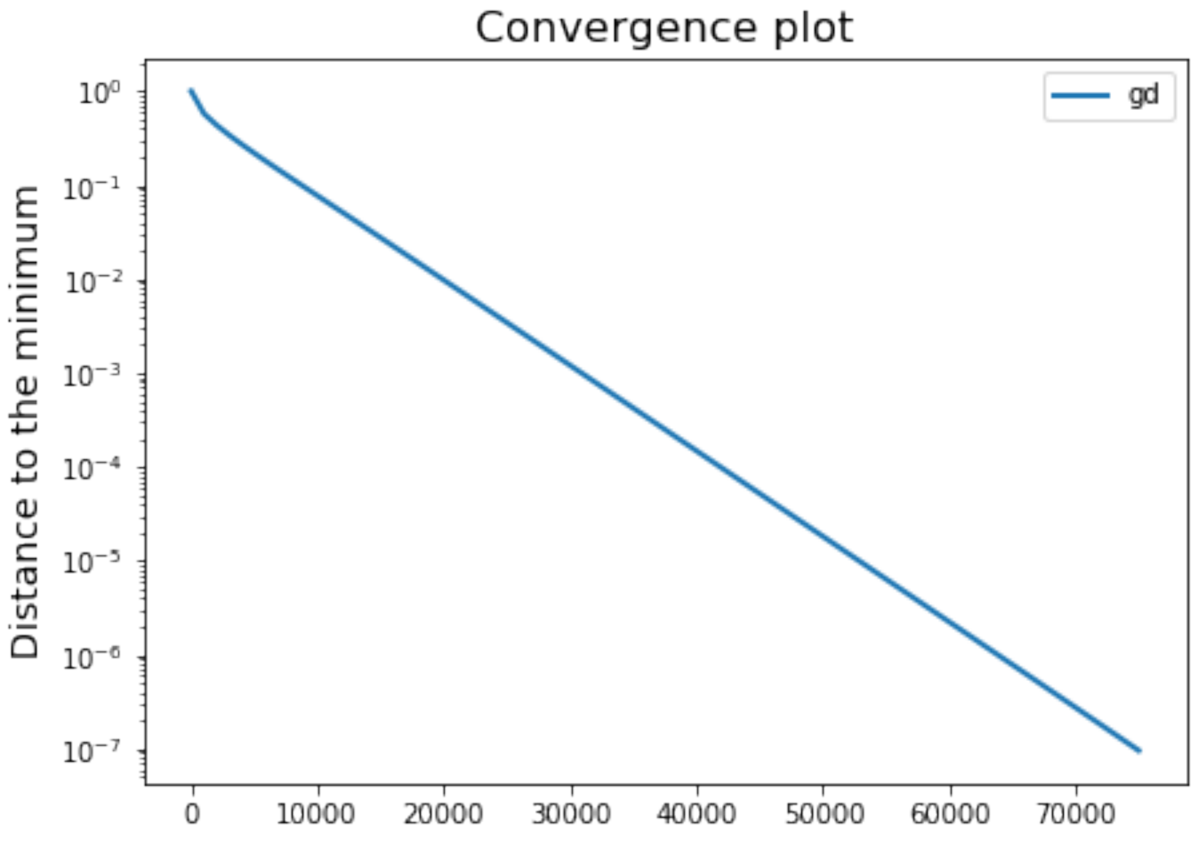
$$L = \sigma_{\max}(A)$$

$$\mu = \sigma_{\min}(A)$$

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

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

Gradient Descent Example: logistic

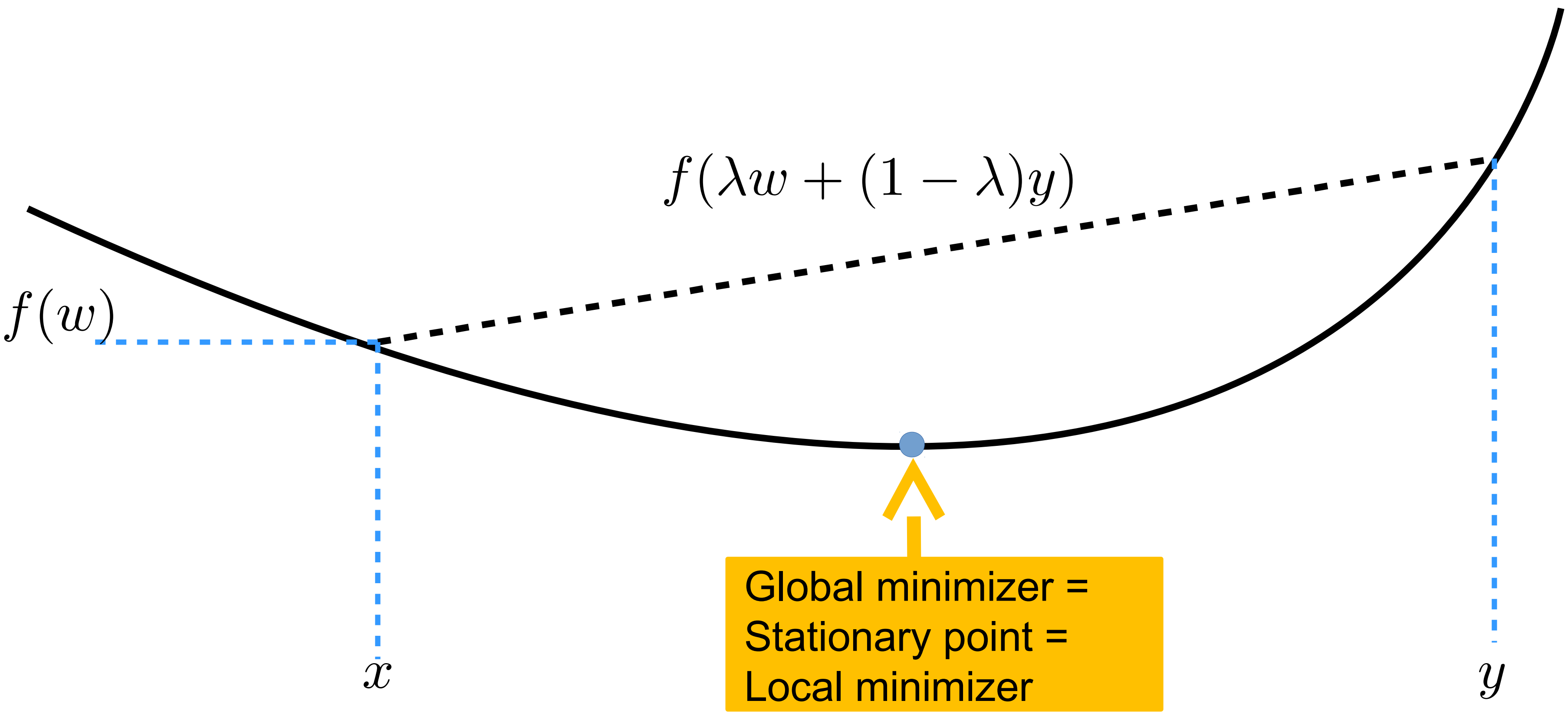


$$y\text{–axis} = \frac{||w^t - w^*||_2^2}{||w^1 - w^*||_2^2} \quad \Rightarrow \quad \log\left(\frac{||w^t - w^*||_2^2}{||w^1 - w^*||_2^2}\right) \leq t \log\left(1 - \frac{\mu}{L}\right)$$

Convexity

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

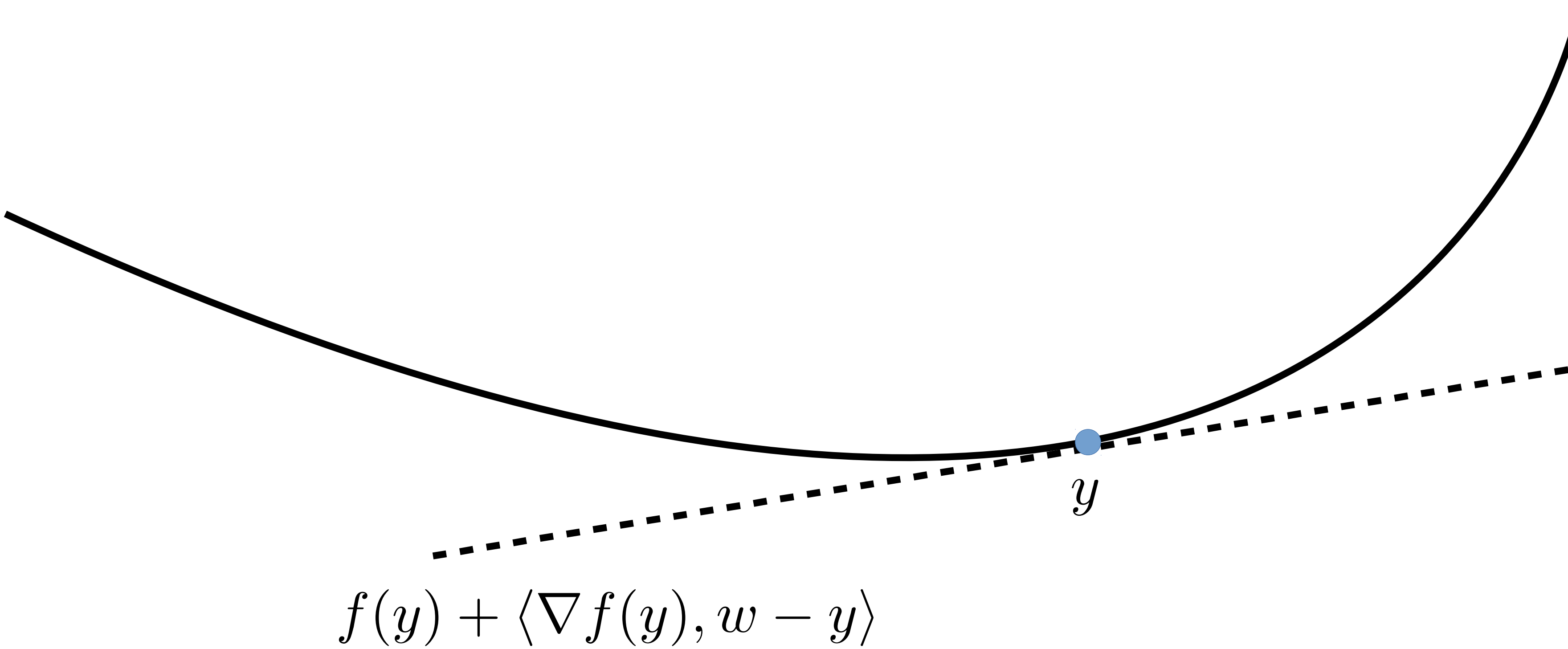
$$f(\lambda w + (1-\lambda)y) \leq \lambda f(w) + (1-\lambda) f(y), \quad \forall w, y \in C, \lambda \in [0,\, 1]$$



Convexity: First derivative

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

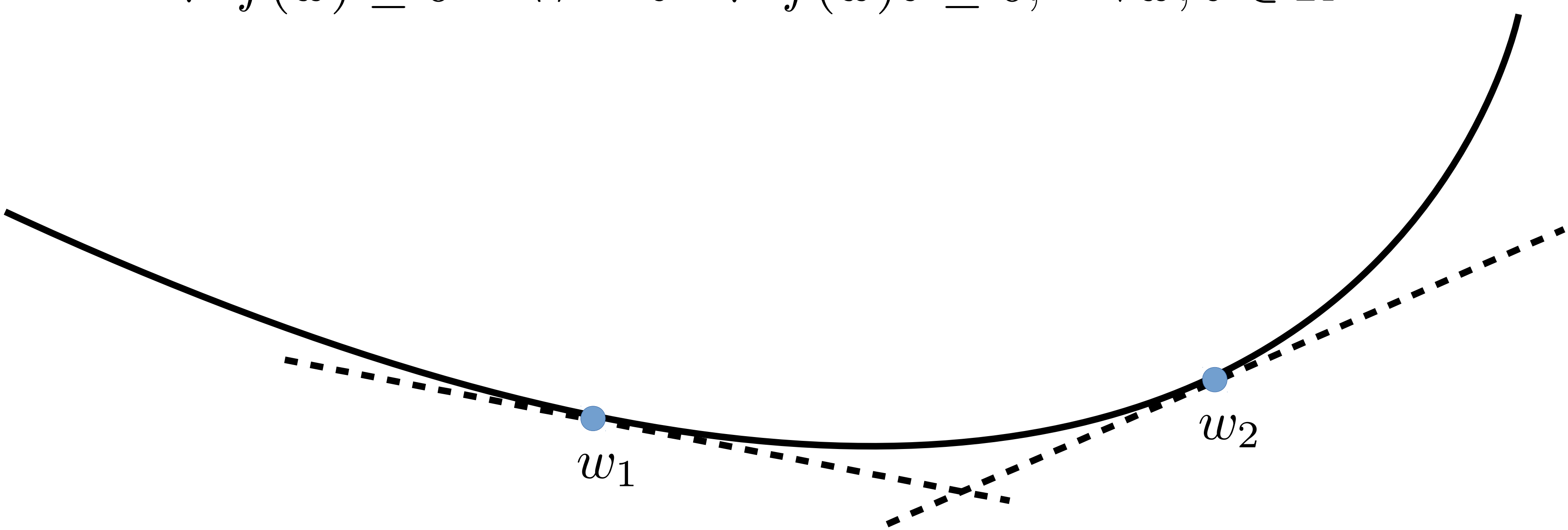
$$f(w) \geq f(y) + \langle \nabla f(y), w - y \rangle$$



Convexity: Second derivative

A twice differential 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 an exercise!

Negative log and logistic:

$$x \mapsto -\log(x)$$

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

Hinge loss

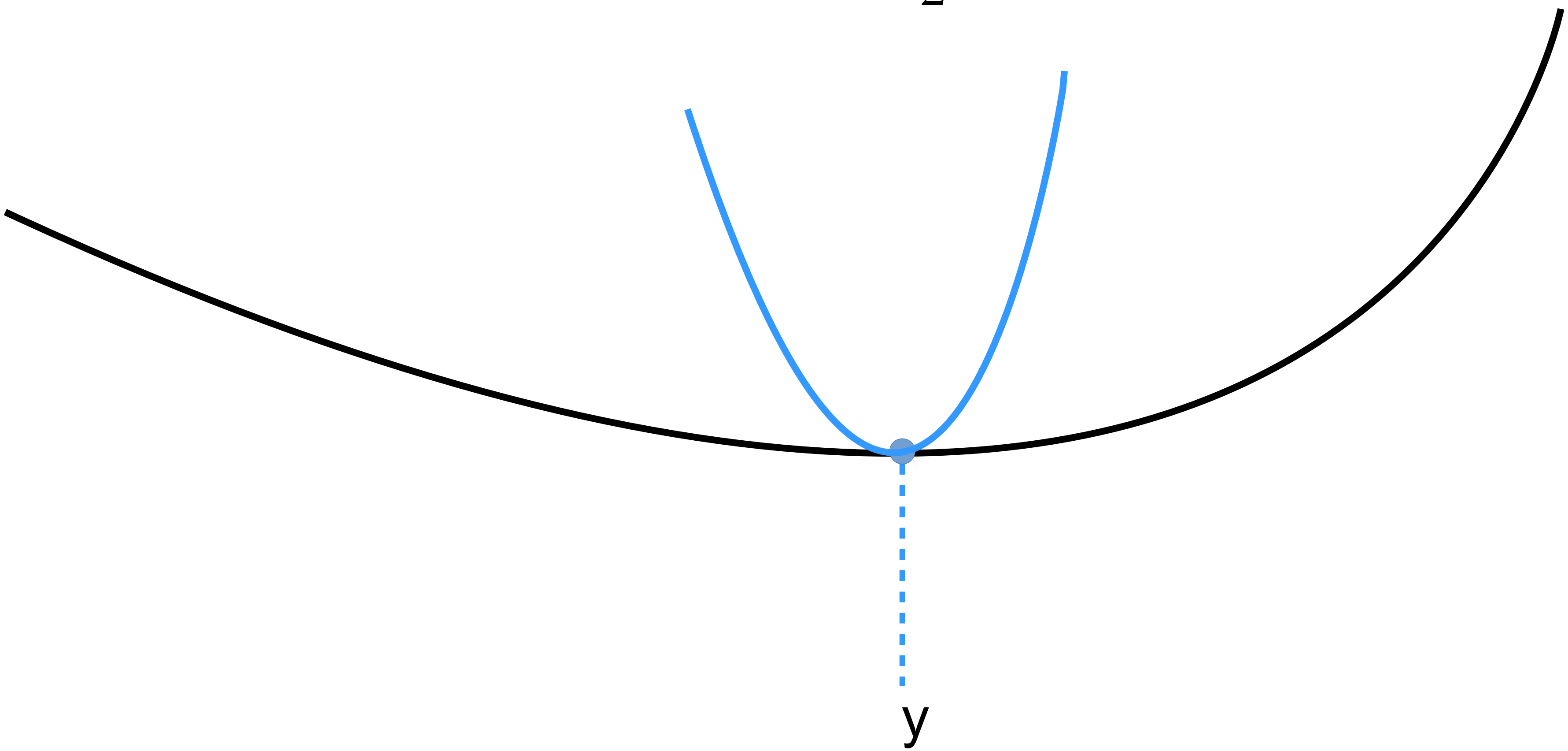
$$x \mapsto \max\{0, 1 - yx\}$$

Negatives log determinant, exponentiation … etc

Smoothness

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

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



Smoothness: Examples

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

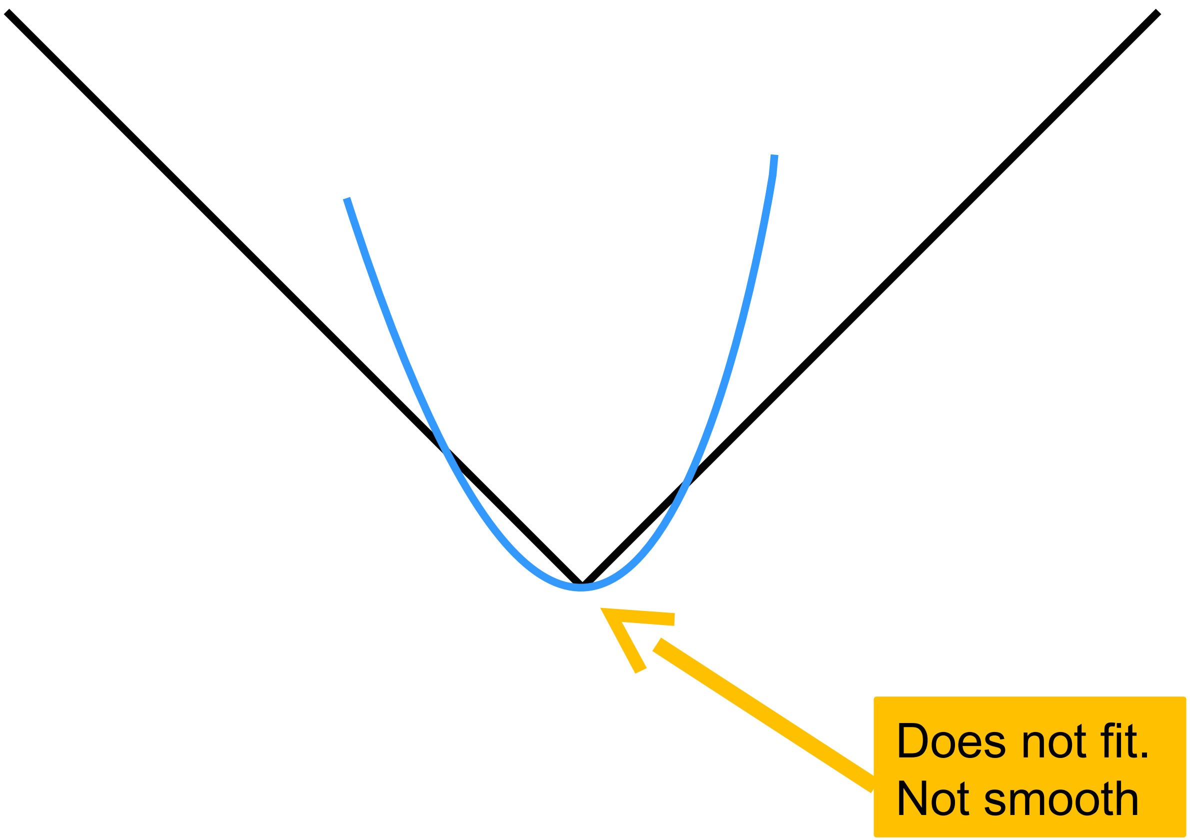
Logistic: $x \mapsto \log\left(1 + e^{-y\langle a, x\rangle}\right)$

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

Proof is an exercise!

Smoothness: Convex counter-example

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



Smoothness Equivalence

A twice differentiable $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is L–smooth if either

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

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

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

EXE: Using that

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

Show that

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

Insight into Gradient Descent

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

Minimizing the upper bound in *w* we get:

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

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

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

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

A gradient descent step !

Smoothness Properties

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

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

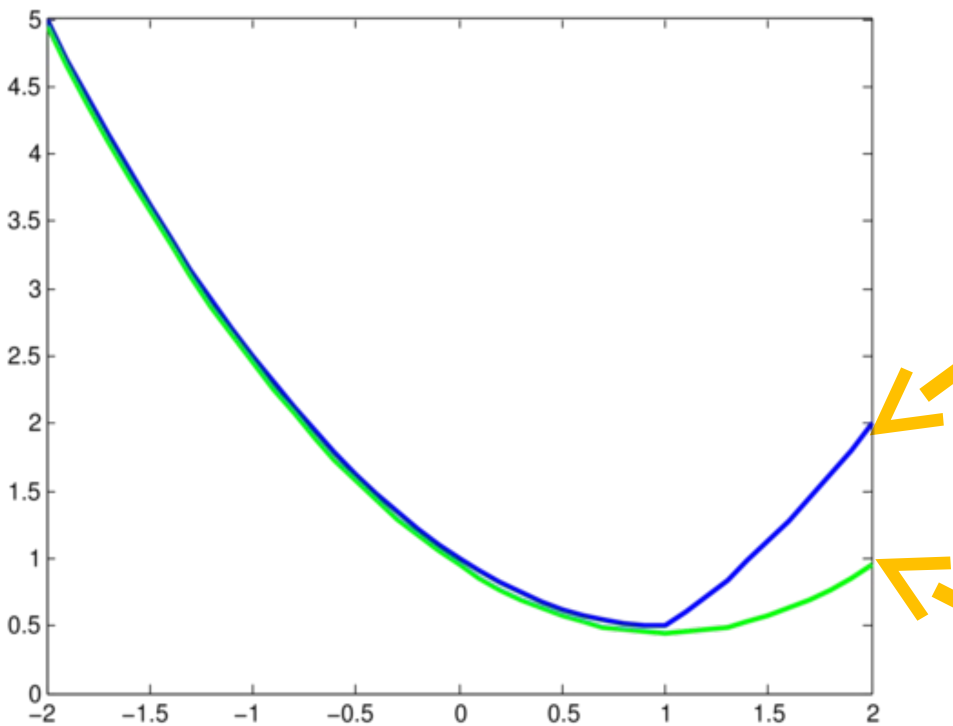
$$f(w^*) - f(w) \leq -\frac{1}{2L}||\nabla f(w)||_2^2, \quad \forall w \in \mathbb{R}^n$$

Proof on board

Strong convexity

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

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



Hinge loss + L2

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

Quadratic lower bound

Strong convexity

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

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

$$d^\top \nabla^2 f(w) d \geq \mu ||d||^2, \quad \forall d \in \mathbb{R}^n$$

EXE: Using that

$$\sigma_{\min}(X)^2 ||d||_2^2 \leq ||X^\top d||_2^2$$

Show that

$$\tfrac{1}{2}||X^\top w - b||_2^2 \text{ is } \sigma_{\min}(X)^2\text{–strongly convex}$$

Convergence GD

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

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

Proof on board

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

Convergence GD I

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

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

Proof on board

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

Strong Convexity Properties

If $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is μ–strongly convex then

$$||\nabla f(w)||_2^2 \geq 2\mu(f(w) - f(w^*)), \quad \forall w \in \mathbb{R}^n$$

This property is known as the *Polyak-Lojasiewicz* inequality

Proof on board

Convex and Smooth Properties

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

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

Co-coercivity

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

Proof on board

Acceleration and lower bouds

The Accelerated gradient method

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

Accelerated gradient

Set $w^1 = 0 = y^1, \kappa = L/\mu$

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

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

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

Output w^{T+1}

Weird extrapolation, but it works

Convergence lower bounds strongly 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 μ–strongly convex such that

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

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

Accelerated gradient has this rate

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

Convergence lower bounds convex

Theorem (Nesterov)

For any optimization algorithm where

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

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

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

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