

# Statistical machine learning and convex optimization

**Francis Bach - Aymeric Dieuleveut**

Mastère M2 - Paris-Sud - Spring 2022

Slides available: [www.di.ens.fr/~fbach/fbach\\_orsay\\_2022.pdf](http://www.di.ens.fr/~fbach/fbach_orsay_2022.pdf)

# Statistical machine learning and convex optimization

- **Six classes** (lecture notes and slides online), Gotomeeting/live
  1. FB: Monday January 24, 2pm to 5pm
  2. FB: Monday January 31, 2pm to 5pm
  3. AD: Monday February 07, 2pm to 5pm
  4. AD: Monday February 14, 2pm to 5pm
  5. AD: Monday February 21, 2pm to 5pm
  6. FB: Monday March 07, 2pm to 5pm
- **Evaluation**
  1. Basic implementations (Matlab / Python / R)
  2. Attending 4 out of 6 classes is mandatory
  3. Short exam (Monday March 28, 2pm to 4/5pm)
- **Register online** (<https://www.di.ens.fr/~fbach/orsay2022.html>)
- Book in preparation: [https://www.di.ens.fr/~fbach/ltfp\\_book.pdf](https://www.di.ens.fr/~fbach/ltfp_book.pdf)

# “Big data” revolution?

## A new scientific context

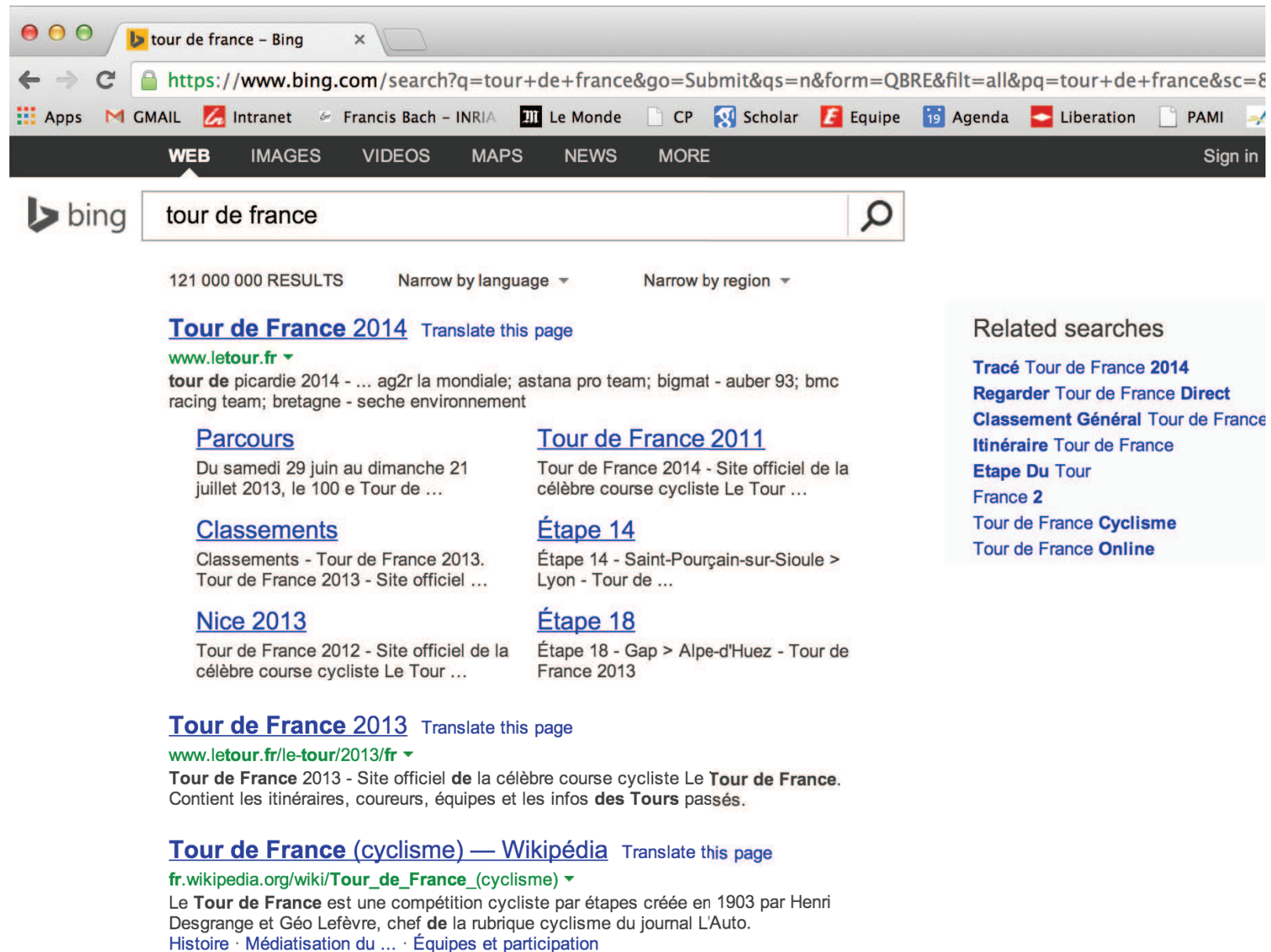
- **Data everywhere:** size does not (always) matter
- **Science and industry**
- **Size and variety**
- **Learning from examples**
  - $n$  observations in dimension  $d$

# Search engines - Advertising

The image shows a screenshot of a Google search results page. The browser's address bar shows the URL: [https://www.google.fr/search?hl=fr&safe=active&q=fete+de+la+science&oq=fete+de+la+sci&gs\\_l=serp.3.0.0i...](https://www.google.fr/search?hl=fr&safe=active&q=fete+de+la+science&oq=fete+de+la+sci&gs_l=serp.3.0.0i...). The search bar contains the text "fete de la science". Below the search bar, the word "Recherche" is displayed in red, followed by the text "Environ 561 000 000 résultats (0,20 secondes)". On the left side, there is a vertical navigation menu with the following items: "Web", "Images", "Maps", "Vidéos", "Actualités", "Shopping", and "Plus". The main content area displays several search results:

- Web**
  - [Accueil - Fête de la science \(site internet\)](#)  
[www.fetedelascience.fr/](http://www.fetedelascience.fr/)  
**Fête de la science 2012**, du 10 au 14 octobre. La science vient à votre rencontre !  
Manipulez, jouez, expérimentez, visitez des laboratoires, dialoguez avec des ...
  - [Les programmes régionaux](#)  
... imprimable. Quel que soit votre choix, toutes les animations ...
  - [Fête de la science 2012](#)  
Villages des sciences, opérations d'envergure, manifestations ...
  - [Déposer un projet ? Le mode ...](#)  
Déposer un projet ? Le mode d'emploi. Bienvenue aux futurs ...
  - [20e édition en 2011](#)  
20e édition en 2011. La Fête de la science se déroule du 12 au 16 ...
  - [Tout savoir sur la Fête de la ...](#)
  - [Les lauréats nationaux](#)

# Search engines - Advertising



The image shows a screenshot of a web browser displaying a Bing search results page for the query "tour de france". The browser's address bar shows the URL: <https://www.bing.com/search?q=tour+de+france&go=Submit&qsn=n&form=QBRE&filt=all&pq=tour+de+france&sc=8>. The search bar contains the text "tour de france". Below the search bar, the results show 121,000,000 results. The first result is for "Tour de France 2014" from [www.letour.fr](http://www.letour.fr), with a snippet: "tour de picardie 2014 - ... ag2r la mondiale; astana pro team; bigmat - auber 93; bmc racing team; bretagne - seche environnement". Other results include "Parcours", "Classements", "Nice 2013", "Tour de France 2011", "Étape 14", and "Étape 18". A "Related searches" sidebar on the right lists: "Tracé Tour de France 2014", "Regarder Tour de France Direct", "Classement Général Tour de France", "Itinéraire Tour de France", "Étape Du Tour", "France 2", "Tour de France Cyclisme", and "Tour de France Online". At the bottom, there is a result for "Tour de France (cyclisme) — Wikipédia" with a snippet: "Le Tour de France est une compétition cycliste par étapes créée en 1903 par Henri Desgrange et Géo Lefèvre, chef de la rubrique cyclisme du journal L'Auto." The browser's taskbar at the top shows various open applications like Gmail, Intranet, Francis Bach - INRIA, Le Monde, CP, Scholar, Equipe, Agenda, Liberation, and PAMI.

tour de france - Bing

<https://www.bing.com/search?q=tour+de+france&go=Submit&qsn=n&form=QBRE&filt=all&pq=tour+de+france&sc=8>

Apps GMAIL Intranet Francis Bach - INRIA Le Monde CP Scholar Equipe 19 Agenda Liberation PAMI

WEB IMAGES VIDEOS MAPS NEWS MORE Sign in

bing tour de france

121 000 000 RESULTS Narrow by language Narrow by region

**Tour de France 2014** [Translate this page](#)  
[www.letour.fr](http://www.letour.fr)  
tour de picardie 2014 - ... ag2r la mondiale; astana pro team; bigmat - auber 93; bmc racing team; bretagne - seche environnement

**Parcours**  
Du samedi 29 juin au dimanche 21 juillet 2013, le 100 e Tour de ...

**Classements**  
Classements - Tour de France 2013.  
Tour de France 2013 - Site officiel ...

**Nice 2013**  
Tour de France 2012 - Site officiel de la célèbre course cycliste Le Tour ...

**Tour de France 2011**  
Tour de France 2014 - Site officiel de la célèbre course cycliste Le Tour ...

**Étape 14**  
Étape 14 - Saint-Pourçain-sur-Sioule > Lyon - Tour de ...

**Étape 18**  
Étape 18 - Gap > Alpe-d'Huez - Tour de France 2013

**Related searches**

- Tracé Tour de France 2014
- Regarder Tour de France Direct
- Classement Général Tour de France
- Itinéraire Tour de France
- Étape Du Tour
- France 2
- Tour de France Cyclisme
- Tour de France Online

**Tour de France 2013** [Translate this page](#)  
[www.letour.fr/le-tour/2013/fr](http://www.letour.fr/le-tour/2013/fr)  
Tour de France 2013 - Site officiel de la célèbre course cycliste Le Tour de France. Contient les itinéraires, coureurs, équipes et les infos des Tours passés.

**Tour de France (cyclisme) — Wikipédia** [Translate this page](#)  
[fr.wikipedia.org/wiki/Tour\\_de\\_France\\_\(cyclisme\)](http://fr.wikipedia.org/wiki/Tour_de_France_(cyclisme))  
Le Tour de France est une compétition cycliste par étapes créée en 1903 par Henri Desgrange et Géo Lefèvre, chef de la rubrique cyclisme du journal L'Auto.  
[Histoire](#) · [Médiatisation du ...](#) · [Équipes et participation](#)

# Advertising

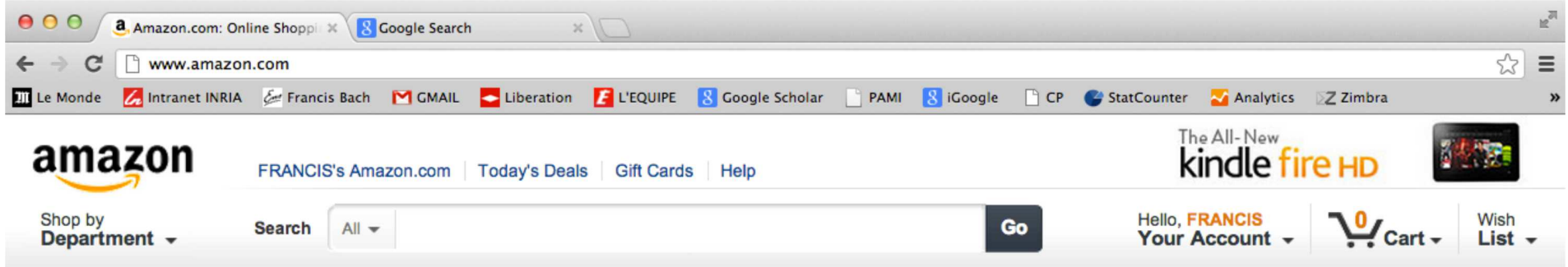
The screenshot shows the Liberation.fr website interface. At the top, there is a browser address bar with the URL www.liberation.fr and a search bar labeled 'Rechercher'. Below the browser bar, the Liberation logo is prominently displayed, along with social media icons for Twitter and Facebook. A navigation menu is visible on the left, and a search icon, an infinity symbol, and the number 100 are on the right.

A large blue banner advertisement for 'PARIS MÔMES' is featured, with the text 'le guide des sorties culturelles pour les 0-12 ans' and an image of a book cover.

Below the banner, there are three main content areas:

- Left Article:** A photograph of a man is shown above a red 'RÉCIT' label. The headline reads: 'Budget : les socialistes pointent un «retour au Moyen Age fiscal»'.
- Middle Article:** A dark background with a red 'DÉCRYPTAGE' label. The headline reads: 'Macron, Robin des bois pour le Trésor, président des riches pour l'OFCE'.
- Right Article:** A 'TOP 100' list with four items:
  - 1 INTERVIEW** Edouard Philippe : «Si ma politique crée des tensions, c'est normal»
  - 2 RÉCIT** Burger King : «On est face à du travail partiellement dissimulé»
  - 3 SANTÉ** Perturbateurs endocriniens: le Parlement européen invalide la définition de la Commission
  - 4 ECONOMIE** Le CICE n'a pas vraiment aidé l'emploi

# Marketing - Personalized recommendation



 Achetez-vous depuis la France? Shopping from France? Essayez **amazon.fr** > [Cliquez ici](#)

 **Get the Free Amazon Mobile App**  
Search & buy millions of products on the go  
[Learn more](#)

Instant Video MP3 Store Cloud Player **Kindle** Cloud Drive Appstore for Android Digital Games & Software Audible Audiobooks

## The All-New Kindle Family

Kindle Paperwhite \$119  
Kindle Fire HD \$199  
Kindle Fire HD 8.9" \$299



Bikes with Street Cred **Clothing Trends** Amazon Prime

THE AMAZON CLOTHING STORE

# Color Theory

Bright outerwear by Nicole Miller, Calvin Klein, Diesel, and more.

[View Looks](#)  
[Shop All Clothing](#)



Understand what the **Zeros and Ones** are telling you.



[Learn more](#)

Advertisement

**3M Streaming Projector Powered by Roku**

Pre-order now for \$20 Amazon Instant Video credit [Learn more](#)

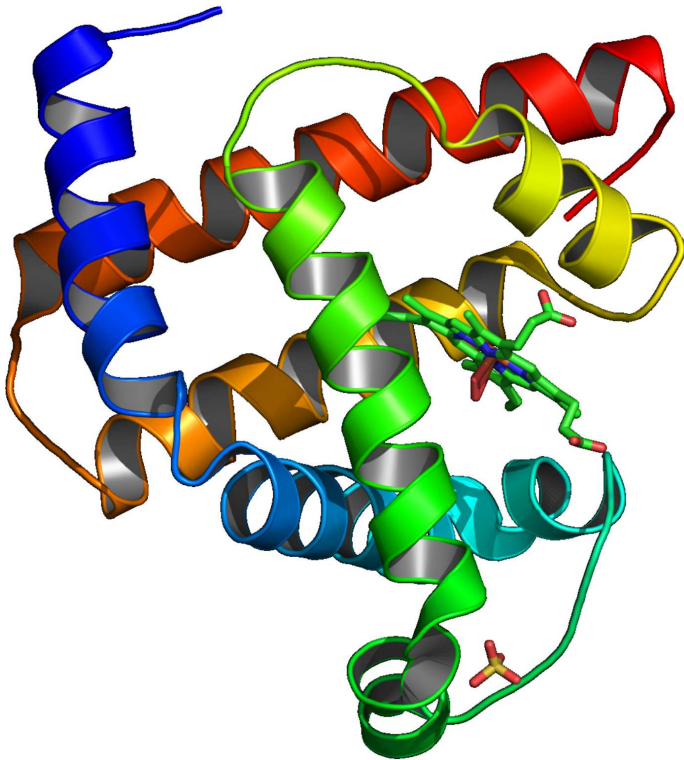


# Visual object recognition





# Bioinformatics



- **Protein:** Crucial elements of cell life
- **Massive data:** 2 millions for humans
- **Complex data**

# Context

## Machine learning for “big data”

- **Large-scale machine learning:** **large  $d$ , large  $n$** 
  - $d$  : dimension of each observation (input)
  - $n$  : number of observations
- **Examples:** computer vision, bioinformatics, advertising

# Context

## Machine learning for “big data”

- **Large-scale machine learning:** **large  $d$ , large  $n$** 
  - $d$  : dimension of each observation (input)
  - $n$  : number of observations
- **Examples:** computer vision, bioinformatics, advertising
- **Ideal running-time complexity:**  $O(dn)$

# Context

## Machine learning for “big data”

- **Large-scale machine learning:** **large  $d$ , large  $n$** 
  - $d$  : dimension of each observation (input)
  - $n$  : number of observations
- **Examples:** computer vision, bioinformatics, advertising
- **Ideal running-time complexity:**  $O(dn)$
- **Going back to simple methods**
  - Stochastic gradient methods (Robbins and Monro, 1951b)
  - Mixing statistics and optimization

# Scaling to large problems

## “Retour aux sources”

- **1950's:** Computers not powerful enough



IBM “1620”, 1959

CPU frequency: 50 KHz

Price > 100 000 dollars

- **2010's:** Data too massive

# Scaling to large problems

## “Retour aux sources”

- **1950's**: Computers not powerful enough



IBM “1620”, 1959

CPU frequency: 50 KHz

Price > 100 000 dollars

- **2010's**: Data too massive
- **Stochastic gradient methods** (Robbins and Monro, 1951a)
  - Going back to simple methods

# Outline - I

## 1. Introduction

- Large-scale machine learning and optimization
- Classes of functions (convex, smooth, etc.)
- Traditional statistical analysis through Rademacher complexity

## 2. Classical methods for convex optimization

- Smooth optimization (gradient descent, Newton method)
- Non-smooth optimization (subgradient descent)
- Proximal methods

## 3. Non-smooth stochastic approximation

- Stochastic (sub)gradient and averaging
- Non-asymptotic results and lower bounds
- Strongly convex vs. non-strongly convex

# Outline - II

## 4. **Classical stochastic approximation**

- Asymptotic analysis
- Robbins-Monro algorithm
- Polyak-Rupert averaging

## 5. **Smooth stochastic approximation algorithms**

- Non-asymptotic analysis for smooth functions
- Logistic regression
- Least-squares regression without decaying step-sizes

## 6. **Finite data sets**

- Gradient methods with exponential convergence rates
- Convex duality
- (Dual) stochastic coordinate descent - Frank-Wolfe



# Supervised machine learning

- **Data:**  $n$  observations  $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ ,  $i = 1, \dots, n$ , **i.i.d.**
- Prediction as a linear function  $\theta^\top \Phi(x)$  of features  $\Phi(x) \in \mathbb{R}^d$
- **(regularized) empirical risk minimization:** find  $\hat{\theta}$  solution of

$$\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \ell(y_i, \theta^\top \Phi(x_i)) + \mu \Omega(\theta)$$

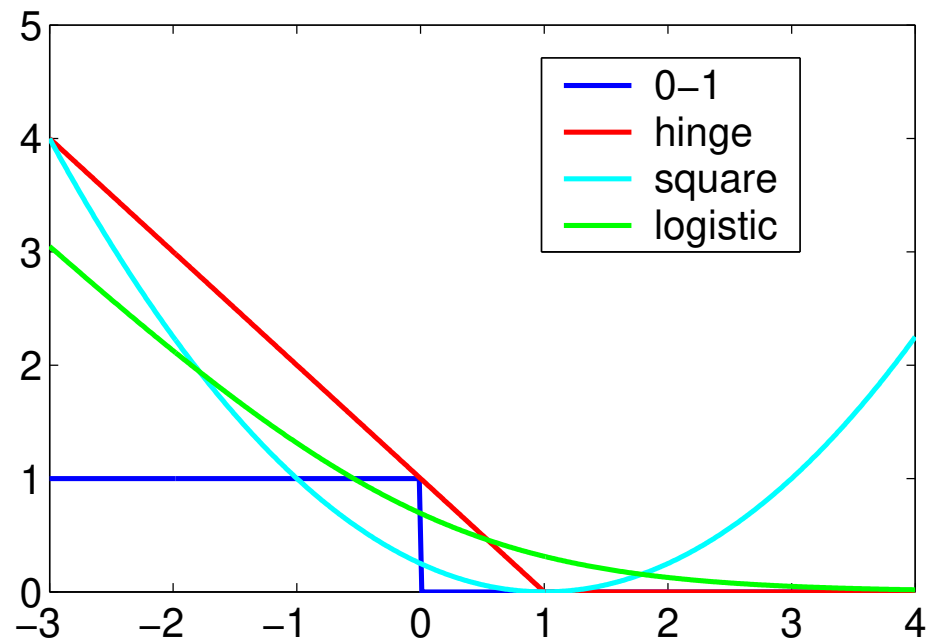
convex data fitting term + regularizer

# Usual losses

- **Regression:**  $y \in \mathbb{R}$ , prediction  $\hat{y} = \theta^\top \Phi(x)$ 
  - quadratic loss  $\frac{1}{2}(y - \hat{y})^2 = \frac{1}{2}(y - \theta^\top \Phi(x))^2$

# Usual losses

- **Regression:**  $y \in \mathbb{R}$ , prediction  $\hat{y} = \theta^\top \Phi(x)$ 
  - quadratic loss  $\frac{1}{2}(y - \hat{y})^2 = \frac{1}{2}(y - \theta^\top \Phi(x))^2$
- **Classification :**  $y \in \{-1, 1\}$ , prediction  $\hat{y} = \text{sign}(\theta^\top \Phi(x))$ 
  - loss of the form  $\ell(y \theta^\top \Phi(x))$
  - “True” **0-1** loss:  $\ell(y \theta^\top \Phi(x)) = 1_{y \theta^\top \Phi(x) < 0}$
  - Usual **convex** losses:



# Main motivating examples

- **Support vector machine** (hinge loss): **non-smooth**

$$\ell(Y, \theta^\top \Phi(X)) = \max\{1 - Y\theta^\top \Phi(X), 0\}$$

- **Logistic regression**: **smooth**

$$\ell(Y, \theta^\top \Phi(X)) = \log(1 + \exp(-Y\theta^\top \Phi(X)))$$

- **Least-squares regression**

$$\ell(Y, \theta^\top \Phi(X)) = \frac{1}{2}(Y - \theta^\top \Phi(X))^2$$

- **Structured output regression**

– See Tsochantaridis et al. (2005); Lacoste-Julien et al. (2013)

# Usual regularizers

- **Main goal:** avoid overfitting
- **(squared) Euclidean norm:**  $\|\theta\|_2^2 = \sum_{j=1}^d |\theta_j|^2$ 
  - Numerically well-behaved
  - Representer theorem and kernel methods :  $\theta = \sum_{i=1}^n \alpha_i \Phi(x_i)$
  - See, e.g., Schölkopf and Smola (2001); Shawe-Taylor and Cristianini (2004)

# Usual regularizers

- **Main goal:** avoid overfitting
- **(squared) Euclidean norm:**  $\|\theta\|_2^2 = \sum_{j=1}^d |\theta_j|^2$ 
  - Numerically well-behaved
  - Representer theorem and kernel methods :  $\theta = \sum_{i=1}^n \alpha_i \Phi(x_i)$
  - See, e.g., Schölkopf and Smola (2001); Shawe-Taylor and Cristianini (2004)
- **Sparsity-inducing norms**
  - Main example:  $\ell_1$ -norm  $\|\theta\|_1 = \sum_{j=1}^d |\theta_j|$
  - Perform model selection as well as regularization
  - Non-smooth optimization and structured sparsity
  - See, e.g., Bach, Jenatton, Mairal, and Obozinski (2012b,a)

# Supervised machine learning

- **Data:**  $n$  observations  $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ ,  $i = 1, \dots, n$ , **i.i.d.**
- Prediction as a linear function  $\theta^\top \Phi(x)$  of features  $\Phi(x) \in \mathbb{R}^d$
- **(regularized) empirical risk minimization:** find  $\hat{\theta}$  solution of

$$\min_{\theta \in \mathbb{R}^d} \quad \frac{1}{n} \sum_{i=1}^n \ell(y_i, \theta^\top \Phi(x_i)) \quad + \quad \mu \Omega(\theta)$$

convex data fitting term + regularizer

# Supervised machine learning

- **Data:**  $n$  observations  $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ ,  $i = 1, \dots, n$ , **i.i.d.**
- Prediction as a linear function  $\theta^\top \Phi(x)$  of features  $\Phi(x) \in \mathbb{R}^d$
- **(regularized) empirical risk minimization:** find  $\hat{\theta}$  solution of

$$\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \ell(y_i, \theta^\top \Phi(x_i)) + \mu \Omega(\theta)$$

convex data fitting term + regularizer

- Empirical risk:  $\hat{f}(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \theta^\top \Phi(x_i))$  **training cost**
- Expected risk:  $f(\theta) = \mathbb{E}_{(x,y)} \ell(y, \theta^\top \Phi(x))$  **testing cost**
- **Two fundamental questions:** (1) computing  $\hat{\theta}$  and (2) analyzing  $\hat{\theta}$



# Supervised machine learning

- **Data:**  $n$  observations  $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ ,  $i = 1, \dots, n$ , **i.i.d.**
- Prediction as a linear function  $\theta^\top \Phi(x)$  of features  $\Phi(x) \in \mathbb{R}^d$
- **(regularized) empirical risk minimization:** find  $\hat{\theta}$  solution of

$$\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \ell(y_i, \theta^\top \Phi(x_i)) + \mu \Omega(\theta)$$

convex data fitting term + regularizer

- Empirical risk:  $\hat{f}(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \theta^\top \Phi(x_i))$  **training cost**
- Expected risk:  $f(\theta) = \mathbb{E}_{(x,y)} \ell(y, \theta^\top \Phi(x))$  **testing cost**
- **Two fundamental questions:** (1) computing  $\hat{\theta}$  and (2) analyzing  $\hat{\theta}$ 
  - **May be tackled simultaneously**

# Supervised machine learning

- **Data:**  $n$  observations  $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ ,  $i = 1, \dots, n$ , **i.i.d.**
- Prediction as a linear function  $\theta^\top \Phi(x)$  of features  $\Phi(x) \in \mathbb{R}^d$
- **(regularized) empirical risk minimization:** find  $\hat{\theta}$  solution of

$$\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \ell(y_i, \theta^\top \Phi(x_i)) \text{ such that } \Omega(\theta) \leq D$$

convex data fitting term + constraint

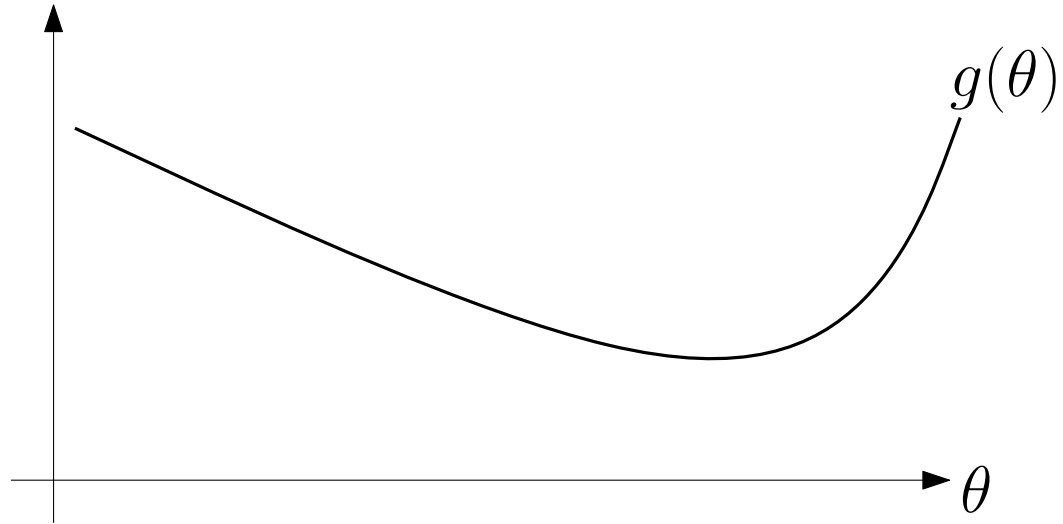
- Empirical risk:  $\hat{f}(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \theta^\top \Phi(x_i))$  **training cost**
- Expected risk:  $f(\theta) = \mathbb{E}_{(x,y)} \ell(y, \theta^\top \Phi(x))$  **testing cost**
- **Two fundamental questions:** (1) computing  $\hat{\theta}$  and (2) analyzing  $\hat{\theta}$ 
  - **May be tackled simultaneously**

# General assumptions

- **Data:**  $n$  observations  $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ ,  $i = 1, \dots, n$ , **i.i.d.**
- Bounded features  $\Phi(x) \in \mathbb{R}^d$ :  $\|\Phi(x)\|_2 \leq R$
- Empirical risk:  $\hat{f}(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \theta^\top \Phi(x_i))$  **training cost**
- Expected risk:  $f(\theta) = \mathbb{E}_{(x,y)} \ell(y, \theta^\top \Phi(x))$  **testing cost**
- Loss for a single observation:  $f_i(\theta) = \ell(y_i, \theta^\top \Phi(x_i))$   
 $\Rightarrow \forall i, f(\theta) = \mathbb{E} f_i(\theta)$
- **Properties of  $f_i, f, \hat{f}$** 
  - **Convex** on  $\mathbb{R}^d$
  - Additional regularity assumptions: Lipschitz-continuity, smoothness and strong convexity

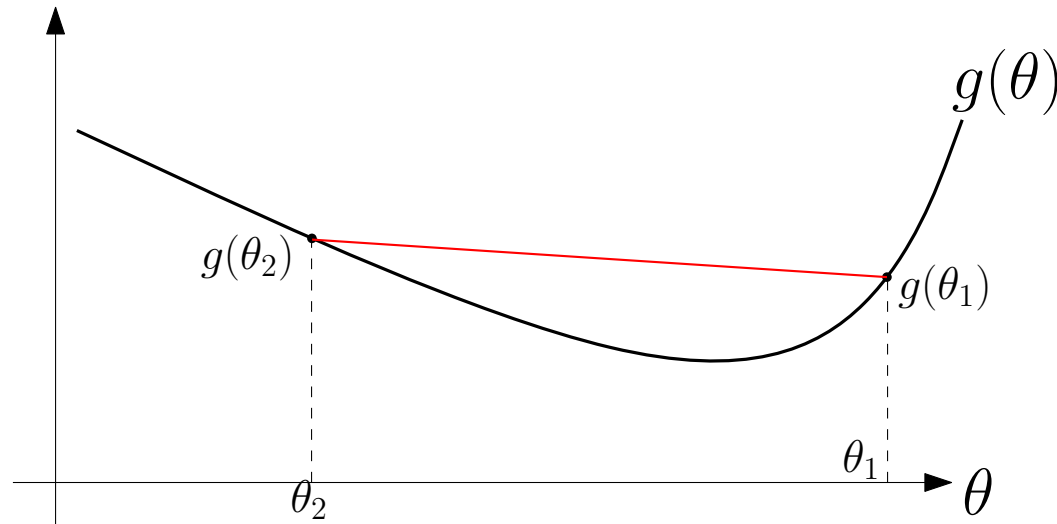
# Convexity

- **Global definitions**



# Convexity

- Global definitions (full domain)

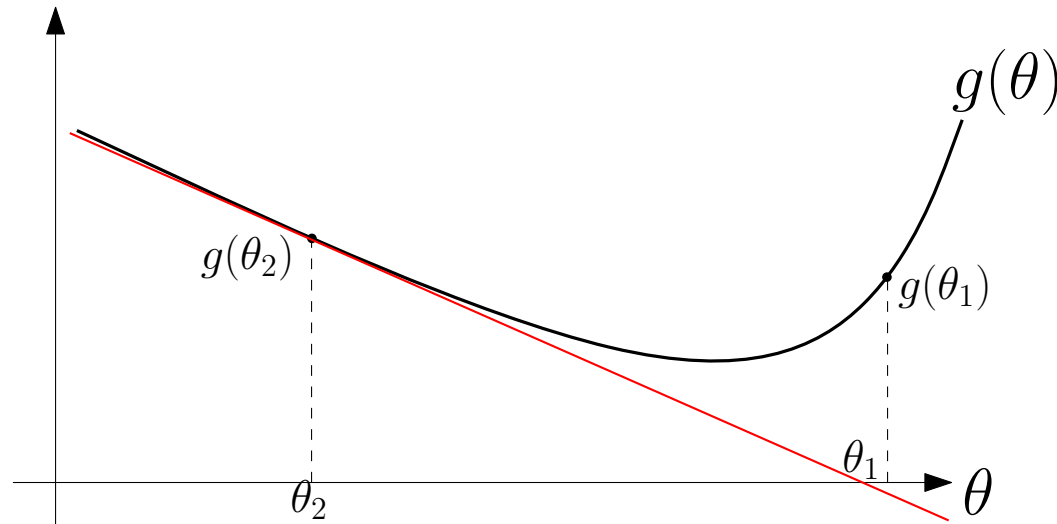


– Not assuming differentiability:

$$\forall \theta_1, \theta_2, \alpha \in [0, 1], \quad g(\alpha\theta_1 + (1 - \alpha)\theta_2) \leq \alpha g(\theta_1) + (1 - \alpha)g(\theta_2)$$

# Convexity

- **Global definitions (full domain)**



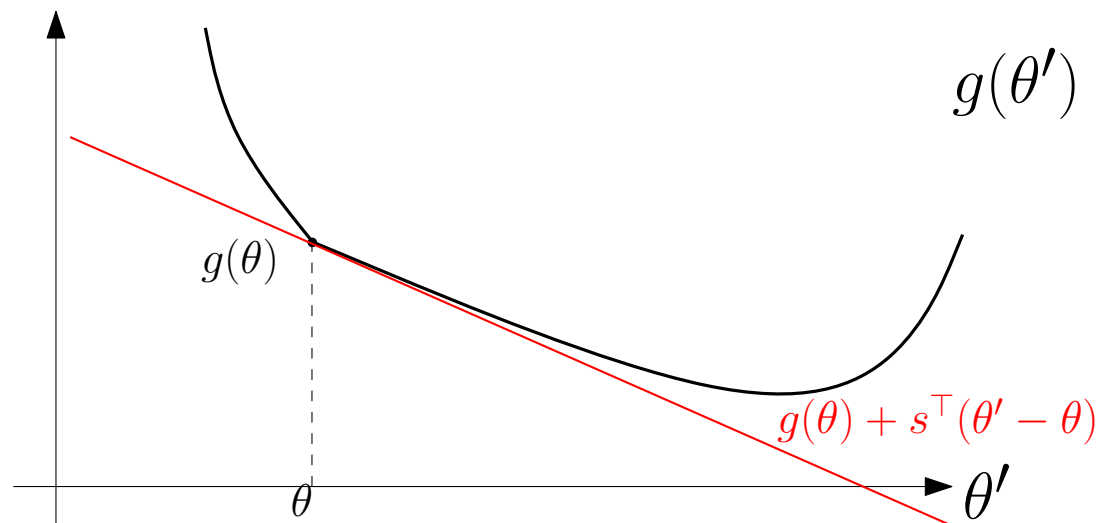
– Assuming differentiability:

$$\forall \theta_1, \theta_2, \quad g(\theta_1) \geq g(\theta_2) + g'(\theta_2)^\top (\theta_1 - \theta_2)$$

- **Extensions to all functions with subgradients / subdifferential**

# Subgradients and subdifferentials

- Given  $g : \mathbb{R}^d \rightarrow \mathbb{R}$  convex



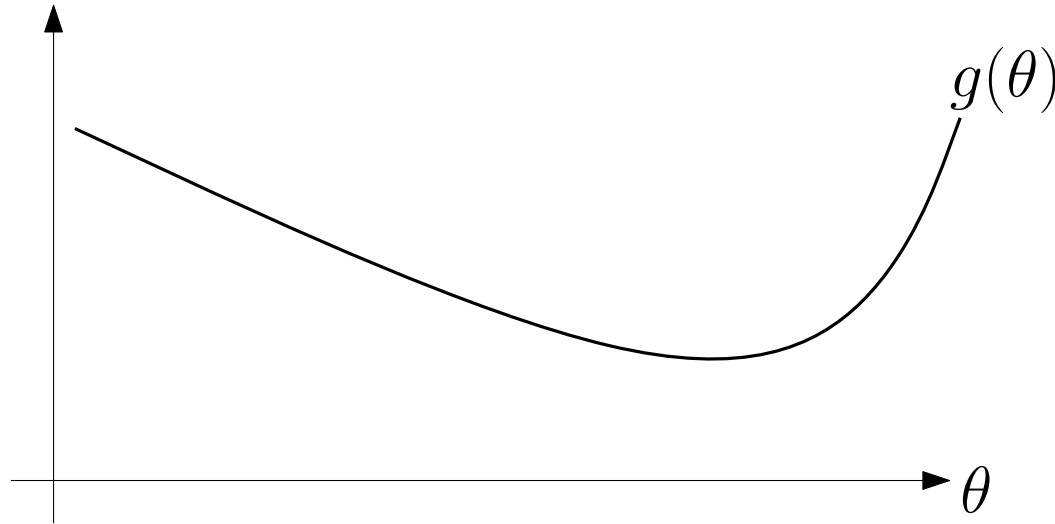
- $s \in \mathbb{R}^d$  is a **subgradient** of  $g$  at  $\theta$  if and only if

$$\forall \theta' \in \mathbb{R}^d, g(\theta') \geq g(\theta) + s^\top (\theta' - \theta)$$

- **Subdifferential**  $\partial g(\theta) =$  set of all subgradients at  $\theta$
  - If  $g$  is differentiable at  $\theta$ , then  $\partial g(\theta) = \{g'(\theta)\}$
  - Example: absolute value
- **The subdifferential is never empty!** See Rockafellar (1997)

# Convexity

- **Global definitions (full domain)**



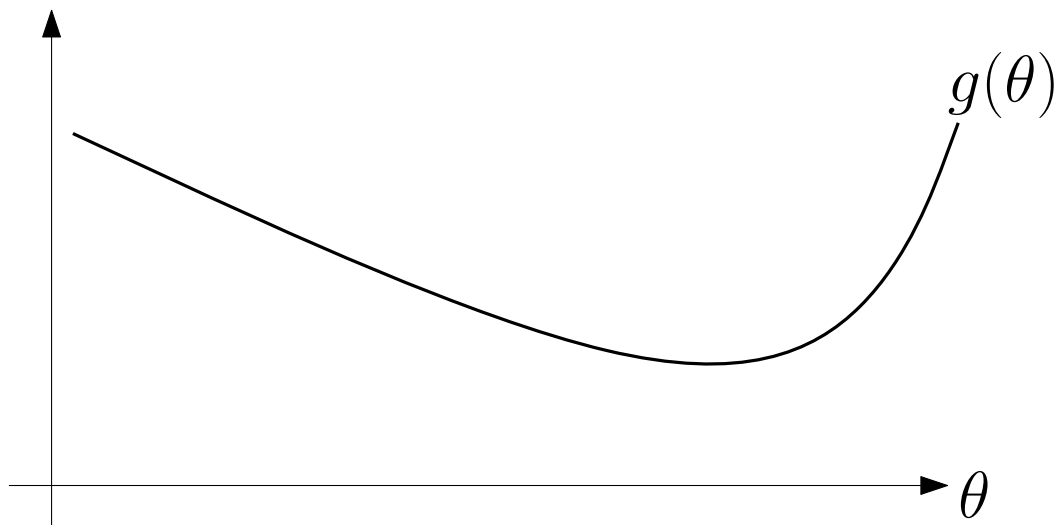
- **Local definitions**

- Twice differentiable functions
- $\forall \theta, g''(\theta) \succcurlyeq 0$  (positive semi-definite Hessians)



# Convexity

- **Global definitions (full domain)**



- **Local definitions**

- Twice differentiable functions
- $\forall \theta, g''(\theta) \succcurlyeq 0$  (positive semi-definite Hessians)

- **Why convexity?**

# Why convexity?

- **Local minimum = global minimum**
  - Optimality condition (non-smooth):  $0 \in \partial g(\theta)$
  - Optimality condition (smooth):  $g'(\theta) = 0$
- **Convex duality**
  - See Boyd and Vandenberghe (2003)
- **Recognizing convex problems**
  - See Boyd and Vandenberghe (2003)

# Lipschitz continuity

- **Bounded gradients of  $g$  ( $\Leftrightarrow$  Lipschitz-continuity):** the function  $g$  if convex, differentiable and has (sub)gradients uniformly bounded by  $B$  on the ball of center 0 and radius  $D$ :

$$\forall \theta \in \mathbb{R}^d, \|\theta\|_2 \leq D \Rightarrow \|g'(\theta)\|_2 \leq B$$

$\Leftrightarrow$

$$\forall \theta, \theta' \in \mathbb{R}^d, \|\theta\|_2, \|\theta'\|_2 \leq D \Rightarrow |g(\theta) - g(\theta')| \leq B\|\theta - \theta'\|_2$$

- **Machine learning**

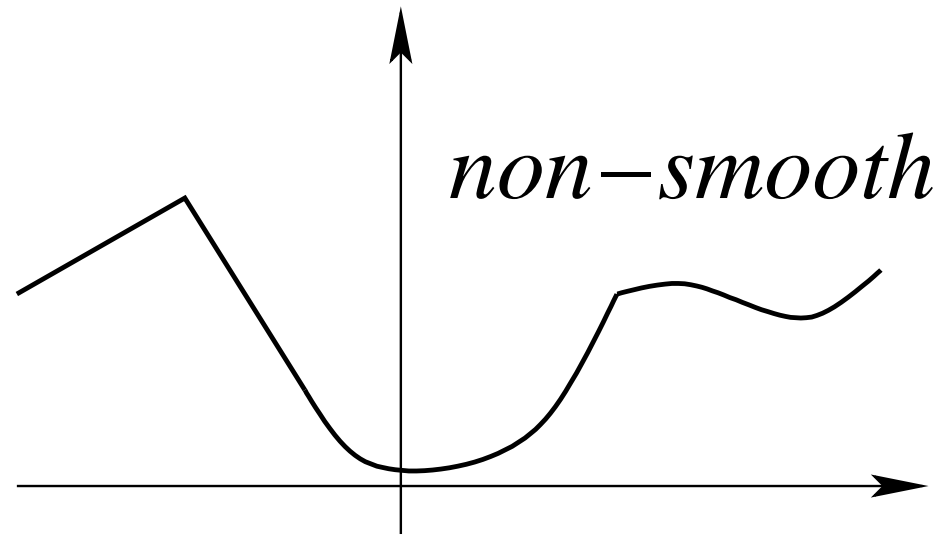
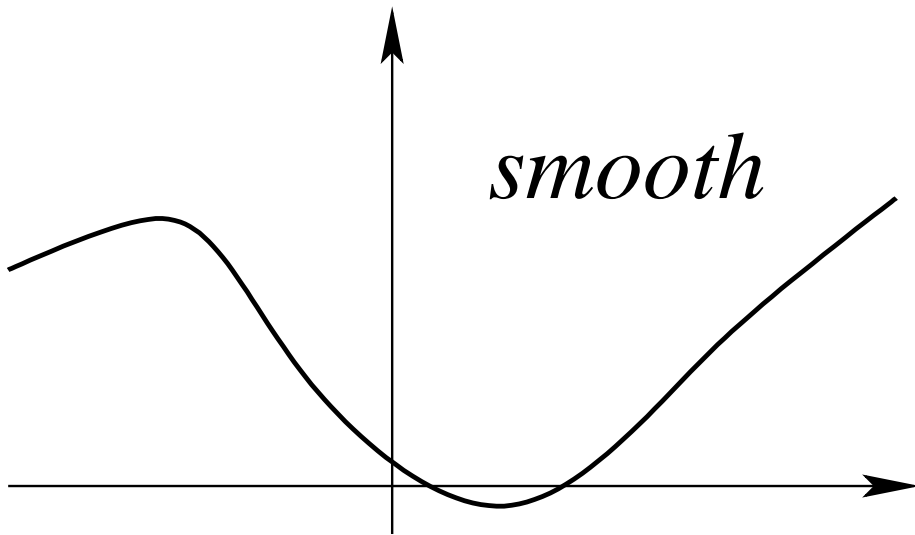
- with  $g(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \theta^\top \Phi(x_i))$
- $G$ -Lipschitz loss and  $R$ -bounded data:  $B = GR$

# Smoothness and strong convexity

- A function  $g : \mathbb{R}^d \rightarrow \mathbb{R}$  is  **$L$ -smooth** if and only if it is differentiable and its gradient is  $L$ -Lipschitz-continuous

$$\forall \theta_1, \theta_2 \in \mathbb{R}^d, \|g'(\theta_1) - g'(\theta_2)\|_2 \leq L \|\theta_1 - \theta_2\|_2$$

- If  $g$  is twice differentiable:  $\forall \theta \in \mathbb{R}^d, g''(\theta) \preceq L \cdot Id$



# Smoothness and strong convexity

- A function  $g : \mathbb{R}^d \rightarrow \mathbb{R}$  is  **$L$ -smooth** if and only if it is differentiable and its gradient is  $L$ -Lipschitz-continuous

$$\forall \theta_1, \theta_2 \in \mathbb{R}^d, \quad \|g'(\theta_1) - g'(\theta_2)\|_2 \leq L \|\theta_1 - \theta_2\|_2$$

- If  $g$  is twice differentiable:  $\forall \theta \in \mathbb{R}^d, \quad g''(\theta) \preceq L \cdot Id$
- **Machine learning**

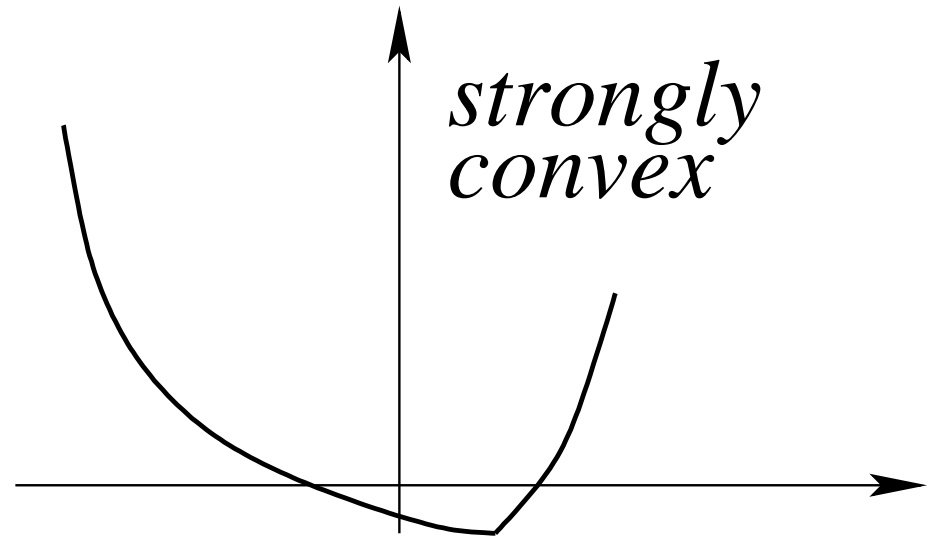
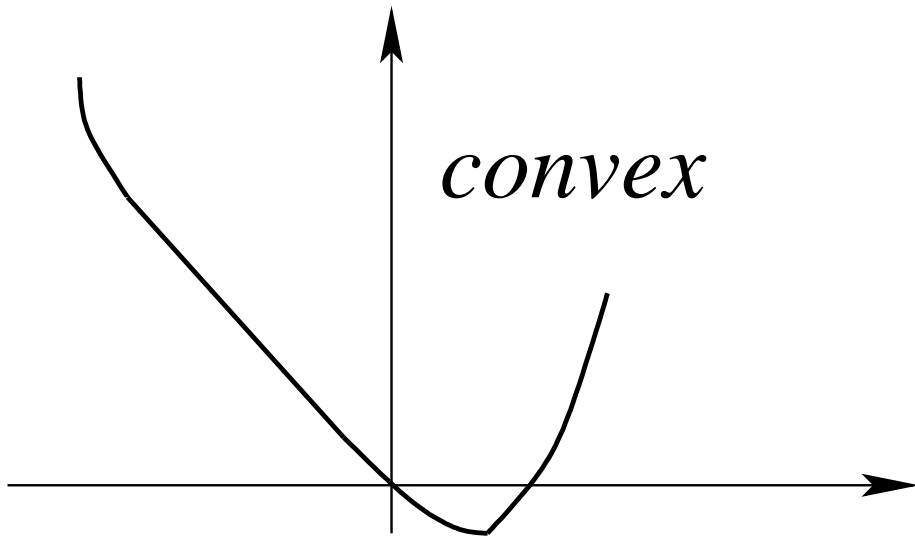
- with  $g(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \theta^\top \Phi(x_i))$
- Hessian  $\approx$  covariance matrix  $\frac{1}{n} \sum_{i=1}^n \Phi(x_i) \Phi(x_i)^\top$
- **$L_{\text{loss}}$ -smooth loss and  $R$ -bounded data:  $L = L_{\text{loss}} R^2$**

# Smoothness and strong convexity

- A function  $g : \mathbb{R}^d \rightarrow \mathbb{R}$  is  $\mu$ -strongly convex if and only if

$$\forall \theta_1, \theta_2 \in \mathbb{R}^d, g(\theta_1) \geq g(\theta_2) + g'(\theta_2)^\top (\theta_1 - \theta_2) + \frac{\mu}{2} \|\theta_1 - \theta_2\|_2^2$$

- If  $g$  is twice differentiable:  $\forall \theta \in \mathbb{R}^d, g''(\theta) \succeq \mu \cdot \text{Id}$



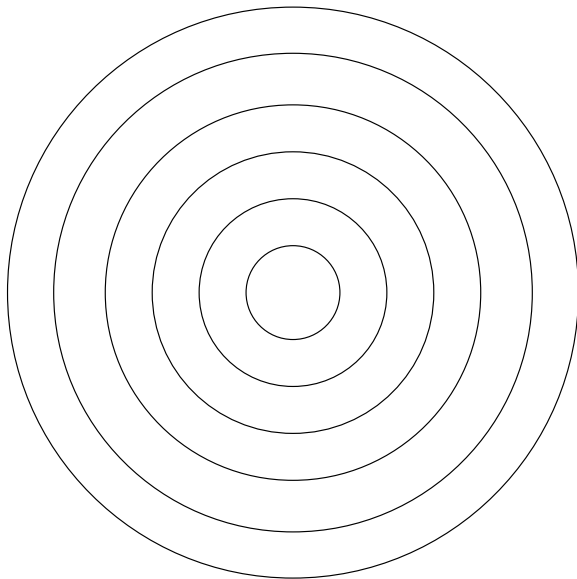
- If  $g$  is convex, then  $g + \frac{\mu}{2} \|\cdot\|_2^2$  is  $\mu$ -strongly convex

# Smoothness and strong convexity

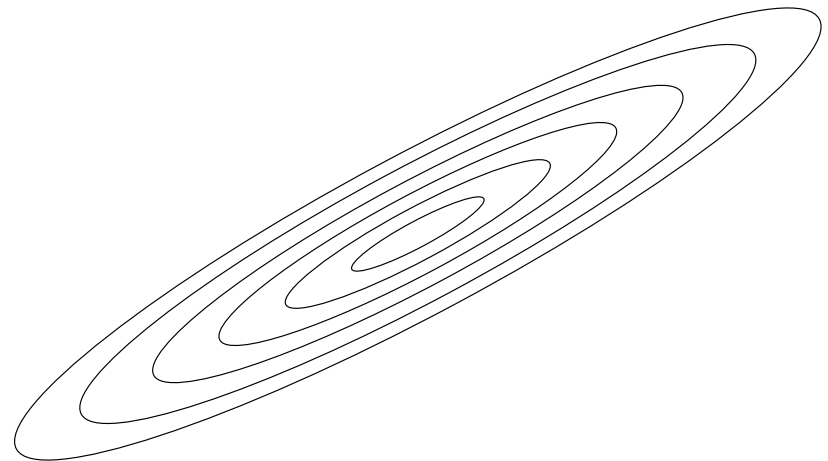
- A function  $g : \mathbb{R}^d \rightarrow \mathbb{R}$  is  $\mu$ -strongly convex if and only if

$$\forall \theta_1, \theta_2 \in \mathbb{R}^d, g(\theta_1) \geq g(\theta_2) + g'(\theta_2)^\top (\theta_1 - \theta_2) + \frac{\mu}{2} \|\theta_1 - \theta_2\|_2^2$$

- If  $g$  is twice differentiable:  $\forall \theta \in \mathbb{R}^d, g''(\theta) \succeq \mu \cdot \text{Id}$



(large  $\mu/L$ )



(small  $\mu/L$ )

# Smoothness and strong convexity

- A function  $g : \mathbb{R}^d \rightarrow \mathbb{R}$  is  $\mu$ -strongly convex if and only if

$$\forall \theta_1, \theta_2 \in \mathbb{R}^d, \quad g(\theta_1) \geq g(\theta_2) + g'(\theta_2)^\top (\theta_1 - \theta_2) + \frac{\mu}{2} \|\theta_1 - \theta_2\|_2^2$$

- If  $g$  is twice differentiable:  $\forall \theta \in \mathbb{R}^d, \quad g''(\theta) \succeq \mu \cdot \text{Id}$

- **Machine learning**

- with  $g(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \theta^\top \Phi(x_i))$
- Hessian  $\approx$  covariance matrix  $\frac{1}{n} \sum_{i=1}^n \Phi(x_i) \Phi(x_i)^\top$
- **Data with invertible covariance matrix** (low correlation/dimension)



# Smoothness and strong convexity

- A function  $g : \mathbb{R}^d \rightarrow \mathbb{R}$  is  $\mu$ -strongly convex if and only if

$$\forall \theta_1, \theta_2 \in \mathbb{R}^d, \quad g(\theta_1) \geq g(\theta_2) + g'(\theta_2)^\top (\theta_1 - \theta_2) + \frac{\mu}{2} \|\theta_1 - \theta_2\|_2^2$$

- If  $g$  is twice differentiable:  $\forall \theta \in \mathbb{R}^d, \quad g''(\theta) \succeq \mu \cdot \text{Id}$

- **Machine learning**

- with  $g(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \theta^\top \Phi(x_i))$
- Hessian  $\approx$  covariance matrix  $\frac{1}{n} \sum_{i=1}^n \Phi(x_i) \Phi(x_i)^\top$
- **Data with invertible covariance matrix** (low correlation/dimension)

- **Adding regularization by  $\frac{\mu}{2} \|\theta\|^2$**

- **creates additional bias unless  $\mu$  is small**

# Summary of smoothness/convexity assumptions

- **Bounded gradients of  $g$  (Lipschitz-continuity):** the function  $g$  is convex, differentiable and has (sub)gradients uniformly bounded by  $B$  on the ball of center 0 and radius  $D$ :

$$\forall \theta \in \mathbb{R}^d, \|\theta\|_2 \leq D \Rightarrow \|g'(\theta)\|_2 \leq B$$

- **Smoothness of  $g$ :** the function  $g$  is convex, differentiable with  $L$ -Lipschitz-continuous gradient  $g'$  (e.g., bounded Hessians):

$$\forall \theta_1, \theta_2 \in \mathbb{R}^d, \|g'(\theta_1) - g'(\theta_2)\|_2 \leq L\|\theta_1 - \theta_2\|_2$$

- **Strong convexity of  $g$ :** The function  $g$  is strongly convex with respect to the norm  $\|\cdot\|$ , with convexity constant  $\mu > 0$ :

$$\forall \theta_1, \theta_2 \in \mathbb{R}^d, g(\theta_1) \geq g(\theta_2) + g'(\theta_2)^\top (\theta_1 - \theta_2) + \frac{\mu}{2}\|\theta_1 - \theta_2\|_2^2$$

# Analysis of empirical risk minimization

- **Approximation and estimation errors:**  $\Theta = \{\theta \in \mathbb{R}^d, \Omega(\theta) \leq D\}$

$$f(\hat{\theta}) - \min_{\theta \in \mathbb{R}^d} f(\theta) = \underbrace{\left[ f(\hat{\theta}) - \min_{\theta \in \Theta} f(\theta) \right]}_{\text{Estimation error}} + \underbrace{\left[ \min_{\theta \in \Theta} f(\theta) - \min_{\theta \in \mathbb{R}^d} f(\theta) \right]}_{\text{Approximation error}}$$

- NB: may replace  $\min_{\theta \in \mathbb{R}^d} f(\theta)$  by best (non-linear) predictions

# Analysis of empirical risk minimization

- **Approximation and estimation errors:**  $\Theta = \{\theta \in \mathbb{R}^d, \Omega(\theta) \leq D\}$

$$f(\hat{\theta}) - \min_{\theta \in \mathbb{R}^d} f(\theta) = \underbrace{\left[ f(\hat{\theta}) - \min_{\theta \in \Theta} f(\theta) \right]}_{\text{Estimation error}} + \underbrace{\left[ \min_{\theta \in \Theta} f(\theta) - \min_{\theta \in \mathbb{R}^d} f(\theta) \right]}_{\text{Approximation error}}$$

**1. Uniform deviation bounds, with**  $\hat{\theta} \in \arg \min_{\theta \in \Theta} \hat{f}(\theta)$

$$\begin{aligned} f(\hat{\theta}) - \min_{\theta \in \Theta} f(\theta) &= [f(\hat{\theta}) - \hat{f}(\hat{\theta})] + [\hat{f}(\hat{\theta}) - \hat{f}((\theta_*)_{\Theta})] + [\hat{f}((\theta_*)_{\Theta}) - f((\theta_*)_{\Theta})] \\ &\leq \sup_{\theta \in \Theta} f(\theta) - \hat{f}(\theta) + 0 + \sup_{\theta \in \Theta} \hat{f}(\theta) - f(\theta) \end{aligned}$$

# Analysis of empirical risk minimization

- **Approximation and estimation errors:**  $\Theta = \{\theta \in \mathbb{R}^d, \Omega(\theta) \leq D\}$

$$f(\hat{\theta}) - \min_{\theta \in \mathbb{R}^d} f(\theta) = \underbrace{\left[ f(\hat{\theta}) - \min_{\theta \in \Theta} f(\theta) \right]}_{\text{Estimation error}} + \underbrace{\left[ \min_{\theta \in \Theta} f(\theta) - \min_{\theta \in \mathbb{R}^d} f(\theta) \right]}_{\text{Approximation error}}$$

1. **Uniform deviation bounds**, with  $\hat{\theta} \in \arg \min_{\theta \in \Theta} \hat{f}(\theta)$

$$f(\hat{\theta}) - \min_{\theta \in \Theta} f(\theta) \leq \underbrace{\sup_{\theta \in \Theta} f(\theta) - \hat{f}(\theta)}_{\text{red}} + \underbrace{\sup_{\theta \in \Theta} \hat{f}(\theta) - f(\theta)}_{\text{red}}$$

– Typically slow rate  $O(1/\sqrt{n})$

2. **More refined concentration results** with faster rates  $O(1/n)$

# Analysis of empirical risk minimization

- **Approximation and estimation errors:**  $\Theta = \{\theta \in \mathbb{R}^d, \Omega(\theta) \leq D\}$

$$f(\hat{\theta}) - \min_{\theta \in \mathbb{R}^d} f(\theta) = \underbrace{\left[ f(\hat{\theta}) - \min_{\theta \in \Theta} f(\theta) \right]}_{\text{Estimation error}} + \underbrace{\left[ \min_{\theta \in \Theta} f(\theta) - \min_{\theta \in \mathbb{R}^d} f(\theta) \right]}_{\text{Approximation error}}$$

1. **Uniform deviation bounds**, with

$$\hat{\theta} \in \arg \min_{\theta \in \Theta} \hat{f}(\theta)$$

$$f(\hat{\theta}) - \min_{\theta \in \Theta} f(\theta) \leq 2 \cdot \sup_{\theta \in \Theta} |f(\theta) - \hat{f}(\theta)|$$

– Typically slow rate  $O(1/\sqrt{n})$

2. **More refined concentration results** with faster rates  $O(1/n)$

# Motivation from least-squares

- For least-squares, we have  $\ell(y, \theta^\top \Phi(x)) = \frac{1}{2}(y - \theta^\top \Phi(x))^2$ , and

$$\begin{aligned}\hat{f}(\theta) - f(\theta) &= \frac{1}{2}\theta^\top \left( \frac{1}{n} \sum_{i=1}^n \Phi(x_i)\Phi(x_i)^\top - \mathbb{E}\Phi(X)\Phi(X)^\top \right) \theta \\ &\quad - \theta^\top \left( \frac{1}{n} \sum_{i=1}^n y_i \Phi(x_i) - \mathbb{E}Y\Phi(X) \right) + \frac{1}{2} \left( \frac{1}{n} \sum_{i=1}^n y_i^2 - \mathbb{E}Y^2 \right),\end{aligned}$$

$$\begin{aligned}\sup_{\|\theta\|_2 \leq D} |f(\theta) - \hat{f}(\theta)| &\leq \frac{D^2}{2} \left\| \frac{1}{n} \sum_{i=1}^n \Phi(x_i)\Phi(x_i)^\top - \mathbb{E}\Phi(X)\Phi(X)^\top \right\|_{\text{op}} \\ &\quad + D \left\| \frac{1}{n} \sum_{i=1}^n y_i \Phi(x_i) - \mathbb{E}Y\Phi(X) \right\|_2 + \frac{1}{2} \left| \frac{1}{n} \sum_{i=1}^n y_i^2 - \mathbb{E}Y^2 \right|,\end{aligned}$$

$$\sup_{\|\theta\|_2 \leq D} |f(\theta) - \hat{f}(\theta)| \leq O(1/\sqrt{n}) \text{ with high probability from 3 concentrations}$$

# Slow rate for supervised learning

- **Assumptions** ( $f$  is the expected risk,  $\hat{f}$  the empirical risk)
  - $\Omega(\theta) = \|\theta\|_2$  (Euclidean norm)
  - “Linear” predictors:  $\theta(x) = \theta^\top \Phi(x)$ , with  $\|\Phi(x)\|_2 \leq R$  a.s.
  - $G$ -Lipschitz loss:  $f$  and  $\hat{f}$  are  $GR$ -Lipschitz on  $\Theta = \{\|\theta\|_2 \leq D\}$
  - **No assumptions regarding convexity**



# Slow rate for supervised learning

- **Assumptions** ( $f$  is the expected risk,  $\hat{f}$  the empirical risk)
  - $\Omega(\theta) = \|\theta\|_2$  (Euclidean norm)
  - “Linear” predictors:  $\theta(x) = \theta^\top \Phi(x)$ , with  $\|\Phi(x)\|_2 \leq R$  a.s.
  - $G$ -Lipschitz loss:  $f$  and  $\hat{f}$  are  $GR$ -Lipschitz on  $\Theta = \{\|\theta\|_2 \leq D\}$
  - **No assumptions regarding convexity**

- With probability greater than  $1 - \delta$

$$\sup_{\theta \in \Theta} |\hat{f}(\theta) - f(\theta)| \leq \frac{\ell_0 + GRD}{\sqrt{n}} \left[ 2 + \sqrt{2 \log \frac{2}{\delta}} \right]$$

- Expected estimation error:  $\mathbb{E} \left[ \sup_{\theta \in \Theta} |\hat{f}(\theta) - f(\theta)| \right] \leq \frac{4\ell_0 + 4GRD}{\sqrt{n}}$

- Using Rademacher averages (see, e.g., Boucheron et al., 2005)

- **Lipschitz functions  $\Rightarrow$  slow rate**

# Symmetrization with Rademacher variables

- Let  $\mathcal{D}' = \{x'_1, y'_1, \dots, x'_n, y'_n\}$  an **independent copy** of the data  $\mathcal{D} = \{x_1, y_1, \dots, x_n, y_n\}$ , with corresponding loss functions  $f'_i(\theta)$

$$\begin{aligned}\mathbb{E}\left[\sup_{\theta \in \Theta} f(\theta) - \hat{f}(\theta)\right] &= \mathbb{E}\left[\sup_{\theta \in \Theta} \left(f(\theta) - \frac{1}{n} \sum_{i=1}^n f_i(\theta)\right)\right] \\ &= \mathbb{E}\left[\sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \mathbb{E}(f'_i(\theta) - f_i(\theta) | \mathcal{D})\right] \\ &\leq \mathbb{E}\left[\mathbb{E}\left[\sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n (f'_i(\theta) - f_i(\theta)) \mid \mathcal{D}\right]\right] \\ &= \mathbb{E}\left[\sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n (f'_i(\theta) - f_i(\theta))\right] \\ &= \mathbb{E}\left[\sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \varepsilon_i (f'_i(\theta) - f_i(\theta))\right] \text{ with } \varepsilon_i \text{ uniform in } \{-1, 1\} \\ &\leq 2\mathbb{E}\left[\sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f_i(\theta)\right] = \text{Rademacher complexity}\end{aligned}$$

# Rademacher complexity

- Rademacher complexity of the class of functions  $(X, Y) \mapsto \ell(Y, \theta^\top \Phi(X))$

$$R_n = \mathbb{E} \left[ \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f_i(\theta) \right]$$

– with  $f_i(\theta) = \ell(x_i, \theta^\top \Phi(x_i))$ ,  $(x_i, y_i)$ , i.i.d

- NB 1: **two** expectations, with respect to  $\mathcal{D}$  and with respect to  $\varepsilon$ 
  - “Empirical” Rademacher average  $\hat{R}_n$  by conditioning on  $\mathcal{D}$

- NB 2: sometimes defined as  $\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f_i(\theta) \right|$

- **Main property:**

$$\mathbb{E} \left[ \sup_{\theta \in \Theta} f(\theta) - \hat{f}(\theta) \right] \text{ and } \mathbb{E} \left[ \sup_{\theta \in \Theta} \hat{f}(\theta) - f(\theta) \right] \leq 2R_n$$

# From Rademacher complexity to uniform bound

- Let  $Z = \sup_{\theta \in \Theta} |f(\theta) - \hat{f}(\theta)|$
- By changing the pair  $(x_i, y_i)$ ,  $Z$  may only change by

$$\frac{2}{n} \sup |\ell(Y, \theta^\top \Phi(X))| \leq \frac{2}{n} (\sup |\ell(Y, 0)| + GRD) \leq \frac{2}{n} (\ell_0 + GRD) = c$$

with  $\sup |\ell(Y, 0)| = \ell_0$

- **MacDiarmid inequality:** with probability greater than  $1 - \delta$ ,

$$Z \leq \mathbb{E}Z + \sqrt{\frac{n}{2}}c \cdot \sqrt{\log \frac{1}{\delta}} \leq 2R_n + \frac{\sqrt{2}}{\sqrt{n}}(\ell_0 + GRD) \sqrt{\log \frac{1}{\delta}}$$

# Bounding the Rademacher average - I

- We have, with  $\varphi_i(u) = \ell(y_i, u) - \ell(y_i, 0)$  is almost surely  $G$ -Lipschitz:

$$\begin{aligned}\hat{R}_n &= \mathbb{E}_\varepsilon \left[ \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f_i(\theta) \right] \\ &= \mathbb{E}_\varepsilon \left[ \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f_i(0) \right] + \mathbb{E}_\varepsilon \left[ \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \varepsilon_i [f_i(\theta) - f_i(0)] \right] \\ &= 0 + \mathbb{E}_\varepsilon \left[ \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \varepsilon_i [f_i(\theta) - f_i(0)] \right] \\ &= 0 + \mathbb{E}_\varepsilon \left[ \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \varphi_i(\theta^\top \Phi(x_i)) \right]\end{aligned}$$

- Using Ledoux-Talagrand contraction results for Rademacher averages (since  $\varphi_i$  is  $G$ -Lipschitz), we get (Meir and Zhang, 2003):

$$\hat{R}_n \leq G \cdot \mathbb{E}_\varepsilon \left[ \sup_{\|\theta\|_2 \leq D} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \theta^\top \Phi(x_i) \right]$$

# Proof of Ledoux-Talagrand lemma (Meir and Zhang, 2003, Lemma 5)

- Given any  $b, a_i : \Theta \rightarrow \mathbb{R}$  (no assumption) and  $\varphi_i : \mathbb{R} \rightarrow \mathbb{R}$  any 1-Lipschitz-functions,  $i = 1, \dots, n$

$$\mathbb{E}_\varepsilon \left[ \sup_{\theta \in \Theta} b(\theta) + \sum_{i=1}^n \varepsilon_i \varphi_i(a_i(\theta)) \right] \leq \mathbb{E}_\varepsilon \left[ \sup_{\theta \in \Theta} b(\theta) + \sum_{i=1}^n \varepsilon_i a_i(\theta) \right]$$

- **Proof by induction on  $n$**

–  $n = 0$ : trivial

- From  $n$  to  $n + 1$ : see next slide

# From $n$ to $n + 1$

$$\begin{aligned}
& \mathbb{E}_{\varepsilon_1, \dots, \varepsilon_{n+1}} \left[ \sup_{\theta \in \Theta} b(\theta) + \sum_{i=1}^{n+1} \varepsilon_i \varphi_i(a_i(\theta)) \right] \\
&= \mathbb{E}_{\varepsilon_1, \dots, \varepsilon_n} \left[ \sup_{\theta, \theta' \in \Theta} \frac{b(\theta) + b(\theta')}{2} + \sum_{i=1}^n \varepsilon_i \frac{\varphi_i(a_i(\theta)) + \varphi_i(a_i(\theta'))}{2} + \frac{\varphi_{n+1}(a_{n+1}(\theta)) - \varphi_{n+1}(a_{n+1}(\theta'))}{2} \right] \\
&= \mathbb{E}_{\varepsilon_1, \dots, \varepsilon_n} \left[ \sup_{\theta, \theta' \in \Theta} \frac{b(\theta) + b(\theta')}{2} + \sum_{i=1}^n \varepsilon_i \frac{\varphi_i(a_i(\theta)) + \varphi_i(a_i(\theta'))}{2} + \frac{|\varphi_{n+1}(a_{n+1}(\theta)) - \varphi_{n+1}(a_{n+1}(\theta'))|}{2} \right] \\
&\leq \mathbb{E}_{\varepsilon_1, \dots, \varepsilon_n} \left[ \sup_{\theta, \theta' \in \Theta} \frac{b(\theta) + b(\theta')}{2} + \sum_{i=1}^n \varepsilon_i \frac{\varphi_i(a_i(\theta)) + \varphi_i(a_i(\theta'))}{2} + \frac{|a_{n+1}(\theta) - a_{n+1}(\theta')|}{2} \right] \\
&= \mathbb{E}_{\varepsilon_1, \dots, \varepsilon_n} \mathbb{E}_{\varepsilon_{n+1}} \left[ \sup_{\theta \in \Theta} b(\theta) + \varepsilon_{n+1} a_{n+1}(\theta) + \sum_{i=1}^n \varepsilon_i \varphi_i(a_i(\theta)) \right] \\
&\leq \mathbb{E}_{\varepsilon_1, \dots, \varepsilon_n, \varepsilon_{n+1}} \left[ \sup_{\theta \in \Theta} b(\theta) + \varepsilon_{n+1} a_{n+1}(\theta) + \sum_{i=1}^n \varepsilon_i a_i(\theta) \right] \text{ by recursion}
\end{aligned}$$

# Bounding the Rademacher average - II

- We have:

$$\begin{aligned} R_n &\leq 2G\mathbb{E}\left[\sup_{\|\theta\|_2 \leq D} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \theta^\top \Phi(x_i)\right] \\ &= 2G\mathbb{E}\left\|D \frac{1}{n} \sum_{i=1}^n \varepsilon_i \Phi(x_i)\right\|_2 \\ &\leq 2GD \sqrt{\mathbb{E}\left\|\frac{1}{n} \sum_{i=1}^n \varepsilon_i \Phi(x_i)\right\|_2^2} \text{ by Jensen's inequality} \\ &\leq \frac{2GRD}{\sqrt{n}} \text{ by using } \|\Phi(x)\|_2 \leq R \text{ and independence} \end{aligned}$$

- Overall, we get, with probability  $1 - \delta$ :

$$\sup_{\theta \in \Theta} |f(\theta) - \hat{f}(\theta)| \leq \frac{1}{\sqrt{n}} (\ell_0 + GRD) \left(4 + \sqrt{2 \log \frac{1}{\delta}}\right)$$



## Putting it all together

- We have, with probability  $1 - \delta$ 
  - For exact minimizer  $\hat{\theta} \in \arg \min_{\theta \in \Theta} \hat{f}(\theta)$ , we have

$$\begin{aligned} f(\hat{\theta}) - \min_{\theta \in \Theta} f(\theta) &\leq \sup_{\theta \in \Theta} \hat{f}(\theta) - f(\theta) + \sup_{\theta \in \Theta} f(\theta) - \hat{f}(\theta) \\ &\leq \frac{2}{\sqrt{n}}(\ell_0 + GRD)(4 + \sqrt{2 \log \frac{1}{\delta}}) \end{aligned}$$

- For inexact minimizer  $\eta \in \Theta$

$$f(\eta) - \min_{\theta \in \Theta} f(\theta) \leq 2 \cdot \sup_{\theta \in \Theta} |\hat{f}(\theta) - f(\theta)| + [\hat{f}(\eta) - \hat{f}(\hat{\theta})]$$

- **Only need to optimize with precision  $\frac{2}{\sqrt{n}}(\ell_0 + GRD)$**

## Putting it all together

- We have, with probability  $1 - \delta$ 
  - For exact minimizer  $\hat{\theta} \in \arg \min_{\theta \in \Theta} \hat{f}(\theta)$ , we have

$$\begin{aligned} f(\hat{\theta}) - \min_{\theta \in \Theta} f(\theta) &\leq 2 \cdot \sup_{\theta \in \Theta} |\hat{f}(\theta) - f(\theta)| \\ &\leq \frac{2}{\sqrt{n}} (\ell_0 + GRD) (4 + \sqrt{2 \log \frac{1}{\delta}}) \end{aligned}$$

- For inexact minimizer  $\eta \in \Theta$

$$f(\eta) - \min_{\theta \in \Theta} f(\theta) \leq 2 \cdot \sup_{\theta \in \Theta} |\hat{f}(\theta) - f(\theta)| + [\hat{f}(\eta) - \hat{f}(\hat{\theta})]$$

- **Only need to optimize with precision  $\frac{2}{\sqrt{n}}(\ell_0 + GRD)$**

# Slow rate for supervised learning (summary)

- **Assumptions** ( $f$  is the expected risk,  $\hat{f}$  the empirical risk)
  - $\Omega(\theta) = \|\theta\|_2$  (Euclidean norm)
  - “Linear” predictors:  $\theta(x) = \theta^\top \Phi(x)$ , with  $\|\Phi(x)\|_2 \leq R$  a.s.
  - $G$ -Lipschitz loss:  $f$  and  $\hat{f}$  are  $GR$ -Lipschitz on  $\Theta = \{\|\theta\|_2 \leq D\}$
  - **No assumptions regarding convexity**

- With probability greater than  $1 - \delta$

$$\sup_{\theta \in \Theta} |\hat{f}(\theta) - f(\theta)| \leq \frac{(\ell_0 + GRD)}{\sqrt{n}} \left[ 2 + \sqrt{2 \log \frac{2}{\delta}} \right]$$

- Expected estimation error:  $\mathbb{E} \left[ \sup_{\theta \in \Theta} |\hat{f}(\theta) - f(\theta)| \right] \leq \frac{4(\ell_0 + GRD)}{\sqrt{n}}$

- Using Rademacher averages (see, e.g., Boucheron et al., 2005)

- **Lipschitz functions  $\Rightarrow$  slow rate**

# Motivation from mean estimation

- Estimator  $\hat{\theta} = \frac{1}{n} \sum_{i=1}^n z_i = \arg \min_{\theta \in \mathbb{R}} \frac{1}{2n} \sum_{i=1}^n (\theta - z_i)^2 = \hat{f}(\theta)$ 
  - $\theta_* = \mathbb{E}z = \arg \min_{\theta \in \mathbb{R}} \frac{1}{2} \mathbb{E}(\theta - z)^2 = f(\theta)$
  - From before (estimation error):  $f(\hat{\theta}) - f(\theta_*) = O(1/\sqrt{n})$

# Motivation from mean estimation

- Estimator  $\hat{\theta} = \frac{1}{n} \sum_{i=1}^n z_i = \arg \min_{\theta \in \mathbb{R}} \frac{1}{2n} \sum_{i=1}^n (\theta - z_i)^2 = \hat{f}(\theta)$ 
  - $\theta_* = \mathbb{E}z = \arg \min_{\theta \in \mathbb{R}} \frac{1}{2} \mathbb{E}(\theta - z)^2 = f(\theta)$
  - From before (estimation error):  $f(\hat{\theta}) - f(\theta_*) = O(1/\sqrt{n})$

- Direct computation:

$$- f(\theta) = \frac{1}{2} \mathbb{E}(\theta - z)^2 = \frac{1}{2}(\theta - \mathbb{E}z)^2 + \frac{1}{2} \text{var}(z)$$

- More refined/direct bound:

$$\begin{aligned} f(\hat{\theta}) - f(\mathbb{E}z) &= \frac{1}{2}(\hat{\theta} - \mathbb{E}z)^2 \\ \mathbb{E}[f(\hat{\theta}) - f(\mathbb{E}z)] &= \frac{1}{2} \mathbb{E} \left( \frac{1}{n} \sum_{i=1}^n z_i - \mathbb{E}z \right)^2 = \frac{1}{2n} \text{var}(z) \end{aligned}$$

- **Bound only at  $\hat{\theta}$  + strong convexity** (instead of uniform bound)

# Fast rate for supervised learning

- **Assumptions** ( $f$  is the expected risk,  $\hat{f}$  the empirical risk)
  - Same as before (bounded features, Lipschitz loss)
  - Regularized risks:  $f^\mu(\theta) = f(\theta) + \frac{\mu}{2}\|\theta\|_2^2$  and  $\hat{f}^\mu(\theta) = \hat{f}(\theta) + \frac{\mu}{2}\|\theta\|_2^2$
  - **Convexity**
- For any  $a > 0$ , with probability greater than  $1 - \delta$ , for all  $\theta \in \mathbb{R}^d$ ,
$$f^\mu(\hat{\theta}) - \min_{\eta \in \mathbb{R}^d} f^\mu(\eta) \leq \frac{8G^2 R^2 (32 + \log \frac{1}{\delta})}{\mu n}$$
- Results from Sridharan, Srebro, and Shalev-Shwartz (2008)
  - see also Boucheron and Massart (2011) and references therein
- **Strongly convex functions  $\Rightarrow$  fast rate**
  - Warning:  $\mu$  should decrease with  $n$  to reduce approximation error