Concentration of measure in probability and high-dimensional statistical learning

Guillaume Aubrun, Aurélien Garivier, Rémi Gribonval

remi.gribonval@inria.fr

http://perso.ens-lyon.fr/remi.gribonval

Last week - CS only

- **Deviations for the averages of random variables**
  - ✓ Weak law of large numbers
  - ✓ Central limit theorem
  - ✓ Markov, Chebyshev, Hoeffding’s inequality
  - ✓ Chernoff’s bounding technique
- **Conditional expectation and martingales**
  - ✓ Reminders on measure theory
  - ✓ Martingales and stopping times
  - ✓ Doob’s maximal inequality
  - ✓ Azuma-Hoeffding’s inequality
    - ✦ application to missing mass estimation: to be continued by A. Garivier

Last week - CS only

- **Deviations for the averages of random variables**
  - ✓ Weak law of large numbers
  - ✓ Central limit theorem
  - ✓ Markov, Chebyshev, Hoeffding's inequality
  - ✓ Chernoff's bounding technique
- **Conditional expectation and martingales**
  - ✓ Reminders on measure theory
  - ✓ Martingales and stopping times
  - ✓ Doob's maximal inequality
  - ✓ Azuma-Hoeffding's inequality
    - ✦ application to missing mass estimation: to be continued by A. Garivier

M2 Maths Avancées: see A. Garivier's course

This week

- **Bounded difference (McDiarmid’s) inequality**
- **The PAC framework for statistical learning**
- **Sub-Gaussianity / sub-exponential variables**

McDiarmid’s inequality

Motivation

- **Concentration of the empirical mean**
  - ✓ **n i.i.d. samples** $X_1, \ldots, X_n$
  - ✓ **empirical mean** $\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i = f(X_1, \ldots, X_n)$
  - ✓ **(under assumptions) concentration around**

$$\mathbb{E}[f(X_1, \ldots, X_n)] = \mathbb{E}[X]$$

Motivation

- **Concentration of the empirical mean**
  - ✓ **n i.i.d. samples** $X_1, \ldots, X_n$
  - ✓ **empirical mean** $\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i = f(X_1, \ldots, X_n)$
  - ✓ **(under assumptions) concentration around**

$$\mathbb{E}[f(X_1, \ldots, X_n)] = \mathbb{E}[X]$$

- **Going further**
  - ✓ What if samples not identically distributed ?
  - ✓ What about other functions of the samples ?

$$f(X_1, \ldots, X_n) := \sup_{h \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} h(X_i)$$

McDiarmid's inequality *aka* bounded difference inequality

- **Theorem (McDiarmid's inequality)**
  - ✓ Consider *independent* random variables $X_1, \ldots, X_n$ and $f : \mathcal{X}^n \to \mathbb{R}$
  - ✓ Assume that $\forall 1 \leq i \leq n, \forall (x_1, \ldots, x_n) \in \mathcal{X}^n$

$$|f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c_i$$

McDiarmid's inequality *aka* bounded difference inequality

- **Theorem (McDiarmid's inequality)**
  - ✓ Consider *independent* random variables $X_1, \ldots, X_n$ and $f : \mathcal{X}^n \to \mathbb{R}$
  - ✓ Assume that $\forall 1 \leq i \leq n, \forall (x_1, \ldots, x_n) \in \mathcal{X}^n$

$$|f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c_i$$

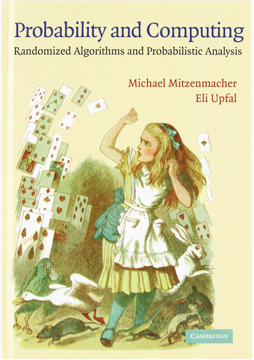
  - ✓ Then, for each t>0

$$\mathbb{P}(f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)] \geq t) \leq e^{-\frac{2t^2}{\sum_{i=1}^n c_i^2}}$$

$$\mathbb{P}(f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)] \leq -t) \leq e^{-\frac{2t^2}{\sum_{i=1}^n c_i^2}}$$

Proof sketch & examples

- **Proof sketch**
  - ✓ build a martingale $Z = f(X)$ $Z_j = \mathbb{E}[Z|X_1, \ldots, X_j]$
  - ✓ use Azuma's inequality *(cf last course by A. Garivier)*
- **Details**
  - ✓ Probability & Computing section 12.5
    - ✦ (the name « McDiarmid » does not appear)
  - ✓ Foundations of Machine Learning, Annex D



- **Home practice: sanity check**
  - ✓ retrieve Hoeffding's inequality using $f(x) = \sum_i x_i$

The PAC learning framework

High dimensional statistical learning

- **Goal**
  - use **training data** to infer parameters θ to achieve a certain **task**
  - **avoid overfitting**: ensure **generalization to unseen data** of similar type
- **Training collection = large point cloud $\mathcal{X}$**
  - signals, images, …
  - feature vectors, labels, …

Digit recognition (MNIST)

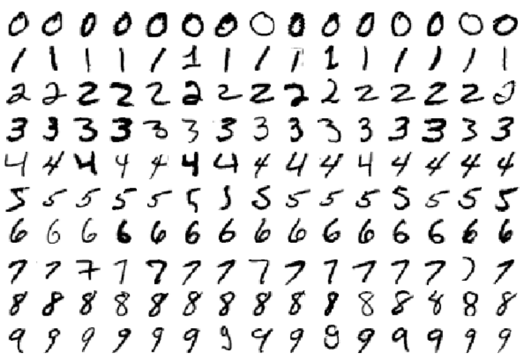
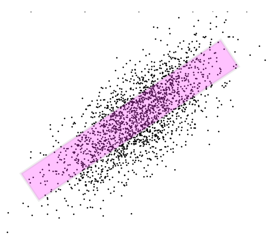


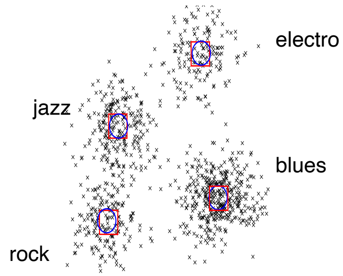
Image classification

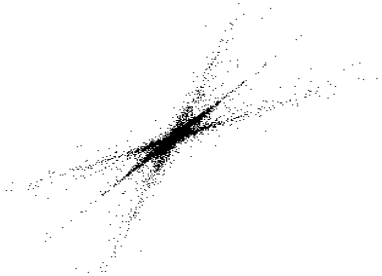
Sound classification

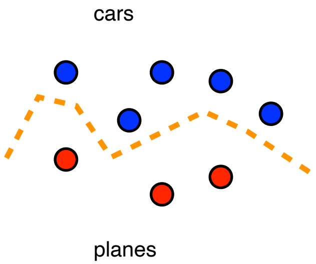
High dimensional statistical learning

- **Goal**
  - use **training data** to infer parameters θ to achieve a certain **task**
  - **avoid overfitting**: ensure **generalization to unseen data** of similar type
- **Training collection = large point cloud** $\mathcal{X}$
  - signals, images, …
  - feature vectors, labels, …
- **Examples of tasks & parameters**









- PCA
  - θ = principal subspace
- Clustering
  - θ = centroids
- Dictionary learning
  - θ = dictionary atoms
- Classification
  - θ = classifier parameters (e.g. support vectors)

Vocabulary - binary classification

- **Training samples & labels** $x_i \in \mathcal{X}$
  $y_i \in \{0,1\},\ 1 \le i \le n$

$$z_i = (x_i, y_i) \in \mathcal{Z} = \mathcal{X} \times \{0,1\}$$

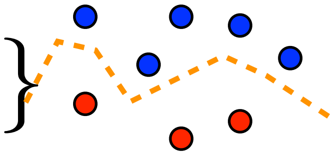
Vocabulary - binary classification

- **Training samples & labels** $x_i \in \mathcal{X}$
  $y_i \in \{0,1\},\ 1 \leq i \leq n$

$$z_i = (x_i, y_i) \in \mathcal{Z} = \mathcal{X} \times \{0,1\}$$

- **Hypothesis class: family of classifiers**

$$\mathcal{H} \subset \{0,1\}^{\mathcal{X}} = \{h : \mathcal{X} \to \{0,1\}\}$$



  ✓ *typically a parametric family* $\mathcal{H} = \{h_\theta : \theta \in \Theta\}$

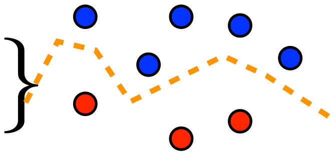
Vocabulary - binary classification

- **Training samples & labels** $x_i \in \mathcal{X}$
  $y_i \in \{0,1\},\ 1 \leq i \leq n$

$$z_i = (x_i, y_i) \in \mathcal{Z} = \mathcal{X} \times \{0,1\}$$

- **Hypothesis class: family of classifiers**

$$\mathcal{H} \subset \{0,1\}^{\mathcal{X}} = \{h : \mathcal{X} \to \{0,1\}\}$$



  - ✓ *typically a parametric family* $\mathcal{H} = \{h_\theta : \theta \in \Theta\}$

- **Loss function**

$$\ell : \mathcal{Z} \times \mathcal{H} \to \mathbb{R}$$

  - ✓ Scalar $\ell(z,h)$ = relevance of hypothesis *h* for sample *z (smaller=better)*

Vocabulary - generic framework

Also with more « abstract »

- sample space (measurable space) $\mathcal{Z}$
- hypothesis class $\mathcal{H}$

- **Training samples & labels** $x_i \in \mathcal{X}$, $y_i \in \{0,1\},\ 1 \leq i \leq n$

$$z_i = (x_i, y_i) \in \mathcal{Z} = \mathcal{X} \times \{0,1\}$$

- **Hypothesis class: family of classifiers**

$$\mathcal{H} \subset \{0,1\}^{\mathcal{X}} = \{h : \mathcal{X} \to \{0,1\}\}$$

  - ✓ *typically a parametric family* $\mathcal{H} = \{h_\theta : \theta \in \Theta\}$

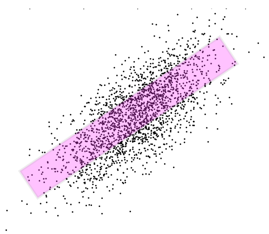
- **Loss function**

$$\ell : \mathcal{Z} \times \mathcal{H} \to \mathbb{R}$$

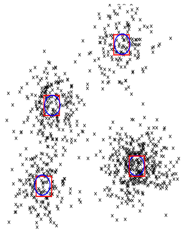
  - ✓ Scalar $\ell(z,h)$ = relevance of hypothesis *h* for sample *z (smaller=better)*

Unsupervised learning examples

- **Principal Component Analysis**



- **K-means clustering**



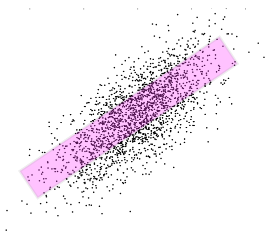
- **Maximum likelihood density fitting**
  parametric density modeling

Exercise: suggest possible
-sample space
-hypothesis class
-loss function ?

Unsupervised learning examples

- **Principal Component Analysis**

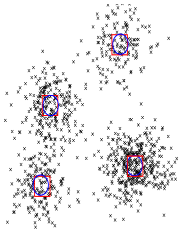
$$z_i = x_i \in \mathbb{R}^d$$



$$\mathcal{H} = \{h \text{ subsp. of } \mathbb{R}^d,\ \mathtt{dim}(h) = k\}$$

$$\ell(z,h) = \mathtt{dist}^2(z,h) = \|z - P_h z\|^2$$

- **K-means clustering**



- **Maximum likelihood density fitting**

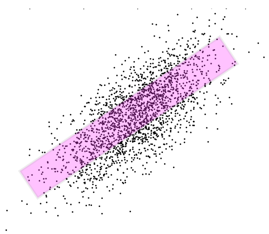
  parametric density modeling

Exercise: suggest possible
-sample space
-hypothesis class
-loss function ?

Unsupervised learning examples

- **Principal Component Analysis**

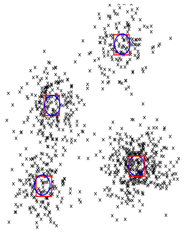
$$z_i = x_i \in \mathbb{R}^d$$



$$\mathcal{H} = \{h \text{ subsp. of } \mathbb{R}^d,\ \mathtt{dim}(h) = k\}$$

$$\ell(z,h) = \mathtt{dist}^2(z,h) = \|z - P_h z\|^2$$

- **K-means clustering**



$$\mathcal{H} = \{h = \{c_1, \ldots, c_k\}, c_j \in \mathbb{R}^d\}$$

$$\ell(z,h) = \mathtt{dist}^2(z,h) = \min_j \|z - c_j\|^2$$

- **Maximum likelihood density fitting**

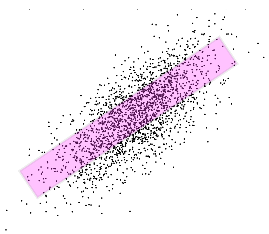
  parametric density modeling

Exercise: suggest possible
-sample space
-hypothesis class
-loss function ?

Unsupervised learning examples

- **Principal Component Analysis**

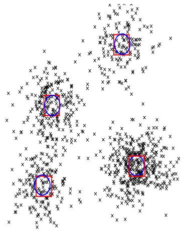
$$z_i = x_i \in \mathbb{R}^d$$



$$\mathcal{H} = \{h \text{ subsp. of } \mathbb{R}^d,\ \mathtt{dim}(h) = k\}$$

$$\ell(z,h) = \mathtt{dist}^2(z,h) = \|z - P_h z\|^2$$

- **K-means clustering**



$$\mathcal{H} = \{h = \{c_1, \ldots, c_k\}, c_j \in \mathbb{R}^d\}$$

$$\ell(z,h) = \mathtt{dist}^2(z,h) = \min_j \|z - c_j\|^2$$

- **Maximum likelihood density fitting**

parametric density modeling

$$\{p_h(x), h \in \mathcal{H}\}$$

$$\ell(z,h) = -\log p_h(z)$$

Exercise: suggest possible
-sample space
-hypothesis class
-loss function ?

Empirical distribution - empirical risk

- **Empirical distribution of the training set**

$$\hat{\mathbb{P}}_n = \frac{1}{n} \sum_i \delta_{z_i}$$

- **Empirical risk**
  - ✓ smaller = better

$$\hat{\mathcal{R}}_n(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(z_i, h)$$

- … only measures relevance of *h* for training samples, **what about generalization to other samples ?**

Notion of generalization - « true » risk

- **Standard model:** training set = *n* **i.i.d.** samples from an ***unknown but fixed*** probability distribution

$$z_i \sim \mathbb{P}_Z$$

Notion of generalization - « true » risk

- **Standard model:** training set = *n* **i.i.d.** samples from an ***unknown but fixed*** probability distribution

$$z_i \sim \mathbb{P}_Z$$

- **True risk** = expectation over « future » samples drawn from the same distribution

$$\mathcal{R}(h) := \mathbb{E}_{Z \sim \mathbb{P}_Z} \ell(Z, h)$$

Notion of generalization - « true » risk

- **Standard model:** training set = *n* **i.i.d.** samples from an ***unknown but fixed*** probability distribution

$$z_i \sim \mathbb{P}_Z$$

- **True risk** = expectation over « future » samples drawn from the same distribution

$$\mathcal{R}(h) := \mathbb{E}_{Z \sim \mathbb{P}_Z} \ell(Z, h)$$

- **Best hypothesis**: one that minimizes the true risk

$$h^\star \in \arg\min_{h \in \mathcal{H}} \mathcal{R}(h)$$

Notion of generalization - « true » risk

- **Standard model:** training set = *n* **i.i.d.** samples from an ***unknown but fixed*** probability distribution

$$z_i \sim \mathbb{P}_Z$$

- **True risk** = expectation over « future » samples drawn from the same distribution

$$\mathcal{R}(h) := \mathbb{E}_{Z \sim \mathbb{P}_Z} \ell(Z, h)$$

- **Best hypothesis**: one that minimizes the true risk

$$h^\star \in \arg\min_{h \in \mathcal{H}} \mathcal{R}(h)$$

unreachable in practice !

Learning algorithms

- **« Learning algorithm »:** $\mathcal{A} : \mathcal{Z}^n \to \mathcal{H}$
  - ✓ input: a training set $S_n = (z_1, \ldots, z_n)$
  - ✓ output: an hypothesis $\hat{h} = \mathcal{A}(S_n)$

Learning algorithms

- **« Learning algorithm »:** $\mathcal{A} : \mathcal{Z}^n \to \mathcal{H}$
  - ✓ input: a training set $S_n = (z_1, \ldots, z_n)$
  - ✓ output: an hypothesis $\hat{h} = \mathcal{A}(S_n)$
  - ✓ **More precisely**
    - ✦ Sequence of algorithms $\mathcal{A}_n : \mathcal{Z}^n \to \mathcal{H}, n \geq 1$
    - ✦ Deterministic or randomized

Learning algorithms

- **« Learning algorithm »:** $\mathcal{A} : \mathcal{Z}^n \to \mathcal{H}$
  - ✓ input: a training set $S_n = (z_1, \ldots, z_n)$
  - ✓ output: an hypothesis $\hat{h} = \mathcal{A}(S_n)$
  - ✓ **More precisely**
    - ✦ Sequence of algorithms $\mathcal{A}_n : \mathcal{Z}^n \to \mathcal{H}, n \geq 1$
    - ✦ Deterministic or randomized
- **Comput. tractability ? Statistical guarantees?**

Examples ?

Learning principle *vs* learning algorithm

- **Empirical risk minimization (ERM)**

$$\hat{h}_n = \mathcal{A}(S_n) := \arg\min_{h \in \mathcal{H}} \hat{\mathcal{R}}_n(h)$$

$$= \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(z_i, h)$$

✓ is the minimum achieved ?
✓ can it be computed in polynomial time ?

Learning principle *vs* learning algorithm

- **Empirical risk minimization (ERM)**

$$\hat{h}_n = \mathcal{A}(S_n) := \arg\min_{h\in\mathcal{H}} \hat{\mathcal{R}}_n(h)$$
$$= \arg\min_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \ell(z_i, h)$$

  - ✓ is the minimum achieved ?
  - ✓ can it be computed in polynomial time ?
- … rather a learning ***principle*** than a learning ***algorithm*** here

Statistical guarantees: objectives

- **Goal**: control the risk $\mathcal{R}(\hat{h}_n)$
  - ✓ *with hypothesis defined by a learning algorithm (or principle)*

Statistical guarantees: objectives

- **Goal**: control the risk $\mathcal{R}(\hat{h}_n)$
  - ✓ *with hypothesis defined by a learning algorithm (or principle)*
- **Baseline**: best possible risk $\mathcal{R}^\star := \inf_{h \in \mathcal{H}} \mathcal{R}(h)$
  - ✓ notion of excess risk

$$\Delta\mathcal{R}(h) = \mathcal{R}(h) - \mathcal{R}^\star$$

Statistical guarantees: objectives

- **Goal**: control the risk $\mathcal{R}(\hat{h}_n)$
  - ✓ *with hypothesis defined by a learning algorithm (or principle)*
- **Baseline**: best possible risk $\mathcal{R}^\star := \inf_{h \in \mathcal{H}} \mathcal{R}(h)$
  - ✓ notion of excess risk

$$\Delta\mathcal{R}(h) = \mathcal{R}(h) - \mathcal{R}^\star$$

- Can we ensure to ***approximate*** the true best hypothesis up to some accuracy ?

$$\Delta\mathcal{R}(\hat{h}_n) \leq \epsilon$$

Statistical guarantees: objectives

statistical model: ***random*** training set $S_n = (Z_1, \ldots, Z_n)$

- **Goal**: control the risk $\mathcal{R}(\hat{h}_n)$
  - ✓ *with hypothesis defined by a learning algorithm or principle*
- **Baseline**: best possible risk $\mathcal{R}^\star := \inf_{h \in \mathcal{H}} \mathcal{R}(h)$
  - ✓ notion of excess risk

$$\Delta\mathcal{R}(h) = \mathcal{R}(h) - \mathcal{R}^\star$$

- Can we ensure to ***approximate*** the true best hypothesis up to some accuracy ?

$$\Delta\mathcal{R}(\hat{h}_n) \leq \epsilon$$

Statistical guarantees: objectives

statistical model: ***random*** training set $S_n = (Z_1, \ldots, Z_n)$

- **Goal**: control the risk $\mathcal{R}(\hat{h}_n)$ $\hat{h}_n = \mathcal{A}(S_n)$
  - ✓ *with hypothesis defined by a learning algorithm or principle*
- **Baseline**: best possible risk $\mathcal{R}^\star := \inf_{h \in \mathcal{H}} \mathcal{R}(h)$
  - ✓ notion of excess risk

$$\Delta\mathcal{R}(h) = \mathcal{R}(h) - \mathcal{R}^\star$$

- Can we ensure to ***approximate*** the true best hypothesis up to some accuracy ?

$$\Delta\mathcal{R}(\hat{h}_n) \leq \epsilon$$

Statistical guarantees: objectives

statistical model: ***random*** training set $S_n = (Z_1, \ldots, Z_n)$

- **Goal**: control the risk $\mathcal{R}(\hat{h}_n)$ $\hat{h}_n = \mathcal{A}(S_n)$
  - ✓ *with hypothesis defined by a learning algorithm or principle*
- **Baseline**: best possible risk $\mathcal{R}^\star := \inf_{h \in \mathcal{H}} \mathcal{R}(h)$
  - ✓ notion of excess risk

$$\Delta\mathcal{R}(h) = \mathcal{R}(h) - \mathcal{R}^\star$$

- Can we ensure to ***approximate*** the true best hypothesis up to some accuracy with high probability ?

$$P(\Delta\mathcal{R}(\hat{h}_n) \leq \epsilon) \geq 1 - \delta$$

Probably Approximately Correct guarantees

- **PAC bounds:** in probability or in expectation

$$P(\Delta\mathcal{R}(\hat{h}_n) \leq \epsilon) \geq 1 - \delta \qquad \mathbb{E}[\Delta\mathcal{R}(\hat{h}_n)] \leq \epsilon$$

✓ given a task (=loss+hypothesis class), bounds depend on
  - ✦ algorithm/principle
  - ✦ ***and data distribution***

Probably Approximately Correct guarantees

- **PAC bounds:** in probability or in expectation

$$P(\Delta\mathcal{R}(\hat{h}_n) \leq \epsilon) \geq 1 - \delta \qquad\qquad \mathbb{E}[\Delta\mathcal{R}(\hat{h}_n)] \leq \epsilon$$

  - ✓ given a task (=loss+hypothesis class), bounds depend on
    - ✦ algorithm/principle
    - ✦ ***and data distribution***

- ***Agnostic* PAC bounds:** when no assumption needed on data distribution

Probably Approximately Correct guarantees

- **PAC bounds:** in probability or in expectation

$$P(\Delta\mathcal{R}(\hat{h}_n) \leq \epsilon) \geq 1 - \delta \qquad\qquad \mathbb{E}[\Delta\mathcal{R}(\hat{h}_n)] \leq \epsilon$$

  - ✓ given a task (=loss+hypothesis class), bounds depend on
    - ✦ algorithm/principle
    - ✦ ***and data distribution***

- ***Agnostic* PAC bounds:** when no assumption needed on data distribution

- Notion of sample complexity (sharp or not) $\quad n(\epsilon, \delta)$

Agnostic PAC bounds for empirical risk minimization

Case study / exercice

- « Application » scenario
  - ✓ several vendors provide a spam detection tool
  - ✓ training set: mails correctly labeled as spam / non-spam
  - ✓ approach: select the tool with the least error
  - ✓ goal: predict how accurate it will be
- Exercice
  - ✓ formalize the problem
  - ✓ propose PAC bounds

Reminders and hints

- **Empirical risk minimization**

$$\hat{\mathcal{R}}_n(h) := \frac{1}{n}\sum_{i=1}^{n} \ell(z_i, h).$$

$$\hat{h}_n = \arg\min_{h\in\mathcal{H}} \hat{\mathcal{R}}_n(h)$$

- **Use Hoeffding's inequality and the union bound**