Statistical Learning and Empirical Risk Minimization

TSE Team

March 10, 2023

Objectives

- Reminders on supervised learning
- Generative approaches
- Discriminative approaches - Empirical Risk Minimization

$\mapsto$ Definition, examples and trade-off (over-fitting)

$\mapsto$ Estimation error upper bounds (concentration inequalities)

Motivation: Statistical Learning Basics !

Supervised learning framework (reminders)

Phenomenon $(\mathbf{X}, Y) \sim \mathbb{P}$.
Task : Predict Y with $\mathbf{X}$.

- Decision rule (classifier, predictor) $g : \mathsf{X} \to \mathsf{Y}$ measurable
- Loss: $\ell(Y, g(\mathbf{X}))$
- Risk: $R(g) = \mathbb{E}[\ell(Y, g(\mathbf{X}))]$

Supervised learning framework (reminders)

Phenomenon $(\mathbf{X}, Y) \sim \mathbb{P}$.
Task : Predict Y with $\mathbf{X}$.

- Decision rule (classifier, predictor) $g : \mathsf{X} \to \mathsf{Y}$ measurable
- Loss: $\ell(Y, g(\mathbf{X}))$
- Risk: $R(g) = \mathbb{E}[\ell(Y, g(\mathbf{X}))]$

Bayesian classifier: $R(g^{Bayes}) \leq R(g), \forall g$

$\mapsto$ Previous slides: computation of the Bayesian classifier:
regression, quadratic loss ,
classification, 0-1 loss

Supervised learning framework (reminders)

Phenomenon $(\mathbf{X}, Y) \sim \mathbb{P}$.
Task : Predict Y with $\mathbf{X}$.

- Decision rule (classifier, predictor) $g : \mathsf{X} \to \mathsf{Y}$ measurable
- Loss: $\ell(Y, g(\mathbf{X}))$
- Risk: $R(g) = \mathbb{E}[\ell(Y, g(\mathbf{X}))]$

Bayesian classifier: $R(g^{Bayes}) \leq R(g), \forall g$

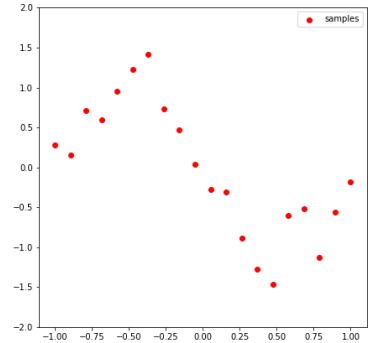
$\mapsto$ Previous slides: computation of the Bayesian classifier: regression, quadratic loss ,
classification, 0-1 loss

Problem: g^{Bayes} depends of $\mathbb{P}$, unknown

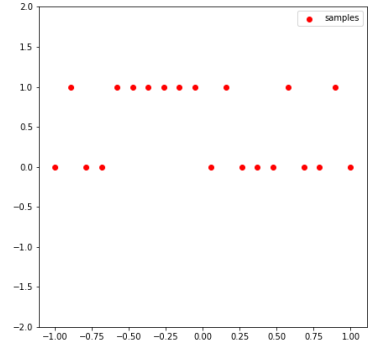
- We observe a Learning Set :
  $D_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ (i.i.d. $\sim \mathbb{P}$)
  $\mapsto$ We build a classifier $\hat{g}_n$.

Goal: Upper bound the excess risk $\mathcal{E}(\hat{g}_n) = R(\hat{g}_n) - R(g^{Bayes})$.

Regression



Classification



Generative approach (Reminder)

Approach

1. We **model** the joint law $(\mathbf{X}, Y)$ with a (parametric) model: $\{\mathbb{P}_\theta, \theta \in \Theta\}$
2. We compute the Bayes' classifier g^*_θ for any value of θ and $\mathbb{P}_\theta$.
3. We estimate $\theta^\star$ with $\hat{\theta}_n$.
4. We define $\hat{g}_n = g^*_{\hat{\theta}_n}$ (plug-in).

Result (Tutorials) Under $\mathbb{P}_{n\theta}$, if $\hat{\theta}_n$ is consistent, under nice continuity properties, the excess risk goes to 0:

$$\mathbb{E}^\theta_n \left[R(g^*_{\hat{\theta}_n}) - R(g^{Bayes}) \right] \longrightarrow 0$$

$\mapsto$ Statistical Approach: we use a statistical model and we obtain some results for this model.

Methods: LDA, QDA, Logistic classification, Gaussian regression, K nearest neighbour...

Generative approach - classification examples

LDA/QDA

Model: $\mathbf{X}|(Y = i) \sim \mathcal{N}(\mu_i, \Sigma_i)$, $Y \sim \mathcal{B}(p)$

Parameters: $(p, (\mu_i), (\Sigma_i))$

Decision: linear or quadratic

Typical estimator: MLE

Logistic Regression

Model : $Y|\mathbf{X} \sim \mathcal{B}(\sigma(\beta^T \mathbf{X}))$

Parameter: β

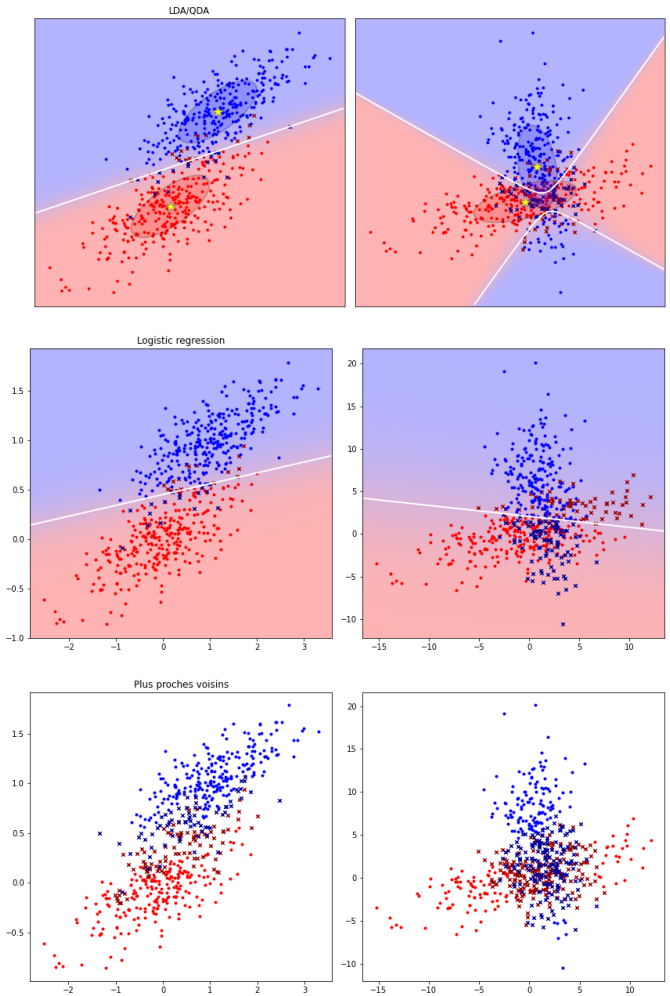
Decision : Linear

Typical estimator: MLE (conditionnally to the $\mathbf{X}_i$)

Naive Bayes / KNN

Models: non-parametric

Decision: non linear



Generative approach - classification examples

LDA/QDA

Model: $\mathbf{X}|(Y = i) \sim \mathcal{N}(\mu_i, \Sigma_i)$, $Y \sim \mathcal{B}(p)$

Parameters: $(p, (\mu_i), (\Sigma_i))$

Decision: linear or quadratic

Typical estimator: MLE

Logistic Regression

Model : $Y|\mathbf{X} \sim \mathcal{B}(\sigma(\beta^T \mathbf{X}))$

Parameter: β

Decision : Linear

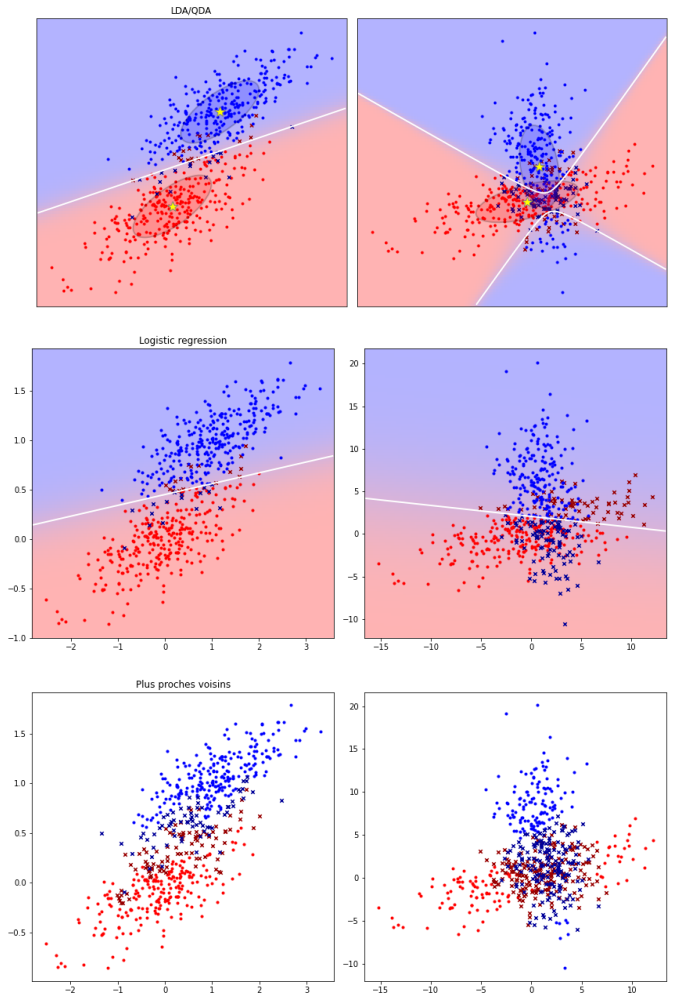
Typical estimator: MLE (conditionnally to the $\mathbf{X}_i$)

Naive Bayes / KNN

Models: non-parametric

Decision: non linear

Problem: Model choice!



Discriminative approach, Empirical Risk Minimization

Idea: Choose a good predictor for the learning set $\{(\mathbf{X}_i, Y_i)_{1 \leq i \leq n}\}$.

- We can compute the empirical risk $\widehat{\mathrm{R}}_n(g)$ ...

$$\widehat{\mathrm{R}}_n(g) = \frac{1}{n} \sum_{i=1}^{n} \ell(g(\mathbf{X}_i), Y_i),$$

Discriminative approach, Empirical Risk Minimization

Idea: Choose a good predictor for the learning set $\{(\mathbf{X}_i, Y_i)_{1 \leq i \leq n}\}$.

- We can compute the empirical risk $\widehat{\mathrm{R}}_n(g)$ ... and minimize it !

$$\widehat{\mathrm{R}}_n(g) = \frac{1}{n} \sum_{i=1}^{n} \ell(g(\mathbf{X}_i), Y_i), \qquad \hat{g}^*_{n,\mathcal{C}} = \arg\min_{g \in \mathcal{C}} \widehat{\mathrm{R}}_n(g)$$

Discriminative approach, Empirical Risk Minimization

Idea: Choose a good predictor for the learning set $\{(\mathbf{X}_i, Y_i)_{1 \leq i \leq n}\}$.

- We can compute the empirical risk $\widehat{\mathrm{R}}_n(g)$ ... and minimize it !

$$\widehat{\mathrm{R}}_n(g) = \frac{1}{n} \sum_{i=1}^{n} \ell(g(\mathbf{X}_i), Y_i), \qquad \hat{g}^*_{n,\mathcal{C}} = \arg\min_{g \in \mathcal{C}} \widehat{\mathrm{R}}_n(g)$$

- Minimize $\widehat{\mathrm{R}}_n$ (ERM): $\hat{g}^*_{n,\mathcal{C}}$

Discriminative approach, Empirical Risk Minimization

Idea: Choose a good predictor for the learning set $\{(\mathbf{X}_i, Y_i)_{1 \leq i \leq n}\}$.

- We can compute the empirical risk $\widehat{\mathrm{R}}_n(g)$ ... and minimize it !

$$\widehat{\mathrm{R}}_n(g) = \frac{1}{n} \sum_{i=1}^{n} \ell(g(\mathbf{X}_i), Y_i), \qquad \hat{g}^*_{n,\mathcal{C}} = \arg\min_{g \in \mathcal{C}} \widehat{\mathrm{R}}_n(g)$$

- Minimize $\widehat{\mathrm{R}}_n$ (ERM): $\hat{g}^*_{n,\mathcal{C}}$

Hope: For any g, since $R(g) = \mathbb{E}[\ell(\mathbf{X}, Y)]$,

- $\widehat{\mathrm{R}}_n(g)$ is an unbiased estimator of $R(g)$
- The LLN shows that $\widehat{\mathrm{R}}_n(g) \overset{\mathbb{P}-\text{prob}}{\longrightarrow} R(g)$

Discriminative approach, Empirical Risk Minimization

Idea: Choose a good predictor for the learning set $\{(\mathbf{X}_i, Y_i)_{1 \leq i \leq n}\}$.

- We can compute the empirical risk $\widehat{\mathrm{R}}_n(g)$ ... and minimize it !

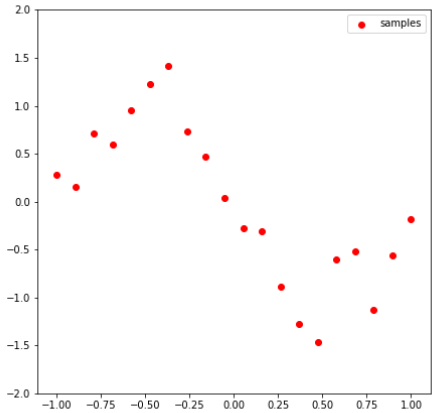
$$\widehat{\mathrm{R}}_n(g) = \frac{1}{n} \sum_{i=1}^{n} \ell(g(\mathbf{X}_i), Y_i), \qquad \hat{g}^*_{n,\mathcal{C}} = \arg\min_{g \in \mathcal{C}} \widehat{\mathrm{R}}_n(g)$$

- Minimize $\widehat{\mathrm{R}}_n$ (ERM): $\hat{g}^*_{n,\mathcal{C}}$

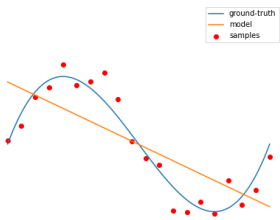
Choice of $\mathcal{C}$?

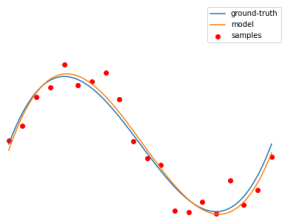
- We must restrict the class of functions. Indeed, if $\mathcal{C} = \mathbb{R}^X$, since $\ell(y, y) = 0$, it is enough to choose $\hat{g}(X_i) = Y_i$ for any $i \in \{1, \ldots, n\}$ and $\hat{g}(x) = 0$ if $x \notin \{x_1, \ldots, x_n\}$.
- Problem: **Generalization** this classifier does not generalize anything.

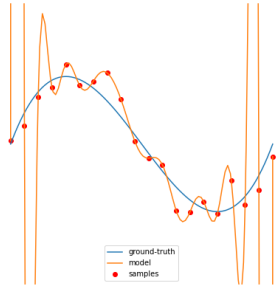
$$\widehat{\mathrm{R}}_n(\hat{g}^*_{n,\mathcal{C}}) \ll R(\hat{g}^*_{n,\mathcal{C}})$$

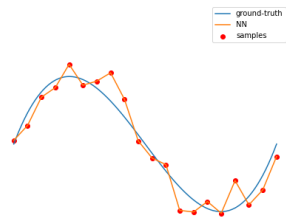


Choice of $\mathcal{C}$ - Examples in regression, quadratic loss.







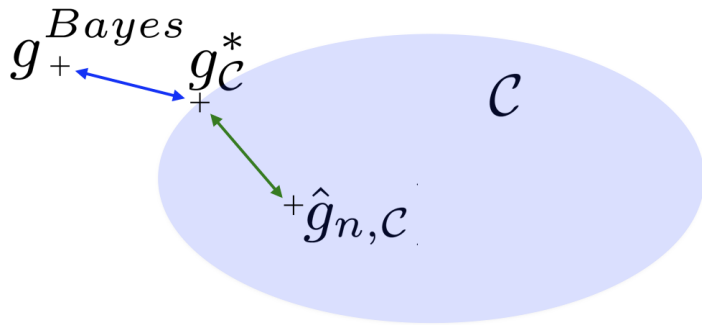


$\mathcal{C} =$	Affine functions	$\mathbb{R}_3[X]$	$\mathbb{R}_{20}[X]$	Dense NN
d				
$\widehat{\mathrm{R}}_n(\hat{g}^*_{n,\mathcal{C}})$				
$R(\hat{g}^*_{n,\mathcal{C}})$				
Problem				

Demo Colab

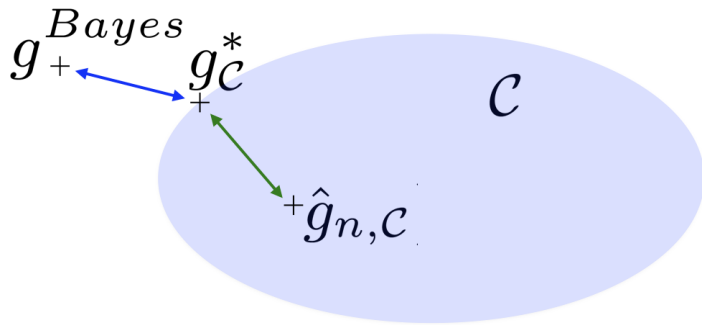
Decomposition of the excess risk for a given $\mathcal{C}$

- Let $g^*_{\mathcal{C}}$
  s.t. $R(g^*_{\mathcal{C}}) = \inf_{g \in \mathcal{C}} R(g)$.
- Let $\hat{g}_{n,\mathcal{C}}$ an estimator in $\mathcal{C}$.



Decomposition of the excess risk for a given $\mathcal{C}$

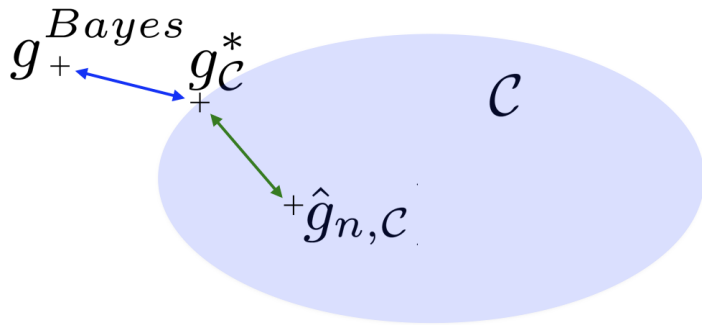
- Let $g^*_{\mathcal{C}}$
  s.t. $R(g^*_{\mathcal{C}}) = \inf_{g \in \mathcal{C}} R(g)$.
- Let $\hat{g}_{n,\mathcal{C}}$ an estimator in $\mathcal{C}$.



$$\mathcal{E}(\hat{g}_{n,\mathcal{C}}) = R(\hat{g}_{n,\mathcal{C}}) - R(g^{Bayes}) = \underbrace{R(\hat{g}_{n,\mathcal{C}}) - R(g^*_{\mathcal{C}})}_{\text{Estimation Error}} + \underbrace{R(g^*_{\mathcal{C}}) - R(g^{Bayes})}_{\text{Approximation Error}}$$

Decomposition of the excess risk for a given $\mathcal{C}$

- Let $g^*_{\mathcal{C}}$
  s.t. $R(g^*_{\mathcal{C}}) = \inf_{g \in \mathcal{C}} R(g)$.
- Let $\hat{g}_{n,\mathcal{C}}$ an estimator in $\mathcal{C}$.



$$\mathcal{E}(\hat{g}_{n,\mathcal{C}}) = R(\hat{g}_{n,\mathcal{C}}) - R(g^{Bayes}) = \underbrace{R(\hat{g}_{n,\mathcal{C}}) - R(g^*_{\mathcal{C}})}_{\text{Estimation Error}} + \underbrace{R(g^*_{\mathcal{C}}) - R(g^{Bayes})}_{\text{Approximation Error}}$$

	Approx. Err.	Estim. Err.
$\mathcal{C}$ increases		
n increases		

- The estimation error translates our partial knowledge of the distribution $(\mathbf{X}, Y)$.
- The approximation error depends on the choice of $\mathcal{C}$.

Summary 1: Minimization of the Empirical Risk Minimization

Discriminative approach. We minimize the ERM on $\mathcal{C}$.

$$\hat{g}^*_{n,\mathcal{C}} = \arg\min_{g \in \mathcal{C}} \left\{ \widehat{\mathrm{R}}_n(g) = n^{-1} \sum_{i=1}^{n} \ell(g(\mathbf{X}_i), Y_i) \right\}$$

3 inter-connected questions:

1. How should we choose $\mathcal{C}$?
   - Examples in régression
   - Decomposition in approximation error/ estimation
   - Overfitting
   - In pratice, cross validation

Summary 1: Minimization of the Empirical Risk Minimization

Discriminative approach. We minimize the ERM on $\mathcal{C}$.

$$\hat{g}^*_{n,\mathcal{C}} = \arg\min_{g \in \mathcal{C}} \left\{ \widehat{\mathrm{R}}_n(g) = n^{-1} \sum_{i=1}^{n} \ell(g(\mathbf{X}_i), Y_i) \right\}$$

3 inter-connected questions:

1. How should we choose $\mathcal{C}$?
   - Examples in régression
   - Decomposition in approximation error/ estimation
   - Overfitting
   - In pratice, cross validation
2. Upper bound of the estimation error $R(\hat{g}^*_{n,\mathcal{C}}) - R(g^*_{\mathcal{C}})$?
3. How to compute $\hat{g}^*_{n,\mathcal{C}}$? Accuracy?

Estimation error of ERM over $\mathcal{C}$

Goal : upper bound $\underbrace{R(\hat{g}^*_{n,\mathcal{C}}) - R(g^*_{\mathcal{C}})}_{\text{Estimation Error}}$, with

$$\begin{cases} \hat{g}^*_{n,\mathcal{C}} = \arg\min_{g \in \mathcal{C}} \widehat{\mathrm{R}}_n(g) \\ g^*_{\mathcal{C}} = \arg\min_{g \in \mathcal{C}} R(g) \end{cases}$$

Remarks: - we consider *a* minimiser - solely the value of the min is important: we compute $\inf_{g \in \mathcal{C}} R(g)$ instead of $R(g^*_{\mathcal{C}})$.

Estimation error of ERM over $\mathcal{C}$

Goal : upper bound $\underbrace{R(\hat{g}^*_{n,\mathcal{C}}) - R(g^*_{\mathcal{C}})}_{\text{Estimation Error}}$, with

$$\begin{cases} \hat{g}^*_{n,\mathcal{C}} = \arg\min_{g\in\mathcal{C}} \widehat{\mathrm{R}}_n(g) \\ g^*_{\mathcal{C}} = \arg\min_{g\in\mathcal{C}} R(g) \end{cases}$$

Remarks: - we consider *a* minimiser - solely the value of the min is important: we compute $\inf_{g\in\mathcal{C}} R(g)$ instead of $R(g^*_{\mathcal{C}})$.

Theorem

We have:

$$R(\hat{g}^*_{n,\mathcal{C}}) - \inf_{g\in\mathcal{C}} R(g) \leq 2 \sup_{g\in\mathcal{C}} |\widehat{\mathrm{R}}_n(g) - R(g)| \text{ , a.s.}$$

$$\mathbb{E}\left[R(\hat{g}^*_{n,\mathcal{C}}) - \inf_{g\in\mathcal{C}} R(g)\right] \leq \mathbb{E}\left[\sup_{g\in\mathcal{C}} |\widehat{\mathrm{R}}_n(g) - R(g)|\right] .$$

The excess risk of ERM is upper bounded by the deviations between the risk and the ERM over the class.

Proof (completed)

Observations:

1. $\widehat{\mathrm{R}}_n(\hat{g}^*_{n,\mathcal{C}}) \leq \widehat{\mathrm{R}}_n(g)$ for any $g \in \mathcal{C}$.
2. $|\widehat{\mathrm{R}}_n(\hat{g}^*_{n,\mathcal{C}}) - R(\hat{g}^*_{n,\mathcal{C}})| \leq \sup_{g\in\mathcal{C}} |\widehat{\mathrm{R}}_n(g) - R(g)|$.

- We use $\widehat{\mathrm{R}}_n(\hat{g}^*_{n,\mathcal{C}}) \leq \widehat{\mathrm{R}}_n(g^*_{\mathcal{C}})$

$$\begin{aligned} R(\hat{g}^*_{n,\mathcal{C}}) - \inf_{g\in\mathcal{C}} R(g) = & \; R(\hat{g}^*_{n,\mathcal{C}}) - \widehat{\mathrm{R}}_n(\hat{g}^*_{n,\mathcal{C}}) + \widehat{\mathrm{R}}_n(\hat{g}^*_{n,\mathcal{C}}) - \inf_{g\in\mathcal{C}} R(g) \\ & \leq R(\hat{g}^*_{n,\mathcal{C}}) - \widehat{\mathrm{R}}_n(\hat{g}^*_{n,\mathcal{C}}) + \widehat{\mathrm{R}}_n(g^*_{\mathcal{C}}) - R(g^*_{\mathcal{C}}) \; . \end{aligned}$$

- We deduce that:

$$R(\hat{g}^*_{n,\mathcal{C}}) - \inf_{g\in\mathcal{C}} R(g) \leq \; 2 \sup_{g\in\mathcal{C}} |\widehat{\mathrm{R}}_n(g) - R(g)| \; .$$

$$\mathbb{E}\left[R(\hat{g}^*_{n,\mathcal{C}}) - \inf_{g\in\mathcal{C}} R(g)\right] \leq \; \mathbb{E}\left[\sup_{g\in\mathcal{C}} |\widehat{\mathrm{R}}_n(g) - R(g)|\right] \; . \qquad \square$$

Upper bound of the estimation error on $\mathcal{C}$

Theorem (Reminder)

We have:

$$R(\hat{g}^*_{n,\mathcal{C}}) - \inf_{g\in\mathcal{C}} R(g) \leq 2 \sup_{g\in\mathcal{C}} |\widehat{\mathrm{R}}_n(g) - R(g)| \text{ , a.s.}$$

$$\mathbb{E}\left[R(\hat{g}^*_{n,\mathcal{C}}) - \inf_{g\in\mathcal{C}} R(g)\right] \leq \mathbb{E}\left[\sup_{g\in\mathcal{C}} |\widehat{\mathrm{R}}_n(g) - R(g)|\right] .$$

The excess risk of the ERM is upper bounded by the deviations between the risk and the ERM over the class.

Type of upper bound? $R(\hat{g}^*_{n,\mathcal{C}})$ is a random variable !

1. Expectation upper bound?
2. With overwhelming overwhelming probability, :
   $\mathbb{P}(R(\hat{g}^*_{n,\mathcal{C}}) - R(g^*_{\mathcal{C}}) \leq \Delta_{n,\delta}(\mathcal{C})) \geq 1 - \delta$
   where $\Delta_{n,\delta}(\mathcal{C})$ depends on the n ($\#\mathcal{D}_n$), δ (the confidence level).

$\mapsto$ we need to derive some upper bounds on the uniform deviations of the random variables relatively to their means.

Uniform deviation upper bounds, 0-1 loss

We only deal with the 0-1 loss: for any $g \in \mathcal{C}$, $n\,\widehat{\mathrm{R}}_n(g) = \sum_{k=1}^n \mathbb{1}_{\{g(X_k) \neq Y_k\}}$ is distributed according to a Binomial distribution of parameters $Bin(n, R(g))$.

We will upper bound $\sup_{g\in\mathcal{C}} \left| n^{-1} \sum_{i=1}^n \{\mathbb{1}_{\{g(\mathbf{X}_i)\neq Y_i\}} - R(g)\} \right|$.

1. Over $\mathcal{C}$ a finite class
2. With overwhelming probability and in expectation.

Approach:

1. For any fixed g, $\mathbb{P}(|\,\widehat{\mathrm{R}}_n(g) - R(g)| > ...)$ is small
   $\mapsto$ Concentration inequality.
2. Union bound:
   $\mathbb{P}(\sup_{g\in\mathcal{C}} |\,\widehat{\mathrm{R}}_n(g) - R(g)| > ...) \leq \sum_{g\in\mathcal{C}} \mathbb{P}(|\,\widehat{\mathrm{R}}_n(g) - R(g)| > ...)$

Concentration inequality: from Markov to Hoeffding

- If $Z \geq 0$, then $\mathbb{P}(Z > t) \leq \frac{\mathbb{E}[Z]}{t}$. which implies the Chebyshev inequality: if Z has a bounded variance: $\mathbb{P}(|Z - \mathbb{E}[Z]| \geq \varepsilon) \leq \frac{\mathrm{Var}(Z)}{\varepsilon^2}$
- If $X_1, \ldots, X_n$ i.i.d. and $Z = \frac{1}{n}\sum_{i=1}^n X_i$, we have $\mathrm{Var}(Z) = \frac{\mathrm{Var}(X_1)}{n}$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}[X_1]\right| > \varepsilon\right) \leq \frac{\mathrm{Var}(X_1)}{n\varepsilon^2}$$

Theorem (Hoeffding inequality)

Let $X_1, \ldots, X_n$ n independent random variables in $[0,1]$. For any $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \{X_i - \mathbb{E}[X_i]\}\right| \geq \varepsilon\right) \leq 2\mathrm{e}^{-2n\varepsilon^2} \ .$$

- No i.i.d., just independence!
- Extension to $X_i \in [a_i, b_i]$, upper bound $2\mathrm{e}^{-2n^2\varepsilon^2 / \sum_{i=1}^n (b_i - a_i)^2}$

Upper bound on the risk

$$\mathbb{P}\left(\sup_{g\in\mathcal{C}} |\,\widehat{\mathrm{R}}_n(g) - R(g)| > \varepsilon\right)$$

Upper bound on the risk (completed)

Theorem

Assume that $|\mathcal{C}| < \infty$. *Then, for any* $n \in \mathbb{N}$ *and* $\varepsilon > 0$,

$$\mathbb{P}\left(\sup_{g \in \mathcal{C}} \widehat{\mathrm{R}}_n(g) - R(g) > \varepsilon \right) \leq |\mathcal{C}| \, \mathrm{e}^{-2n\varepsilon^2} \ ,$$

$$\mathbb{P}\left(\sup_{g \in \mathcal{C}} |\, \widehat{\mathrm{R}}_n(g) - R(g)| > \varepsilon \right) \leq 2\,|\mathcal{C}| \, \mathrm{e}^{-2n\varepsilon^2} \ .$$

Proof.

- Denote $\mathcal{C} = \{g_i\}_{i=1}^{|\mathcal{C}|}$. The union bound shows that:

$$\mathbb{P}\left(\sup_{g \in \mathcal{C}} |\, \widehat{\mathrm{R}}_n(g) - R(g)| > \varepsilon \right) \leq \sum_{i=1}^{|\mathcal{C}|} \mathbb{P}(|\, \widehat{\mathrm{R}}_n(g_i) - R(g_i)| > \varepsilon) \ .$$

- The union bounds come from the Hoeffding inequality using that $\left\{ \mathbb{1}_{\{g(X_i) \neq Y_i\}} \right\}_{i=1}^n$ are distributed according to a Bernoulli distribution.

Conclusion: upper bound in probability

Estimation error upper bound, finite class $\mathcal{C}$, 0-1 loss.

1. We have:

$$R(\hat{g}^*_{n,\mathcal{C}}) - R(g^*_{\mathcal{C}}) \leq 2 \sup_{g \in \mathcal{C}} |\widehat{\mathrm{R}}_n(g) - R(g)|$$

2. Furthermore, with probability $1 - \delta$,

$$\sup_{g \in \mathcal{C}} |\widehat{\mathrm{R}}_n(g) - R(g)| \leq \sqrt{\frac{\log(2|\mathcal{C}|/\delta)}{2n}}$$

3. As a consequence, with a probability larger than $1 - \delta$

$$R(\hat{g}^*_{n,\mathcal{C}}) - R(g^*_{\mathcal{C}}) \leq 2\sqrt{\frac{\log(2|\mathcal{C}|/\delta)}{2n}}$$

Remarks:

1. depends on $|\mathcal{C}|$
2. depends on n

Conclusion: upper bound in expectation

We have $\mathbb{E}\left[R(\hat{g}^*_{n,\mathcal{C}}) - R(g^*_{\mathcal{C}})\right] \leq \mathbb{E}\left[\max_{g\in\mathcal{C}} |\widehat{\mathrm{R}}_n(g) - R(g)|\right]$.

Theorem (Pisier Inequality)

Let $Z_1, \ldots, Z_m$ m sub-gaussian r.v., i.e. s.t., $\forall \lambda \in \mathbb{R}$

$$\mathbb{E}[\mathrm{e}^{\lambda Z_i}] \leq \mathrm{e}^{\lambda^2\sigma^2/2}$$

Then,

$$\mathbb{E}\left[\max_{1\leq i\leq m} |Z_i|\right] \leq \sigma\sqrt{2\log(2m)} \, .$$

Denote $Z_j = \widehat{\mathrm{R}}_n(g_j) - R(g_j)$, $j \in \{1, \ldots, |\mathcal{C}|\}$. For any $\lambda \geq 0$,

$$\begin{aligned}\mathbb{E}[\mathrm{e}^{\lambda Z_j}] &= \mathbb{E}\left[\mathrm{e}^{(\lambda/n)\sum_{i=1}^n \{\mathbb{1}_{\{g_j(X_i)\neq Y_i\}} - R(g_j)\}}\right] \\ &= \left(\mathbb{E}\left[\mathrm{e}^{(\lambda/n)\{\mathbb{1}_{\{g_j(X_1)\neq Y_1\}} - R(g_j)\}}\right]\right)^n \leq \mathrm{e}^{\lambda^2/8n}\end{aligned}$$

using the Hoeffding Lemma, $\mathbb{E}\left[\mathrm{e}^{(\lambda/n)\{\mathbb{1}_{\{g_j(X_1)\neq Y_1\}} - R(g_j)\}}\right] \leq \mathrm{e}^{\lambda^2/(8n^2)}$

Bound for the integrated risk

Theorem

If $\hat{g}^*_{n,\mathcal{C}}$ minimizes the ERM over $\mathcal{C}$, then:

$$\mathbb{E}[R(\hat{g}^*_{n,\mathcal{C}}) - R(g^*_{\mathcal{C}})] \leq \sqrt{\frac{\log(2\,|\mathcal{C}|)}{2n}}$$