General oracle inequalities for ERM, regularized ERM and penalized ERM with applications to High-Dimensional data analysis

Guillaume Lecué

CNRS, Laboratoire d'analyse mathématiques appliquées, Université Paris-Est Marne-la-vallée
Joint works with **Stéphane Gaïffas** and **Shahar Mendelson**.

Princeton University - ORFE. Friday May 25, 2012

A quick example : an oracle inequality for the "squared LASSO"

Data : $(X_i, Y_i)_{i=1}^n$ i.i.d. $\sim (X, Y) \in \mathbb{R}^d \times \mathbb{R}$.

A quick example : an oracle inequality for the "squared LASSO"

Data : $(X_i, Y_i)_{i=1}^n$ i.i.d. $\sim (X, Y) \in \mathbb{R}^d \times \mathbb{R}$.
Assumption : Y and $\|X\|_{\ell_\infty^d}$ are subgaussian.

A quick example : an oracle inequality for the "squared LASSO"

Data : $(X_i, Y_i)_{i=1}^n$ i.i.d. $\sim (X, Y) \in \mathbb{R}^d \times \mathbb{R}$.
Assumption : Y and $\|X\|_{\ell_\infty^d}$ are subgaussian.
The regularized empirical risk minimization (ERM) estimator

$$\hat{\beta}_n \in \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \Big(\frac{1}{n} \sum_{i=1}^n (Y_i - \langle X_i, \beta \rangle)^2 + \lambda \frac{\|\beta\|_{\ell_1}^2}{n\epsilon^2} \Big)$$

where $\lambda = \lambda(n, d) = \mathit{polylog}(n, d)$ and $\epsilon > 0$

A quick example : an oracle inequality for the “squared LASSO”

Data : $(X_i, Y_i)_{i=1}^n$ i.i.d. $\sim (X, Y) \in \mathbb{R}^d \times \mathbb{R}$.
Assumption : Y and $\|X\|_{\ell_\infty^d}$ are subgaussian.
The regularized empirical risk minimization (ERM) estimator

$$\hat{\beta}_n \in \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \left(\frac{1}{n} \sum_{i=1}^n (Y_i - \langle X_i, \beta \rangle)^2 + \lambda \frac{\|\beta\|_{\ell_1}^2}{n\epsilon^2} \right)$$

where $\lambda = \lambda(n, d) = polylog(n, d)$ and $\epsilon > 0$ satisfies, with large probability,

$$\mathbb{E}(Y - \langle \hat{\beta}_n, X \rangle)^2 \leq \inf_{\beta \in \mathbb{R}^d} \left((1+\epsilon)\mathbb{E}(Y - \langle \beta, X \rangle)^2 + c_1 \lambda \frac{(1 + \|\beta\|_{\ell_1^d}^2)}{n\epsilon^2} \right).$$

A quick example : an oracle inequality for the “squared LASSO”

Data : $(X_i, Y_i)_{i=1}^n$ i.i.d. $\sim (X, Y) \in \mathbb{R}^d \times \mathbb{R}$.
Assumption : Y and $\|X\|_{\ell_\infty^d}$ are subgaussian.
The regularized empirical risk minimization (ERM) estimator

$$\hat{\beta}_n \in \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \Big(\frac{1}{n} \sum_{i=1}^n (Y_i - \langle X_i, \beta \rangle)^2 + \lambda \frac{\|\beta\|_{\ell_1}^2}{n\epsilon^2} \Big)$$

where $\lambda = \lambda(n, d) = polylog(n, d)$ and $\epsilon > 0$ satisfies, with large probability,

$$\mathbb{E}(Y - \langle \hat{\beta}_n, X \rangle)^2 \leq \inf_{\beta \in \mathbb{R}^d} \Big((1+\epsilon)\mathbb{E}(Y - \langle \beta, X \rangle)^2 + c_1 \lambda \frac{(1 + \|\beta\|_{\ell_1^d}^2)}{n\epsilon^2} \Big).$$

Question 1 : What is the reason for penalizing by $\| \cdot \|_{\ell_1^d}^2$?

A quick example : an oracle inequality for the "squared LASSO"

Data : $(X_i, Y_i)_{i=1}^n$ i.i.d. $\sim (X, Y) \in \mathbb{R}^d \times \mathbb{R}$.
Assumption : Y and $\|X\|_{\ell_\infty^d}$ are subgaussian.
The regularized empirical risk minimization (ERM) estimator

$$\hat{\beta}_n \in \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \Big(\frac{1}{n} \sum_{i=1}^n (Y_i - \langle X_i, \beta \rangle)^2 + \lambda \frac{\|\beta\|_{\ell_1}^2}{n\epsilon^2} \Big)$$

where $\lambda = \lambda(n, d) = polylog(n, d)$ and $\epsilon > 0$ satisfies, with large probability,

$$\mathbb{E}(Y - \langle \hat{\beta}_n, X \rangle)^2 \leq \inf_{\beta \in \mathbb{R}^d} \Big((1+\epsilon)\mathbb{E}(Y - \langle \beta, X \rangle)^2 + c_1 \lambda \frac{(1 + \|\beta\|_{\ell_1^d}^2)}{n\epsilon^2} \Big).$$

Question 1 : What is the reason for penalizing by $\| \cdot \|_{\ell_1^d}^2$?
Question 2 : Why is it possible to achieve a fast $1/n$-residual term without any "RIP -type" assumption ?

General model in learning theory

- $Z_1, \ldots, Z_n$: n i.i.d.$\sim Z$ random variables in $\mathcal{Z}$

General model in learning theory

- $Z_1, \ldots, Z_n$: n i.i.d.$\sim Z$ random variables in $\mathcal{Z}$
- $\ell : (f, z) \longmapsto \ell(f, z) := \ell_f(z) \in \mathbb{R}$: loss function of $f : \mathcal{Z} \to \mathbb{R}$

General model in learning theory

- $Z_1, \ldots, Z_n$: n i.i.d.$\sim Z$ random variables in $\mathcal{Z}$

- $\ell : (f, z) \longmapsto \ell(f, z) := \ell_f(z) \in \mathbb{R}$: loss function of $f : \mathcal{Z} \to \mathbb{R}$

- $R(f) = \mathbb{E}\ell_f(Z)$: risk of f

General model in learning theory

- $Z_1, \dots, Z_n$: n i.i.d.$\sim Z$ random variables in $\mathcal{Z}$
- $\ell : (f, z) \longmapsto \ell(f, z) := \ell_f(z) \in \mathbb{R}$: loss function of $f : \mathcal{Z} \to \mathbb{R}$
- $R(f) = \mathbb{E}\ell_f(Z)$: risk of f
- risk of a statistics $\hat{f}_n$ is

$$R(\hat{f}_n) = \mathbb{E}\big[\ell_{\hat{f}_n}(Z) | \mathcal{D}\big]$$

where $\mathcal{D} := (Z_1, \cdots, Z_n)$.

General model in learning theory

- $Z_1, \ldots, Z_n$: n i.i.d.$\sim Z$ random variables in $\mathcal{Z}$ $(Z = (X, Y))$

- $\ell : (f, z) \longmapsto \ell(f, z) := \ell_f(z) \in \mathbb{R}$: loss function of $f : \mathcal{Z} \to \mathbb{R}$

- $R(f) = \mathbb{E}\ell_f(Z)$: risk of f

- risk of a statistics $\hat{f}_n$ is

$$R(\hat{f}_n) = \mathbb{E}\big[\ell_{\hat{f}_n}(Z) | \mathcal{D}\big]$$

  where $\mathcal{D} := (Z_1, \cdots, Z_n)$.

General model in learning theory

- $Z_1, \dots, Z_n$: n i.i.d.$\sim Z$ random variables in $\mathcal{Z}$ $(Z = (X, Y))$

- $\ell : (f, z) \longmapsto \ell(f, z) := \ell_f(z) \in \mathbb{R}$: loss function of $f : \mathcal{Z} \to \mathbb{R}$
  $(\ell_f(z) := \ell_f(x, y) = (y - f(x))^2)$

- $R(f) = \mathbb{E}\ell_f(Z)$: risk of f

- risk of a statistics $\hat{f}_n$ is

$$R(\hat{f}_n) = \mathbb{E}\big[\ell_{\hat{f}_n}(Z)|\mathcal{D}\big]$$

  where $\mathcal{D} := (Z_1, \cdots, Z_n)$.

General model in learning theory

- $Z_1, \ldots, Z_n$: n i.i.d.$\sim Z$ random variables in $\mathcal{Z}$ $(Z = (X, Y))$

- $\ell : (f, z) \longmapsto \ell(f, z) := \ell_f(z) \in \mathbb{R}$: loss function of $f : \mathcal{Z} \to \mathbb{R}$
  $(\ell_f(z) := \ell_f(x, y) = (y - f(x))^2)$

- $R(f) = \mathbb{E}\ell_f(Z)$: risk of f $(R(f) = \mathbb{E}(Y - f(X))^2)$

- risk of a statistics $\hat{f}_n$ is

$$R(\hat{f}_n) = \mathbb{E}\big[\ell_{\hat{f}_n}(Z)|\mathcal{D}\big]$$

  where $\mathcal{D} := (Z_1, \cdots, Z_n)$.

General model in learning theory

- Assumption : We don't want to assume any particular model (i.e. we don't assume that $Y = f^*(X) + \sigma g$ etc...). No assumption on the model (only tail assumption on $\ell_f(Z), f \in F$).

General model in learning theory

- Assumption : We don't want to assume any particular model (i.e. we don't assume that $Y = f^*(X) + \sigma g$ etc...). No assumption on the model (only tail assumption on $\ell_f(Z), f \in F$).
- Aim : construct procedures satisfying some oracle inequalities (no control of the approximation term - we focus on the stochastic term...) – three types of oracle inequalities.

General oracle inequalities for Empirial Risk Minimization

Empirical Risk minimization

1. a model F is a class of functions $f : \mathcal{Z} \to \mathbb{R}$

Empirical Risk minimization

1. a model F is a class of functions $f : \mathcal{Z} \to \mathbb{R}$
2. the empirical risk is

$$R_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f, Z_i)$$

Empirical Risk minimization

1. a model F is a class of functions $f : \mathcal{Z} \to \mathbb{R}$
2. the empirical risk is

$$R_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f, Z_i)$$

3. the Empirical Risk Minimization procedure is

$$\hat{f}_n^{(ERM)} \in \operatorname*{argmin}_{f \in F} R_n(f)$$

Three different oracle inequalities. Exemple in aggregation theory.

The ERM over a finite model F w.r.t. the square loss is

$$\hat{f}_n^{(ERM)} \in \mathrm{Arg}\min_{f \in F} R_n(f) \text{ where } R_n(f) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2,$$

Three different oracle inequalities. Exemple in aggregation theory.

The ERM over a finite model F w.r.t. the square loss is

$$\hat{f}_n^{(ERM)} \in \mathrm{Arg}\min_{f \in F} R_n(f) \text{ where } R_n(f) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2,$$

Assume $|Y|, \max_{f\in F}|f(X)| \leq b$ a.s.. For every $x, \epsilon > 0$, with probability greater than $1 - 4\exp(-x)$,

$$R(\hat{f}_n^{(ERM)}) \leq \min_{f\in F} R(f) + c_0\sqrt{\frac{x + \log|F|}{n}}$$

Three different oracle inequalities. Exemple in aggregation theory.

The ERM over a finite model F w.r.t. the square loss is

$$\hat{f}_n^{(ERM)} \in \mathrm{Arg}\min_{f \in F} R_n(f) \text{ where } R_n(f) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2,$$

Assume $|Y|, \max_{f \in F}|f(X)| \leq b$ a.s.. For every $x, \epsilon > 0$, with probability greater than $1 - 4\exp(-x)$,

$$R(\hat{f}_n^{(ERM)}) \leq \min_{f \in F} R(f) + c_0\sqrt{\frac{x + \log|F|}{n}}$$

$$R(\hat{f}_n^{(ERM)}) \leq (1+\epsilon)\min_{f \in F} R(f) + c_0\frac{x + \log|F|}{n\epsilon}$$

Three different oracle inequalities. Exemple in aggregation theory.

The ERM over a finite model F w.r.t. the square loss is

$$\hat{f}_n^{(ERM)} \in \mathrm{Arg}\min_{f\in F} R_n(f) \text{ where } R_n(f) = \frac{1}{n}\sum_{i=1}^n (Y_i - f(X_i))^2,$$

Assume $|Y|, \max_{f\in F}|f(X)| \leq b$ a.s.. For every $x, \epsilon > 0$, with probability greater than $1 - 4\exp(-x)$,

$$R(\hat{f}_n^{(ERM)}) \leq \min_{f\in F} R(f) + c_0\sqrt{\frac{x+\log|F|}{n}}$$

$$R(\hat{f}_n^{(ERM)}) \leq (1+\epsilon)\min_{f\in F} R(f) + c_0\frac{x+\log|F|}{n\epsilon}$$

3. and for $f^*(X) = \mathbb{E}[Y|X]$

$$R(\hat{f}_n^{(ERM)}) - R(f^*) \leq (1+\epsilon)\min_{f\in F}\left(R(f) - R(f^*)\right) + c_0\frac{x+\log|F|}{n\epsilon}$$

Three different oracle inequalities. Exemple in aggregation theory.

The ERM over a finite model F w.r.t. the square loss is

$$\hat{f}_n^{(ERM)} \in \mathrm{Arg}\min_{f\in F} R_n(f) \text{ where } R_n(f) = \frac{1}{n}\sum_{i=1}^n (Y_i - f(X_i))^2,$$

Assume $|Y|, \max_{f\in F}|f(X)| \leq b$ a.s.. For every $x, \epsilon > 0$, with probability greater than $1 - 4\exp(-x)$,

1. Exact Oracle Inequality

$$R(\hat{f}_n^{(ERM)}) \leq \min_{f\in F} R(f) + c_0\sqrt{\frac{x+\log|F|}{n}}$$

$$R(\hat{f}_n^{(ERM)}) \leq (1+\epsilon)\min_{f\in F} R(f) + c_0\frac{x+\log|F|}{n\epsilon}$$

3. and for $f^*(X) = \mathbb{E}[Y|X]$

$$R(\hat{f}_n^{(ERM)}) - R(f^*) \leq (1+\epsilon)\min_{f\in F}\left(R(f) - R(f^*)\right) + c_0\frac{x+\log|F|}{n\epsilon}$$

Three different oracle inequalities. Exemple in aggregation theory.

The ERM over a finite model F w.r.t. the square loss is

$$\hat{f}_n^{(ERM)} \in \operatorname{Arg}\min_{f\in F} R_n(f) \text{ where } R_n(f) = \frac{1}{n}\sum_{i=1}^n (Y_i - f(X_i))^2,$$

Assume $|Y|, \max_{f\in F}|f(X)| \leq b$ a.s.. For every $x, \epsilon > 0$, with probability greater than $1 - 4\exp(-x)$,

1. Exact Oracle Inequality

$$R(\hat{f}_n^{(ERM)}) \leq \min_{f\in F} R(f) + c_0\sqrt{\frac{x+\log|F|}{n}}$$

2. Non-Exact Oracle Inequality

$$R(\hat{f}_n^{(ERM)}) \leq (1+\epsilon)\min_{f\in F} R(f) + c_0\frac{x+\log|F|}{n\epsilon}$$

3. and for $f^*(X) = \mathbb{E}[Y|X]$

$$R(\hat{f}_n^{(ERM)}) - R(f^*) \leq (1+\epsilon)\min_{f\in F}\left(R(f) - R(f^*)\right) + c_0\frac{x+\log|F|}{n\epsilon}$$

Three different oracle inequalities. Exemple in aggregation theory.

The ERM over a finite model F w.r.t. the square loss is

$$\hat{f}_n^{(ERM)} \in \operatorname{Arg}\min_{f\in F} R_n(f) \text{ where } R_n(f) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2,$$

Assume $|Y|, \max_{f\in F}|f(X)| \leq b$ a.s.. For every $x, \epsilon > 0$, with probability greater than $1 - 4\exp(-x)$,

1. Exact Oracle Inequality

$$R(\hat{f}_n^{(ERM)}) \leq \min_{f\in F} R(f) + c_0\sqrt{\frac{x + \log|F|}{n}}$$

2. Non-Exact Oracle Inequality

$$R(\hat{f}_n^{(ERM)}) \leq (1+\epsilon)\min_{f\in F} R(f) + c_0\frac{x + \log|F|}{n\epsilon}$$

3. and for $f^*(X) = \mathbb{E}[Y|X]$ Non-Exact Oracle Inequality for the estimation problem

$$R(\hat{f}_n^{(ERM)}) - R(f^*) \leq (1+\epsilon)\min_{f\in F}\left(R(f) - R(f^*)\right) + c_0\frac{x + \log|F|}{n\epsilon}$$

Three different oracle inequalities. Exemple in aggregation theory.

Three oracle inequalities with two different residual terms :

Three different oracle inequalities. Exemple in aggregation theory.

Three oracle inequalities with two different residual terms :

- fast decaying residual term for the "non-exact oracle inequality" and "non-exact oracle inequality for the estimation problem" :

$$\frac{x + \log |F|}{n} \sim \frac{\mathrm{comp}(F)}{n}$$

Three different oracle inequalities. Exemple in aggregation theory.

Three oracle inequalities with two different residual terms :

- fast decaying residual term for the "non-exact oracle inequality" and "non-exact oracle inequality for the estimation problem" :

$$\frac{x + \log |F|}{n} \sim \frac{\mathrm{comp}(F)}{n}$$

- slow decaying residual term (non-improvable : there exists lower bounds) for the "exact oracle inequality" :

$$\sqrt{\frac{x + \log |F|}{n}} \sim \sqrt{\frac{\mathrm{comp}(F)}{n}}$$

Three different oracle inequalities. Exemple in aggregation theory.

Three oracle inequalities with two different residual terms :

- fast decaying residual term for the "non-exact oracle inequality" and "non-exact oracle inequality for the estimation problem" :

$$\frac{x + \log |F|}{n} \sim \frac{\mathrm{comp}(F)}{n}$$

- slow decaying residual term (non-improvable : there exists lower bounds) for the "exact oracle inequality" :

$$\sqrt{\frac{x + \log |F|}{n}} \sim \sqrt{\frac{\mathrm{comp}(F)}{n}}$$

Question : why is there such a difference between the three oracle inequalities (exact, non-exact, non-exact for estimation) ?

Three different oracle inequalities. Exemple in aggregation theory.

Three oracle inequalities with two different residual terms :

- fast decaying residual term for the "non-exact oracle inequality" and "non-exact oracle inequality for the estimation problem" :

$$\frac{x + \log |F|}{n} \sim \frac{\mathrm{comp}(F)}{n}$$

- slow decaying residual term (non-improvable : there exists lower bounds) for the "exact oracle inequality" :

$$\sqrt{\frac{x + \log |F|}{n}} \sim \sqrt{\frac{\mathrm{comp}(F)}{n}}$$

Question : why is there such a difference between the three oracle inequalities (exact, non-exact, non-exact for estimation) ?(Fundamental reasons ? Geometry - complexity -concentration)

Exact and non-exact oracle inequalities in a general framework

- loss functions class :

$$\ell_F := \{\ell_f : f \in F\}$$

Exact and non-exact oracle inequalities in a general framework

- loss functions class :

$$\ell_F := \{\ell_f : f \in F\}$$

- excess loss functions class : for $f_F^* \in \operatorname{argmin}_{f \in F} R(f)$

$$\mathcal{L}_F := \{\ell_f - \ell_{f_F^*} : f \in F\} = \ell_F - \ell_{f_F^*}$$

Exact and non-exact oracle inequalities in a general framework

- loss functions class :

$$\ell_F := \{\ell_f : f \in F\}$$

- excess loss functions class : for $f_F^* \in \mathrm{argmin}_{f \in F} R(f)$

$$\mathcal{L}_F := \{\ell_f - \ell_{f_F^*} : f \in F\} = \ell_F - \ell_{f_F^*}$$

- excess loss functions class for the estimation problem : for $f^* \in \mathrm{argmin}_f R(f)$

$$\mathcal{E}_F := \{\ell_f - \ell_{f^*} : f \in F\} = \ell_F - \ell_{f^*}$$

Exact and non-exact oracle inequalities in a general framework

- loss functions class :

$$\ell_F := \{\ell_f : f \in F\}$$

- excess loss functions class : for $f_F^* \in \operatorname{argmin}_{f\in F} R(f)$

$$\mathcal{L}_F := \{\ell_f - \ell_{f_F^*} : f \in F\} = \ell_F - \ell_{f_F^*}$$

- excess loss functions class for the estimation problem : for $f^* \in \operatorname{argmin}_f R(f)$

$$\mathcal{E}_F := \{\ell_f - \ell_{f^*} : f \in F\} = \ell_F - \ell_{f^*}$$

For every functions class H, the *star-shaped hull of H in* 0 is

$$V(H) = \mathrm{star}(H, 0) := \{\theta h : 0 \le \theta \le 1, h \in H\}$$

Exact and non-exact oracle inequalities in a general framework

- loss functions class :

$$\ell_F := \{\ell_f : f \in F\}$$

- excess loss functions class : for $f_F^* \in \operatorname{argmin}_{f \in F} R(f)$

$$\mathcal{L}_F := \{\ell_f - \ell_{f_F^*} : f \in F\} = \ell_F - \ell_{f_F^*}$$

- excess loss functions class for the estimation problem : for $f^* \in \operatorname{argmin}_f R(f)$

$$\mathcal{E}_F := \{\ell_f - \ell_{f^*} : f \in F\} = \ell_F - \ell_{f^*}$$

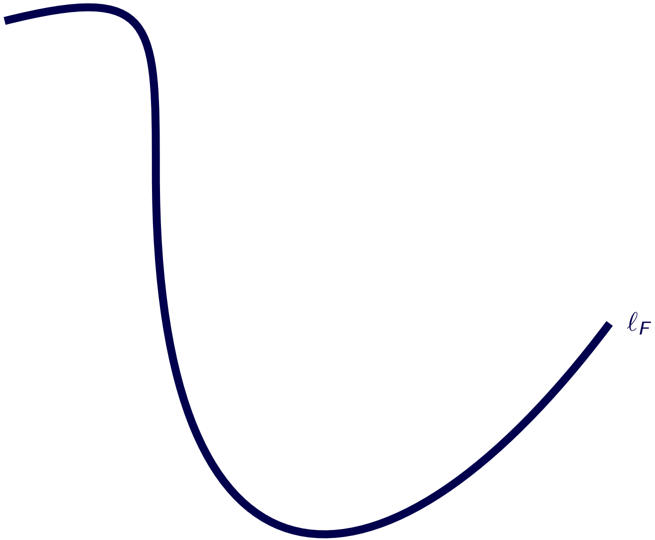
For every functions class H, the *star-shaped hull of H in* 0 is

$$V(H) = \mathrm{star}(H, 0) := \{\theta h : 0 \le \theta \le 1, h \in H\}$$

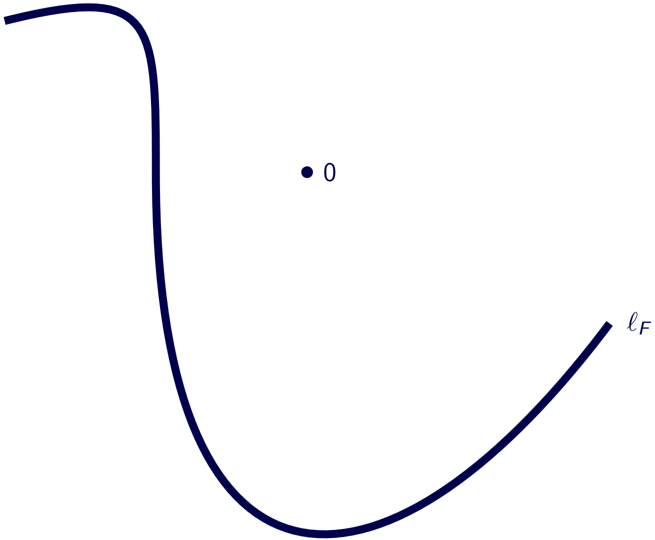
and its *localized set at level* $\lambda > 0$ is

$$V(H)_\lambda := \{g \in V(H) : \mathbb{E} g \le \lambda\}$$

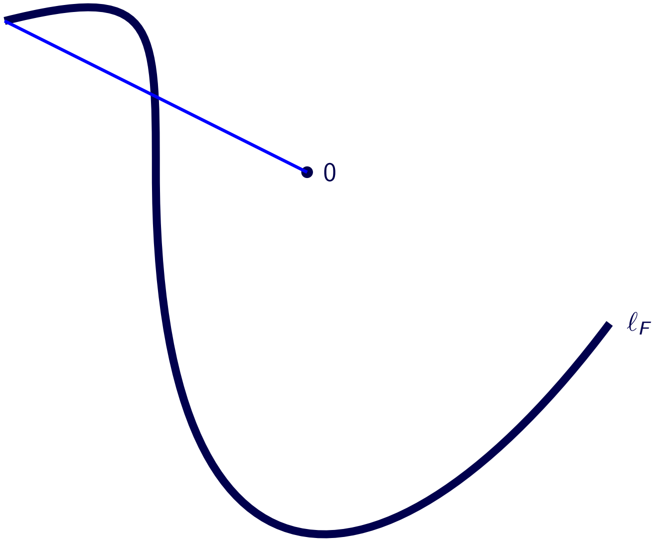
Exact and non-exact oracle inequalities in a general framework



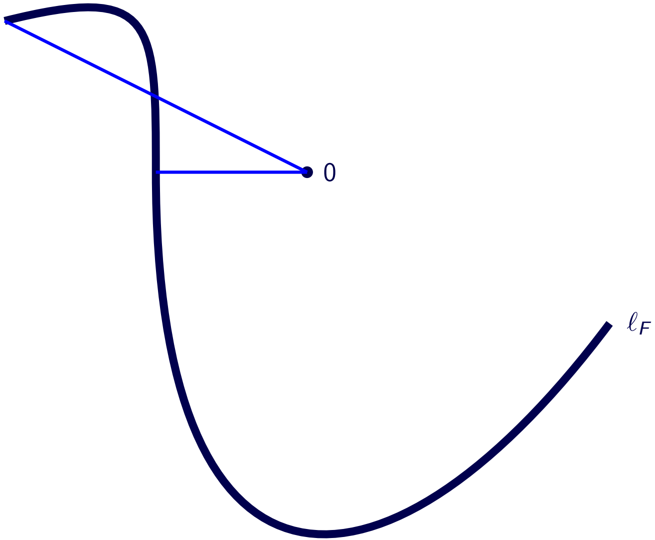
Exact and non-exact oracle inequalities in a general framework



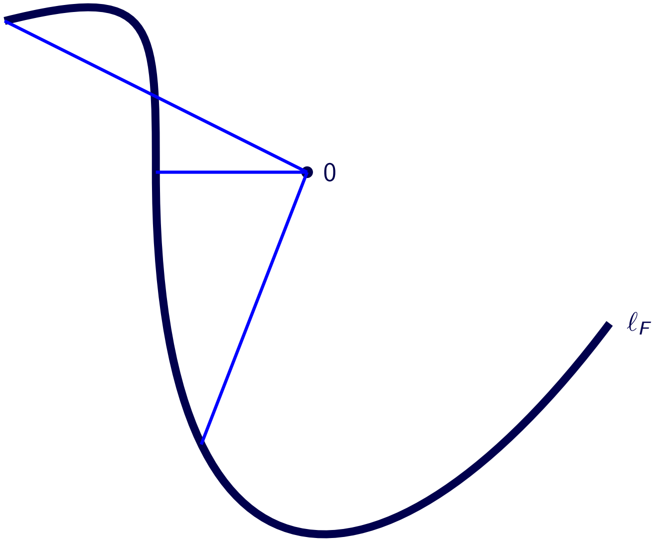
Exact and non-exact oracle inequalities in a general framework



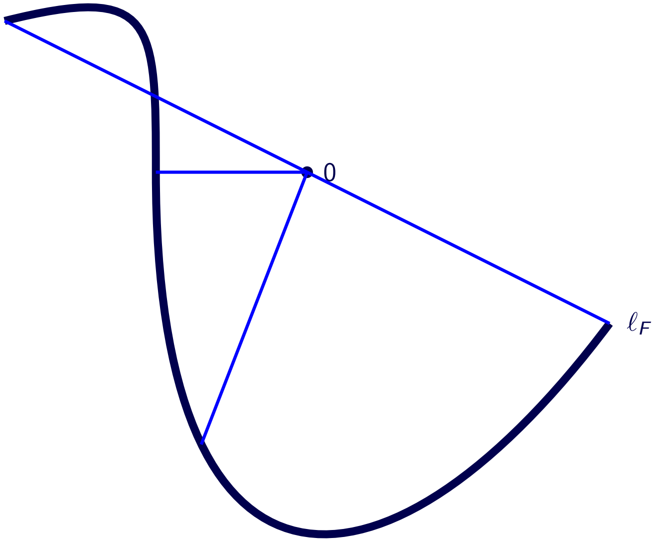
Exact and non-exact oracle inequalities in a general framework



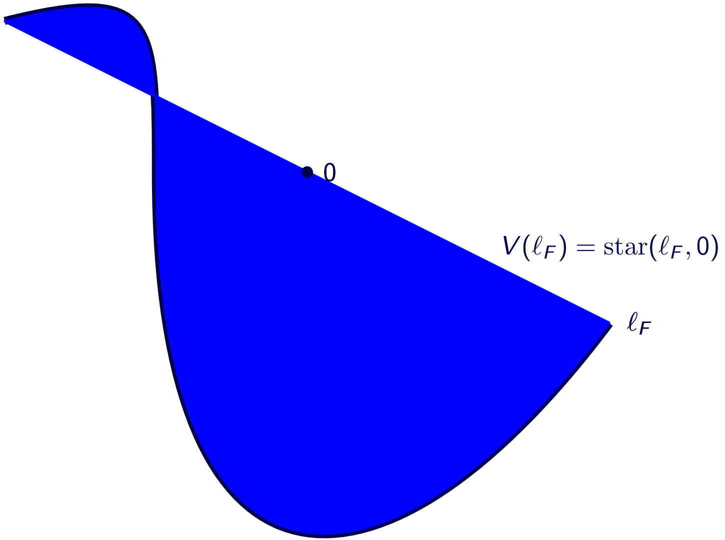
Exact and non-exact oracle inequalities in a general framework



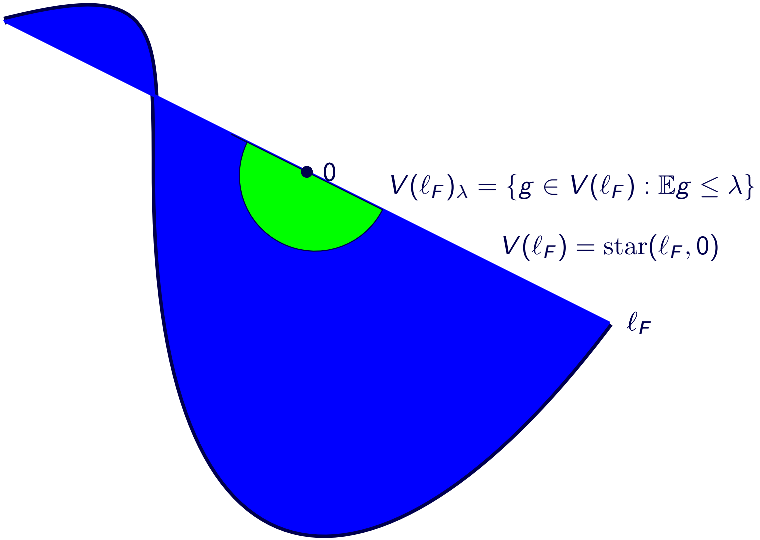
Exact and non-exact oracle inequalities in a general framework



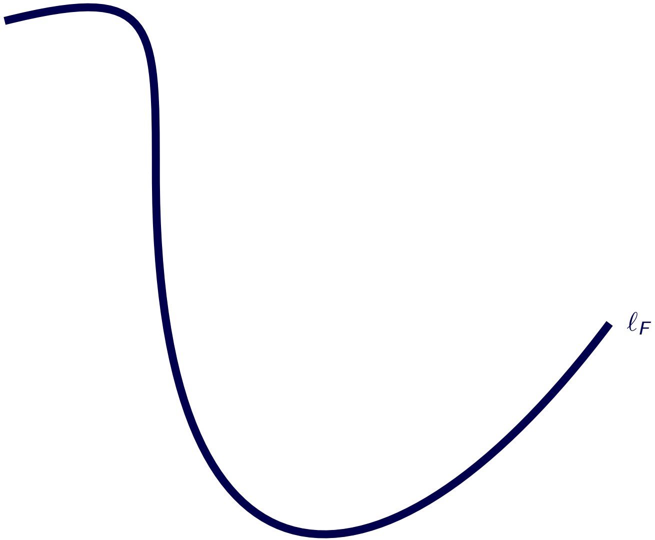
Exact and non-exact oracle inequalities in a general framework



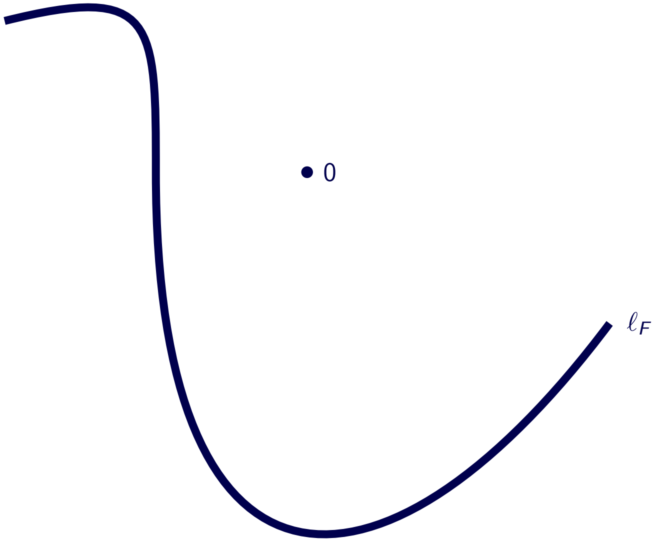
Exact and non-exact oracle inequalities in a general framework



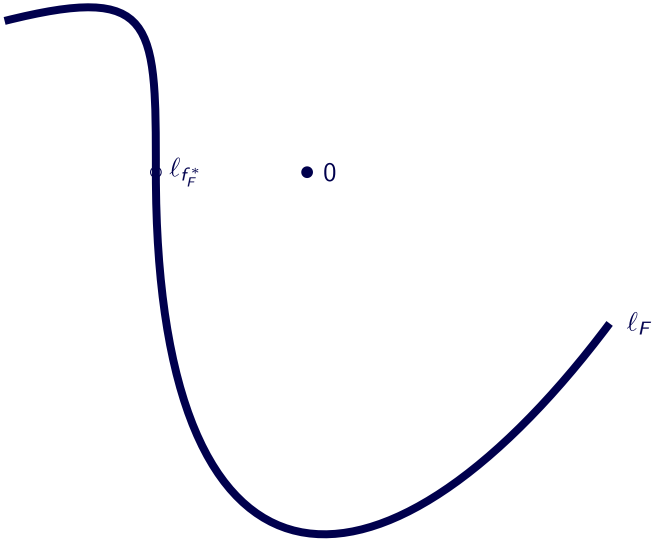
Exact and non-exact oracle inequalities in a general framework



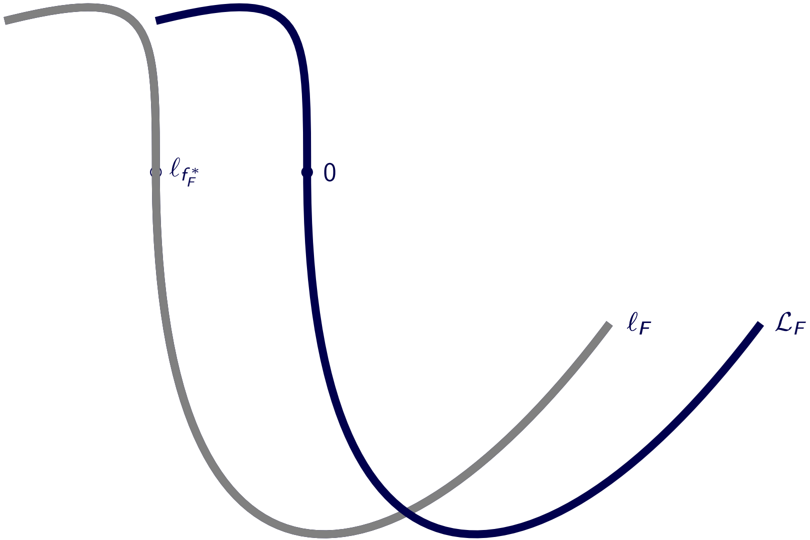
Exact and non-exact oracle inequalities in a general framework



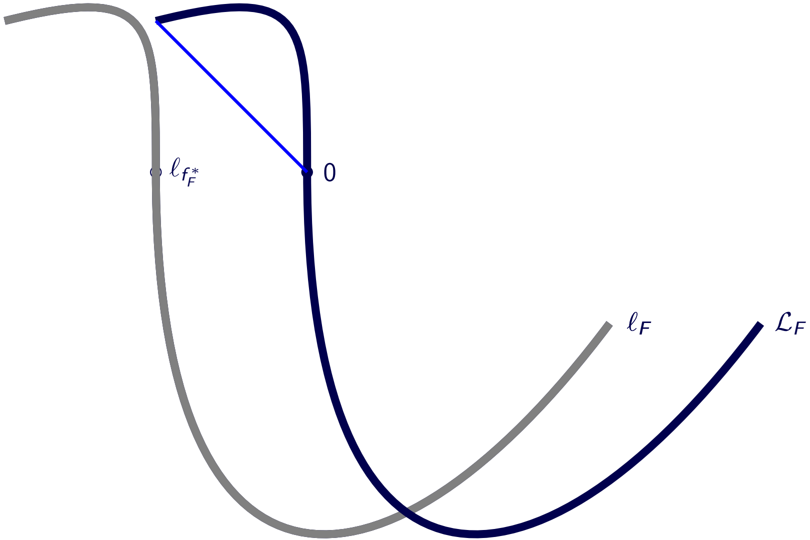
Exact and non-exact oracle inequalities in a general framework



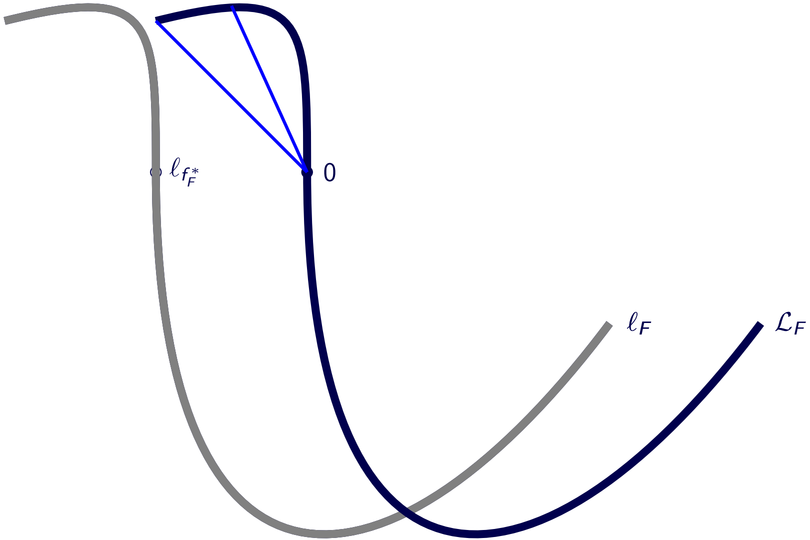
Exact and non-exact oracle inequalities in a general framework



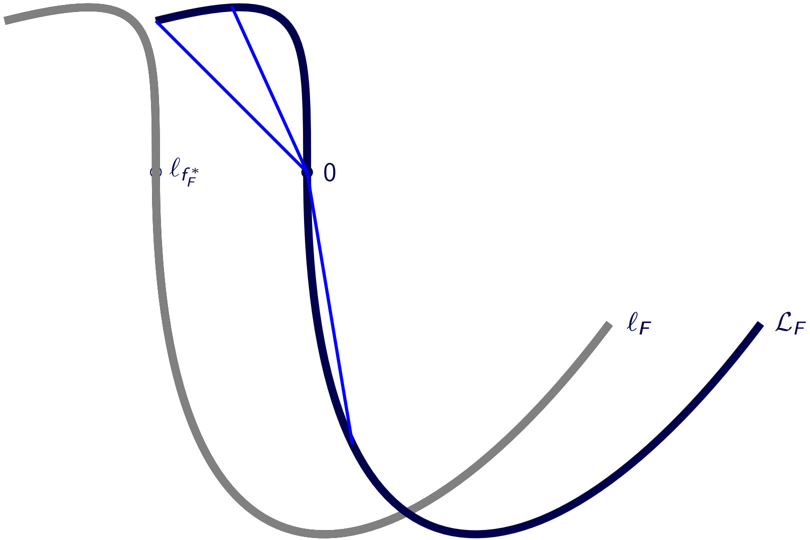
Exact and non-exact oracle inequalities in a general framework



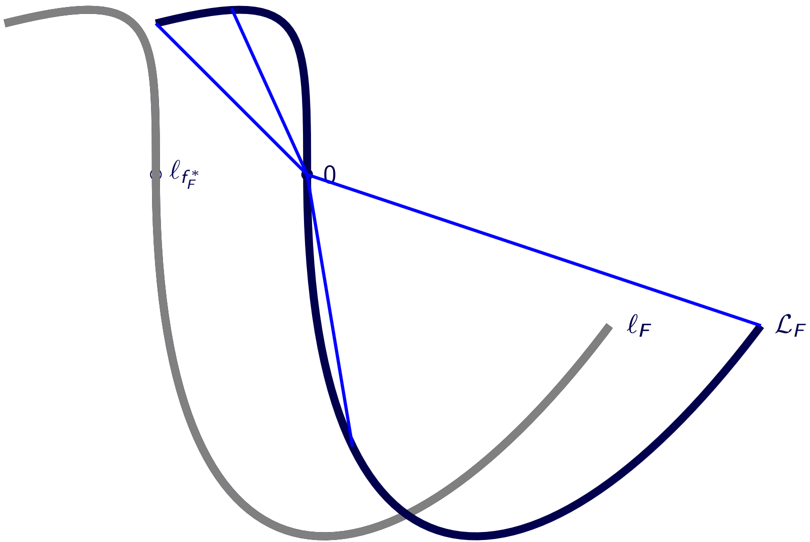
Exact and non-exact oracle inequalities in a general framework



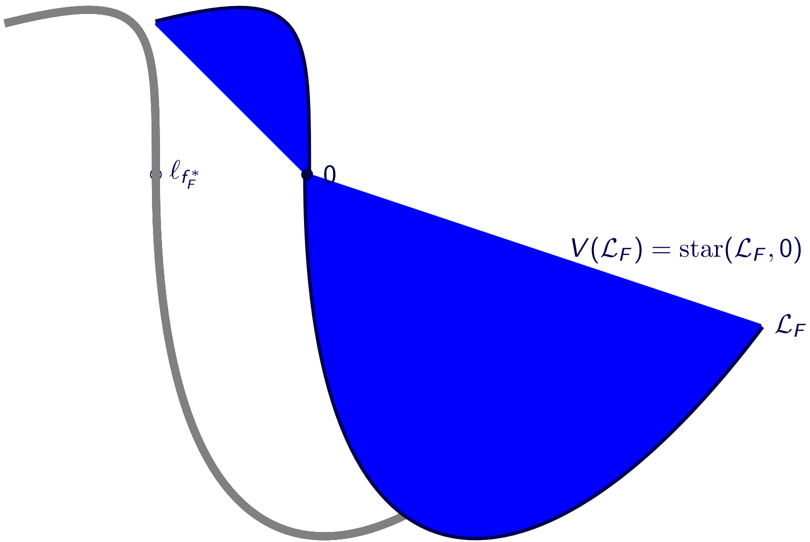
Exact and non-exact oracle inequalities in a general framework



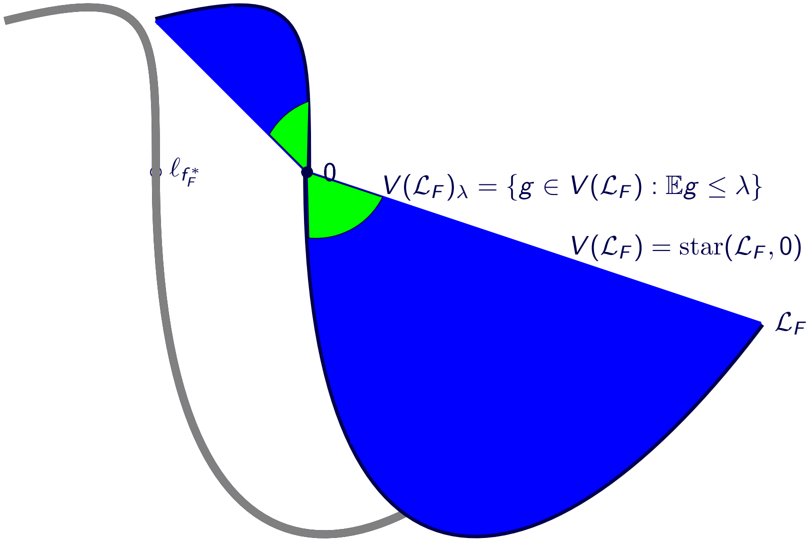
Exact and non-exact oracle inequalities in a general framework



Exact and non-exact oracle inequalities in a general framework

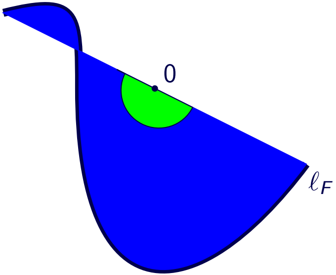


Exact and non-exact oracle inequalities in a general framework

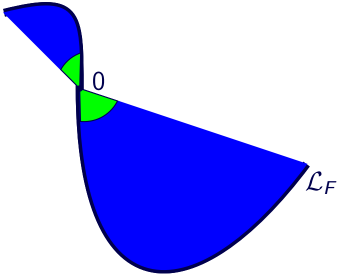


Exact and non-exact oracle inequalities in a general framework

$$V(\ell_F)_\lambda = \{g \in V(\ell_F) : \mathbb{E}g \leq \lambda\}$$

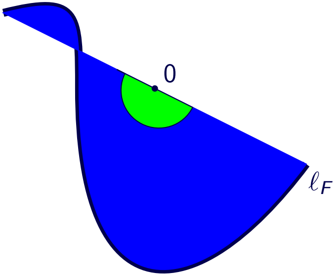


$$V(\mathcal{L}_F)_\lambda = \{g \in V(\mathcal{L}_F) : \mathbb{E}g \leq \lambda\}$$



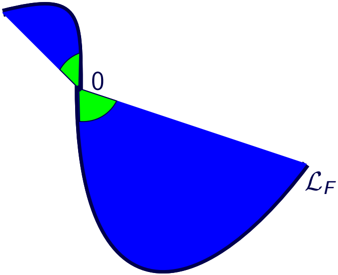
Exact and non-exact oracle inequalities in a general framework

$$V(\ell_F)_\lambda = \{g \in V(\ell_F) : \mathbb{E} g \le \lambda\}$$



Non-exact oracle inequalities

$$V(\mathcal{L}_F)_\lambda = \{g \in V(\mathcal{L}_F) : \mathbb{E} g \le \lambda\}$$



Exact oracle inequalities

Exact and non-exact oracle inequalities in a general framework

$$\|P - P_n\|_H := \sup_{h \in H} |Ph - P_n h|$$

where

$$Ph := \mathbb{E}h(Z) \text{ and } P_n h := \frac{1}{n}\sum_{i=1}^{n} h(Z_i)$$

Exact and non-exact oracle inequalities in a general framework

$$\|P - P_n\|_H := \sup_{h \in H} |Ph - P_n h|$$

where

$$Ph := \mathbb{E}h(Z) \text{ and } P_n h := \frac{1}{n}\sum_{i=1}^{n} h(Z_i)$$

Two important fixed points driving exact and non-exact oracle inequalities :

Exact and non-exact oracle inequalities in a general framework

$$\|P - P_n\|_H := \sup_{h \in H} |Ph - P_n h|$$

where

$$Ph := \mathbb{E}h(Z) \text{ and } P_n h := \frac{1}{n}\sum_{i=1}^{n} h(Z_i)$$

Two important fixed points driving exact and non-exact oracle inequalities :

- for exact oracle inequalities :

$$\mu^* := \inf\big(\mu > 0 : \mathbb{E}\|P - P_n\|_{V(\mathcal{L}_F)_\mu} \leq \mu/8\big)$$

Exact and non-exact oracle inequalities in a general framework

$$\|P - P_n\|_H := \sup_{h \in H} |Ph - P_n h|$$

where

$$Ph := \mathbb{E}h(Z) \text{ and } P_n h := \frac{1}{n}\sum_{i=1}^{n} h(Z_i)$$

Two important fixed points driving exact and non-exact oracle inequalities :

- for exact oracle inequalities :

$$\mu^* := \inf\left(\mu > 0 : \mathbb{E}\|P - P_n\|_{V(\mathcal{L}_F)_\mu} \leq \mu/8\right)$$

- non-exact oracle inequalities :

$$\lambda_\epsilon^* := \inf\left(\lambda > 0 : \mathbb{E}\|P - P_n\|_{V(\ell_F)_\lambda} \leq (\epsilon/4)\lambda\right)$$

Exact oracle inequality

Theorem (Bartlett and Mendelson)

Let F be a class of functions and assume that there exists $B > 0$ such that for every $f \in F$,

$$P\mathcal{L}_f^2 \leq BP\mathcal{L}_f$$

Let $\mu^* > 0$ be s.t. $\mathbb{E}\|P_n - P\|_{V(\mathcal{L}_F)_{\mu^*}} \leq \mu^*/8$

Exact oracle inequality

Theorem (Bartlett and Mendelson)

Let F be a class of functions and assume that there exists $B > 0$ such that for every $f \in F$,

$$P\mathcal{L}_f^2 \le BP\mathcal{L}_f$$

Let $\mu^* > 0$ be s.t. $\mathbb{E}\|P_n - P\|_{V(\mathcal{L}_F)_{\mu^*}} \le \mu^*/8$
Then, for every $x > 0$, with probability greater than $1 - 8\exp(-x)$,

$$R(\hat{f}_n^{ERM}) \le \inf_{f \in F} R(f) + \rho_n(x)$$

Exact oracle inequality

Theorem (Bartlett and Mendelson)

Let F be a class of functions and assume that there exists $B > 0$ such that for every $f \in F$,

$$P\mathcal{L}_f^2 \leq BP\mathcal{L}_f$$

Let $\mu^* > 0$ be s.t. $\mathbb{E}\|P_n - P\|_{V(\mathcal{L}_F)_{\mu^*}} \leq \mu^*/8$
Then, for every $x > 0$, with probability greater than $1 - 8\exp(-x)$,

$$R(\hat{f}_n^{ERM}) \leq \inf_{f \in F} R(f) + \rho_n(x)$$

where ρ_n is an increasing function s.t. for every $x > 0$,

$$\rho_n(x) \geq \max\left(\mu^*, c_0 \frac{(\|\mathcal{L}_F\|_\infty + B)x}{n}\right).$$

Exact oracle inequality

Theorem (Bartlett and Mendelson)

Let F be a class of functions and assume that there exists $B > 0$ such that for every $f \in F$,

$$P\mathcal{L}_f^2 \leq BP\mathcal{L}_f$$

Let $\mu^* > 0$ be s.t. $\mathbb{E}\|P_n - P\|_{V(\mathcal{L}_F)_{\mu^*}} \leq \mu^*/8$
Then, for every $x > 0$, with probability greater than $1 - 8\exp(-x)$,

$$R(\hat{f}_n^{ERM}) \leq \inf_{f \in F} R(f) + \rho_n(x)$$

where ρ_n is an increasing function s.t. for every $x > 0$,

$$\rho_n(x) \geq \max\left(\mu^*, c_0 \frac{(\|\mathcal{L}_F\|_\infty + B)x}{n}\right).$$

cf. similar results in [Massart and Nédélec], [Koltchinskii],..

Non-exact oracle inequality

Theorem (L. and Mendelson)

Let F be a class of functions and assume that there exists $B \geq 0$ such that for every $f \in F$,

$$P\ell_f^2 \leq BP\ell_f + B^2/n$$

Let $0 < \epsilon < 1$ and consider $\lambda_\epsilon^* > 0$ for which

$$\mathbb{E}\|P_n - P\|_{V(\ell_F)_{\lambda_\epsilon^*}} \leq (\epsilon/4)\lambda_\epsilon^*$$

Non-exact oracle inequality

Theorem (L. and Mendelson)

Let F be a class of functions and assume that there exists $B \geq 0$ such that for every $f \in F$,

$$P\ell_f^2 \leq BP\ell_f + B^2/n$$

Let $0 < \epsilon < 1$ and consider $\lambda_\epsilon^* > 0$ for which

$$\mathbb{E}\|P_n - P\|_{V(\ell_F)_{\lambda_\epsilon^*}} \leq (\epsilon/4)\lambda_\epsilon^*$$

Then, for every $x > 0$, with probability greater than $1 - 8\exp(-x)$,

$$R(\hat{f}_n^{ERM}) \leq (1 + 2\epsilon) \inf_{f \in F} R(f) + \tilde{\rho}_n(x)$$

Non-exact oracle inequality

Theorem (L. and Mendelson)

Let F be a class of functions and assume that there exists $B \geq 0$ such that for every $f \in F$,

$$P\ell_f^2 \leq BP\ell_f + B^2/n$$

Let $0 < \epsilon < 1$ and consider $\lambda_\epsilon^* > 0$ for which

$$\mathbb{E}\|P_n - P\|_{V(\ell_F)_{\lambda_\epsilon^*}} \leq (\epsilon/4)\lambda_\epsilon^*$$

Then, for every $x > 0$, with probability greater than $1 - 8\exp(-x)$,

$$R(\hat{f}_n^{ERM}) \leq (1 + 2\epsilon) \inf_{f \in F} R(f) + \tilde{\rho}_n(x)$$

where ρ_n is an increasing function s.t. for every $x > 0$

$$\tilde{\rho}_n(x) \geq \max\left(\lambda_\epsilon^*, c_0 \frac{(\|\ell_F\|_\infty + B/\epsilon)x}{n\epsilon}\right).$$

The Bernstein Condition

1. Exact oracle inequality : $\forall f \in F, P\mathcal{L}_f^2 \leq BP\mathcal{L}_f$;

The Bernstein Condition

1. Exact oracle inequality : $\forall f \in F, P\mathcal{L}_f^2 \leq BP\mathcal{L}_f$;
2. Non-exact oracle inequality : $\forall f \in F, P\ell_f^2 \leq BP\ell_f + B^2/n$.

The Bernstein Condition

1. Exact oracle inequality : $\forall f \in F, P\mathcal{L}_f^2 \leq BP\mathcal{L}_f$;
2. Non-exact oracle inequality : $\forall f \in F, P\ell_f^2 \leq BP\ell_f + B^2/n$.

Lemma

For every function f s.t. $\ell_f \geq 0$ a.s. and $\|\ell_f(Z)\|_{\psi_1} \leq D$ for some $D \geq 1$, we have, for every n,

$$P\ell_f^2 \leq \big(c_0 D \log(en)\big) P\ell_f + \frac{\big(c_0 D \log(en)\big)^2}{n}.$$

The Bernstein Condition

1. Exact oracle inequality : $\forall f \in F, P\mathcal{L}_f^2 \le BP\mathcal{L}_f$;
2. Non-exact oracle inequality : $\forall f \in F, P\ell_f^2 \le BP\ell_f + B^2/n$.

Lemma

For every function f s.t. $\ell_f \ge 0$ a.s. and $\|\ell_f(Z)\|_{\psi_1} \le D$ for some $D \ge 1$, we have, for every n,

$$P\ell_f^2 \le \big(c_0 D \log(en)\big) P\ell_f + \frac{\big(c_0 D \log(en)\big)^2}{n}.$$

Conclusion 1 : In the case of non-exact oracle inequalities, the Bernstein condition for ℓ_F is almost trivially satisfied.

The Bernstein condition of the excess loss class $\mathcal{L}_F$

$\bullet f_2(X)$

$\bullet f_1(X)$

The Bernstein condition of the excess loss class $\mathcal{L}_F$

$\bullet f_2(X)$

$F = \{f_1, f_2\}$

$\bullet f_1(X)$

The Bernstein condition of the excess loss class $\mathcal{L}_F$

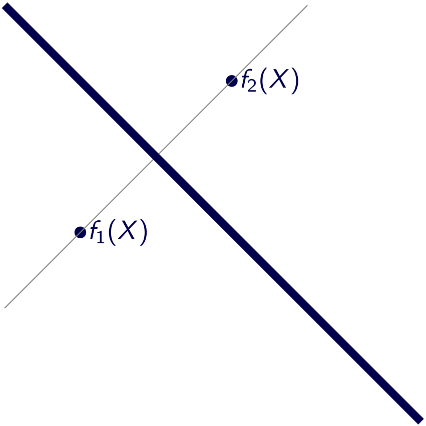
$\bullet f_2(X)$

$$F = \{f_1, f_2\}$$

$$R(f) = \mathbb{E}(Y - f(X))^2$$

$\bullet f_1(X)$

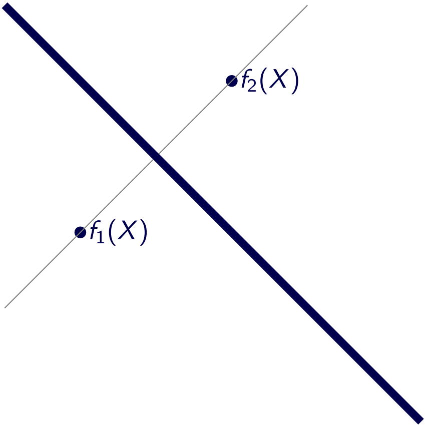
The Bernstein condition of the excess loss class $\mathcal{L}_F$



$$F = \{f_1, f_2\}$$

$$R(f) = \mathbb{E}(Y - f(X))^2$$

The Bernstein condition of the excess loss class $\mathcal{L}_F$

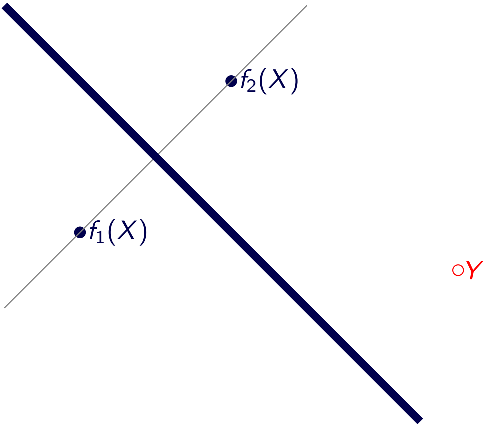


$$F = \{f_1, f_2\}$$

$$R(f) = \mathbb{E}(Y - f(X))^2$$

$$M_F := \{Y : \mathrm{Card}\{f \in F : R(f) = \min_{f \in F} R(f)\} \geq 2\}$$

The Bernstein condition of the excess loss class $\mathcal{L}_F$

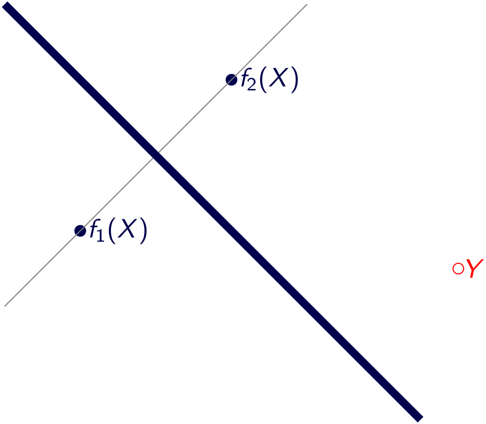


$$F = \{f_1, f_2\}$$

$$R(f) = \mathbb{E}(Y - f(X))^2$$

$$M_F := \{Y : \mathrm{Card}\{f \in F : R(f) = \min_{f\in F} R(f)\} \geq 2\}$$

The Bernstein condition of the excess loss class $\mathcal{L}_F$



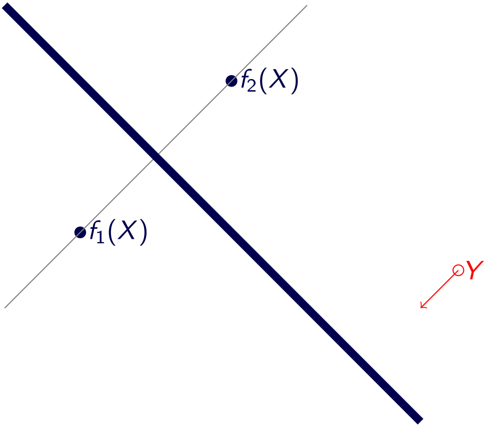
$$F = \{f_1, f_2\}$$

$$R(f) = \mathbb{E}(Y - f(X))^2$$

$$P\mathcal{L}_f^2 \leq BP\mathcal{L}_f, \; B \sim const$$

$$M_F := \{Y : \mathrm{Card}\{f \in F : R(f) = \min_{f \in F} R(f)\} \geq 2\}$$

The Bernstein condition of the excess loss class $\mathcal{L}_F$



$F = \{f_1, f_2\}$

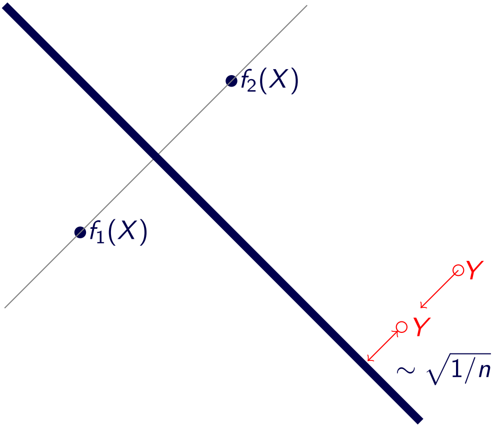
$R(f) = \mathbb{E}(Y - f(X))^2$

$P\mathcal{L}_f^2 \leq BP\mathcal{L}_f$, $B \sim const$

B increases

$M_F := \{Y : \mathrm{Card}\{f \in F : R(f) = \min_{f \in F} R(f)\} \geq 2\}$

The Bernstein condition of the excess loss class $\mathcal{L}_F$



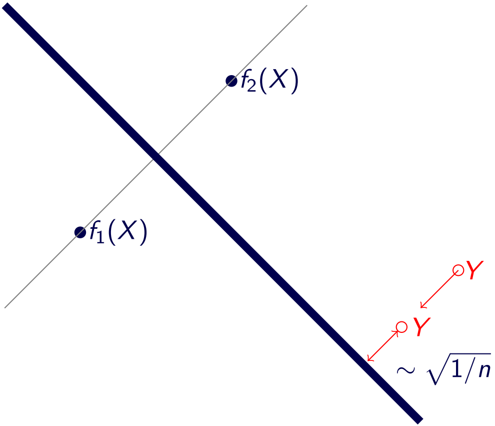
$$F = \{f_1, f_2\}$$

$$R(f) = \mathbb{E}(Y - f(X))^2$$

$$P\mathcal{L}_f^2 \leq B P\mathcal{L}_f, \ B \sim const$$

$$M_F := \{Y : \mathrm{Card}\{f \in F : R(f) = \min_{f \in F} R(f)\} \geq 2\}$$

The Bernstein condition of the excess loss class $\mathcal{L}_F$



$F = \{f_1, f_2\}$

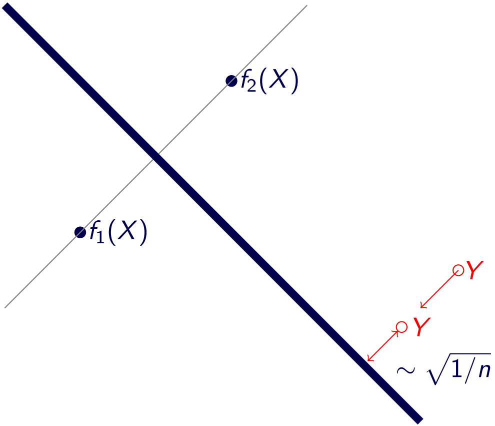
$R(f) = \mathbb{E}(Y - f(X))^2$

$P\mathcal{L}_f^2 \leq BP\mathcal{L}_f$, $B \sim const$

$B \sim \sqrt{n}$

$M_F := \{Y : \mathrm{Card}\{f \in F : R(f) = \min_{f \in F} R(f)\} \geq 2\}$

The Bernstein condition of the excess loss class $\mathcal{L}_F$



$F = \{f_1, f_2\}$

$R(f) = \mathbb{E}(Y - f(X))^2$

$P\mathcal{L}_f^2 \leq BP\mathcal{L}_f$, $B \sim const$

$B \sim \sqrt{n}$

residual term $\sim 1/\sqrt{n}$

$M_F := \{Y : \mathrm{Card}\{f \in F : R(f) = \min_{f \in F} R(f)\} \geq 2\}$

The Bernstein condition

Conclusion 2 : In the case of exact oracle inequalities, the Bernstein condition depends in a very strong way of the geometry of the couple (F, Y).

The Bernstein condition

Conclusion 2 : In the case of exact oracle inequalities, the Bernstein condition depends in a very strong way of the geometry of the couple (F, Y).

This explains the gap in the aggregation problem : for this problem, the set of multiple minimizer M_F is never empty. So it is always possible to find a target Y in a "bad" position leading to an excess loss class $\mathcal{L}_F$ with a trivial Bernstein constant ($B \sim \sqrt{n}$) and thus a slow residual term $\sim \sqrt{\mathrm{Comp}(F)/n}$.

The Bernstein condition

Conclusion 2 : In the case of exact oracle inequalities, the Bernstein condition depends in a very strong way of the geometry of the couple (F, Y).

This explains the gap in the aggregation problem : for this problem, the set of multiple minimizer M_F is never empty. So it is always possible to find a target Y in a "bad" position leading to an excess loss class $\mathcal{L}_F$ with a trivial Bernstein constant ($B \sim \sqrt{n}$) and thus a slow residual term $\sim \sqrt{\mathrm{Comp}(F)/n}$.

1. When the class F is convex : the Bernstein condition of $\mathcal{L}_F$ is always satisfied (quadratic loss).

The Bernstein condition

Conclusion 2 : In the case of exact oracle inequalities, the Bernstein condition depends in a very strong way of the geometry of the couple (F, Y).

This explains the gap in the aggregation problem : for this problem, the set of multiple minimizer M_F is never empty. So it is always possible to find a target Y in a "bad" position leading to an excess loss class $\mathcal{L}_F$ with a trivial Bernstein constant ($B \sim \sqrt{n}$) and thus a slow residual term $\sim \sqrt{\mathrm{Comp}(F)/n}$.

1. When the class F is convex : the Bernstein condition of $\mathcal{L}_F$ is always satisfied (quadratic loss).
2. When the class F is not convex : the ERM is likely to be a suboptimal procedure but there are some possibilities to "improve the geometry" of F : by "starification" (Audibert) or "pre-selection-convexification" (L. and Mendelson).

The complexity terms : μ^* and λ^*_ϵ - Part 1

The fixed points μ^* and λ^* characterize the isomorphic properties of $\mathcal{L}_F$ and ℓ_F respectively :

The complexity terms : μ^* and λ^*_ϵ - Part 1

The fixed points μ^* and λ^* characterize the isomorphic properties of $\mathcal{L}_F$ and ℓ_F respectively :

Theorem (Bartlett and Mendelson)

If H is a class of functions s.t.

$$Ph^2 \leq BPh, \forall h \in H,$$

The complexity terms : μ^* and λ^*_ϵ - Part 1

The fixed points μ^* and λ^* characterize the isomorphic properties of $\mathcal{L}_F$ and ℓ_F respectively :

Theorem (Bartlett and Mendelson)

If H is a class of functions s.t.

$$Ph^2 \leq BPh, \forall h \in H,$$

then for every $x > 0$, with probability greater than $1 - 4\exp(-x)$,

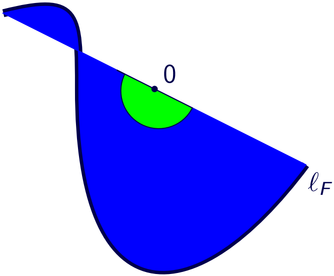
$$(1/2)P_n h \leq Ph \leq (3/2)P_n h$$

for every $h \in H$ s.t. $Ph \geq \max(\kappa^*, x/n)$ where

$$\kappa^* := \inf\big(\kappa > 0 : \mathbb{E}\|P - P_n\|_{V(H)_\kappa} \leq \kappa/8\big).$$

Exact and non-exact oracle inequalities in a general framework - Part 4

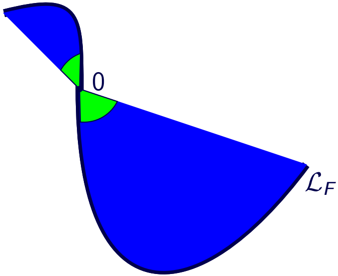
$V(\ell_F)_\lambda = \{g \in V(\ell_F) : \mathbb{E}g \leq \lambda\}$



Non-exact oracle inequalities

$\mathbb{E}\|P - P_n\|_{V(\ell_F)_{\lambda_\epsilon^*}} \leq (\epsilon/4)\lambda_\epsilon^*$

$V(\mathcal{L}_F)_\lambda = \{g \in V(\mathcal{L}_F) : \mathbb{E}g \leq \lambda\}$



Exact oracle inequalities

$\mathbb{E}\|P - P_n\|_{V(\mathcal{L}_F)_{\mu^*}} \leq \mu^*/8$

An example of computation of the fixed points λ_ϵ^* and μ^*

[Peeling argument :]

An example of computation of the fixed points λ_ϵ^* and μ^*

[Peeling argument :]H a class of functions s.t. $Ph \geq 0, \forall h \in H$:

$$V(H)_\lambda \subset \bigcup_{i=0}^{\infty} \{\theta h : 0 \leq \theta \leq 2^{-i}, h \in H, Ph \leq 2^{i+1}\lambda\}.$$

Therefore, setting $H_\lambda = \{h \in H : Ph \leq \lambda\}$,

$$\mathbb{E}\|P - P_n\|_{V(H)_\lambda} \leq \sum_{i=0}^{\infty} 2^{-i} \mathbb{E}\|P - P_n\|_{H_{2^{i+1}\lambda}}$$

An example of computation of the fixed points λ_ϵ^* and μ^*

[Peeling argument :]H a class of functions s.t. $Ph \geq 0, \forall h \in H$:

$$V(H)_\lambda \subset \bigcup_{i=0}^{\infty} \{\theta h : 0 \leq \theta \leq 2^{-i}, h \in H, Ph \leq 2^{i+1}\lambda\}.$$

Therefore, setting $H_\lambda = \{h \in H : Ph \leq \lambda\}$,

$$\mathbb{E}\|P - P_n\|_{V(H)_\lambda} \leq \sum_{i=0}^{\infty} 2^{-i} \mathbb{E}\|P - P_n\|_{H_{2^{i+1}\lambda}}$$

Different ways of computing $\mathbb{E}\|P - P_n\|_{H_\mu}$:

1. Symetrization+Contraction principle+Dudley entropy integrale ;

An example of computation of the fixed points λ_ϵ^* and μ^*

[Peeling argument :]H a class of functions s.t. $Ph \geq 0, \forall h \in H$:

$$V(H)_\lambda \subset \bigcup_{i=0}^{\infty} \{\theta h : 0 \leq \theta \leq 2^{-i}, h \in H, Ph \leq 2^{i+1}\lambda\}.$$

Therefore, setting $H_\lambda = \{h \in H : Ph \leq \lambda\}$,

$$\mathbb{E}\|P - P_n\|_{V(H)_\lambda} \leq \sum_{i=0}^{\infty} 2^{-i} \mathbb{E}\|P - P_n\|_{H_{2^{i+1}\lambda}}$$

Different ways of computing $\mathbb{E}\|P - P_n\|_{H_\mu}$:

1. Symetrization+Contraction principle+Dudley entropy integrale ;
2. Some particular chaining methods ;

An example of computation of the fixed points λ_ϵ^* and μ^*

[Peeling argument :]H a class of functions s.t. $Ph \geq 0, \forall h \in H$:

$$V(H)_\lambda \subset \bigcup_{i=0}^{\infty} \{\theta h : 0 \leq \theta \leq 2^{-i}, h \in H, Ph \leq 2^{i+1}\lambda\}.$$

Therefore, setting $H_\lambda = \{h \in H : Ph \leq \lambda\}$,

$$\mathbb{E}\|P - P_n\|_{V(H)_\lambda} \leq \sum_{i=0}^{\infty} 2^{-i}\mathbb{E}\|P - P_n\|_{H_{2^{i+1}\lambda}}$$

Different ways of computing $\mathbb{E}\|P - P_n\|_{H_\mu}$:

1. Symetrization+Contraction principle+Dudley entropy integrale ;
2. Some particular chaining methods ;
3. Gaussian complexities ;

An example of computation of the fixed points λ_ϵ^* and μ^*

[Peeling argument :]H a class of functions s.t. $Ph \geq 0, \forall h \in H$:

$$V(H)_\lambda \subset \bigcup_{i=0}^{\infty} \{\theta h : 0 \leq \theta \leq 2^{-i}, h \in H, Ph \leq 2^{i+1}\lambda\}.$$

Therefore, setting $H_\lambda = \{h \in H : Ph \leq \lambda\}$,

$$\mathbb{E}\|P - P_n\|_{V(H)_\lambda} \leq \sum_{i=0}^{\infty} 2^{-i} \mathbb{E}\|P - P_n\|_{H_{2^{i+1}\lambda}}$$

Different ways of computing $\mathbb{E}\|P - P_n\|_{H_\mu}$:

1. Symetrization+Contraction principle+Dudley entropy integrale ;
2. Some particular chaining methods ;
3. Gaussian complexities ;
4. Bourgain a priori method (in particular Rudelson method),..

An example of computation of the fixed points λ_ϵ^* and μ^*

Computation of λ_ϵ^* and μ^* in the case of the Regression model with quadratic loss :

$$\ell_F := \{\ell_f : (y,x) \longmapsto (y - f(x))^2 : f \in F\}$$

and

$$\mathcal{L}_F := \{\mathcal{L}_f : (y,x) \longmapsto (y - f(x))^2 - (y - f_F^*(x))^2 : f \in F\}.$$

An example of computation of the fixed points λ_ϵ^* and μ^*

Computation of λ_ϵ^* and μ^* in the case of the Regression model with quadratic loss :

$$\ell_F := \{\ell_f : (y,x) \longmapsto (y - f(x))^2 : f \in F\}$$

and

$$\mathcal{L}_F := \{\mathcal{L}_f : (y,x) \longmapsto (y - f(x))^2 - (y - f_F^*(x))^2 : f \in F\}.$$

Complexity measure of F :

An example of computation of the fixed points λ_ϵ^* and μ^*

Computation of λ_ϵ^* and μ^* in the case of the Regression model with quadratic loss :

$$\ell_F := \{\ell_f : (y,x) \longmapsto (y - f(x))^2 : f \in F\}$$

and

$$\mathcal{L}_F := \{\mathcal{L}_f : (y,x) \longmapsto (y - f(x))^2 - (y - f_F^*(x))^2 : f \in F\}.$$

Complexity measure of F :

$P_\sigma F := \{(f(X_1), \cdots, f(X_n)) : f \in F\}$

An example of computation of the fixed points λ_ϵ^* and μ^*

Computation of λ_ϵ^* and μ^* in the case of the Regression model with quadratic loss :

$$\ell_F := \{\ell_f : (y, x) \longmapsto (y - f(x))^2 : f \in F\}$$

and

$$\mathcal{L}_F := \{\mathcal{L}_f : (y, x) \longmapsto (y - f(x))^2 - (y - f_F^*(x))^2 : f \in F\}.$$

Complexity measure of F :

$P_\sigma F := \{(f(X_1), \cdots, f(X_n)) : f \in F\}$ and $U_n(F^{(\mu)}) := \mathbb{E}\gamma_2(\widetilde{P_\sigma F^{(\mu)}}, \ell_\infty^n)^2$

An example of computation of the fixed points λ_ϵ^* and μ^*

Computation of λ_ϵ^* and μ^* in the case of the Regression model with quadratic loss :

$$\ell_F := \{\ell_f : (y,x) \longmapsto (y - f(x))^2 : f \in F\}$$

and

$$\mathcal{L}_F := \{\mathcal{L}_f : (y,x) \longmapsto (y - f(x))^2 - (y - f_F^*(x))^2 : f \in F\}.$$

Complexity measure of F :

$P_\sigma F := \{(f(X_1), \cdots, f(X_n)) : f \in F\}$ and $U_n(F^{(\mu)}) := \mathbb{E}\gamma_2(\widetilde{P_\sigma F^{(\mu)}}, \ell_\infty^n)^2$

where $\gamma_2(T,d) := \inf_{(T_s)} \sup_{t\in T} \sum_{s=0}^\infty 2^{-s/2} d(t, T_s)$ and $|T_s| \leq 2^{2^s}$ and $\tilde{A} = A - A$ and $F^{(\mu)} := \{f \in F : P\ell_f \leq \mu\}$

An example of computation of the fixed points λ_ϵ^* and μ^*

Computation of λ_ϵ^* and μ^* in the case of the Regression model with quadratic loss :

$$\ell_F := \{\ell_f : (y,x) \longmapsto (y - f(x))^2 : f \in F\}$$

and

$$\mathcal{L}_F := \{\mathcal{L}_f : (y,x) \longmapsto (y - f(x))^2 - (y - f_F^*(x))^2 : f \in F\}.$$

Complexity measure of F :

$P_\sigma F := \{(f(X_1), \cdots, f(X_n)) : f \in F\}$ and $U_n(F^{(\mu)}) := \mathbb{E}\gamma_2(\widetilde{P_\sigma F^{(\mu)}}, \ell_\infty^n)^2$

where $\gamma_2(T,d) := \inf_{(T_s)} \sup_{t \in T} \sum_{s=0}^{\infty} 2^{-s/2} d(t, T_s)$ and $|T_s| \leq 2^{2^s}$ and $\tilde{A} = A - A$ and $F^{(\mu)} := \{f \in F : P\ell_f \leq \mu\}$

Lemma

1. $\mathbb{E}\|P - P_n\|_{(\ell_F)_\mu} \lesssim \sqrt{\mu \frac{U_n(F^{(\mu)})}{n}}$;

An example of computation of the fixed points λ_ϵ^* and μ^*

Computation of λ_ϵ^* and μ^* in the case of the Regression model with quadratic loss :

$$\ell_F := \{\ell_f : (y,x) \longmapsto (y - f(x))^2 : f \in F\}$$

and

$$\mathcal{L}_F := \{\mathcal{L}_f : (y,x) \longmapsto (y - f(x))^2 - (y - f_F^*(x))^2 : f \in F\}.$$

Complexity measure of F :

$P_\sigma F := \{(f(X_1), \cdots, f(X_n)) : f \in F\}$ and $U_n(F^{(\mu)}) := \mathbb{E}\gamma_2(\widetilde{P_\sigma F^{(\mu)}}, \ell_\infty^n)^2$

where $\gamma_2(T,d) := \inf_{(T_s)} \sup_{t \in T} \sum_{s=0}^\infty 2^{-s/2} d(t, T_s)$ and $|T_s| \leq 2^{2^s}$ and $\tilde{A} = A - A$ and $F^{(\mu)} := \{f \in F : P\ell_f \leq \mu\}$

Lemma

1. $\mathbb{E}\|P - P_n\|_{(\ell_F)_\mu} \lesssim \sqrt{\mu \frac{U_n(F^{(\mu)})}{n}}$;
2. $\mathbb{E}\|P - P_n\|_{(\mathcal{L}_F)_\mu} \lesssim \sqrt{(\mu + R^*)\frac{U_n(F^{(\mu)})}{n}}$ *where* $R^* = \inf_{f \in F} R(f)$.

An example of computation of the fixed points λ_ϵ^* and μ^*

Then combining

1. $\mathbb{E}\|P - P_n\|_{V(H)_\lambda} \leq \sum_{i=0}^{\infty} 2^{-i}\mathbb{E}\|P - P_n\|_{H_{2^{i+1}\lambda}}$ for $H = \ell_F, \mathcal{L}_F$

An example of computation of the fixed points λ_ϵ^* and μ^*

Then combining

1. $\mathbb{E}\|P-P_n\|_{V(H)_\lambda} \leq \sum_{i=0}^\infty 2^{-i}\mathbb{E}\|P-P_n\|_{H_{2^{i+1}\lambda}}$ for $H=\ell_F,\mathcal{L}_F$
2. $\mathbb{E}\|P-P_n\|_{(\ell_F)_\mu} \lesssim \sqrt{\mu\frac{U_n(F^{(\mu)})}{n}}$ and
   $\mathbb{E}\|P-P_n\|_{(\mathcal{L}_F)_\mu} \lesssim \sqrt{(\mu+R^*)\frac{U_n(F^{(\mu)})}{n}}$

An example of computation of the fixed points λ_ϵ^* and μ^*

Then combining

1. $\mathbb{E}\|P - P_n\|_{V(H)_\lambda} \leq \sum_{i=0}^\infty 2^{-i}\mathbb{E}\|P - P_n\|_{H_{2^{i+1}\lambda}}$ for $H = \ell_F, \mathcal{L}_F$
2. $\mathbb{E}\|P - P_n\|_{(\ell_F)_\mu} \lesssim \sqrt{\mu \frac{U_n(F^{(\mu)})}{n}}$ and
   $\mathbb{E}\|P - P_n\|_{(\mathcal{L}_F)_\mu} \lesssim \sqrt{(\mu + R^*)\frac{U_n(F^{(\mu)})}{n}}$

roughly, we obtain

An example of computation of the fixed points λ_ϵ^* and μ^*

Then combining

1. $\mathbb{E}\|P - P_n\|_{V(H)_\lambda} \leq \sum_{i=0}^\infty 2^{-i}\mathbb{E}\|P - P_n\|_{H_{2^{i+1}\lambda}}$ for $H = \ell_F, \mathcal{L}_F$
2. $\mathbb{E}\|P - P_n\|_{(\ell_F)_\mu} \lesssim \sqrt{\mu\frac{U_n(F^{(\mu)})}{n}}$ and
   $\mathbb{E}\|P - P_n\|_{(\mathcal{L}_F)_\mu} \lesssim \sqrt{(\mu + R^*)\frac{U_n(F^{(\mu)})}{n}}$

roughly, we obtain

1. $\lambda_\epsilon^* \lesssim U_n(F^{(\lambda_\epsilon^*)})/(\epsilon n)$;
2. $\mu^* \lesssim \sqrt{U_n(F^{(\mu^*)})/n}$.

An example of computation of the fixed points λ_ϵ^* and μ^*

Then combining

1. $\mathbb{E}\|P - P_n\|_{V(H)_\lambda} \leq \sum_{i=0}^{\infty} 2^{-i}\mathbb{E}\|P - P_n\|_{H_{2^{i+1}\lambda}}$ for $H = \ell_F, \mathcal{L}_F$
2. $\mathbb{E}\|P - P_n\|_{(\ell_F)_\mu} \lesssim \sqrt{\mu \frac{U_n(F^{(\mu)})}{n}}$ and
   $\mathbb{E}\|P - P_n\|_{(\mathcal{L}_F)_\mu} \lesssim \sqrt{(\mu + R^*)\frac{U_n(F^{(\mu)})}{n}}$

roughly, we obtain

1. $\lambda_\epsilon^* \lesssim U_n(F^{(\lambda_\epsilon^*)})/(\epsilon n)$;
2. $\mu^* \lesssim \sqrt{U_n(F^{(\mu^*)})/n}$.

Because $R^* = \inf_{f\in F} R(f) \neq 0$ in general, λ_ϵ^* will be the square of μ^* (of course in some particular cases, we can obtain fast rates for exact oracle inequalities).

From this point of view, the differences between exact and non-exact oracle inequalities have two sources :

1. The geometry of F is very important for Exact-oracle inequalities and has no particular effects on non-exact oracle inequality : Bernstein condition ;

From this point of view, the differences between exact and non-exact oracle inequalities have two sources :

1. The geometry of F is very important for Exact-oracle inequalities and has no particular effects on non-exact oracle inequality : Bernstein condition ;
2. The complexities of $V(\mathcal{L}_F)_\lambda$ and $V(\ell_F)_\lambda$ are very different.

Applications to classification

Classification model

- $(X_1, Y_1), \dots, (X_n, Y_n)$: n i.i.d.$\sim (X, Y)$ random variables in $\mathcal{X} \times \{0, 1\}$

Classification model

- $(X_1, Y_1), \dots, (X_n, Y_n)$: n i.i.d.$\sim (X, Y)$ random variables in $\mathcal{X} \times \{0, 1\}$

- $\ell : (f, (x, y)) \longmapsto \mathbb{1}_{f(x) \neq y}$: $0 - 1$-loss function of $f : \mathcal{X} \to \{0, 1\}$

Classification model

- $(X_1, Y_1), \dots, (X_n, Y_n)$: n i.i.d.$\sim (X, Y)$ random variables in $\mathcal{X} \times \{0, 1\}$
- $\ell : (f, (x, y)) \longmapsto \mathbb{I}_{f(x) \neq y}$: $0-1$-loss function of $f : \mathcal{X} \to \{0, 1\}$
- $R(f) = \mathbb{P}[f(X) \neq Y]$: risk of f

Classification model

- $(X_1, Y_1), \dots, (X_n, Y_n)$: n i.i.d.$\sim (X, Y)$ random variables in $\mathcal{X} \times \{0, 1\}$
- $\ell : (f, (x, y)) \longmapsto \mathbb{1}_{f(x) \neq y}$: $0 - 1$-loss function of $f : \mathcal{X} \to \{0, 1\}$
- $R(f) = \mathbb{P}[f(X) \neq Y]$: risk of f
- F a class of $\{0, 1\}$-valued functions ; $f_F^* \in \operatorname{argmin}_{f \in F} R(f)$; $f^* \in \operatorname{argmin}_f R(f)$ (Bayes rule).

Classification model

- $(X_1, Y_1), \ldots, (X_n, Y_n)$: n i.i.d.$\sim (X, Y)$ random variables in $\mathcal{X} \times \{0, 1\}$
- $\ell : (f, (x, y)) \longmapsto \mathbb{I}_{f(x) \neq y}$: $0 - 1$-loss function of $f : \mathcal{X} \to \{0, 1\}$
- $R(f) = \mathbb{P}[f(X) \neq Y]$: risk of f
- F a class of $\{0, 1\}$-valued functions ; $f_F^* \in \operatorname{argmin}_{f \in F} R(f)$; $f^* \in \operatorname{argmin}_f R(f)$ (Bayes rule).

$$\mathcal{L}_f = \ell_f - \ell_{f_F^*} \text{ and } \mathcal{E}_f = \ell_f - \ell_{f^*}.$$

Oracle inequalities in classification

The VC dimension of a class F of $\{0,1\}$-valued functions is

$$V = \max\left(N : max_{x_1,\cdots,x_N \in \mathcal{X}} \mathrm{Card}\big\{(f(x_1),\ldots,f(x_N)) : f \in F\big\} = 2^N\right).$$

Oracle inequalities in classification

The VC dimension of a class F of $\{0,1\}$-valued functions is

$$V = \max\Big(N : max_{x_1,\cdots,x_N \in \mathcal{X}} \mathrm{Card}\big\{(f(x_1),\dots,f(x_N)) : f \in F\big\} = 2^N\Big).$$

M.&N. If $P\mathcal{L}_f^2 \leq B\big(P\mathcal{L}_f\big)^\beta, \forall f \in F$ $(0 \leq \beta \leq 1)$ Bernstein condition then $\forall x \geq 1$, w.p. $\geq 1 - 4e^{-x}$,

$$R(\hat{f}_n^{(ERM)}) \leq \inf_{f \in F} R(f) + c_0\Big(\frac{xV \log(enB^{1/\beta}/V)}{n}\Big)^{\frac{1}{2-\beta}}.$$

Oracle inequalities in classification

The VC dimension of a class F of $\{0,1\}$-valued functions is

$$V = \max\Big(N : max_{x_1,\cdots,x_N\in\mathcal{X}}\mathrm{Card}\big\{(f(x_1),\ldots,f(x_N)) : f\in F\big\} = 2^N\Big).$$

M.&N. If $P\mathcal{L}_f^2 \le B\big(P\mathcal{L}_f\big)^\beta, \forall f\in F$ $(0\le\beta\le 1)$ Bernstein condition then $\forall x\ge 1$, w.p. $\ge 1-4e^{-x}$,

$$R(\hat{f}_n^{(ERM)}) \le \inf_{f\in F} R(f) + c_0\Big(\frac{xV\log(enB^{1/\beta}/V)}{n}\Big)^{\frac{1}{2-\beta}}.$$

M.&N. If $P\mathcal{E}_f^2 \le B\big(P\mathcal{E}_f\big)^\beta, \forall f\in F$ $(0\le\beta\le 1)$ Margin assumption then $\forall x\ge 1$, w.p. $\ge 1-4e^{-x}$,

$$R(\hat{f}_n^{(ERM)})-R(f^*) \le (1+\epsilon)\inf_{f\in F}\big(R(f)-R(f^*)\big)+c_0\Big(\frac{xV\log(enB^{1/\beta}/V)}{n\epsilon}\Big)^{\frac{1}{2-\beta}}.$$

Oracle inequalities in classification

The VC dimension of a class F of $\{0,1\}$-valued functions is

$$V = \max\Big(N : max_{x_1,\cdots,x_N\in\mathcal{X}}\mathrm{Card}\big\{(f(x_1),\dots,f(x_N)) : f\in F\big\} = 2^N\Big).$$

M.&N. If $P\mathcal{L}_f^2 \leq B\big(P\mathcal{L}_f\big)^\beta, \forall f\in F$ $(0\leq\beta\leq 1)$ Bernstein condition then $\forall x\geq 1$, w.p. $\geq 1-4e^{-x}$,

$$R(\hat{f}_n^{(ERM)}) \leq \inf_{f\in F} R(f) + c_0\Big(\frac{xV\log(enB^{1/\beta}/V)}{n}\Big)^{\frac{1}{2-\beta}}.$$

M.&N. If $P\mathcal{E}_f^2 \leq B\big(P\mathcal{E}_f\big)^\beta, \forall f\in F$ $(0\leq\beta\leq 1)$ Margin assumption then $\forall x\geq 1$, w.p. $\geq 1-4e^{-x}$,

$$R(\hat{f}_n^{(ERM)}) - R(f^*) \leq (1+\epsilon)\inf_{f\in F}\big(R(f)-R(f^*)\big) + c_0\Big(\frac{xV\log(enB^{1/\beta}/V)}{n\epsilon}\Big)^{\frac{1}{2-\beta}}.$$

L. Since $P\ell_f^2 \leq BP\ell_f, \forall f\in F$ is always true then $\forall x\geq 1$, w.p. $\geq 1-4e^{-x}$,

$$R(\hat{f}_n^{(ERM)}) \leq (1+\epsilon)\inf_{f\in F} R(f) + c_0\frac{xV\log(enB^{1/\beta}/V)}{n\epsilon}.$$

The Margin-Bernstein conditions

1. for exact oracle inequalities : $P\mathcal{L}_f^2 \leq B\big(P\mathcal{L}_f\big)^\beta, \forall f \in F$ $(0 \leq \beta \leq 1)$.

The Margin-Bernstein conditions

1. for exact oracle inequalities : $P\mathcal{L}_f^2 \leq B\big(P\mathcal{L}_f\big)^\beta, \forall f \in F$ $(0 \leq \beta \leq 1)$.

$$\mathbb{E}(\ell_f - \ell_{f_F^*})^2 \leq B\big(\mathbb{E}(\ell_f - \ell_{f_F^*})\big)^\beta.$$

(hard to characterize from a geometrical point of view because the loss is not convex).

The Margin-Bernstein conditions

1. for exact oracle inequalities : $P\mathcal{L}_f^2 \leq B\big(P\mathcal{L}_f\big)^\beta, \forall f \in F$ $(0 \leq \beta \leq 1)$.

$$\mathbb{E}(\ell_f - \ell_{f_F^*})^2 \leq B\big(\mathbb{E}(\ell_f - \ell_{f_F^*})\big)^\beta.$$

   (hard to characterize from a geometrical point of view because the loss is not convex).

2. for non-exact oracle inequalities for the estimation problem : $P\mathcal{E}_f^2 \leq B\big(P\mathcal{E}_f\big)^\beta, \forall f \in F$ $(0 \leq \beta \leq 1)$.

The Margin-Bernstein conditions

1. for exact oracle inequalities : $P\mathcal{L}_f^2 \leq B\big(P\mathcal{L}_f\big)^\beta, \forall f \in F$ $(0 \leq \beta \leq 1)$.

$$\mathbb{E}(\ell_f - \ell_{f_F^*})^2 \leq B\big(\mathbb{E}(\ell_f - \ell_{f_F^*})\big)^\beta.$$

(hard to characterize from a geometrical point of view because the loss is not convex).

2. for non-exact oracle inequalities for the estimation problem : $P\mathcal{E}_f^2 \leq B\big(P\mathcal{E}_f\big)^\beta, \forall f \in F$ $(0 \leq \beta \leq 1)$.

$$\mathbb{E}(\ell_f - \ell_{f^*})^2 \leq B\big(\mathbb{E}(\ell_f - \ell_{f^*})\big)^\beta.$$

Statistical condition on the model :
$P\mathcal{E}_f^2 \leq B\big(P\mathcal{E}_f\big), \forall f \Leftrightarrow \exists c > 0, \mathbb{P}[|f^*(X) - 1/2| \geq c] = 1$ (where $f^*(X) = \mathbb{E}[Y|X] = \mathbb{P}[Y = 1|X]$).

The Margin-Bernstein conditions

1. for exact oracle inequalities : $P\mathcal{L}_f^2 \leq B\big(P\mathcal{L}_f\big)^\beta, \forall f \in F \ (0 \leq \beta \leq 1)$.

$$\mathbb{E}(\ell_f - \ell_{f_F^*})^2 \leq B\big(\mathbb{E}(\ell_f - \ell_{f_F^*})\big)^\beta.$$

(hard to characterize from a geometrical point of view because the loss is not convex).

2. for non-exact oracle inequalities for the estimation problem : $P\mathcal{E}_f^2 \leq B\big(P\mathcal{E}_f\big)^\beta, \forall f \in F \ (0 \leq \beta \leq 1)$.

$$\mathbb{E}(\ell_f - \ell_{f^*})^2 \leq B\big(\mathbb{E}(\ell_f - \ell_{f^*})\big)^\beta.$$

Statistical condition on the model :
$P\mathcal{E}_f^2 \leq B\big(P\mathcal{E}_f\big), \forall f \Leftrightarrow \exists c > 0, \mathbb{P}[|f^*(X) - 1/2| \geq c] = 1$ (where $f^*(X) = \mathbb{E}[Y|X] = \mathbb{P}[Y = 1|X]$).

3. for non-exact oracle inequalities : $P\ell_f^2 = P\ell_f \leq BP\ell_f, \forall f$.

Oracle inequalities for regularized ERM

Regularized Empirical risk minimization - Part 1

A problem in learning theory is given by

Regularized Empirical risk minimization - Part 1

A problem in learning theory is given by

1. Observations : $Z_1, \ldots, Z_n$: n i.i.d.$\sim Z$ random variables in $\mathcal{Z}$;

Regularized Empirical risk minimization - Part 1

A problem in learning theory is given by

1. Observations : $Z_1, \ldots, Z_n$: n i.i.d.$\sim Z$ random variables in $\mathcal{Z}$;
2. Loss function : $\ell : (f, z) \longmapsto \ell_f(z) \in \mathbb{R}$;

Regularized Empirical risk minimization - Part 1

A problem in learning theory is given by

1. Observations : $Z_1, \ldots, Z_n$: n i.i.d.$\sim Z$ random variables in $\mathcal{Z}$;
2. Loss function : $\ell : (f, z) \longmapsto \ell_f(z) \in \mathbb{R}$;
3. model : $F \subset L_2(P_Z)$.

Regularized Empirical risk minimization - Part 1

A problem in learning theory is given by

1. Observations : $Z_1, \ldots, Z_n$: n i.i.d.$\sim Z$ random variables in $\mathcal{Z}$;
2. Loss function : $\ell : (f, z) \longmapsto \ell_f(z) \in \mathbb{R}$;
3. model : $F \subset L_2(P_Z)$.

Choosing a particular F means that we believe that an oracle f_F^* in F ($R(f_F^*) = \min_{f \in F} R(f)$) is close to the best element f^* minimizing the risk $\min_f R(f)$ (over $L_2(P_Z)$ or other large class of functions).

Regularized Empirical risk minimization - Part 1

A problem in learning theory is given by

1. Observations : $Z_1, \ldots, Z_n$: n i.i.d.$\sim Z$ random variables in $\mathcal{Z}$;
2. Loss function : $\ell : (f, z) \longmapsto \ell_f(z) \in \mathbb{R}$;
3. model : $F \subset L_2(P_Z)$.

Choosing a particular F means that we believe that an oracle f_F^* in F ($R(f_F^*) = \min_{f \in F} R(f)$) is close to the best element f^* minimizing the risk $\min_f R(f)$ (over $L_2(P_Z)$ or other large class of functions).

Example in regression : when we construct

$$\hat{f}_n^{ERM} \in \operatorname{Arg}\min_{f \in F} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2,$$

Regularized Empirical risk minimization - Part 1

A problem in learning theory is given by

1. Observations : $Z_1, \ldots, Z_n$: n i.i.d.$\sim Z$ random variables in $\mathcal{Z}$;
2. Loss function : $\ell : (f, z) \longmapsto \ell_f(z) \in \mathbb{R}$;
3. model : $F \subset L_2(P_Z)$.

Choosing a particular F means that we believe that an oracle f_F^* in F ($R(f_F^*) = \min_{f \in F} R(f)$) is close to the best element f^* minimizing the risk $\min_f R(f)$ (over $L_2(P_Z)$ or other large class of functions).

Example in regression : when we construct

$$\hat{f}_n^{ERM} \in \operatorname{Arg}\min_{f \in F} \frac{1}{n} \sum_{i=1}^n (Y_i - f(X_i))^2,$$

we hope that F will be chosen in such a way that $\hat{f}_n^{ERM}$ will be close to the oracle

$$f_F^* \in \operatorname{Arg}\min_{f \in F} \mathbb{E}(Y - f(X))^2$$

Regularized Empirical risk minimization - Part 1

A problem in learning theory is given by

1. Observations : $Z_1, \ldots, Z_n$: n i.i.d.$\sim Z$ random variables in $\mathcal{Z}$;
2. Loss function : $\ell : (f, z) \longmapsto \ell_f(z) \in \mathbb{R}$;
3. model : $F \subset L_2(P_Z)$.

Choosing a particular F means that we believe that an oracle f_F^* in F ($R(f_F^*) = \min_{f \in F} R(f)$) is close to the best element f^* minimizing the risk $\min_f R(f)$ (over $L_2(P_Z)$ or other large class of functions).

Example in regression : when we construct

$$\hat{f}_n^{ERM} \in \operatorname{Arg}\min_{f \in F} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2,$$

we hope that F will be chosen in such a way that $\hat{f}_n^{ERM}$ will be close to the oracle

$$f_F^* \in \operatorname{Arg}\min_{f \in F} \mathbb{E}(Y - f(X))^2$$

(=> Oracle inequalities)

Regularized Empirical risk minimization - Part 1

A problem in learning theory is given by

1. Observations : $Z_1, \dots, Z_n$: n i.i.d.$\sim Z$ random variables in $\mathcal{Z}$;
2. Loss function : $\ell : (f, z) \longmapsto \ell_f(z) \in \mathbb{R}$;
3. model : $F \subset L_2(P_Z)$.

Choosing a particular F means that we believe that an oracle f_F^* in F ($R(f_F^*) = \min_{f \in F} R(f)$) is close to the best element f^* minimizing the risk $\min_f R(f)$ (over $L_2(P_Z)$ or other large class of functions).

Example in regression : when we construct

$$\hat{f}_n^{ERM} \in \operatorname{Arg}\min_{f \in F} \frac{1}{n}\sum_{i=1}^n (Y_i - f(X_i))^2,$$

we hope that F will be chosen in such a way that $\hat{f}_n^{ERM}$ will be close to the oracle

$$f_F^* \in \operatorname{Arg}\min_{f \in F} \mathbb{E}(Y - f(X))^2$$

(=> Oracle inequalities) And, we hope that f_F^* will be close to the regression function f^* :

$$f^* \in \operatorname{Arg}\min_{f \in L^2(P_X)} \mathbb{E}(Y - f(X))^2.$$

Regularized Empirical risk minimization - Part 2

Idea : By choosing F, it is implicitly said that we believe that f^* has some properties so that f^* is close to F.

Regularized Empirical risk minimization - Part 2

Idea : By choosing F, it is implicitly said that we believe that f^* has some properties so that f^* is close to F. But, for a given property on f^* (for instance, smoothness or low-dimensional structure), it is not always possible to construct a class F (with a "reasonable complexity") so that, thanks to this property, f^* will be close to F.

Regularized Empirical risk minimization - Part 2

Idea : By choosing F, it is implicitly said that we believe that f^* has some properties so that f^* is close to F. But, for a given property on f^* (for instance, smoothness or low-dimensional structure), it is not always possible to construct a class F (with a "reasonable complexity") so that, thanks to this property, f^* will be close to F. In this situation, it is common to introduce a function

$$\mathrm{crit} : \mathcal{F} \subset L_2(P_Z) \longmapsto \mathbb{R}$$

called a criterion. So that

$$\mathrm{crit}(f) \text{ is small } \Rightarrow f \text{ has this property.}$$

Regularized Empirical risk minimization - Part 2

Idea : By choosing F, it is implicitly said that we believe that f^* has some properties so that f^* is close to F. But, for a given property on f^* (for instance, smoothness or low-dimensional structure), it is not always possible to construct a class F (with a "reasonable complexity") so that, thanks to this property, f^* will be close to F. In this situation, it is common to introduce a function

$$\mathrm{crit} : \mathcal{F} \subset L_2(P_Z) \longmapsto \mathbb{R}$$

called a criterion. So that

$$\mathrm{crit}(f) \text{ is small } \Rightarrow f \text{ has this property.}$$

Ex.1 : $\mathrm{crit}(f) = \int (f')^2$;

Regularized Empirical risk minimization - Part 2

Idea : By choosing F, it is implicitly said that we believe that f^* has some properties so that f^* is close to F. But, for a given property on f^* (for instance, smoothness or low-dimensional structure), it is not always possible to construct a class F (with a “reasonable complexity”) so that, thanks to this property, f^* will be close to F. In this situation, it is common to introduce a function

$$\mathrm{crit} : \mathcal{F} \subset L_2(P_Z) \longmapsto \mathbb{R}$$

called a criterion. So that

$$\mathrm{crit}(f) \text{ is small } \Rightarrow f \text{ has this property.}$$

Ex.1 : $\mathrm{crit}(f) = \int (f')^2$; $\mathrm{crit}(f)$ small $\Rightarrow f$ is smooth.

Regularized Empirical risk minimization - Part 2

Idea : By choosing F, it is implicitly said that we believe that f^* has some properties so that f^* is close to F. But, for a given property on f^* (for instance, smoothness or low-dimensional structure), it is not always possible to construct a class F (with a "reasonable complexity") so that, thanks to this property, f^* will be close to F. In this situation, it is common to introduce a function

$$\mathrm{crit} : \mathcal{F} \subset L_2(P_Z) \longmapsto \mathbb{R}$$

called a criterion. So that

$$\mathrm{crit}(f) \text{ is small } \Rightarrow f \text{ has this property.}$$

Ex.1 : $\mathrm{crit}(f) = \int (f')^2$; $\mathrm{crit}(f)$ small $\Rightarrow f$ is smooth.
Ex.2 : $\mathcal{F} := \{f_\beta = \langle \cdot, \beta \rangle : \beta \in \mathbb{R}^d\}$ and $\mathrm{crit}(f_\beta) = |\mathrm{Supp}(\beta)|$;

Regularized Empirical risk minimization - Part 2

Idea : By choosing F, it is implicitly said that we believe that f^* has some properties so that f^* is close to F. But, for a given property on f^* (for instance, smoothness or low-dimensional structure), it is not always possible to construct a class F (with a "reasonable complexity") so that, thanks to this property, f^* will be close to F. In this situation, it is common to introduce a function

$$\mathrm{crit} : \mathcal{F} \subset L_2(P_Z) \longmapsto \mathbb{R}$$

called a criterion. So that

$$\mathrm{crit}(f) \text{ is small } \Rightarrow f \text{ has this property.}$$

Ex.1 : $\mathrm{crit}(f) = \int (f')^2$; $\mathrm{crit}(f)$ small $\Rightarrow f$ is smooth.
Ex.2 : $\mathcal{F} := \{f_\beta = \langle \cdot, \beta \rangle : \beta \in \mathbb{R}^d\}$ and $\mathrm{crit}(f_\beta) = |\mathrm{Supp}(\beta)|$;
$\mathrm{crit}(f_\beta)$ small $\Rightarrow f_\beta$ has a low-dimensional structure.

Regularized Empirical risk minimization procedure - Part 3

Model :

- $Z_1, \ldots, Z_n$: n i.i.d.$\sim Z$ random variables in $\mathcal{Z}$ (observations) ;

Regularized Empirical risk minimization procedure - Part 3

Model :

- $Z_1, \ldots, Z_n$: n i.i.d.$\sim Z$ random variables in $\mathcal{Z}$ (observations) ;
- $\ell : (f, z) \longmapsto \ell_f(z) \in \mathbb{R}$: a loss function

Regularized Empirical risk minimization procedure - Part 3

Model :

- $Z_1, \ldots, Z_n$: n i.i.d.$\sim Z$ random variables in $\mathcal{Z}$ (observations) ;
- $\ell : (f, z) \longmapsto \ell_f(z) \in \mathbb{R}$: a loss function
- $\mathcal{F}$ and $\mathrm{crit} : \mathcal{F} \longmapsto \mathbb{R}$

Regularized Empirical risk minimization procedure - Part 3

Model :

- $Z_1, \ldots, Z_n$: n i.i.d.$\sim Z$ random variables in $\mathcal{Z}$ (observations) ;
- $\ell : (f, z) \longmapsto \ell_f(z) \in \mathbb{R}$: a loss function
- $\mathcal{F}$ and $\mathrm{crit} : \mathcal{F} \longmapsto \mathbb{R}$

Aim : We want to construct $\hat{f}_n$ having a small criterion and having a good empirical behaviour :

Regularized Empirical risk minimization procedure - Part 3

Model :

- $Z_1, \dots, Z_n$: n i.i.d.$\sim Z$ random variables in $\mathcal{Z}$ (observations) ;
- $\ell : (f, z) \longmapsto \ell_f(z) \in \mathbb{R}$: a loss function
- $\mathcal{F}$ and $\mathrm{crit} : \mathcal{F} \longmapsto \mathbb{R}$

Aim : We want to construct $\hat{f}_n$ having a small criterion and having a good empirical behaviour : Regularized Empirical Risk Minimization (RERM) :

$$\hat{f}_n^{RERM} \in \mathrm{Arg}\min_{f\in\mathcal{F}} \big(R_n(f) + \mathrm{reg}(f)\big),$$

(for instance, $\mathrm{reg}(f) = \lambda \mathrm{crit}^{\alpha}(f)$;

Regularized Empirical risk minimization procedure - Part 3

Model :

- $Z_1, \ldots, Z_n$: n i.i.d.$\sim Z$ random variables in $\mathcal{Z}$ (observations) ;
- $\ell : (f, z) \longmapsto \ell_f(z) \in \mathbb{R}$: a loss function
- $\mathcal{F}$ and $\mathrm{crit} : \mathcal{F} \longmapsto \mathbb{R}$

Aim : We want to construct $\hat{f}_n$ having a small criterion and having a good empirical behaviour : Regularized Empirical Risk Minimization (RERM) :

$$\hat{f}_n^{RERM} \in \operatorname{Arg}\min_{f \in \mathcal{F}} \big(R_n(f) + \mathrm{reg}(f)\big),$$

(for instance, $\mathrm{reg}(f) = \lambda \mathrm{crit}^{\alpha}(f)$; λ (regularization parameter), α : parameters to be chosen).

Regularized Empirical risk minimization procedure - Part 3

Model :

- $Z_1, \dots, Z_n$: n i.i.d.$\sim Z$ random variables in $\mathcal{Z}$ (observations) ;
- $\ell : (f, z) \longmapsto \ell_f(z) \in \mathbb{R}$: a loss function
- $\mathcal{F}$ and $\mathrm{crit} : \mathcal{F} \longmapsto \mathbb{R}$

Aim : We want to construct $\hat{f}_n$ having a small criterion and having a good empirical behaviour : Regularized Empirical Risk Minimization (RERM) :

$$\hat{f}_n^{RERM} \in \mathrm{Arg}\min_{f \in \mathcal{F}} \big(R_n(f) + \mathrm{reg}(f) \big),$$

(for instance, $\mathrm{reg}(f) = \lambda \mathrm{crit}^{\alpha}(f)$; λ (regularization parameter), α : parameters to be chosen). We hope that w.h.p.

$$R(\hat{f}_n^{RERM}) + \mathrm{reg}(\hat{f}_n^{RERM}) \leq (1 + \epsilon) \min_{f \in \mathcal{F}} \big(R(f) + \mathrm{reg}(f) \big).$$

Regularized Empirical risk minimization procedure - Part 3

Model :

- $Z_1, \ldots, Z_n$: n i.i.d.$\sim Z$ random variables in $\mathcal{Z}$ (observations) ;
- $\ell : (f, z) \longmapsto \ell_f(z) \in \mathbb{R}$: a loss function
- $\mathcal{F}$ and $\mathrm{crit} : \mathcal{F} \longmapsto \mathbb{R}$

Aim : We want to construct $\hat{f}_n$ having a small criterion and having a good empirical behaviour : Regularized Empirical Risk Minimization (RERM) :

$$\hat{f}_n^{RERM} \in \mathrm{Arg}\min_{f \in \mathcal{F}} \big(R_n(f) + \mathrm{reg}(f)\big),$$

(for instance, $\mathrm{reg}(f) = \lambda \mathrm{crit}^\alpha(f)$; λ (regularization parameter), α : parameters to be chosen). We hope that w.h.p.

$$R(\hat{f}_n^{RERM}) + \mathrm{reg}(\hat{f}_n^{RERM}) \leq (1+\epsilon) \min_{f \in \mathcal{F}} \big(R(f) + \mathrm{reg}(f)\big).$$

1. $\epsilon = 0$: Exact oracle inequality ;

Regularized Empirical risk minimization procedure - Part 3

Model :

- $Z_1, \dots, Z_n$: n i.i.d.$\sim Z$ random variables in $\mathcal{Z}$ (observations) ;
- $\ell : (f, z) \longmapsto \ell_f(z) \in \mathbb{R}$: a loss function
- $\mathcal{F}$ and $\mathrm{crit} : \mathcal{F} \longmapsto \mathbb{R}$

Aim : We want to construct $\hat{f}_n$ having a small criterion and having a good empirical behaviour : Regularized Empirical Risk Minimization (RERM) :

$$\hat{f}_n^{RERM} \in \mathrm{Arg}\min_{f \in \mathcal{F}} \big(R_n(f) + \mathrm{reg}(f)\big),$$

(for instance, $\mathrm{reg}(f) = \lambda \mathrm{crit}^{\alpha}(f)$; λ (regularization parameter), α : parameters to be chosen). We hope that w.h.p.

$$R(\hat{f}_n^{RERM}) + \mathrm{reg}(\hat{f}_n^{RERM}) \leq (1+\epsilon) \min_{f \in \mathcal{F}} \big(R(f) + \mathrm{reg}(f)\big).$$

1. $\epsilon = 0$: Exact oracle inequality ;
2. $\epsilon > 0$: Non-exact oracle inequality.

Exact and non-exact oracle inequalities for RERM - Part 1

The choice of the regularizing function $\mathrm{reg}(f) = \lambda \mathrm{crit}^{\alpha}(f)$ is dictated by the complexity of the sequence of models $(F_r)_{r \geq 0}$ where

$$F_r := \{f \in \mathcal{F} : \mathrm{crit}(f) \leq r\}.$$

Exact and non-exact oracle inequalities for RERM - Part 1

The choice of the regularizing function $\mathrm{reg}(f) = \lambda \mathrm{crit}^{\alpha}(f)$ is dictated by the complexity of the sequence of models $(F_r)_{r\geq 0}$ where

$$F_r := \{f \in \mathcal{F} : \mathrm{crit}(f) \leq r\}.$$

For every $r \geq 0$:

- loss functions classes :

$$\ell_{F_r} := \{\ell_f : f \in F_r\} \text{ and } \mathbb{E}\|P_n - P\|_{V(\ell_{F_r})_{\lambda_\epsilon^*(r)}} \leq (\epsilon/4)\lambda_\epsilon^*(r)$$

Exact and non-exact oracle inequalities for RERM - Part 1

The choice of the regularizing function $\mathrm{reg}(f) = \lambda \mathrm{crit}^{\alpha}(f)$ is dictated by the complexity of the sequence of models $(F_r)_{r\geq 0}$ where

$$F_r := \{f \in \mathcal{F} : \mathrm{crit}(f) \leq r\}.$$

For every $r \geq 0$:

- loss functions classes :

$$\ell_{F_r} := \{\ell_f : f \in F_r\} \text{ and } \mathbb{E}\|P_n - P\|_{V(\ell_{F_r})_{\lambda^*_\epsilon(r)}} \leq (\epsilon/4)\lambda^*_\epsilon(r)$$

- excess loss functions classes :

$$\mathcal{L}_{F_r} := \{\mathcal{L}_{r,f} := \ell_f - \ell_{f^*_{F_r}} : f \in F_r\} \text{ and } \mathbb{E}\|P_n - P\|_{V(\mathcal{L}_{F_r})_{\mu^*(r)}} \leq \mu^*(r)/8$$

(where $R(f^*_{F_r}) = \min_{f\in F_r} R(f)$).

Theorem (L. and Mendelson)

Assume that there are non-decreasing functions ϕ_n and B such that

1. $\|\max_{1 \leq i \leq n} \sup_{f \in F_r} f(Z_i)\|_{\psi_1} := b_n(\ell_{F_r}) \leq \phi_n(r)$

Theorem (L. and Mendelson)

Assume that there are non-decreasing functions ϕ_n and B such that

1. $\| \max_{1\leq i\leq n} \sup_{f\in F_r} f(Z_i)\|_{\psi_1} := b_n(\ell_{F_r}) \leq \phi_n(r)$
2. $P\ell_f^2 \leq B(r)P\ell_f^2 + B^2(r)/n, \forall r \geq 0, f \in F_r.$

Theorem (L. and Mendelson)

Assume that there are non-decreasing functions ϕ_n and B such that

1. $\| \max_{1\leq i\leq n} \sup_{f\in F_r} f(Z_i)\|_{\psi_1} := b_n(\ell_{F_r}) \leq \phi_n(r)$
2. $P\ell_f^2 \leq B(r)P\ell_f^2 + B^2(r)/n, \forall r \geq 0, f \in F_r.$

Let $0 < \epsilon < 1/2$ and assume that for every $(r,x) \in \mathbb{R}_+ \times \mathbb{R}_+^*$,

$$\rho_n(r,x) \geq \max\left(\lambda_\epsilon^*(r), c_0 \frac{(\phi_n(r) + B(r)/\epsilon)(x+1)}{n\epsilon}\right).$$

Theorem (L. and Mendelson)

Assume that there are non-decreasing functions ϕ_n and B such that

1. $\| \max_{1\leq i\leq n} \sup_{f\in F_r} f(Z_i)\|_{\psi_1} := b_n(\ell_{F_r}) \leq \phi_n(r)$
2. $P\ell_f^2 \leq B(r)P\ell_f^2 + B^2(r)/n, \forall r \geq 0, f \in F_r.$

Let $0 < \epsilon < 1/2$ and assume that for every $(r,x) \in \mathbb{R}_+ \times \mathbb{R}_+^*$,

$$\rho_n(r,x) \geq \max\Big(\lambda_\epsilon^*(r), c_0\frac{(\phi_n(r)+B(r)/\epsilon)(x+1)}{n\epsilon}\Big).$$

Let $x > 0$ and set

$$\hat{f}_n^{RERM} \in \mathrm{Arg}\min_{f\in\mathcal{F}}\Big(R_n(f) + \frac{1}{1+\epsilon}\rho_n(\mathrm{crit}(f)+1,x)\Big).$$

Theorem (L. and Mendelson)

Assume that there are non-decreasing functions ϕ_n and B such that

1. $\| \max_{1\leq i\leq n} \sup_{f\in F_r} f(Z_i)\|_{\psi_1} := b_n(\ell_{F_r}) \leq \phi_n(r)$
2. $P\ell_f^2 \leq B(r)P\ell_f^2 + B^2(r)/n, \forall r \geq 0, f \in F_r.$

Let $0 < \epsilon < 1/2$ and assume that for every $(r,x) \in \mathbb{R}_+ \times \mathbb{R}_+^*$,

$$\rho_n(r,x) \geq \max\Big(\lambda_\epsilon^*(r), c_0 \frac{(\phi_n(r) + B(r)/\epsilon)(x+1)}{n\epsilon}\Big).$$

Let $x > 0$ and set

$$\hat{f}_n^{RERM} \in \mathrm{Arg}\min_{f\in\mathcal{F}} \Big(R_n(f) + \frac{1}{1+\epsilon}\rho_n(\mathrm{crit}(f)+1,x)\Big).$$

Then, with probability greater than $1 - 10\exp(-x)$,

$$R(\hat{f}_n^{RERM}) + \rho_n(\mathrm{crit}(\hat{f}_n^{RERM}),x) \leq \inf_{f\in\mathcal{F}} \Big[(1+2\epsilon)R(f) + 2\rho_n(\mathrm{crit}(f)+1,x)\Big].$$

Theorem (L. and Mendelson)

Assume that there are non-decreasing functions ϕ_n and B such that

1. $\| \max_{1\leq i\leq n} \sup_{f\in F_r} f(Z_i)\|_{\psi_1} := b_n(\ell_{F_r}) \leq \phi_n(r)$
2. $P\ell_f{}^2 \leq B(r)P\ell_{r,f}{}^2 + B^2(r)/n, \forall r \geq 0, f \in F_r$.

Let $0 < \epsilon < 1/2$ and assume that for every $(r,x) \in \mathbb{R}_+ \times \mathbb{R}_+^*$,

$$\rho_n(r,x) \geq \max\left(\lambda_\epsilon^*(r), c_0 \frac{(\phi_n(r) + B(r)/\epsilon)(x+1)}{n\epsilon}\right).$$

Let $x > 0$ and set

$$\hat{f}_n^{RERM} \in \mathrm{Arg}\min_{f\in\mathcal{F}}\left(R_n(f) + \frac{1}{1+\epsilon}\rho_n(\mathrm{crit}(f)+1,x)\right).$$

Then, with probability greater than $1 - 10\exp(-x)$,

$$R(\hat{f}_n^{RERM}) + \rho_n(\mathrm{crit}(\hat{f}_n^{RERM}),x) \leq \inf_{f\in\mathcal{F}}\left[(1+2\epsilon)R(f) + 2\rho_n(\mathrm{crit}(f)+1,x)\right].$$

Theorem (Bartlett, Neeman and Mendelson)

Assume that there are non-decreasing functions ϕ_n and B such that

1. $\| \max_{1\leq i\leq n} \sup_{f\in F_r} f(Z_i)\|_{\psi_1} := b_n(\ell_{F_r}) \leq \phi_n(r)$
2. $P\mathcal{L}_f{}^2 \leq B(r)P\mathcal{L}_{r,f}{}^2 + B^2(r)/n, \forall r \geq 0, f \in F_r$.

Let $0 < \epsilon < 1/2$ and assume that for every $(r,x) \in \mathbb{R}_+ \times \mathbb{R}_+^*$,

$$\rho_n(r,x) \geq \max\Big(\mu^*(r), c_0 \frac{(\phi_n(r) + B(r)/\epsilon)(x+1)}{n\epsilon}\Big).$$

Let $x > 0$ and set

$$\hat{f}_n^{RERM} \in \operatorname{Arg}\min_{f\in\mathcal{F}} \Big(R_n(f) + \frac{1}{1+\epsilon}\rho_n(\mathrm{crit}(f)+1,x)\Big).$$

Then, with probability greater than $1 - 10\exp(-x)$,

$$R(\hat{f}_n^{RERM}) + \rho_n(\mathrm{crit}(\hat{f}_n^{RERM}),x) \leq \inf_{f\in\mathcal{F}} \Big[1\times R(f) + 2\rho_n(\mathrm{crit}(f)+1,x)\Big].$$

Conclusion on Exact and Non-exact oracle inequalities for RERM

We are given $\mathcal{F}$ and $\mathrm{crit} : \mathcal{F} \longmapsto \mathbb{R}$. We consider the models $(F_r)_{r\geq 0}$:

$$F_r := \{f \in \mathcal{F} : \mathrm{crit}(f) \leq r\}.$$

Conclusion on Exact and Non-exact oracle inequalities for RERM

We are given $\mathcal{F}$ and $\mathrm{crit} : \mathcal{F} \longmapsto \mathbb{R}$. We consider the models $(F_r)_{r \geq 0}$:

$$F_r := \{f \in \mathcal{F} : \mathrm{crit}(f) \leq r\}.$$

- loss functions classes : for all $r > 0$,

$$\ell_{F_r} := \{\ell_f : f \in F_r\} \text{ and } \mathbb{E}\|P_n - P\|_{V(\ell_{F_r})_{\lambda^*_\epsilon(r)}} \leq (\epsilon/4)\lambda^*_\epsilon(r)$$

Conclusion on Exact and Non-exact oracle inequalities for RERM

We are given $\mathcal{F}$ and $\mathrm{crit} : \mathcal{F} \longmapsto \mathbb{R}$. We consider the models $(F_r)_{r \geq 0}$:

$$F_r := \{f \in \mathcal{F} : \mathrm{crit}(f) \leq r\}.$$

- loss functions classes : for all $r > 0$,

$$\ell_{F_r} := \{\ell_f : f \in F_r\} \text{ and } \mathbb{E}\|P_n - P\|_{V(\ell_{F_r})_{\lambda^*_\epsilon(r)}} \leq (\epsilon/4)\lambda^*_\epsilon(r)$$

- excess loss functions classes : for all $r > 0$,

$$\mathcal{L}_{F_r} := \{\ell_f - \ell_{f^*_{F_r}} : f \in F_r\} \text{ and } \mathbb{E}\|P_n - P\|_{V(\mathcal{L}_{F_r})_{\mu^*(r)}} \leq \mu^*(r)/8.$$

Conclusion on Exact and Non-exact oracle inequalities for RERM

We are given $\mathcal{F}$ and $\mathrm{crit}: \mathcal{F} \longmapsto \mathbb{R}$. We consider the models $(F_r)_{r\geq 0}$:

$$F_r := \{f \in \mathcal{F} : \mathrm{crit}(f) \leq r\}.$$

- loss functions classes : for all $r > 0$,

$$\ell_{F_r} := \{\ell_f : f \in F_r\} \text{ and } \mathbb{E}\|P_n - P\|_{V(\ell_{F_r})_{\lambda^*_\epsilon(r)}} \leq (\epsilon/4)\lambda^*_\epsilon(r)$$

- excess loss functions classes : for all $r > 0$,

$$\mathcal{L}_{F_r} := \{\ell_f - \ell_{f^*_{F_r}} : f \in F_r\} \text{ and } \mathbb{E}\|P_n - P\|_{V(\mathcal{L}_{F_r})_{\mu^*(r)}} \leq \mu^*(r)/8.$$

1. RERM with regularizing function $\mathrm{reg}(f) \gtrsim \lambda^*_\epsilon(\mathrm{crit}(f))$

Conclusion on Exact and Non-exact oracle inequalities for RERM

We are given $\mathcal{F}$ and $\mathrm{crit} : \mathcal{F} \longmapsto \mathbb{R}$. We consider the models $(F_r)_{r\geq 0}$:

$$F_r := \{f \in \mathcal{F} : \mathrm{crit}(f) \leq r\}.$$

- loss functions classes : for all $r > 0$,

$$\ell_{F_r} := \{\ell_f : f \in F_r\} \text{ and } \mathbb{E}\|P_n - P\|_{V(\ell_{F_r})_{\lambda^*_\epsilon(r)}} \leq (\epsilon/4)\lambda^*_\epsilon(r)$$

- excess loss functions classes : for all $r > 0$,

$$\mathcal{L}_{F_r} := \{\ell_f - \ell_{f^*_{F_r}} : f \in F_r\} \text{ and } \mathbb{E}\|P_n - P\|_{V(\mathcal{L}_{F_r})_{\mu^*(r)}} \leq \mu^*(r)/8.$$

1. RERM with regularizing function $\mathrm{reg}(f) \gtrsim \lambda^*_\epsilon(\mathrm{crit}(f)) \Longrightarrow$ Non-exact oracle inequality ;

Conclusion on Exact and Non-exact oracle inequalities for RERM

We are given $\mathcal{F}$ and $\mathrm{crit} : \mathcal{F} \longmapsto \mathbb{R}$. We consider the models $(F_r)_{r\geq 0}$:

$$F_r := \{f \in \mathcal{F} : \mathrm{crit}(f) \leq r\}.$$

- loss functions classes : for all $r > 0$,

$$\ell_{F_r} := \{\ell_f : f \in F_r\} \text{ and } \mathbb{E}\|P_n - P\|_{V(\ell_{F_r})_{\lambda^*_\epsilon(r)}} \leq (\epsilon/4)\lambda^*_\epsilon(r)$$

- excess loss functions classes : for all $r > 0$,

$$\mathcal{L}_{F_r} := \{\ell_f - \ell_{f^*_{F_r}} : f \in F_r\} \text{ and } \mathbb{E}\|P_n - P\|_{V(\mathcal{L}_{F_r})_{\mu^*(r)}} \leq \mu^*(r)/8.$$

1. RERM with regularizing function $\mathrm{reg}(f) \gtrsim \lambda^*_\epsilon(\mathrm{crit}(f)) \Longrightarrow$ Non-exact oracle inequality ;
2. RERM with regularizing function $\mathrm{reg}(f) \gtrsim \mu^*(\mathrm{crit}(f))$

Conclusion on Exact and Non-exact oracle inequalities for RERM

We are given $\mathcal{F}$ and $\mathrm{crit} : \mathcal{F} \longmapsto \mathbb{R}$. We consider the models $(F_r)_{r \geq 0}$:

$$F_r := \{f \in \mathcal{F} : \mathrm{crit}(f) \leq r\}.$$

- loss functions classes : for all $r > 0$,

$$\ell_{F_r} := \{\ell_f : f \in F_r\} \text{ and } \mathbb{E}\|P_n - P\|_{V(\ell_{F_r})_{\lambda_\epsilon^*(r)}} \leq (\epsilon/4)\lambda_\epsilon^*(r)$$

- excess loss functions classes : for all $r > 0$,

$$\mathcal{L}_{F_r} := \{\ell_f - \ell_{f^*_{F_r}} : f \in F_r\} \text{ and } \mathbb{E}\|P_n - P\|_{V(\mathcal{L}_{F_r})_{\mu^*(r)}} \leq \mu^*(r)/8.$$

1. RERM with regularizing function $\mathrm{reg}(f) \gtrsim \lambda_\epsilon^*(\mathrm{crit}(f)) \Longrightarrow$ Non-exact oracle inequality ;
2. RERM with regularizing function $\mathrm{reg}(f) \gtrsim \mu^*(\mathrm{crit}(f)) \Longrightarrow$ exact oracle inequality.

Conclusion on Exact and Non-exact oracle inequalities for RERM

We are given $\mathcal{F}$ and $\mathrm{crit} : \mathcal{F} \longmapsto \mathbb{R}$. We consider the models $(F_r)_{r\geq 0}$:

$$F_r := \{f \in \mathcal{F} : \mathrm{crit}(f) \leq r\}.$$

- loss functions classes : for all $r > 0$,

$$\ell_{F_r} := \{\ell_f : f \in F_r\} \text{ and } \mathbb{E}\|P_n - P\|_{V(\ell_{F_r})_{\lambda^*_\epsilon(r)}} \leq (\epsilon/4)\lambda^*_\epsilon(r)$$

- excess loss functions classes : for all $r > 0$,

$$\mathcal{L}_{F_r} := \{\ell_f - \ell_{f^*_{F_r}} : f \in F_r\} \text{ and } \mathbb{E}\|P_n - P\|_{V(\mathcal{L}_{F_r})_{\mu^*(r)}} \leq \mu^*(r)/8.$$

1. RERM with regularizing function $\mathrm{reg}(f) \gtrsim \lambda^*_\epsilon(\mathrm{crit}(f)) \Longrightarrow$ Non-exact oracle inequality ;
2. RERM with regularizing function $\mathrm{reg}(f) \gtrsim \mu^*(\mathrm{crit}(f)) \Longrightarrow$ exact oracle inequality.

Remark : Usually, we have to regularize more to get an exact oracle inequality than for a non-exact oracle inequality.

Conclusion on Exact and Non-exact oracle inequalities for RERM

We are given $\mathcal{F}$ and $\mathrm{crit} : \mathcal{F} \longmapsto \mathbb{R}$. We consider the models $(F_r)_{r\geq 0}$:

$$F_r := \{f \in \mathcal{F} : \mathrm{crit}(f) \leq r\}.$$

- loss functions classes : for all $r > 0$,

$$\ell_{F_r} := \{\ell_f : f \in F_r\} \text{ and } \mathbb{E}\|P_n - P\|_{V(\ell_{F_r})_{\lambda_\epsilon^*(r)}} \leq (\epsilon/4)\lambda_\epsilon^*(r)$$

- excess loss functions classes : for all $r > 0$,

$$\mathcal{L}_{F_r} := \{\ell_f - \ell_{f^*_{F_r}} : f \in F_r\} \text{ and } \mathbb{E}\|P_n - P\|_{V(\mathcal{L}_{F_r})_{\mu^*(r)}} \leq \mu^*(r)/8.$$

1. RERM with regularizing function $\mathrm{reg}(f) \gtrsim \lambda_\epsilon^*(\mathrm{crit}(f)) \Longrightarrow$ Non-exact oracle inequality ;
2. RERM with regularizing function $\mathrm{reg}(f) \gtrsim \mu^*(\mathrm{crit}(f)) \Longrightarrow$ exact oracle inequality.

Remark : Usually, we have to regularize more to get an exact oracle inequality than for a non-exact oracle inequality.
Ex. : [Bousquet, Blanchard, Massart] : regularization by $\|\cdot\|_{\mathcal{H}}$ or in [Bartlett, Neeman, Mendelson] : regularization by $\log\|\cdot\|_{\mathcal{H}}$ up to $\|\cdot\|_{\mathcal{H}}^2$.

Applications in matrix completion

Example in matrix completion

Model :

- $Z_1 = (Y_1, X_1), \dots, Z_n = (Y_n, X_n)$: n i.i.d. random variables in $\mathbb{R} \times \mathbb{R}^{m \times T}$;

Example in matrix completion

Model :

- $Z_1 = (Y_1, X_1), \dots, Z_n = (Y_n, X_n)$: n i.i.d. random variables in $\mathbb{R} \times \mathbb{R}^{m \times T}$;
- $\ell^{(q)} : \mathbb{R}^{m \times T} \times \mathbb{R} \times \mathbb{R}^{m \times T} \longmapsto \mathbb{R}$ such that $\ell_A^{(q)}(Y, X) = |Y - \langle A, X \rangle|^q$ where $\langle A, X \rangle = \mathrm{Tr}(A^\top X)$ and $q \geq 2$.

Example in matrix completion

Model :

- $Z_1 = (Y_1, X_1), \dots, Z_n = (Y_n, X_n)$: n i.i.d. random variables in $\mathbb{R} \times \mathbb{R}^{m\times T}$;
- $\ell^{(q)} : \mathbb{R}^{m\times T} \times \mathbb{R} \times \mathbb{R}^{m\times T} \longmapsto \mathbb{R}$ such that $\ell_A^{(q)}(Y, X) = |Y - \langle A, X\rangle|^q$ where $\langle A, X\rangle = \mathrm{Tr}(A^\top X)$ and $q \geq 2$.

Notation :

- $\ell_A^{(q)}$: L_q-loss function of a matrix $A \in \mathbb{R}^{m\times T}$

Example in matrix completion

Model :

- $Z_1 = (Y_1, X_1), \dots, Z_n = (Y_n, X_n)$: n i.i.d. random variables in $\mathbb{R} \times \mathbb{R}^{m \times T}$;
- $\ell^{(q)} : \mathbb{R}^{m \times T} \times \mathbb{R} \times \mathbb{R}^{m \times T} \longmapsto \mathbb{R}$ such that $\ell_A^{(q)}(Y, X) = |Y - \langle A, X \rangle|^q$ where $\langle A, X \rangle = \mathrm{Tr}(A^\top X)$ and $q \geq 2$.

Notation :

- $\ell_A^{(q)}$: L_q-loss function of a matrix $A \in \mathbb{R}^{m \times T}$
- $R^{(q)}(A) = \mathbb{E}|Y - \langle A, X \rangle|^q$: L_q-risk of a matrix $A \in \mathbb{R}^{m \times T}$

Example in matrix completion

Model :

- $Z_1 = (Y_1, X_1), \dots, Z_n = (Y_n, X_n)$: n i.i.d. random variables in $\mathbb{R} \times \mathbb{R}^{m \times T}$;
- $\ell^{(q)} : \mathbb{R}^{m \times T} \times \mathbb{R} \times \mathbb{R}^{m \times T} \longmapsto \mathbb{R}$ such that $\ell_A^{(q)}(Y, X) = |Y - \langle A, X \rangle|^q$ where $\langle A, X \rangle = \mathrm{Tr}(A^\top X)$ and $q \geq 2$.

Notation :

- $\ell_A^{(q)}$: L_q-loss function of a matrix $A \in \mathbb{R}^{m \times T}$
- $R^{(q)}(A) = \mathbb{E}|Y - \langle A, X \rangle|^q$: L_q-risk of a matrix $A \in \mathbb{R}^{m \times T}$
- The L_q-risk of a statistic $\hat{f}_n = \langle \cdot, \hat{A}_n \rangle$ is $R^{(q)}(\hat{A}_n) = \mathbb{E}\big[|Y - \langle \hat{A}_n, X \rangle|^q \big| \mathcal{D}\big]$.

Example in matrix completion

Model :

- $Z_1 = (Y_1, X_1), \dots, Z_n = (Y_n, X_n)$: n i.i.d. random variables in $\mathbb{R} \times \mathbb{R}^{m \times T}$;
- $\ell^{(q)} : \mathbb{R}^{m \times T} \times \mathbb{R} \times \mathbb{R}^{m \times T} \longmapsto \mathbb{R}$ such that $\ell_A^{(q)}(Y, X) = |Y - \langle A, X \rangle|^q$ where $\langle A, X \rangle = \mathrm{Tr}(A^\top X)$ and $q \geq 2$.

Notation :

- $\ell_A^{(q)}$: L_q-loss function of a matrix $A \in \mathbb{R}^{m \times T}$
- $R^{(q)}(A) = \mathbb{E}|Y - \langle A, X \rangle|^q$: L_q-risk of a matrix $A \in \mathbb{R}^{m \times T}$
- The L_q-risk of a statistic $\hat{f}_n = \langle \cdot, \hat{A}_n \rangle$ is $R^{(q)}(\hat{A}_n) = \mathbb{E}\big[|Y - \langle \hat{A}_n, X \rangle|^q \big| \mathcal{D}\big]$.

Problem : $mT >> n$ (more variables than observations) but we believe that $Y \approx \langle A_0, X \rangle$ where A_0 is of low rank ($\mathrm{rank}(A_0) < n$) (This is not an assumption !)

Example in matrix completion

Model :

- $Z_1 = (Y_1, X_1), \dots, Z_n = (Y_n, X_n)$: n i.i.d. random variables in $\mathbb{R} \times \mathbb{R}^{m\times T}$;
- $\ell^{(q)} : \mathbb{R}^{m\times T} \times \mathbb{R} \times \mathbb{R}^{m\times T} \longmapsto \mathbb{R}$ such that $\ell_A^{(q)}(Y, X) = |Y - \langle A, X\rangle|^q$ where $\langle A, X\rangle = \mathrm{Tr}(A^\top X)$ and $q \geq 2$.

Notation :

- $\ell_A^{(q)}$: L_q-loss function of a matrix $A \in \mathbb{R}^{m\times T}$
- $R^{(q)}(A) = \mathbb{E}|Y - \langle A, X\rangle|^q$: L_q-risk of a matrix $A \in \mathbb{R}^{m\times T}$
- The L_q-risk of a statistic $\hat{f}_n = \langle \cdot, \hat{A}_n\rangle$ is $R^{(q)}(\hat{A}_n) = \mathbb{E}\big[|Y - \langle \hat{A}_n, X\rangle|^q \big| \mathcal{D}\big]$.

Problem : $mT >> n$ (more variables than observations) but we believe that $Y \approx \langle A_0, X\rangle$ where A_0 is of low rank ($\mathrm{rank}(A_0) < n$) (This is not an assumption !)
$\mathcal{F} := \{\langle \cdot, A\rangle : A \in \mathbb{R}^{m\times T}\}$ and $\mathrm{crit}(A) = \mathrm{rank}(A)$.

Matrix Completion - Convexification

$A \longmapsto \mathrm{rank}(A)$ is not convex $\Longrightarrow$ not possible to use it in practice as a regularizing function.

Matrix Completion - Convexification

$A \longmapsto \mathrm{rank}(A)$ is not convex $\Longrightarrow$ not possible to use it in practice as a regularizing function.

Convexification : The convex envelope of $\mathrm{rank}(\cdot)$ on $\{A \in \mathbb{R}^{m\times T} : \|A\|_{S_\infty} \leq 1\}$ is the nuclear norm ($\|A\|_{S_1} = \|\mathrm{spec}(A)\|_{\ell_1^{m\wedge T}}$).

Matrix Completion - Convexification

$A \longmapsto \mathrm{rank}(A)$ is not convex $\Longrightarrow$ not possible to use it in practice as a regularizing function.

Convexification : The convex envelope of $\mathrm{rank}(\cdot)$ on $\{A \in \mathbb{R}^{m \times T} : \|A\|_{S_\infty} \leq 1\}$ is the nuclear norm ($\|A\|_{S_1} = \|\mathrm{spec}(A)\|_{\ell_1^{m \wedge T}}$).

$\Longrightarrow$ We use the nuclear norm as a criterion : $\mathrm{crit}(A) = \|A\|_{S_1}$.

Matrix Completion - Convexification

$A \longmapsto \mathrm{rank}(A)$ is not convex $\Longrightarrow$ not possible to use it in practice as a regularizing function.

Convexification : The convex envelope of $\mathrm{rank}(\cdot)$ on $\{A \in \mathbb{R}^{m\times T} : \|A\|_{S_\infty} \leq 1\}$ is the nuclear norm $(\|A\|_{S_1} = \|\mathrm{spec}(A)\|_{\ell_1^{m\wedge T}})$.

$\Longrightarrow$ We use the nuclear norm as a criterion : $\mathrm{crit}(A) = \|A\|_{S_1}$.

bibliography :

1. Candés, Tao, Romberg, Plan, Recht, Fazel, Parillo, Gross,... (Exact reconstruction problem : $Y = \langle X, A_0 \rangle$ and often $X \sim \mathrm{Unif}(e_i e_j^\top : 1 \leq i \leq m, 1 \leq j \leq T))$;

Matrix Completion - Convexification

$A \longmapsto \mathrm{rank}(A)$ is not convex $\Longrightarrow$ not possible to use it in practice as a regularizing function.

Convexification : The convex envelope of $\mathrm{rank}(\cdot)$ on $\{A \in \mathbb{R}^{m\times T} : \|A\|_{S_\infty} \leq 1\}$ is the nuclear norm ($\|A\|_{S_1} = \|\mathrm{spec}(A)\|_{\ell_1^{m\wedge T}}$).

$\Longrightarrow$ We use the nuclear norm as a criterion : $\mathrm{crit}(A) = \|A\|_{S_1}$.

bibliography :

1. Candés, Tao, Romberg, Plan, Recht, Fazel, Parillo, Gross,... (Exact reconstruction problem : $Y = \langle X, A_0 \rangle$ and often $X \sim \mathrm{Unif}(e_i e_j^\top : 1 \leq i \leq m, 1 \leq j \leq T)$) ;
2. Tsybakov, Rohde, Koltchinskii, Lounici, Negahban, Wainright, Bach,... (statistical point of view).

Matrix Completion - Application of the general result

$$F_r := \{A \in \mathbb{R}^{m\times T} : \mathrm{crit}(A) \leq r\} = rB_{S_1}$$

Matrix Completion - Application of the general result

$$F_r := \{A \in \mathbb{R}^{m\times T} : \mathrm{crit}(A) \leq r\} = rB_{S_1}$$

For non-exact oracle inequalities for RERM :

$$\lambda_\epsilon^*(r) := \inf\Big(\lambda > 0 : \mathbb{E}\|P - P_n\|_{V(\ell_{F_r}^{(q)})_\lambda} \leq (\epsilon/4)\lambda\Big).$$

where $\ell_{F_r}^{(q)} := \{\ell_A^{(q)} : \|A\|_{S_1} \leq r\}$ and $\ell_A^{(q)}(y,x) = |y - \langle x, A\rangle|^q$.

Matrix Completion - Application of the general result

$$F_r := \{A \in \mathbb{R}^{m\times T} : \mathrm{crit}(A) \leq r\} = rB_{S_1}$$

For non-exact oracle inequalities for RERM :

$$\lambda_\epsilon^*(r) := \inf\Big(\lambda > 0 : \mathbb{E}\|P - P_n\|_{V(\ell_{F_r}^{(q)})_\lambda} \leq (\epsilon/4)\lambda\Big).$$

where $\ell_{F_r}^{(q)} := \{\ell_A^{(q)} : \|A\|_{S_1} \leq r\}$ and $\ell_A^{(q)}(y,x) = |y - \langle x, A\rangle|^q$.

For exact oracle inequalities for RERM :

$$\mu^*(r) := \inf\Big(\mu > 0 : \mathbb{E}\|P - P_n\|_{V(\mathcal{L}_{F_r}^{(q)})_\mu} \leq \mu/8\Big).$$

where $\mathcal{L}_{F_r}^{(q)} = \ell_{F_r}^{(q)} - \ell_{A_r^*}^{(q)}$ and $R^{(q)}(A_r^*) = \min_{A\in F_r} R^{(q)}(A)$.

Computation of the fixed point

Computation of the fixed point

Lemma (L. and Mendelson)

$U_n = \mathbb{E}\gamma_2^2(\widetilde{P_\sigma F}, \ell_\infty^n)$ *where* $P_\sigma F = \{(f(X_1), \cdots, f(X_n)) : f \in F\}$.

Computation of the fixed point

Lemma (L. and Mendelson)

$U_n = \mathbb{E}\gamma_2^2(\widetilde{P_\sigma F}, \ell_\infty^n)$ *where* $P_\sigma F = \{(f(X_1), \cdots, f(X_n)) : f \in F\}$.

q=2 $\mathbb{E}\|P - P_n\|_{(\ell_F^{(2)})_\mu} \leq \max\left[\sqrt{\mu \frac{U_n}{n}}, \frac{U_n}{n}\right]$

Computation of the fixed point

Lemma (L. and Mendelson)

$U_n = \mathbb{E}\gamma_2^2(\widetilde{P_\sigma F}, \ell_\infty^n)$ *where* $P_\sigma F = \{(f(X_1), \cdots, f(X_n)) : f \in F\}$.

q=2 $\mathbb{E}\|P - P_n\|_{(\ell_F^{(2)})_\mu} \leq \max\left[\sqrt{\mu \frac{U_n}{n}}, \frac{U_n}{n}\right]$

q>2 $\mathbb{E}\|P - P_n\|_{(\ell_F^{(q)})_\mu} \leq$

$$\max\left[\sqrt{\mu \frac{U_n}{n}}\sqrt{\left(M \log n\right)^{1-2/q}}, \frac{U_n}{n}\left(M \log n\right)^{1-2/q}, \frac{M \log n}{n}\right]$$

where $M = \|\sup_{\ell \in \ell_F^{(q)}} |\ell|\|_{\psi_1}$.

Theorem (L. and Mendelson)

Assume that $q \geq 2$, $\|Y\|_{\psi_q}, \| \|X\|_{S_2} \|_{\psi_q} \leq K(mT)$ *for some constant* $K(mT)$ *which depends only on the product* mT.

Theorem (L. and Mendelson)

Assume that $q \geq 2$, $\|Y\|_{\psi_q}, \|\|X\|_{S_2}\|_{\psi_q} \leq K(mT)$ for some constant $K(mT)$ which depends only on the product mT. Let $x > 0$ and $0 < \epsilon < 1/2$, and put $\lambda(n, mT, x) = c_0 K(mT)^q (\log n)^{(4q-2)/q}(x + \log n)$.

Theorem (L. and Mendelson)

Assume that $q \geq 2$, $\|Y\|_{\psi_q}, \|\|X\|_{S_2}\|_{\psi_q} \leq K(mT)$ *for some constant* $K(mT)$ *which depends only on the product* mT. *Let* $x > 0$ *and* $0 < \epsilon < 1/2$, *and put* $\lambda(n, mT, x) = c_0 K(mT)^q (\log n)^{(4q-2)/q}(x + \log n)$. *Consider the RERM procedure*

$$\hat{A}_n \in \mathrm{Arg} \min_{A \in \mathcal{M}_{m \times T}} \Big(R_n^{(q)}(A) + \lambda(n, mT, x) \frac{\|A\|_{S_1}^q}{n\epsilon^2} \Big)$$

Theorem (L. and Mendelson)

Assume that $q \geq 2$, $\|Y\|_{\psi_q}, \|\|X\|_{S_2}\|_{\psi_q} \leq K(mT)$ *for some constant* $K(mT)$ *which depends only on the product* mT. *Let* $x > 0$ *and* $0 < \epsilon < 1/2$, *and put* $\lambda(n, mT, x) = c_0 K(mT)^q (\log n)^{(4q-2)/q}(x + \log n)$. *Consider the RERM procedure*

$$\hat{A}_n \in \mathrm{Arg}\min_{A \in \mathcal{M}_{m\times T}} \Big(R_n^{(q)}(A) + \lambda(n, mT, x)\frac{\|A\|_{S_1}^q}{n\epsilon^2}\Big)$$

Then, with probability greater than $1 - 10\exp(-x)$,

$$R^{(q)}(\hat{A}_n) \leq \inf_{A \in \mathcal{M}_{m\times T}} \Big((1+2\epsilon)R^{(q)}(A) + c_1\lambda(n, mT, x)\frac{(1 + \|A\|_{S_1}^q)}{n\epsilon^2}\Big).$$

Matrix Completion - Part 5

Remarks :

1. Almost no assumption (no RIP type of assumption, we don't need to assume that $\mathbb{E}(Y|X) = \langle X, A_0 \rangle$,etc..). Assumptions only on the tails of Y and $\|X\|_{S_2}$.

Matrix Completion - Part 5

Remarks :

1. Almost no assumption (no RIP type of assumption, we don't need to assume that $\mathbb{E}(Y|X) = \langle X, A_0 \rangle$,etc..). Assumptions only on the tails of Y and $\|X\|_{S_2}$.
2. For $q = 2$, we regularize by the square $\|A\|_{S_1}^2$. We have fast rates $\sim \|A_0\|_{S_1}^2 / n$.

Matrix Completion - Part 5

Remarks :

1. Almost no assumption (no RIP type of assumption, we don't need to assume that $\mathbb{E}(Y|X) = \langle X, A_0 \rangle$,etc..). Assumptions only on the tails of Y and $\|X\|_{S_2}$.
2. For $q = 2$, we regularize by the square $\|A\|_{S_1}^2$. We have fast rates $\sim \|A_0\|_{S_1}^2 / n$.

Imagine that we "know" more : for instance, that $Y \approx \langle X, A_0 \rangle$ where

- A_0 is low-rank

Matrix Completion - Part 5

Remarks :

1. Almost no assumption (no RIP type of assumption, we don't need to assume that $\mathbb{E}(Y|X) = \langle X, A_0 \rangle$,etc..). Assumptions only on the tails of Y and $\|X\|_{S_2}$.
2. For $q = 2$, we regularize by the square $\|A\|_{S_1}^2$. We have fast rates $\sim \|A_0\|_{S_1}^2/n$.

Imagine that we "know" more : for instance, that $Y \approx \langle X, A_0 \rangle$ where

- A_0 is low-rank $\Longrightarrow \mathrm{crit}(A) = \|A\|_{S_1}$;

Matrix Completion - Part 5

Remarks :

1. Almost no assumption (no RIP type of assumption, we don't need to assume that $\mathbb{E}(Y|X) = \langle X, A_0 \rangle$,etc..). Assumptions only on the tails of Y and $\|X\|_{S_2}$.
2. For $q = 2$, we regularize by the square $\|A\|_{S_1}^2$. We have fast rates $\sim \|A_0\|_{S_1}^2 / n$.

Imagine that we "know" more : for instance, that $Y \approx \langle X, A_0 \rangle$ where

- A_0 is low-rank $\Longrightarrow \mathrm{crit}(A) = \|A\|_{S_1}$;
- and, the singular values of A_0 are well-spread

Matrix Completion - Part 5

Remarks :

1. Almost no assumption (no RIP type of assumption, we don't need to assume that $\mathbb{E}(Y|X) = \langle X, A_0 \rangle$,etc..). Assumptions only on the tails of Y and $\|X\|_{S_2}$.
2. For $q = 2$, we regularize by the square $\|A\|_{S_1}^2$. We have fast rates $\sim \|A_0\|_{S_1}^2 / n$.

Imagine that we "know" more : for instance, that $Y \approx \langle X, A_0 \rangle$ where

- A_0 is low-rank $\Longrightarrow \mathrm{crit}(A) = \|A\|_{S_1}$;
- and, the singular values of A_0 are well-spread $\Longrightarrow$ $\mathrm{crit}(A) = r_1 \|A\|_{S_1} + r_2 \|A\|_{S_2}^2$;

Matrix Completion - Part 5

Remarks :

1. Almost no assumption (no RIP type of assumption, we don't need to assume that $\mathbb{E}(Y|X) = \langle X, A_0 \rangle$,etc..). Assumptions only on the tails of Y and $\|X\|_{S_2}$.
2. For $q = 2$, we regularize by the square $\|A\|_{S_1}^2$. We have fast rates $\sim \|A_0\|_{S_1}^2 / n$.

Imagine that we "know" more : for instance, that $Y \approx \langle X, A_0 \rangle$ where

- A_0 is low-rank $\Longrightarrow \mathrm{crit}(A) = \|A\|_{S_1}$;
- and, the singular values of A_0 are well-spread $\Longrightarrow$ $\mathrm{crit}(A) = r_1\|A\|_{S_1} + r_2\|A\|_{S_2}^2$;
- and, A_0 has many zeroes

Matrix Completion - Part 5

Remarks :

1. Almost no assumption (no RIP type of assumption, we don't need to assume that $\mathbb{E}(Y|X) = \langle X, A_0 \rangle$,etc..). Assumptions only on the tails of Y and $\|X\|_{S_2}$.
2. For $q = 2$, we regularize by the square $\|A\|_{S_1}^2$. We have fast rates $\sim \|A_0\|_{S_1}^2 / n$.

Imagine that we "know" more : for instance, that $Y \approx \langle X, A_0 \rangle$ where

- A_0 is low-rank $\Longrightarrow \mathrm{crit}(A) = \|A\|_{S_1}$;
- and, the singular values of A_0 are well-spread $\Longrightarrow$ $\mathrm{crit}(A) = r_1\|A\|_{S_1} + r_2\|A\|_{S_2}^2$;
- and, A_0 has many zeroes $\Longrightarrow$ $\mathrm{crit}(A) = r_1\|A\|_{S_1} + r_2\|A\|_{S_2}^2 + r_3\|A\|_{\ell_1^{mT}}$;

Matrix Completion - Part 5

Remarks :

1. Almost no assumption (no RIP type of assumption, we don't need to assume that $\mathbb{E}(Y|X) = \langle X, A_0 \rangle$,etc..). Assumptions only on the tails of Y and $\|X\|_{S_2}$.
2. For $q = 2$, we regularize by the square $\|A\|_{S_1}^2$. We have fast rates $\sim \|A_0\|_{S_1}^2/n$.

Imagine that we "know" more : for instance, that $Y \approx \langle X, A_0 \rangle$ where

- A_0 is low-rank $\Longrightarrow \mathrm{crit}(A) = \|A\|_{S_1}$;
- and, the singular values of A_0 are well-spread $\Longrightarrow$ $\mathrm{crit}(A) = r_1\|A\|_{S_1} + r_2\|A\|_{S_2}^2$;
- and, A_0 has many zeroes $\Longrightarrow$ $\mathrm{crit}(A) = r_1\|A\|_{S_1} + r_2\|A\|_{S_2}^2 + r_3\|A\|_{\ell_1^{mT}}$;

We can obtain exact and non-exact oracle inequalities for a RERM based on the criterion

$$\mathrm{crit}(A) = r_1\|A\|_{S_1} + r_2\|A\|_{S_2}^2 + r_3\|A\|_{\ell_1^{mT}}$$

Theorem (Gaïffas and L.)

Assume that $\|Y\|_{\psi_2}, \|\|X\|_{S_2}\|_{\psi_2} \leq K(mT)$ *for some constant* $K(mT)$ *which depends only on the product* mT*.*

Theorem (Gaïffas and L.)

Assume that $\|Y\|_{\psi_2}, \|\|X\|_{S_2}\|_{\psi_2} \leq K(mT)$ for some constant $K(mT)$ which depends only on the product mT. Fix any $x, r_1, r_2, r_3 > 0$, and consider

$$\hat{A}_n \in \underset{A \in \mathcal{M}_{m,T}}{\operatorname{argmin}} \Big\{ R_n^{(2)}(A) + \frac{\lambda_{n,mT,x}}{\sqrt{n}} \big(r_1 \|A\|_{S_1} + r_2 \|A\|_{S_2}^2 + r_3 \|A\|_1 \big) \Big\}$$

Theorem (Gaïffas and L.)

Assume that $\|Y\|_{\psi_2}, \| \|X\|_{S_2} \|_{\psi_2} \leq K(mT)$ *for some constant* $K(mT)$ *which depends only on the product* mT. *Fix any* $x, r_1, r_2, r_3 > 0$, *and consider*

$$\hat{A}_n \in \operatorname*{argmin}_{A \in \mathcal{M}_{m,T}} \Big\{ R_n^{(2)}(A) + \frac{\lambda_{n,mT,x}}{\sqrt{n}} \big(r_1 \|A\|_{S_1} + r_2 \|A\|_{S_2}^2 + r_3 \|A\|_1 \big) \Big\}$$

Then, with probability larger than $1 - 5e^{-x}$,

$$R^{(2)}(\hat{A}_n) \leq \inf_{A \in \mathcal{M}_{m,T}} \Big\{ R^{(2)}(A) + \frac{\lambda_{n,mT,x}}{\sqrt{n}} \big(1 + r_1 \|A\|_{S_1} + r_2 \|A\|_{S_2}^2 + r_3 \|A\|_1 \big) \Big\}$$

Applications to ℓ_1-regularization

Regression model

Model :

- $(Y_1, X_1), \ldots, (Y_n, X_n)$: n i.i.d. random variables in $\mathbb{R} \times \mathbb{R}^d$;

Regression model

Model :

- $(Y_1, X_1), \ldots, (Y_n, X_n)$: n i.i.d. random variables in $\mathbb{R} \times \mathbb{R}^d$;
- $\ell^{(q)} : \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d \longmapsto \ell^{(q)}(\beta, (y, x)) = \ell_\beta^{(q)}(y, x) = |y - \langle \beta, x \rangle|^q$.

Regression model

Model :

- $(Y_1, X_1), \dots, (Y_n, X_n)$: n i.i.d. random variables in $\mathbb{R} \times \mathbb{R}^d$;
- $\ell^{(q)} : \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d \longmapsto \ell^{(q)}(\beta, (y, x)) = \ell_\beta^{(q)}(y, x) = |y - \langle \beta, x \rangle|^q$.

Notation :

- $\ell_\beta^{(q)}$: L_q-loss function of a vector $\beta \in \mathbb{R}^d$

Regression model

Model :

- $(Y_1, X_1), \ldots, (Y_n, X_n)$: n i.i.d. random variables in $\mathbb{R} \times \mathbb{R}^d$;
- $\ell^{(q)} : \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d \longmapsto \ell^{(q)}(\beta, (y, x)) = \ell_\beta^{(q)}(y, x) = |y - \langle \beta, x \rangle|^q$.

Notation :

- $\ell_\beta^{(q)}$: L_q-loss function of a vector $\beta \in \mathbb{R}^d$
- $R^{(q)}(\beta) = \mathbb{E}|Y - \langle \beta, X \rangle|^q$: L_q-risk of a vector $\beta \in \mathbb{R}^d$

Regression model

Model :

- $(Y_1, X_1), \ldots, (Y_n, X_n)$: n i.i.d. random variables in $\mathbb{R} \times \mathbb{R}^d$;
- $\ell^{(q)} : \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d \longmapsto \ell^{(q)}(\beta, (y, x)) = \ell_\beta^{(q)}(y, x) = |y - \langle \beta, x \rangle|^q$.

Notation :

- $\ell_\beta^{(q)}$: L_q-loss function of a vector $\beta \in \mathbb{R}^d$
- $R^{(q)}(\beta) = \mathbb{E}|Y - \langle \beta, X \rangle|^q$: L_q-risk of a vector $\beta \in \mathbb{R}^d$

Problem : $d >> n$ (more variables than observations) but we believe that $Y \approx \langle \beta_0, X \rangle$ where β_0 is of short support ($|\mathrm{Supp}(\beta_0)| < n$) (This is not an assumption !)

Regression model

Model :

- $(Y_1, X_1), \ldots, (Y_n, X_n)$: n i.i.d. random variables in $\mathbb{R} \times \mathbb{R}^d$;
- $\ell^{(q)} : \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d \longmapsto \ell^{(q)}(\beta, (y, x)) = \ell_\beta^{(q)}(y, x) = |y - \langle \beta, x \rangle|^q$.

Notation :

- $\ell_\beta^{(q)}$: L_q-loss function of a vector $\beta \in \mathbb{R}^d$
- $R^{(q)}(\beta) = \mathbb{E}|Y - \langle \beta, X \rangle|^q$: L_q-risk of a vector $\beta \in \mathbb{R}^d$

Problem : $d >> n$ (more variables than observations) but we believe that $Y \approx \langle \beta_0, X \rangle$ where β_0 is of short support ($|\mathrm{Supp}(\beta_0)| < n$) (This is not an assumption !)
$\mathcal{F} := \{\langle \cdot, \beta \rangle : \beta \in \mathbb{R}^d\}$ and $\mathrm{crit}(\beta) = |\mathrm{Supp}(\beta)| \Rightarrow$

Regression model

Model :

- $(Y_1, X_1), \dots, (Y_n, X_n)$: n i.i.d. random variables in $\mathbb{R} \times \mathbb{R}^d$;
- $\ell^{(q)} : \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d \longmapsto \ell^{(q)}(\beta, (y, x)) = \ell_\beta^{(q)}(y, x) = |y - \langle \beta, x \rangle|^q$.

Notation :

- $\ell_\beta^{(q)}$: L_q-loss function of a vector $\beta \in \mathbb{R}^d$
- $R^{(q)}(\beta) = \mathbb{E}|Y - \langle \beta, X \rangle|^q$: L_q-risk of a vector $\beta \in \mathbb{R}^d$

Problem : $d >> n$ (more variables than observations) but we believe that $Y \approx \langle \beta_0, X \rangle$ where β_0 is of short support ($|\mathrm{Supp}(\beta_0)| < n$) (This is not an assumption !)
$\mathcal{F} := \{\langle \cdot, \beta \rangle : \beta \in \mathbb{R}^d\}$ and $\mathrm{crit}(\beta) = |\mathrm{Supp}(\beta)| \Rightarrow$ Convexification

$$\mathrm{crit}(\beta) = \|\beta\|_1$$

Oracle inequalitiy for the square LASSO

Let $q \geq 2$. Assume that there exists some constant $c_d > 0$ (which may depend only on d) such that $\|Y\|_{\psi_q}, \|\|X\|_{\ell_\infty^d}\|_{\psi_q} \leq c_d$.

Oracle inequalitiy for the square LASSO

Let $q \geq 2$. Assume that there exists some constant $c_d > 0$ (which may depend only on d) such that $\|Y\|_{\psi_q}, \|\|X\|_{\ell_\infty^d}\|_{\psi_q} \leq c_d$.For $x > 0$ and $0 < \epsilon < 1/2$, let

$$\lambda(n, d, x) = c_0 c_d^q (\log n)^{(4q-2)/q} (\log d)^2 (x + \log n)$$

Oracle inequalitiy for the square LASSO

Let $q \geq 2$. Assume that there exists some constant $c_d > 0$ (which may depend only on d) such that $\|Y\|_{\psi_q}, \|\|X\|_{\ell_\infty^d}\|_{\psi_q} \leq c_d$.For $x > 0$ and $0 < \epsilon < 1/2$, let

$$\lambda(n, d, x) = c_0 c_d^q (\log n)^{(4q-2)/q} (\log d)^2 (x + \log n)$$

and consider the regularized ERM estimator

$$\hat{\beta}_n \in \operatorname*{argmin}_{\beta \in \mathbb{R}^d} \Big(R_n^{(q)}(\beta) + \lambda(n, d, x) \frac{\|\beta\|_{\ell_1}^q}{n\epsilon^2} \Big).$$

Oracle inequalitiy for the square LASSO

Let $q \geq 2$. Assume that there exists some constant $c_d > 0$ (which may depend only on d) such that $\|Y\|_{\psi_q}, \, \| \|X\|_{\ell_\infty^d} \|_{\psi_q} \leq c_d$.For $x > 0$ and $0 < \epsilon < 1/2$, let

$$\lambda(n, d, x) = c_0 c_d^q (\log n)^{(4q-2)/q} (\log d)^2 (x + \log n)$$

and consider the regularized ERM estimator

$$\hat{\beta}_n \in \operatorname*{argmin}_{\beta \in \mathbb{R}^d} \Big(R_n^{(q)}(\beta) + \lambda(n, d, x) \frac{\|\beta\|_{\ell_1}^q}{n\epsilon^2} \Big).$$

Then, with probability greater than $1 - 12\exp(-x)$, the L_q-risk of $\hat{\beta}_n$ satisfies

$$R^{(q)}(\hat{\beta}_n) \leq \inf_{\beta \in \mathbb{R}^d} \Big((1 + 2\epsilon) R^{(q)}(\beta) + c_1 \lambda(n, d, x) \frac{(1 + \|\beta\|_{\ell_1}^q)}{n\epsilon^2} \Big).$$

Oracle inequalities for penalized estimators

Model selection framework

$\mathcal{M}$: a countable collection of models.

Model selection framework

$\mathcal{M}$: a countable collection of models.

1. $\forall m \in \mathcal{M}, \hat{f}_m \in \text{argmin}_{f \in m} R_n(f)$,

Model selection framework

$\mathcal{M}$: a countable collection of models.

1. $\forall m \in \mathcal{M}, \hat{f}_m \in \operatorname{argmin}_{f \in m} R_n(f)$,
2. construct $\mathrm{pen} : \mathcal{M} \to \mathbb{R}_+$ and define

$$\hat{m} \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \big(R_n(\hat{f}_m) + \mathrm{pen}(m)\big).$$

Model selection framework

$\mathcal{M}$: a countable collection of models.

1. $\forall m \in \mathcal{M}, \hat{f}_m \in \operatorname{argmin}_{f \in m} R_n(f)$,
2. construct $\mathrm{pen} : \mathcal{M} \to \mathbb{R}_+$ and define

$$\hat{m} \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \left(R_n(\hat{f}_m) + \mathrm{pen}(m) \right).$$

3. oracle inequalities for the penalized estimator $\hat{f}_{\hat{m}}$.

Model selection framework

$\mathcal{M}$: a countable collection of models.

1. $\forall m \in \mathcal{M}, \hat{f}_m \in \mathrm{argmin}_{f \in m} R_n(f)$,
2. construct $\mathrm{pen} : \mathcal{M} \to \mathbb{R}_+$ and define

$$\hat{m} \in \underset{m \in \mathcal{M}}{\mathrm{argmin}} \left(R_n(\hat{f}_m) + \mathrm{pen}(m) \right).$$

3. oracle inequalities for the penalized estimator $\hat{f}_{\hat{m}}$.

construction of pen depends on the type of oracle inequality that we want to prove :

Model selection framework

$\mathcal{M}$: a countable collection of models.

1. $\forall m \in \mathcal{M}, \hat{f}_m \in \operatorname{argmin}_{f \in m} R_n(f)$,
2. construct $\mathrm{pen} : \mathcal{M} \to \mathbb{R}_+$ and define

$$\hat{m} \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \left(R_n(\hat{f}_m) + \mathrm{pen}(m)\right).$$

3. oracle inequalities for the penalized estimator $\hat{f}_{\hat{m}}$.

construction of pen depends on the type of oracle inequality that we want to prove : for any $m \in \mathcal{M}$

$$\ell_m = \{\ell_f : f \in m\}, \quad \mathcal{L}_m = \{\ell_f - \ell_{f_m^*} : f \in m\} \text{ and } \mathcal{E}_m = \{\ell_f - \ell_{f^*} : f \in m\}$$

where we assume that there exists $f_m^* \in \operatorname{argmin}_{f \in m} R(f)$ for any $m \in \mathcal{M}$ (and $f^* \in \operatorname{argmin}_f R(f)$).

Three fixed points

1. For non-exact oracle inequalities : $\forall m \in \mathcal{M}$, for some $0 < \eta < 1/2$,

$$\mathbb{E}\|P_n - P\|_{V(\ell_m)_{\lambda_\eta^*(m)}} \leq (\eta/4)\lambda_\eta^*(m).$$

Three fixed points

1. For non-exact oracle inequalities : $\forall m \in \mathcal{M}$, for some $0 < \eta < 1/2$,

$$\mathbb{E}\|P_n - P\|_{V(\ell_m)_{\lambda^*_\eta(m)}} \leq (\eta/4)\lambda^*_\eta(m).$$

2. For non-exact oracle inequalities for the estimation problem : $\forall m \in \mathcal{M}$

$$\mathbb{E}\|P_n - P\|_{V(\mathcal{E}_m)_{\nu^*_\eta(m)}} \leq (\eta/4)\nu^*_\eta(m).$$

Three fixed points

1. For non-exact oracle inequalities : $\forall m \in \mathcal{M}$, for some $0 < \eta < 1/2$,
$$\mathbb{E}\|P_n - P\|_{V(\ell_m)_{\lambda_\eta^*(m)}} \leq (\eta/4)\lambda_\eta^*(m).$$

2. For non-exact oracle inequalities for the estimation problem : $\forall m \in \mathcal{M}$
$$\mathbb{E}\|P_n - P\|_{V(\mathcal{E}_m)_{\nu_\eta^*(m)}} \leq (\eta/4)\nu_\eta^*(m).$$

3. for exact oracle inequalities : $\forall m \in \mathcal{M}$
$$\mathbb{E}\|P_n - P\|_{V(\mathcal{L}_m)_{\mu_{1/2}^*(m)}} \leq (1/8)\mu_{1/2}^*(m)$$

where $\mathcal{L}_m = \{\ell_f - \ell_{f_m^*} : f \in m\}$ and $f_m^* \in \operatorname{argmin}_{f\in m} R(f)$.

Non-exact oracle inequalities for the penalized estimator

Assume that there are some functions ϕ_n and B_n such that for every $m \in \mathcal{M}$ and every $f \in m$,

$$\|\max_{1\leq i\leq n}\sup_{f\in m}\ell_f(Z_i))\|_{\psi_1} \leq \phi_n(m) \text{ and } P\ell_f^2 \leq B_n(m)P\ell_f + B_n^2(m)/n.$$

Non-exact oracle inequalities for the penalized estimator

Assume that there are some functions ϕ_n and B_n such that for every $m \in \mathcal{M}$ and every $f \in m$,

$$\|\max_{1\leq i\leq n}\sup_{f\in m}\ell_f(Z_i))\|_{\psi_1} \leq \phi_n(m) \text{ and } P\ell_f^2 \leq B_n(m)P\ell_f + B_n^2(m)/n.$$

Let $0 < \eta < 1/2$ and assume that for every $(m, x) \in \mathcal{M} \times \mathbb{R}_+^*$,

$$\rho_n^\ell(m, x) \geq \max\Big(\lambda_\eta^*(m), c_0\frac{(\phi_n(m) + B_n(m)/\eta)(x+1)}{n\eta}\Big).$$

Non-exact oracle inequalities for the penalized estimator

Assume that there are some functions ϕ_n and B_n such that for every $m \in \mathcal{M}$ and every $f \in m$,

$$\| \max_{1\leq i\leq n} \sup_{f\in m} \ell_f(Z_i))\|_{\psi_1} \leq \phi_n(m) \text{ and } P\ell_f^2 \leq B_n(m)P\ell_f + B_n^2(m)/n.$$

Let $0 < \eta < 1/2$ and assume that for every $(m,x) \in \mathcal{M} \times \mathbb{R}_+^*$,

$$\rho_n^\ell(m,x) \geq \max\Big(\lambda_\eta^*(m), c_0\frac{(\phi_n(m) + B_n(m)/\eta)(x+1)}{n\eta}\Big).$$

Let $(x_m)_{m\in\mathcal{M}}$ be a family of positive numbers such that $\sum_{m\in\mathcal{M}} \exp(-x_m) \leq c_1$. Let $x > 0$ and consider the penalty function $\mathrm{pen}^\ell(m) = \rho_n^\ell(m, x + x_m)$ and the penalized estimator $\hat{f}_{\hat{m}}$ associated with this penalty function.

Non-exact oracle inequalities for the penalized estimator

Assume that there are some functions ϕ_n and B_n such that for every $m \in \mathcal{M}$ and every $f \in m$,

$$\|\max_{1\leq i\leq n}\sup_{f\in m}\ell_f(Z_i))\|_{\psi_1} \leq \phi_n(m) \text{ and } P\ell_f^2 \leq B_n(m)P\ell_f + B_n^2(m)/n.$$

Let $0 < \eta < 1/2$ and assume that for every $(m, x) \in \mathcal{M} \times \mathbb{R}_+^*$,

$$\rho_n^\ell(m,x) \geq \max\Big(\lambda_\eta^*(m), c_0\frac{(\phi_n(m) + B_n(m)/\eta)(x+1)}{n\eta}\Big).$$

Let $(x_m)_{m\in\mathcal{M}}$ be a family of positive numbers such that $\sum_{m\in\mathcal{M}}\exp(-x_m) \leq c_1$. Let $x > 0$ and consider the penalty function $\mathrm{pen}^\ell(m) = \rho_n^\ell(m, x + x_m)$ and the penalized estimator $\hat{f}_{\hat{m}}$ associated with this penalty function. Then, with probability greater than $1 - c_2e^{-x}$,

$$R(\hat{f}_{\hat{m}}) \leq \frac{1+\eta}{1-\eta}\inf_{m\in\mathcal{M}}\big(\inf_{f\in m}P\ell_f + \mathrm{pen}^\ell(m)\big).$$

Non-exact oracle inequalities for the penalized estimator

$$\mathrm{pen}^{\ell}(m) = \max\left(\lambda_\eta^*(m), c_0 \frac{(\phi_n(m) + B_n(m)/\eta)(x + x_m + 1)}{n\eta}\right) \sim \lambda_\eta^*(m)$$

where

$$\mathbb{E}\|P_n - P\|_{V(\ell_m)_{\lambda_\eta^*(m)}} \leq (\eta/4)\lambda_\eta^*(m).$$

Oracle inequality for the estimation problem

Assume that there exists $0 < \beta \leq 1$ and some functions ϕ_n and B_n such that for every $m \in \mathcal{M}$ and every $f \in m$,

$$\| \max_{1 \leq i \leq n} \sup_{f \in m} \mathcal{E}_f(Z_i) \|_{\psi_1} \leq \phi_n(m) \text{ and } P\mathcal{E}_f^2 \leq B_n(m)\big(P\mathcal{E}_f\big)^{\beta} + B_n^2(m)/n.$$

Oracle inequality for the estimation problem

Assume that there exists $0 < \beta \leq 1$ and some functions ϕ_n and B_n such that for every $m \in \mathcal{M}$ and every $f \in m$,

$$\| \max_{1 \leq i \leq n} \sup_{f \in m} \mathcal{E}_f(Z_i) \|_{\psi_1} \leq \phi_n(m) \text{ and } P\mathcal{E}_f^2 \leq B_n(m) \big(P\mathcal{E}_f \big)^{\beta} + B_n^2(m)/n.$$

Let $0 < \eta < 1/2$ and assume that for every $(m, x) \in \mathcal{M} \times \mathbb{R}_+^*$,

$$\rho_n^{\mathcal{E}}(m, x) \geq \max \Big(\nu_\eta^*(m), c_2 (B_n(m) + \phi_n(m)) \Big(\frac{x+1}{n\eta} \Big)^{\frac{1}{2-\beta}} \Big).$$

Oracle inequality for the estimation problem

Assume that there exists $0 < \beta \leq 1$ and some functions ϕ_n and B_n such that for every $m \in \mathcal{M}$ and every $f \in m$,

$$\| \max_{1 \leq i \leq n} \sup_{f \in m} \mathcal{E}_f(Z_i) \|_{\psi_1} \leq \phi_n(m) \text{ and } P\mathcal{E}_f^2 \leq B_n(m) \big(P\mathcal{E}_f \big)^{\beta} + B_n^2(m)/n.$$

Let $0 < \eta < 1/2$ and assume that for every $(m, x) \in \mathcal{M} \times \mathbb{R}_+^*$,

$$\rho_n^{\mathcal{E}}(m,x) \geq \max \Big(\nu_\eta^*(m), c_2 (B_n(m) + \phi_n(m)) \Big(\frac{x+1}{n\eta} \Big)^{\frac{1}{2-\beta}} \Big).$$

Let $(x_m)_{m \in \mathcal{M}}$ be a family of positive numbers such that $\sum_{m \in \mathcal{M}} \exp(-x_m) \leq c_1$. Let $x > 0$ and consider the penalty function $\mathrm{pen}^{\mathcal{E}}(m) = \rho_n^{\mathcal{E}}(m, x + x_m)$ and the penalized estimator $\hat{f}_{\hat{m}}$ associated with this penalty function.

Oracle inequality for the estimation problem

Assume that there exists $0 < \beta \leq 1$ and some functions ϕ_n and B_n such that for every $m \in \mathcal{M}$ and every $f \in m$,

$$\| \max_{1 \leq i \leq n} \sup_{f \in m} \mathcal{E}_f(Z_i) \|_{\psi_1} \leq \phi_n(m) \text{ and } P\mathcal{E}_f^2 \leq B_n(m) \big(P\mathcal{E}_f\big)^{\beta} + B_n^2(m)/n.$$

Let $0 < \eta < 1/2$ and assume that for every $(m, x) \in \mathcal{M} \times \mathbb{R}_+^*$,

$$\rho_n^{\mathcal{E}}(m, x) \geq \max\Big(\nu_\eta^*(m), c_2(B_n(m) + \phi_n(m))\Big(\frac{x+1}{n\eta}\Big)^{\frac{1}{2-\beta}}\Big).$$

Let $(x_m)_{m \in \mathcal{M}}$ be a family of positive numbers such that $\sum_{m \in \mathcal{M}} \exp(-x_m) \leq c_1$. Let $x > 0$ and consider the penalty function $\mathrm{pen}^{\mathcal{E}}(m) = \rho_n^{\mathcal{E}}(m, x + x_m)$ and the penalized estimator $\hat{f}_{\hat{m}}$ associated with this penalty function. Then, with probability greater than $1 - c_3 e^{-x}$,

$$R(\hat{f}_{\hat{m}}) - R(f^*) \leq \frac{1+\eta}{1-\eta} \inf_{m \in \mathcal{M}} \Big(\inf_{f \in m} \big(R(f) - R(f^*)\big) + \mathrm{pen}^{\mathcal{E}}(m)\Big).$$

Oracle inequality for the estimation problem

In the context of the estimation problem, a possible way of penalizing the empirical risk is by the function

$$\mathrm{pen}^{\mathcal{E}}(m) = \max\left(\nu_\eta^*(m), c_2(B_n(m)+\phi_n(m))\left(\frac{x+x_m}{n\eta}\right)^{\frac{1}{2-\beta}}\right)$$

where

$$\mathbb{E}\|P_n - P\|_{V(\mathcal{E}_m)_{\nu_\eta^*(m)}} \leq (\eta/4)\nu_\eta^*(m).$$

Exact oracle inequality for the penalized estimator

Take $\mathcal{M} = \left(m_n\right)_{n\in\mathbb{N}}$ s.t. $m_0 \subset m_1 \subset m_2 \subset \cdots$.

Exact oracle inequality for the penalized estimator

Take $\mathcal{M} = \big(m_n\big)_{n\in\mathbb{N}}$ s.t. $m_0 \subset m_1 \subset m_2 \subset \cdots$. Assume that there exists $0 < \beta \leq 1$ and two non-decreasing functions ϕ_n and B_n such that for every $m \in \mathcal{M}$ and every $f \in m$,

$$\|\max_{1\leq i\leq n}\sup_{f\in m}\mathcal{L}_f(Z_i)\|_{\psi_1} \leq \phi_n(m) \text{ and } P\mathcal{L}_{m,f}^2 \leq B_n(m)\big(P\mathcal{L}_{m,f}\big)^\beta + B_n^2(m)/n$$

where $\mathcal{L}_{m,f} = \ell_f - \ell_{f_m^*}$.

Exact oracle inequality for the penalized estimator

Take $\mathcal{M} = \big(m_n\big)_{n\in\mathbb{N}}$ s.t. $m_0 \subset m_1 \subset m_2 \subset \cdots$. Assume that there exists $0 < \beta \leq 1$ and two non-decreasing functions ϕ_n and B_n such that for every $m \in \mathcal{M}$ and every $f \in m$,

$$\| \max_{1\leq i\leq n} \sup_{f\in m} \mathcal{L}_f(Z_i)\|_{\psi_1} \leq \phi_n(m) \text{ and } P\mathcal{L}_{m,f}^2 \leq B_n(m)\big(P\mathcal{L}_{m,f}\big)^{\beta} + B_n^2(m)/n$$

where $\mathcal{L}_{m,f} = \ell_f - \ell_{f_m^*}$. Let $\rho_n^{\mathcal{L}}$ be an increasing function such that for every $(m,x) \in \mathcal{M} \times \mathbb{R}_+^*$,

$$\rho_n^{\mathcal{L}}(m,x) \geq \max\Big(\mu_{1/2}^*(m), (B_n(m) + \phi_n(m))\Big(\frac{x+1}{n}\Big)^{\frac{1}{2-\beta}}\Big).$$

Exact oracle inequality for the penalized estimator

Take $\mathcal{M} = \big(m_n\big)_{n\in\mathbb{N}}$ s.t. $m_0 \subset m_1 \subset m_2 \subset \cdots$. Assume that there exists $0 < \beta \leq 1$ and two non-decreasing functions ϕ_n and B_n such that for every $m \in \mathcal{M}$ and every $f \in m$,

$$\|\max_{1\leq i\leq n} \sup_{f\in m} \mathcal{L}_f(Z_i)\|_{\psi_1} \leq \phi_n(m) \text{ and } P\mathcal{L}^2_{m,f} \leq B_n(m)\big(P\mathcal{L}_{m,f}\big)^{\beta} + B_n^2(m)/n$$

where $\mathcal{L}_{m,f} = \ell_f - \ell_{f_m^*}$. Let $\rho_n^{\mathcal{L}}$ be an increasing function such that for every $(m,x) \in \mathcal{M} \times \mathbb{R}_+^*$,

$$\rho_n^{\mathcal{L}}(m,x) \geq \max\Big(\mu^*_{1/2}(m), (B_n(m) + \phi_n(m))\Big(\frac{x+1}{n}\Big)^{\frac{1}{2-\beta}}\Big).$$

Let $(x_m)_{m\in\mathcal{M}}$ be a family of positive numbers such that $\sum_{m\in\mathcal{M}} \exp(-x_m) \leq c_1$. Let $x > 0$ and consider the penalty function $\mathrm{pen}^{\mathcal{L}}(m) = (7/2)\rho_n^{\mathcal{L}}(m, x + x_m)$ and the penalized estimator $\hat{f}_{\hat{m}}$ associated with this penalty function.

Exact oracle inequality for the penalized estimator

Take $\mathcal{M} = (m_n)_{n\in\mathbb{N}}$ s.t. $m_0 \subset m_1 \subset m_2 \subset \cdots$. Assume that there exists $0 < \beta \leq 1$ and two non-decreasing functions ϕ_n and B_n such that for every $m \in \mathcal{M}$ and every $f \in m$,

$$\|\max_{1\leq i\leq n} \sup_{f\in m} \mathcal{L}_f(Z_i)\|_{\psi_1} \leq \phi_n(m) \text{ and } P\mathcal{L}_{m,f}^2 \leq B_n(m)\big(P\mathcal{L}_{m,f}\big)^\beta + B_n^2(m)/n$$

where $\mathcal{L}_{m,f} = \ell_f - \ell_{f_m^*}$. Let $\rho_n^{\mathcal{L}}$ be an increasing function such that for every $(m,x) \in \mathcal{M}\times\mathbb{R}_+^*$,

$$\rho_n^{\mathcal{L}}(m,x) \geq \max\Big(\mu_{1/2}^*(m), (B_n(m)+\phi_n(m))\Big(\frac{x+1}{n}\Big)^{\frac{1}{2-\beta}}\Big).$$

Let $(x_m)_{m\in\mathcal{M}}$ be a family of positive numbers such that $\sum_{m\in\mathcal{M}} \exp(-x_m) \leq c_1$. Let $x>0$ and consider the penalty function $\mathrm{pen}^{\mathcal{L}}(m) = (7/2)\rho_n^{\mathcal{L}}(m, x+x_m)$ and the penalized estimator $\hat{f}_{\hat{m}}$ associated with this penalty function. Then, with probability greater than $1 - c_1 e^{-x}$,

$$R(\hat{f}_{\hat{m}}) \leq \inf_{m\in\mathcal{M}}\big(\inf_{f\in m} R(f) + (18/7)\mathrm{pen}^{\mathcal{L}}(m)\big).$$

Exact oracle inequality for the penalized estimator

Therefore, for the exact prediction problem, a possible way of penalizing the empirical risk is by the function

$$\mathrm{pen}^{\mathcal{L}}(m) = c_2 \max\left(\mu^*_{1/2}(m), (B_n(m) + \phi_n(m))\left(\frac{x + x_m}{n}\right)^{\frac{1}{2-\beta}}\right)$$

Exact oracle inequality for the penalized estimator

Therefore, for the exact prediction problem, a possible way of penalizing the empirical risk is by the function

$$\mathrm{pen}^{\mathcal{L}}(m) = c_2 \max\left(\mu^*_{1/2}(m), (B_n(m) + \phi_n(m))\Big(\frac{x + x_m}{n}\Big)^{\frac{1}{2-\beta}}\right)$$

where

$$\mathbb{E}\|P_n - P\|_{V(\mathcal{L}_m)_{\mu^*_{1/2}(m)}} \leq (1/8)\mu^*_{1/2}(m)$$

and $\mathcal{L}_m = \{\ell_f - \ell_{f^*_m} : f \in m\}$ and $f^*_m \in \mathrm{argmin}_{f \in m} R(f)$.

Thanks ! !