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 $\|X\|_{\ell_\infty}$ 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

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

where \(\lambda = \lambda(n, d) = \text{polylog}(n, d)\) 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

\[
\hat{\beta}_n \in \arg\min_{\beta \in \mathbb{R}^d} \left( \frac{1}{n} \sum_{i=1}^n (Y_i - \langle X_i, \beta \rangle)^2 + \lambda \frac{\|\beta\|_{\ell_1^d}^2}{nc^2} \right)
\]

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

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

\[
\hat{\beta}_n \in \arg\min_{\beta \in \mathbb{R}^d} \left( \frac{1}{n} \sum_{i=1}^n (Y_i - \langle X_i, \beta \rangle)^2 + \lambda \frac{\|\beta\|_{\ell_1^d}^2}{nc^2} \right)
\]

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

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

**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

\[
\hat{\beta}_n \in \arg\min_{\beta \in \mathbb{R}^d} \left( \frac{1}{n} \sum_{i=1}^n (Y_i - \langle X_i, \beta \rangle)^2 + \lambda \frac{\|\beta\|_{\ell_1}^2}{nc^2} \right)
\]

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

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

**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

- $Z_1, \ldots, Z_n : n \text{ i.i.d.} \sim Z$ random variables in $Z$
General model in learning theory

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

- \( \ell : (f, z) \mapsto \ell(f, z) := \ell_f(z) \in \mathbb{R} : \) loss function of \( f : Z \to \mathbb{R} \)
General model in learning theory

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

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

- $R(f) = \mathbb{E}\ell_f(Z)$: risk of $f$
General model in learning theory

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

- \( \ell : (f, z) \mapsto \ell(f, z) := \ell_f(z) \in \mathbb{R} : \) 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

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

where \( \mathcal{D} := (Z_1, \ldots, Z_n) \).
General model in learning theory

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

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

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

- Risk of a statistics \( \hat{f}_n \) is

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

where \( \mathcal{D} := (Z_1, \cdots, Z_n) \).
General model in learning theory

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

- $\ell : (f, z) \mapsto \ell(f, z) := \ell_f(z) \in \mathbb{R}$: 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

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

  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 $Z$ ($Z = (X, Y)$)

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

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

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

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

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 Empirical Risk Minimization
Empirical Risk minimization

1. a model $F$ is a class of functions $f : \mathcal{Z} \rightarrow \mathbb{R}$
Empirical Risk minimization

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

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

$$\hat{f}_n^{(ERM)} \in \arg\min_{f \in F} R_n(f)$$
Three different oracle inequalities. Exemple in aggregation theory.

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

$$\hat{f}_n^{(ERM)} \in \arg\min_{f \in F} R_n(f) \quad \text{where} \quad R_n(f) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2,$$
Three different oracle inequalities. Example in aggregation theory.

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

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

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. Example in aggregation theory.

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

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

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

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

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

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

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

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

$$R(\hat{f}_n^{(ERM)}) - R(f^*) \leq (1 + \epsilon) \min_{f \in F} (R(f) - R(f^*)) + c_0 \frac{x + \log |F|}{n\epsilon}$$
Three different oracle inequalities. Exemple in aggregation theory.

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

$$
\hat{f}_n^{(ERM)} \in \text{arg min}_{f \in F} R_n(f) \quad \text{where} \quad R_n(f) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2,
$$

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 \textbf{ERM} over a finite model $F$ w.r.t. the square loss is

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

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

1. \textbf{Exact Oracle Inequality}

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

2. \textbf{Non-Exact Oracle Inequality}

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

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

$$R(\hat{f}_n^{(ERM)}) - R(f^*) \leq (1 + \varepsilon) \min_{f \in F} (R(f) - R(f^*)) + c_0 \frac{x + \log |F|}{n \varepsilon}.$$
Three different oracle inequalities. Example in aggregation theory.

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

\[
\hat{f}_n^{(\text{ERM})} \in \text{Arg min}_{f \in F} R_n(f) \text{ where } R_n(f) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2,
\]

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

1. **Exact Oracle Inequality**

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

2. **Non-Exact Oracle Inequality**

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

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

\[
R(\hat{f}_n^{(\text{ERM})}) - R(f^*) \leq (1 + \epsilon) \min_{f \in F} (R(f) - R(f^*)) + c_0 \frac{x + \log |F|}{n \epsilon}
\]
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”:

  \[
  \frac{x + \log |F|}{n} \sim \frac{\text{comp}(F)}{n}
  \]
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”:
  \[
  \frac{x + \log |F|}{n} \sim \frac{\text{comp}(F)}{n}
  \]

- **Slow** decaying residual term (non-improvable: there exists lower bounds) for the “exact oracle inequality”:
  \[
  \sqrt{\frac{x + \log |F|}{n}} \sim \sqrt{\frac{\text{comp}(F)}{n}}
  \]
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”:

\[
\frac{x + \log |F|}{n} \sim \frac{\text{comp}(F)}{n}
\]

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

\[
\sqrt{\frac{x + \log |F|}{n}} \sim \sqrt{\frac{\text{comp}(F)}{n}}
\]

**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”:
  \[
  \frac{x + \log |F|}{n} \sim \frac{\text{comp}(F)}{n}
  \]

- **slow** decaying residual term (non-improvable: there exists lower bounds) for the “exact oracle inequality”:
  \[
  \sqrt{\frac{x + \log |F|}{n}} \sim \sqrt{\frac{\text{comp}(F)}{n}}
  \]

**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**:\[ l_F := \{l_f : f \in F\} \]

- **excess loss functions class**: for \( f^*_F \in \arg\min_{f \in F} R(f) \)
  \[ L_F := \{l_f - l_{f^*_F} : f \in F\} = l_F - l_{f^*_F} \]
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 \( f^*_F \in \text{argmin}_{f \in F} R(f) \)
  \[ \mathcal{L}_F := \{ \ell_f - \ell_{f^*_F} : f \in F \} = \ell_F - \ell_{f^*_F} \]

- Excess loss functions class for the estimation problem: for \( f^* \in \text{argmin}_f R(f) \)
  \[ \mathcal{E}_F := \{ \ell_f - \ell_{f^*} : f \in F \} = \ell_F - \ell_{f^*} \]
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 \( f^*_F \in \arg\min_{f \in F} R(f) \)
  \[ \mathcal{L}_F := \{ \ell_f - \ell_{f^*_F} : f \in F \} = \ell_F - \ell_{f^*_F} \]

- **excess loss functions class for the estimation problem**: for \( f^* \in \arg\min_{f} R(f) \)
  \[ \mathcal{E}_F := \{ \ell_f - \ell_{f^*} : f \in F \} = \ell_F - \ell_{f^*} \]

For every functions class \( H \), the **star-shaped hull of \( H \) in \( 0 \)** is
\[ V(H) = \text{star}(H, 0) := \{ \theta h : 0 \leq \theta \leq 1, h \in H \} \]
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 \( f_F^* \in \arg\min_{f \in F} R(f) \)
  \[ \mathcal{L}_F := \{ \ell_f - \ell_{f_F^*} : f \in F \} = \ell_F - \ell_{f_F^*} \]

- **excess loss functions class for the estimation problem**: for \( f^* \in \arg\min_{f \in F} R(f) \)
  \[ \mathcal{E}_F := \{ \ell_f - \ell_{f^*} : f \in F \} = \ell_F - \ell_{f^*} \]

For every functions class \( H \), the **star-shaped hull of \( H \) in 0** is
\[ V(H) = \text{star}(H, 0) := \{ \theta h : 0 \leq \theta \leq 1, h \in H \} \]
and its **localized set at level \( \lambda > 0 \)** is
\[ V(H)_\lambda := \{ g \in V(H) : \mathbb{E}g \leq \lambda \} \]
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

\[ V(\ell_F) = \text{star}(\ell_F, 0) \]
Exact and non-exact oracle inequalities in a general framework

\[ V(\ell_F) = \text{star}(\ell_F, 0) \]

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

\[ V(\mathcal{L}_F) = \text{star}(\mathcal{L}_F, 0) \]
Exact and non-exact oracle inequalities in a general framework

\[ V(L_F)_\lambda = \{ g \in V(L_F) : \mathbb{E}g \leq \lambda \} \]

\[ V(L_F) = \text{star}(L_F, 0) \]
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

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

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

Non-exact oracle inequalities

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

\[ \| P - P_n \|_H := \sup_{h \in H} |Ph - P_nh| \]

where

\[ Ph := \mathbb{E}h(Z) \text{ and } P_nh := \frac{1}{n} \sum_{i=1}^{n} h(Z_i) \]
Exact and non-exact oracle inequalities in a general framework

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

where

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

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:

\[ \mu^* := \inf (\mu > 0 : \mathbb{E} \| P - P_n \|_V(\mathcal{L}_F)_\mu \leq \mu/8) \]
Exact and non-exact oracle inequalities in a general framework

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

where

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

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

- for exact oracle inequalities:
  \[ \mu^* := \inf \left( \mu > 0 : \mathbb{E} \| P - P_n \|_{\mathcal{V}(\mathcal{L}_F), \mu} \leq \mu / 8 \right) \]

- non-exact oracle inequalities:
  \[ \lambda^*_\epsilon := \inf \left( \lambda > 0 : \mathbb{E} \| P - P_n \|_{\mathcal{V}(\ell_F), \lambda} \leq (\epsilon / 4) \lambda \right) \]
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 $f \in F$,

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

Let $\mu^* > 0$ be s.t.

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

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

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

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

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 $\rho_n(x)$ is an increasing function s.t. for every $x > 0$,

$$ \rho_n(x) \geq \max\left(\mu^*, c_0 (\|\mathcal{L}_F\|_{\infty} + B) x^n\right). $$
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 $f \in F$,\[ P \mathcal{L}^2_f \leq B P \mathcal{L}_f \]

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

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 $\rho_n$ is an increasing function s.t. for every $x > 0$,

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

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

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

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

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

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

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

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 $f \in F$,

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

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 $1 - 8 \exp(-x)$,

$$R(\hat{f}_n^{\text{ERM}}) \leq (1 + 2\epsilon) \inf_{f \in F} R(f) + \bar{\rho}_n(x)$$
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 $\rho_n$ 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: \( \forall f \in F, P\mathcal{L}_f^2 \leq B\mathcal{L}_f \);
The Bernstein Condition

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

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

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 \( n \),

\[
P \ell_f^2 \leq (c_0 D \log(en)) P \ell_f + \frac{(c_0 D \log(en))^2}{n}.
\]
The Bernstein Condition

1. **Exact oracle inequality**: \( \forall f \in F, P\mathcal{L}_f^2 \leq B\mathcal{L}_f \);

2. **Non-exact oracle inequality**: \( \forall f \in F, P\ell_f^2 \leq B\ell_f + B^2/n \).

**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 \( n \),

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

**Conclusion 1**: In the case of non-exact oracle inequalities, the Bernstein condition for \( \ell_F \) is almost trivially satisfied.
The Bernstein condition of the excess loss class $\mathcal{L}_F$

- $f_1(X)$
- $f_2(X)$
The Bernstein condition of the excess loss class $\mathcal{L}_F$

\[ f_2(X) \quad \quad \quad F = \{ f_1, f_2 \} \]

\[ f_1(X) \]
The Bernstein condition of the excess loss class $\mathcal{L}_F$

- $f_2(X)$
- $f_1(X)$

$F = \{ f_1, f_2 \}$

$R(f) = \mathbb{E}(Y - f(X))^2$
The Bernstein condition of the excess loss class $\mathcal{L}_F$

$F = \{f_1, f_2\}$

$R(f) = \mathbb{E}(Y - f(X))^2$
The Bernstein condition of the excess loss class $\mathcal{L}_F$

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

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

$$M_F := \{Y : \text{Card}\{f \in F : R(f) = \min_{f \in F} R(f)\} \geq 2\}$$
The Bernstein condition of the excess loss class $\mathcal{L}_F$

$F = \{f_1, f_2\}$

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

$M_F := \{Y : \text{Card}\{f \in F : R(f) = \min_{f \in F} R(f)\} \geq 2\}$
The Bernstein condition of the excess loss class $\mathcal{L}_F$

$F = \{f_1, f_2\}$

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

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

$M_F := \{Y : \text{Card}\{f \in F : R(f) = \min_{f \in F} R(f)\} \geq 2\}$
The Bernstein condition of the excess loss class $\mathcal{L}_F$

\[ F = \{ f_1, f_2 \} \]

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

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

\[ B \text{ increases} \]

\[ M_F := \{ Y : \text{Card}\{ f \in F : R(f) = \min_{f \in F} R(f) \} \geq 2 \} \]
The Bernstein condition of the excess loss class $\mathcal{L}_F$

$f_1(X)$

$f_2(X)$

$Y$ \sim \sqrt{1/n}$

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

$F = \{ f_1, f_2 \}$

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

$P \mathcal{L}_f^2 \leq B \mathcal{L}_f$, $B \sim \text{const}$
The Bernstein condition of the excess loss class $\mathcal{L}_F$

\[ F = \{ f_1, f_2 \} \]

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

\[ P \mathcal{L}^2_f \leq B P \mathcal{L}_f, \quad B \sim \text{const} \]

\[ B \sim \sqrt{n} \]

\[ M_F := \{ Y : \text{Card}\{ f \in F : R(f) = \min_{f \in F} R(f) \} \geq 2 \} \]
The Bernstein condition of the excess loss class $\mathcal{L}_F$

\[ F = \{ f_1, f_2 \} \]

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

\[ \mathcal{P} \mathcal{L}_f^2 \leq B \mathcal{P} \mathcal{L}_f, \; B \sim const \]

\[ B \sim \sqrt{n} \]

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

\[ M_F := \{ Y : \text{Card}\{ f \in F : R(f) = \min_{f \in F} R(f) \} \geq 2 \} \]
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 $L_F$ with a trivial Bernstein constant ($B \sim \sqrt{n}$) and thus a slow residual term $\sim \sqrt{\text{Comp}(F)/n}$. 
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 $\sim \sqrt{\text{Comp}(F)/n}$.

When the class $F$ is convex: the Bernstein condition of $\mathcal{L}_F$ is always satisfied (quadratic loss).
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 $\sim \sqrt{\text{Comp}(F)/n}$.

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 fixed points $\mu^*$ and $\lambda^*_\epsilon$ characterize the isomorphic properties of $\mathcal{L}_F$ and $\ell_F$ respectively.
The complexity terms: $\mu^*$ and $\lambda^*_\epsilon$ - Part 1

The fixed points $\mu^*$ and $\lambda^*$ characterize the isomorphic properties of $\mathcal{L}_F$ and $\ell_F$ 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: $\mu^*$ and $\lambda^*_\epsilon$ - Part 1

The fixed points $\mu^*$ and $\lambda^*$ characterize the isomorphic properties of $\mathcal{L}_F$ and $\ell_F$ 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 $1 - 4 \exp(-x)$,*

\[ (1/2)P_nh \leq Ph \leq (3/2)P_nh \]

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

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

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

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

Exact oracle inequalities

\[ \mathbb{E}\|P - P_n\|_{V(\mathcal{L}_F)_{\mu^*}} \leq \mu^*/8 \]
An example of computation of the fixed points $\lambda^*$ and $\mu^*$

[Peeling argument :]

\[ \text{[Peeling argument :]} \]
An example of computation of the fixed points $\lambda^*_\epsilon$ and $\mu^*$

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

$$\mathbb{E} \| P - P_n \|_{V(H)_\lambda} \leq \sum_{i=0}^{\infty} 2^{-i} \mathbb{E} \| P - P_n \|_{H_{2i+1}\lambda}$$
An example of computation of the fixed points $\lambda^*$ and $\mu^*$

[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 $\lambda^*_\epsilon$ and $\mu^*$

[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 $\lambda^*_\epsilon$ and $\mu^*$

[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 $\lambda^*_\epsilon$ and $\mu^*$

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

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

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 $\lambda_\epsilon^*$ and $\mu^*$

Computation of $\lambda_\epsilon^*$ and $\mu^*$ in the case of the Regression model with quadratic loss:

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

and

$$\mathcal{L}_F := \{ \mathcal{L}_f : (y, x) \mapsto (y - f(x))^2 - (y - f_F^*(x))^2 : f \in F \}.$$
An example of computation of the fixed points $\lambda^*_\epsilon$ and $\mu^*$

Computation of $\lambda^*_\epsilon$ and $\mu^*$ in the case of the Regression model with quadratic loss:

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

and

$$\mathcal{L}_F := \{\mathcal{L}_f : (y, x) \mapsto (y - f(x))^2 - (y - f^*_F(x))^2 : f \in F\}.$$
An example of computation of the fixed points $\lambda^*_\epsilon$ and $\mu^*$

Computation of $\lambda^*_\epsilon$ and $\mu^*$ in the case of the Regression model with quadratic loss:

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

and

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

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 $\lambda_\epsilon^*$ and $\mu^*$

Computation of $\lambda_\epsilon^*$ and $\mu^*$ in the case of the Regression model with quadratic loss:

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

and

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

Complexity measure of $F$:

$$P_\sigma F := \{(f(X_1), \ldots, f(X_n)) : f \in F\} \text{ and } U_n(F(\mu)) := \mathbb{E} \gamma_2(P_\sigma F(\mu), \ell_\infty^n)^2$$
An example of computation of the fixed points $\lambda^*_\epsilon$ and $\mu^*$

Computation of $\lambda^*_\epsilon$ and $\mu^*$ in the case of the Regression model with quadratic loss:

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

and

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

Complexity measure of $F$:

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

where $\gamma_i(T, d) := \inf_{(T_s)} \sup_{t \in T} \sum_{s=0}^\infty 2^{-s/2} d(t, T_s)$ and $|T_s| \leq 2^s$ and $\tilde{A} = A - A$ and $F(\mu) := \{f \in F : P\ell_f \leq \mu\}$
An example of computation of the fixed points $\lambda^*_\epsilon$ and $\mu^*$

Computation of $\lambda^*_\epsilon$ and $\mu^*$ in the case of the Regression model with quadratic loss:

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

and

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

Complexity measure of $F$:

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

and

$$U_n(F(\mu)) := \mathbb{E} \gamma_2(\tilde{P}_\sigma F(\mu), \ell_{\infty})^2$$

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

Lemma

1. $\mathbb{E} \left\| P - P_n \right\|_{(\ell_F)^\mu} \lesssim \sqrt{\mu \frac{U_n(F(\mu))}{n}}$.
An example of computation of the fixed points $\lambda_\epsilon^*$ and $\mu^*$

Computation of $\lambda_\epsilon^*$ and $\mu^*$ in the case of the Regression model with quadratic loss:

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

and

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

Complexity measure of $F$:

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

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

**Lemma**

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

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

22 / 56
An example of computation of the fixed points $\lambda^*_\epsilon$ and $\mu^*$

Then combining

1. $\mathbb{E}\|P - P_n\|_{V(H)\lambda} \leq \sum_{i=0}^{\infty} 2^{-i} \mathbb{E}\|P - P_n\|_{H_{2i+1}\lambda}$ for $H = \ell_F, \mathcal{L}_F$
An example of computation of the fixed points $\lambda^*_\epsilon$ and $\mu^*$

Then combining

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

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

$\mathbb{E}\|P - P_n\|_{(L_F)_\mu} \lesssim \sqrt{(\mu + R^*) \frac{U_n(F(\mu))}{n}}$
Then combining
1. $\mathbb{E}\|P - P_n\|_{V(H)_{\lambda}} \leq \sum_{i=0}^{\infty} 2^{-i} \mathbb{E}\|P - P_n\|_{H_{2i+1}\lambda}$ 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 $\lambda^*_\epsilon$ and $\mu^*$

Then combining

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

Then combining
1. $\mathbb{E}\|P - P_n\|\nu(H)_\lambda \leq \sum_{i=0}^{\infty} 2^{-i} \mathbb{E}\|P - P_n\|H_{2i+1,\lambda}$ 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, $\lambda^*_\epsilon$ will be the square of $\mu^*$ (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

\( (X_1, Y_1), \ldots, (X_n, Y_n) : n \text{ i.i.d.} \sim (X, Y) \) random variables in \( \mathcal{X} \times \{0, 1\} \)
Classification model

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

- \(\ell : (f, (x, y)) \mapsto \mathbb{I}_{f(x) \neq y} : 0 - 1\)-loss function of \(f : \mathcal{X} \to \{0, 1\}\)
Classification model

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

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

- $R(f) = \mathbb{P}[f(X) \neq Y] :$ risk of $f$
Classification model

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

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

- \(R(f) = \mathbb{P}[f(X) \neq Y] : \text{risk of } f\)

- \(F\) a class of \(\{0, 1\}\)-valued functions; \(f^*_F \in \arg\min_{f \in F} R(f)\); \(f^* \in \arg\min_{f} R(f)\) (Bayes rule).
Classification model

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

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

- $R(f) = \mathbb{P}[f(X) \neq Y] : \text{risk of } f$

- $F$ a class of $\{0, 1\}$-valued functions; $f^*_F \in \text{argmin}_{f \in F} R(f)$; $f^* \in \text{argmin}_f R(f)$ (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

$$V = \max \left( N : \max_{x_1,\ldots,x_N \in X} \text{Card}\{(f(x_1),\ldots,f(x_N)) : f \in F\} = 2^N \right).$$
Oracle inequalities in classification

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

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

M.&N. If $\mathcal{P}\ell^2_f \leq B(\mathcal{P}\ell_f)^\beta$, $\forall f \in F$ ($0 \leq \beta \leq 1$) Bernstein condition then

$\forall x \geq 1$, w.p. $\geq 1 - 4e^{-x}$,

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

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

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

M.&N. If $P\mathcal{L}_f^2 \leq B(P\mathcal{L}_f)^\beta, \forall f \in F$ ($0 \leq \beta \leq 1$) Bernstein condition then

$$\forall x \geq 1, \text{ w.p. } \geq 1 - 4e^{-x},$$

$$R(\hat{f}_n^{(ERM)}) \leq \inf_{f \in F} R(f) + c_0 \left( \frac{xV \log(enB_1^{1/\beta}/V)}{n} \right)^{\frac{1}{2-\beta}}.$$

M.&N. If $P\mathcal{E}_f^2 \leq B(P\mathcal{E}_f)^\beta, \forall f \in F$ ($0 \leq \beta \leq 1$) Margin assumption then

$$\forall x \geq 1, \text{ w.p. } \geq 1 - 4e^{-x},$$

$$R(\hat{f}_n^{(ERM)}) - R(f^*) \leq (1+\epsilon) \inf_{f \in F} (R(f) - R(f^*)) + c_0 \left( \frac{xV \log(enB_1^{1/\beta}/V)}{n\epsilon} \right)^{\frac{1}{2-\beta}}.$$
Oracle inequalities in classification

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

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

M.&N. If $PL_f^2 \leq B(PL_f)^\beta, \forall f \in F$ ($0 \leq \beta \leq 1$) Bernstein condition then

$$\forall x \geq 1, \text{ w.p. } \geq 1 - 4e^{-x},$$

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

M.&N. If $PE_f^2 \leq B(PE_f)^\beta, \forall f \in F$ ($0 \leq \beta \leq 1$) Margin assumption then

$$\forall x \geq 1, \text{ w.p. } \geq 1 - 4e^{-x},$$

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

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

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

for exact oracle inequalities: \[ P \mathcal{L}_f^2 \leq B (P \mathcal{L}_f)^\beta, \forall f \in F \ (0 \leq \beta \leq 1). \]
The Margin-Bernstein conditions

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

$$(\text{hard to characterize from a geometrical point of view because the loss is not convex}).$$
The Margin-Bernstein conditions

1. for exact oracle inequalities: $\mathbb{P} L_f^2 \leq B \left( \mathbb{P} L_f \right)^\beta, \forall f \in F \ (0 \leq \beta \leq 1)$. 

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

(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: $\mathbb{P} \mathcal{E}_f^2 \leq B \left( \mathbb{P} \mathcal{E}_f \right)^\beta, \forall f \in F \ (0 \leq \beta \leq 1)$. 

The Margin-Bernstein conditions

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

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

(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:
\( P \mathcal{E}_f^2 \leq B(P \mathcal{E}_f)^\beta, \forall f \in F (0 \leq \beta \leq 1). \)

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

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

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

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

(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:
\( P \mathcal{E}_f^2 \leq B (P \mathcal{E}_f)^\beta, \forall f \in F \) (\( 0 \leq \beta \leq 1 \)).

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

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

3. For non-exact oracle inequalities: \( P \mathcal{L}_f^2 = P \ell_f \leq B P \ell_f, \forall f \).
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: \(Z_1, \ldots, Z_n\) i.i.d. \(\sim Z\) random variables in \(\mathcal{Z}\);
2. Loss function: \(\ell : (f, z) \mapsto \ell_f(z) \in \mathbb{R}\);
A problem in learning theory is given by

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

Example in regression: when we construct $\hat{f}_{\text{ERM}} \in \mathrm{Arg\,min}_{f \in F} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2$, we hope that $F$ will be chosen in such a way that $\hat{f}_{\text{ERM}}$ will be close to the oracle $f^* \in \mathrm{Arg\,min}_{f \in F} \mathbb{E}(Y - f(X))^2$. And, we hope that $f^* \in \mathrm{Arg\,min}_{f \in L_2(P_X)} \mathbb{E}(Y - f(X))^2$. 

Oracle inequalities for ERM
Oracle inequalities for RERM
Applications to $S_1$ and $\ell_1$ regularization
Oracle inequalities for PERM
A problem in learning theory is given by

1. Observations: $Z_1, \ldots, Z_n$ i.i.d. $\sim Z$ random variables in $\mathcal{Z}$;
2. Loss function: $\ell: (f, z) \mapsto \ell_f(z) \in \mathbb{R}$;
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).
A problem in learning theory is given by

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

Choosing a particular \( F \) means that we believe that an oracle \( f^*_F \) in \( F \) \( \left( R(f^*_F) = \min_{f \in F} R(f) \right) \) 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

\[
\hat{f}_{n}^{\text{ERM}} \in \text{Arg min}_{f \in F} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2, 
\]
Regularized Empirical risk minimization - Part 1

A problem in learning theory is given by

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

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

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

$$f_F^* \in \text{Arg min}_{f \in F} \mathbb{E}(Y - f(X))^2$$

(\Rightarrow Oracle inequalities)
A problem in learning theory is given by

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

\[ \hat{f}^{ERM}_n \in \text{Arg min}_{f \in F} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2, \]

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

\[ f^*_F \in \text{Arg min}_{f \in F} \mathbb{E}(Y - f(X))^2 \]

(\( \Rightarrow \) Oracle inequalities)

And, we hope that \( f^*_F \) will be close to the regression function \( f^* \):

\[ f^* \in \text{Arg min}_{f \in L^2(P_X)} \mathbb{E}(Y - f(X))^2. \]
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 $F$. 
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

$$\text{crit} : \mathcal{F} \subset L_2(P_Z) \longrightarrow \mathbb{R}$$

called a criterion. So that

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

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

called a criterion. So that

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

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

$$\text{crit} : \mathcal{F} \subset L_2(P_Z) \mapsto \mathbb{R}$$

called a criterion. So that

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

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

\[
\text{crit} : \mathcal{F} \subset L_2(P_Z) \mapsto \mathbb{R}
\]

called a **criterion**. So that

\[
\text{crit}(f) \text{ is small } \Rightarrow f \text{ has this property.}
\]

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

**Ex.2:** \( \mathcal{F} := \{ f_{\beta} = \langle \cdot , \beta \rangle : \beta \in \mathbb{R}^d \} \text{ and } \text{crit}(f_{\beta}) = |\text{Supp}(\beta)| ; \)
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

$$\text{crit} : \mathcal{F} \subset L_2(P_Z) \mapsto \mathbb{R}$$

called a criterion. So that

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

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

- $Z_1, \ldots, Z_n : n \text{ i.i.d.} \sim Z$ 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);
- $\ell : (f, z) \mapsto \ell_f(z) \in \mathbb{R}$ : a loss function
Regularized Empirical risk minimization procedure - Part 3

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

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

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

\[
\hat{f}_{n}^{\text{RERM}} \in \text{Arg min}_{f \in \mathcal{F}} (R_n(f) + \text{reg}(f)),
\]

(for instance, \( \text{reg}(f) = \lambda \text{crit}^\alpha(f) \))
Regularized Empirical risk minimization procedure - Part 3

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

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

\[
\hat{f}^{\text{RERM}}_n \in \text{Arg min }_{f \in \mathcal{F}} \left( R_n(f) + \text{reg}(f) \right),
\]

(for instance, \( \text{reg}(f) = \lambda \text{crit}^\alpha(f) \); \( \lambda \) (regularization parameter), \( \alpha \) : parameters to be chosen).
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);
- $\ell : (f, z) \mapsto \ell_f(z) \in \mathbb{R}$: a loss function
- $\mathcal{F}$ and $\text{crit} : \mathcal{F} \mapsto \mathbb{R}$

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

Regularized Empirical Risk Minimization (RERM):

$$
\hat{f}_n^{RERM} \in \text{Arg min}_{f \in \mathcal{F}} \left( R_n(f) + \text{reg}(f) \right),
$$

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

$$
R(\hat{f}_n^{RERM}) + \text{reg}(\hat{f}_n^{RERM}) \leq (1 + \epsilon) \min_{f \in \mathcal{F}} \left( R(f) + \text{reg}(f) \right).
$$
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);
- \( \ell : (f, z) \mapsto \ell_f(z) \in \mathbb{R} \) : a loss function
- \( \mathcal{F} \) and \( \text{crit} : \mathcal{F} \mapsto \mathbb{R} \)

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

\[
\hat{f}_n^{\text{RERM}} \in \text{Arg min}_{f \in \mathcal{F}} \left( R_n(f) + \text{reg}(f) \right),
\]

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

\[
R(\hat{f}_n^{\text{RERM}}) + \text{reg}(\hat{f}_n^{\text{RERM}}) \leq (1 + \epsilon) \min_{f \in \mathcal{F}} \left( R(f) + \text{reg}(f) \right).
\]

\( \epsilon = 0 \) : Exact oracle inequality;
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);
- $\ell : (f, z) \mapsto \ell_f(z) \in \mathbb{R}$ : a loss function
- $\mathcal{F}$ and $\text{crit} : \mathcal{F} \mapsto \mathbb{R}$

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

$$\hat{f}_n^{\text{RERM}} \in \text{Arg min}_{f \in \mathcal{F}} (R_n(f) + \text{reg}(f)),$$

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

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

1. $\epsilon = 0$ : Exact oracle inequality;
2. $\epsilon > 0$ : Non-exact oracle inequality.
The choice of the regularizing function \( \text{reg}(f) = \lambda \text{crit}^\alpha(f) \) is dictated by the complexity of the sequence of models \((F_r)_{r \geq 0}\) where

\[
F_r := \{ f \in \mathcal{F} : \text{crit}(f) \leq r \}.
\]
Exact and non-exact oracle inequalities for RERM - Part 1

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

\[
F_r := \{ f \in \mathcal{F} : \text{crit}(f) \leq r \}.
\]

For every \( r \geq 0 \):

- loss functions classes :

\[
\mathcal{L}_{F_r} := \{ \ell_f : f \in F_r \} \quad \text{and} \quad \mathbb{E}\| P_n - P \|_{\mathcal{V}(\mathcal{L}_{F_r})} \lambda^*_\epsilon(r) \leq (\epsilon/4)\lambda^*_\epsilon(r)
\]
Exact and non-exact oracle inequalities for RERM - Part 1

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

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

For every $r \geq 0$:

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

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

(Where $R(f_{F_r}^*) = \min_{f \in F_r} R(f)$).
Theorem (L. and Mendelson)

Assume that there are non-decreasing functions $\phi_n$ and $B$ such that

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

1
Theorem (L. and Mendelson)

Assume that there are non-decreasing functions $\phi_n$ 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 $\phi_n$ 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 \( \phi_n \) and \( B \) such that

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

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 \( (r, x) \in \mathbb{R}_+ \times \mathbb{R}_+^* \),

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

Let \( x > 0 \) and set

\[
\hat{f}_n^{RERM} \in \text{Arg min}_{f \in \mathcal{F}} \left( R_n(f) + \frac{1}{1 + \epsilon} \rho_n(\text{crit}(f) + 1, x) \right).
\]
Theorem (L. and Mendelson)

Assume that there are non-decreasing functions $\phi_n$ 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. $P\ell^2_f \leq B(r)P\ell^2_f + B^2(r)/n, \forall r \geq 0, f \in F_r.$

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

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

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

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

Assume that there are non-decreasing functions \( \phi_n \) 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

\[ \hat{f}_n^{\text{RERM}} \in \operatorname{Arg \min}_{f \in \mathcal{F}} \left( R_n(f) + \frac{1}{1 + \epsilon} \rho_n(\text{crit}(f) + 1, x) \right). \]

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

\[ R(\hat{f}_n^{\text{RERM}}) + \rho_n(\text{crit}(\hat{f}_n^{\text{RERM}}), x) \leq \inf_{f \in \mathcal{F}} \left[ (1 + 2\epsilon)R(f) + 2\rho_n(\text{crit}(f) + 1, x) \right]. \]
Theorem (Bartlett, Neeman and Mendelson)

Assume that there are non-decreasing functions $\phi_n$ and $B$ such that

1. $\left\| \max_{1 \leq i \leq n} \sup_{f \in F_r} f(Z_i) \right\|_{\psi_1} := b_n(\ell F_r) \leq \phi_n(r)$
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 $(r, x) \in \mathbb{R}_+ \times \mathbb{R}_+^*$,

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

Let $x > 0$ and set

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

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

$$R(\hat{f}_n^{RERM}) + \rho_n(\text{crit}(\hat{f}_n^{RERM}), x) \leq \inf_{f \in \mathcal{F}} \left[ 1 \times R(f) + 2\rho_n(\text{crit}(f) + 1, x) \right].$$
Conclusion on Exact and Non-exact oracle inequalities for RERM

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

$$F_r := \{ f \in \mathcal{F} : \text{crit}(f) \leq r \}.$$
Conclusion on Exact and Non-exact oracle inequalities for RERM

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

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

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

$$\ell_{F_r} := \{ \ell_f : f \in F_r \} \text{ and } \mathbb{E} \| P_n - P \|_{\nu(\ell_{F_r})} \lambda^*_\epsilon(r) \leq (\epsilon/4) \lambda^*_\epsilon(r)$$
Conclusion on Exact and Non-exact oracle inequalities for RERM

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

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

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

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

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

  $$\mathcal{L}_{F_r} := \{ \ell_f - \ell_{f_{F_r}}^* : f \in F_r \} \text{ and } \mathbb{E} \| P_n - P \|_{\nu(\mathcal{L}_{F_r})} \mu^*_c(r) \leq \mu^*(r)/8.$$
Conclusion on Exact and Non-exact oracle inequalities for RERM

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

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

- **loss functions classes**: for all $r > 0$,
  $$\ell_{F_r} := \{ \ell_f : f \in F_r \} \quad \text{and} \quad \mathbb{E}\|P_n - P\|_V(\ell_{F_r})_{\lambda^*_\epsilon(r)} \leq (\epsilon/4)\lambda^*_\epsilon(r).$$

- **excess loss functions classes**: for all $r > 0$,
  $$\mathcal{L}_{F_r} := \{ \ell_f - \ell_{f_{F_r}}^* : f \in F_r \} \quad \text{and} \quad \mathbb{E}\|P_n - P\|_V(\mathcal{L}_{F_r})_{\mu^*_\epsilon(r)} \leq \mu^*_\epsilon(r)/8.$$

- RERM with regularizing function $\text{reg}(f) \gtrsim \lambda^*_\epsilon(\text{crit}(f))$.
Conclusion on Exact and Non-exact oracle inequalities for RERM

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

\[ F_r := \{ f \in \mathcal{F} : \text{crit}(f) \leq r \}. \]

- **loss functions classes**: for all \( r > 0 \),
  \[ \ell_{F_r} := \{ \ell_f : f \in F_r \} \text{ and } \mathbb{E}\|P_n - P\|_{\mathcal{V}(\ell_{F_r})_{\lambda^*_\epsilon(r)}} \leq (\epsilon/4)\lambda^*_\epsilon(r) \]

- **excess loss functions classes**: for all \( r > 0 \),
  \[ \mathcal{L}_{F_r} := \{ \ell_f - \ell_{F_r}^* : f \in F_r \} \text{ and } \mathbb{E}\|P_n - P\|_{\mathcal{V}(\mathcal{L}_{F_r})_{\mu^*(r)}} \leq \mu^*(r)/8. \]

- RERM with regularizing function \( \text{reg}(f) \gtrsim \lambda^*_\epsilon(\text{crit}(f)) \implies \)
  Non-exact oracle inequality ;
Conclusion on Exact and Non-exact oracle inequalities for RERM

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

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

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

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

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

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

1. RERM with regularizing function $\text{reg}(f) \gtrsim \lambda^*_\varepsilon(\text{crit}(f))$ \implies Non-exact oracle inequality;
2. RERM with regularizing function $\text{reg}(f) \gtrsim \mu^*(\text{crit}(f))$
Conclusion on Exact and Non-exact oracle inequalities for RERM

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

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

- **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 $r > 0$,
  $$\mathcal{L}_{F_r} := \{ \ell_f - \ell_{f^*_{F_r}} : f \in F_r \} \text{ and } \mathbb{E} \| P_n - P \| V(\mathcal{L}_{F_r})_{\mu^*_\epsilon(r)} \leq \mu^*_\epsilon(r)/8.$$

1. RERM with regularizing function $\text{reg}(f) \gtrsim \lambda^*_\epsilon(\text{crit}(f)) \implies$ Non-exact oracle inequality;
2. RERM with regularizing function $\text{reg}(f) \gtrsim \mu^*(\text{crit}(f)) \implies$ 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 $\| \cdot \|_H$ or in [Bartlett, Neeman, Mendelson] : regularization by $\log \| \cdot \|_H$ up to $\| \cdot \|_2_H$. 

35 / 56
Conclusion on Exact and Non-exact oracle inequalities for RERM

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

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

- **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 $r > 0$,

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

1. **RERM with regularizing function** $\text{reg}(f) \gtrsim \lambda^*_\epsilon(\text{crit}(f))$ $\implies$ Non-exact oracle inequality;
2. **RERM with regularizing function** $\text{reg}(f) \gtrsim \mu^*_\epsilon(\text{crit}(f))$ $\implies$ 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 $\text{crit} : \mathcal{F} \mapsto \mathbb{R}$. We consider the models $(F_r)_{r \geq 0}$:

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

- **loss functions classes**: for all $r > 0$,
  $$\ell_{F_r} := \{ \ell_f : f \in F_r \} \text{ and } \mathbb{E}\|P_n - P\|_{\nu(\ell_{F_r}) \lambda_{\epsilon}^*(r)} \leq (\epsilon/4) \lambda_{\epsilon}^*(r)$$

- **excess loss functions classes**: for all $r > 0$,
  $$\mathcal{L}_{F_r} := \{ \ell_f - \ell_{f_r^*} : f \in F_r \} \text{ and } \mathbb{E}\|P_n - P\|_{\nu(\mathcal{L}_{F_r}) \mu_{\epsilon}^*(r)} \leq \mu_{\epsilon}^*(r)/8.$$

1. **RERM with regularizing function** $\text{reg}(f) \gtrless \lambda_{\epsilon}^*(\text{crit}(f)) \implies$ Non-exact oracle inequality;
2. **RERM with regularizing function** $\text{reg}(f) \gtrless \mu_{\epsilon}^*(\text{crit}(f)) \implies$ 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 $\| \cdot \|^2_\mathcal{H}$.
Applications in matrix completion
Example in matrix completion

Model:

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

Model:

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

Model:

- \( Z_1 = (Y_1, X_1), \ldots, Z_n = (Y_n, X_n) : n \) i.i.d. random variables in \( \mathbb{R} \times \mathbb{R}^{m \times T} \);
- \( \ell(q) : \mathbb{R}^{m \times T} \times \mathbb{R} \times \mathbb{R}^{m \times T} \mapsto \mathbb{R} \) such that
  \[ \ell_A^q(Y, X) = |Y - \langle A, X \rangle|^q \] where \( \langle A, X \rangle = \text{Tr}(A^T X) \) 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:

- \( Z_1 = (Y_1, X_1), \ldots, Z_n = (Y_n, X_n) : n \) i.i.d. random variables in \( \mathbb{R} \times \mathbb{R}^{m \times T} \);
- \( \ell(q) : \mathbb{R}^{m \times T} \times \mathbb{R} \times \mathbb{R}^{m \times T} \rightarrow \mathbb{R} \) such that
  \[ \ell_A(Y, X) = |Y - \langle A, X \rangle|^q \] where \( \langle A, X \rangle = \text{Tr}(A^T X) \) 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:
- $Z_1 = (Y_1, X_1), \ldots, Z_n = (Y_n, X_n)$: $n$ i.i.d. random variables in $\mathbb{R} \times \mathbb{R}^{m \times T}$;
- $\ell(q) : \mathbb{R}^{m \times T} \times \mathbb{R} \times \mathbb{R}^{m \times T} \rightarrow \mathbb{R}$ such that
  $\ell_{A}^{(q)}(Y, X) = |Y - \langle A, X \rangle|^q$ where $\langle A, X \rangle = \text{Tr}(A^T X)$ 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
  $R^{(q)}(\hat{A}_n) = \mathbb{E}[|Y - \langle \hat{A}_n, X \rangle|^q | D]$. 
Example in matrix completion

Model:

- \( Z_1 = (Y_1, X_1), \ldots, Z_n = (Y_n, X_n) : n \text{ i.i.d. random variables in } \mathbb{R} \times \mathbb{R}^{m \times T} ; \)
- \( \ell(q) : \mathbb{R}^{m \times T} \times \mathbb{R} \times \mathbb{R}^{m \times T} \to \mathbb{R} \) such that \( \ell_A(Y, X) = |Y - \langle A, X \rangle|^q \) where \( \langle A, X \rangle = \text{Tr}(A^T X) \) 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 \( R^{(q)}(\hat{A}_n) = \mathbb{E}[|Y - \langle \hat{A}_n, X \rangle|^q|\mathcal{D}] \).

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

\[ \text{Model:} \]
\[ Z_1 = (Y_1, X_1), \ldots, Z_n = (Y_n, X_n) : n \text{ i.i.d. random variables in } \mathbb{R} \times \mathbb{R}^{m \times T} ; \]

\[ \ell(q) : \mathbb{R}^{m \times T} \times \mathbb{R} \times \mathbb{R}^{m \times T} \to \mathbb{R} \text{ such that } \ell_A(Y, X) = |Y - \langle A, X \rangle|^q \text{ where } \langle A, X \rangle = \text{Tr}(A^T X) \text{ and } q \geq 2. \]

\[ \text{Notation:} \]

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

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

\[ \text{The } L_q \text{-risk of a statistic } \hat{f}_n = \langle \cdot, \hat{A}_n \rangle \text{ is } R^{(q)}(\hat{A}_n) = \mathbb{E}[|Y - \langle \hat{A}_n, X \rangle|^q|\mathcal{D}]. \]

\[ \text{Problem: } mT \gg n \text{ (more variables than observations) but we believe that } Y \approx \langle A_0, X \rangle \text{ where } A_0 \text{ is of low rank } \text{rank}(A_0) < n \text{ (This is not an assumption!)} \]
Example in matrix completion

Model:
- $Z_1 = (Y_1, X_1), \ldots, Z_n = (Y_n, X_n) : n$ i.i.d. random variables in $\mathbb{R} \times \mathbb{R}^{m \times T}$;
- $\ell(q) : \mathbb{R}^{m \times T} \times \mathbb{R} \times \mathbb{R}^{m \times T} \rightarrow \mathbb{R}$ such that $\ell_A^q(Y, X) = |Y - \langle A, X \rangle|^q$ where $\langle A, X \rangle = \text{Tr}(A^T X)$ 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 $R^{(q)}(\hat{A}_n) = \mathbb{E}[|Y - \langle \hat{A}_n, X \rangle|^q|D]$.

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

$\mathcal{F} := \{\langle \cdot, A \rangle : A \in \mathbb{R}^{m \times T}\}$ and $\text{crit}(A) = \text{rank}(A)$. 
Matrix Completion - Convexification

$A \mapsto \text{rank}(A)$ is not convex $\implies$ not possible to use it in practice as a regularizing function.
Matrix Completion - Convexification

\[ A \mapsto \text{rank}(A) \text{ is not convex} \implies \text{not possible to use it in practice as a regularizing function.} \]

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

\( A \mapsto \text{rank}(A) \) is not convex \( \implies \) not possible to use it in practice as a regularizing function.

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

\( \implies \) We use the nuclear norm as a criterion: \( \text{crit}(A) = \| A \|_{S_{1}} \).
Matrix Completion - Convexification

$A \mapsto \text{rank}(A)$ is not convex $\implies$ not possible to use it in practice as a regularizing function.

**Convexification**: The convex envelope of $\text{rank}(\cdot)$ on

$\{A \in \mathbb{R}^{m \times T} : \|A\|_{S\infty} \leq 1\}$ is the nuclear norm ($\|A\|_{S1} = \|\text{spec}(A)\|_{\ell_{1}^{m \times T}}$).

$\implies$ We use the nuclear norm as a criterion: $\text{crit}(A) = \|A\|_{S1}$.

**Bibliography**:

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

\[ A \mapsto \text{rank}(A) \text{ is not convex} \implies \text{not possible to use it in practice as a regularizing function.} \]

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

\[ \implies \text{We use the nuclear norm as a criterion: } \text{crit}(A) = \|A\|_{S_1}. \]

**Bibliography**:

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

2. Tsybakov, Rohde, Koltchinskii, Lounici, Negahban, Wainright, Bach,\ldots (statistical point of view).
Matrix Completion - Application of the general result

\[ F_r := \{ A \in \mathbb{R}^{m \times T} : \text{crit}(A) \leq r \} = rB_{S_1} \]
Matrix Completion - Application of the general result

\[ Fr := \{ A \in \mathbb{R}^{m \times T} : \text{crit}(A) \leq r \} = rB_{s_1} \]

For non-exact oracle inequalities for RERM:

\[ \lambda^*_\epsilon(r) := \inf \left( \lambda > 0 : \mathbb{E} \| P - P_n \|_{V(\ell^{(q)}_{F_r})_\lambda} \leq (\epsilon/4) \lambda \right). \]

where \( \ell^{(q)}_{F_r} := \{ \ell^{(q)}_A : \| A \|_{s_1} \leq r \} \) and \( \ell^{(q)}_A(y, x) = |y - \langle x, A \rangle|^q \).
Matrix Completion - Application of the general result

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

For non-exact oracle inequalities for RERM:

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

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:

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

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)

\[ U_n = \mathbb{E} \gamma_2^2(\tilde{P}_\sigma F, \ell_\infty^n) \text{ where } P_\sigma F = \{(f(X_1), \cdots, f(X_n)) : f \in F\}. \]
Computation of the fixed point

Lemma (L. and Mendelson)

\[ U_n = \mathbb{E} \gamma_2^2(\widehat{P}_\sigma F, \ell^n) \text{ where } P_\sigma F = \{(f(X_1), \ldots, f(X_n)) : f \in F\}. \]

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

**Lemma (L. and Mendelson)**

\[ U_n = \mathbb{E} \gamma_2^2(\widetilde{P}_\sigma F, \ell_n^\infty) \text{ where } P_\sigma F = \{ (f(X_1), \cdots, f(X_n)) : f \in F \}. \]

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

For \( q>2 \)
\[ \mathbb{E} \| P - P_n \|_{(\ell_F^q)_\mu} \leq \max \left[ \sqrt{\mu \frac{U_n}{n}} \sqrt{(M \log n)^{1-2/q}}, \frac{U_n}{n} (M \log n)^{1-2/q}, \frac{M \log n}{n} \right] \]

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 $q \geq 2$, $\| Y \|_{\psi_q}$, $\| X \|_{s_2} \leq K(mT)$ 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

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

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

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

$$R^{(q)}(\hat{A}_n) \leq \inf_{A \in \mathcal{M}_{m \times T}} \left( (1 + 2\epsilon) R^{(q)}(A) + c_1 \lambda(n, mT, x) \left(1 + \frac{\|A\|_{S_1}^q}{n \epsilon^2} \right) \right).$$
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 $\implies \text{crit}(A) = \|A\|_{S_1}$;
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 $\Rightarrow \text{crit}(A) = \|A\|_{S_1}$;
- 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 $\Rightarrow \text{crit}(A) = \|A\|_{S_1}$;
- and, the singular values of $A_0$ are well-spread $\Rightarrow \text{crit}(A) = r_1\|A\|_{S_1} + r_2\|A\|_{S_2}^2$;
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 $\implies \text{crit}(A) = \|A\|_{S_1}$;
- and, the singular values of $A_0$ are well-spread $\implies \text{crit}(A) = r_1 \|A\|_{S_1} + r_2 \|A\|_{S_2}^2$;
- 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 $\implies \text{crit}(A) = \|A\|_{S_1}$;
- and, the singular values of $A_0$ are well-spread $\implies \text{crit}(A) = r_1 \|A\|_{S_1} + r_2 \|A\|_{S_2}^2$;
- and, $A_0$ has many zeroes $\implies \text{crit}(A) = r_1 \|A\|_{S_1} + r_2 \|A\|_{S_2}^2 + r_3 \|A\|_{\ell_1^m}$;
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 $\Rightarrow \text{crit}(A) = \|A\|_{S_1}$;

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

- and, $A_0$ has many zeroes $\Rightarrow \text{crit}(A) = r_1\|A\|_{S_1} + r_2\|A\|_{S_2}^2 + r_3\|A\|_{\ell_1^m}$;

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

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

$$\hat{A}_n \in \arg\min_{A \in \mathcal{M}_{m,T}} \left\{ R_n^{(2)}(A) + \frac{\lambda_{n,mT,x}}{\sqrt{n}} (r_1 \|A\|_S + r_2 \|A\|^2_S + r_3 \|A\|_1) \right\}$$
Theorem (Gaïffas and L.)

Assume that $\|Y\|_{\psi_2}, \|X\|_{\ell_2} \leq K(mT)$ for some constant $K(mT)$ which depends only on the product $mT$. Fix any $x, r_1, r_2, r_3 > 0$, and consider

$$\hat{A}_n \in \arg\min_{A \in \mathcal{M}_{m,T}} \left\{ R_n^{(2)}(A) + \frac{\lambda_n, mT, x}{\sqrt{n}} (r_1 \|A\|_{S_1} + r_2 \|A\|_{S_2}^2 + r_3 \|A\|_{1}) \right\}$$

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

$$R_n^{(2)}(\hat{A}_n) \leq \inf_{A \in \mathcal{M}_{m,T}} \left\{ R_n^{(2)}(A) + \frac{\lambda_n, mT, x}{\sqrt{n}} (1 + r_1 \|A\|_{S_1} + r_2 \|A\|_{S_2}^2 + r_3 \|A\|_{1}) \right\}$$
Applications to $\ell_1$-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 $\mathbb{R} \times \mathbb{R}^d$;
- $\ell(q) : \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d \mapsto \ell(q)(\beta, (y, x)) = \ell^{(q)}_\beta(y, x) = |y - \langle \beta, x \rangle|^q$. 

$\ell^{(q)}(\beta, (y, x))$ is the $L^q$-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 \(\mathbb{R} \times \mathbb{R}^d\);
- \(\ell(q) : \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d \mapsto \ell(q) (\beta, (y, x)) = \ell_\beta(y, x) = |y - \langle \beta, x \rangle|^q\).

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 \(\mathbb{R} \times \mathbb{R}^d\);
- \(\ell(q) : \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d \mapsto \ell(q)(\beta, (y, x)) = \ell_{\beta}(y, x) = |y - \langle \beta, x \rangle|^q\).

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 $\mathbb{R} \times \mathbb{R}^d$;
- $\ell^{(q)} : \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d \mapsto \ell^{(q)}(\beta, (y, x)) = \ell^{(q)}_\beta(y, x) = |y - \langle \beta, x \rangle|^q$.

Notation:
- $\ell^{(q)}_\beta$: $L_q$-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: $d \gg n$ (more variables than observations) but we believe that $Y \approx \langle \beta_0, X \rangle$ where $\beta_0$ is of short support ($|\text{Supp}(\beta_0)| < n$) (This is not an assumption!)
Oracle inequalities for ERM
Oracle inequalities for RERM
Oracle inequalities for PERM

Applications to $S_1$ and $\ell_1$ regularization

Regression model

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

Notation:
- $\ell^{(q)}_\beta$: $L_q$-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: $d \gg n$ (more variables than observations) but we believe that $Y \approx \langle \beta_0, X \rangle$ where $\beta_0$ is of short support ($|\text{Supp}(\beta_0)| < n$) (This is not an assumption!)

$\mathcal{F} := \{\langle \cdot, \beta \rangle: \beta \in \mathbb{R}^d\}$ and $\text{crit}(\beta) = |\text{Supp}(\beta)| \Rightarrow$
Regression model

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

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 \(\beta_0\) is of short support (\(|\text{Supp}(\beta_0)| < n\)) (This is not an assumption!)
\(\mathcal{F} := \{\langle \cdot, \beta \rangle : \beta \in \mathbb{R}^d\}\) and \(\text{crit}(\beta) = |\text{Supp}(\beta)| \Rightarrow \text{Convexification}\)

\[\text{crit}(\beta) = \|\beta\|_1\]
Oracle inequality 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 \( \| Y \|_{\psi_q}, \| X \|_{\ell_\infty^d} \|_{\psi_q} \leq c_d \).
Let $q \geq 2$. Assume that there exists some constant $c_d > 0$ (which may depend only on $d$) such that $\|Y\|_{\psi_q}, \|\|X\|_{\ell_\infty}\|_{\psi_q} \leq c_d$. 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 inequality 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 \( \| Y \|_{\psi_q}, \| X \|_{\ell_\infty} \leq c_d \). 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

\[
\hat{\beta}_n \in \arg\min_{\beta \in \mathbb{R}^d} \left( R_n^{(q)}(\beta) + \lambda(n, d, x) \frac{\|\beta\|_{\ell_1}^q}{n \epsilon^2} \right).
\]
Oracle inequality 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

$$\hat{\beta}_n \in \arg \min_{\beta \in \mathbb{R}^d} \left( R_n^{(q)}(\beta) + \lambda(n, d, x) \frac{\|\beta\|_{\ell_1}^q}{n\epsilon^2} \right).$$

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

$$R^{(q)}(\hat{\beta}_n) \leq \inf_{\beta \in \mathbb{R}^d} \left( (1 + 2\epsilon) R^{(q)}(\beta) + c_1 \lambda(n, d, x) \frac{(1 + \|\beta\|_{\ell_1}^q)}{n\epsilon^2} \right).$$
Oracle inequalities for penalized estimators
Model selection framework

\( \mathcal{M} : \) a countable collection of models.
Model selection framework

\[ \mathcal{M} : \text{a countable collection of models.} \]

\[ \forall m \in \mathcal{M}, \hat{f}_m \in \arg\min_{f \in m} R_n(f), \]
Model selection framework

\[ M: \text{a countable collection of models.} \]

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

\[ \hat{m} \in \text{argmin}_{m \in M} \left( R_n(\hat{f}_m) + \text{pen}(m) \right). \]
Model selection framework

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

1. $\forall m \in \mathcal{M}, \hat{f}_m \in \arg\min_{f \in m} R_n(f)$,

2. construct $\text{pen} : \mathcal{M} \to \mathbb{R}_+$ and define

$$\hat{m} \in \arg\min_{m \in \mathcal{M}} (R_n(\hat{f}_m) + \text{pen}(m)).$$

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

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

1. \( \forall m \in \mathcal{M}, \hat{f}_m \in \arg \min_{f \in m} R_n(f), \)

2. construct \( \text{pen} : \mathcal{M} \rightarrow \mathbb{R}_+ \) and define

\[
\hat{m} \in \arg \min_{m \in \mathcal{M}} (R_n(\hat{f}_m) + \text{pen}(m)).
\]

3. oracle inequalities for the penalized estimator \( \hat{f}_m \).

construction of \( \text{pen} \) depends on the type of oracle inequality that we want to prove:
Model selection framework

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

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

$$\hat{m} \in \arg\min_{m \in \mathcal{M}} (R_n(\hat{f}_m) + \text{pen}(m)).$$

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

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

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

where we assume that there exists $f^*_m \in \arg\min_{f \in m} R(f)$ for any $m \in \mathcal{M}$ (and $f^* \in \arg\min_f R(f)$).
Four fixed points

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

$$\mathbb{E}\|P_n - P\|_{\nu(\ell_m)\lambda^*_\eta(m)} \leq (\eta/4)\lambda^*_\eta(m).$$
### Three fixed points

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

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

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

\[
\mathbb{E}\|P_n - P\|\nu(\ell_m)\lambda^*_\eta(m) \leq (\eta/4)\lambda^*_\eta(m).
\]
Three fixed points

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

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

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

\[
\mathbb{E} \| P_n - P \| \nu(\epsilon_m) \nu^*_{\eta}(m) \leq (\eta/4) \nu^*_{\eta}(m).
\]

3. For exact oracle inequalities: \( \forall m \in \mathcal{M} \)

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

where \( \mathcal{L}_m = \{ \ell_f - \ell_{f^*_m} : f \in m \} \) and \( f^*_m \in \text{argmin}_{f \in m} R(f) \).
Non-exact oracle inequalities for the penalized estimator

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

$$\| \max_{1 \leq i \leq n} \sup_{f \in m} \ell_f(Z_i) \|_{\psi_1} \leq \phi_n(m)$$

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 $(m, x) \in \mathcal{M} \times \mathbb{R}^*_+$,

$$\rho_n^\ell(m, x) \geq \max \left( \lambda_\eta^*(m), c_0 \frac{(\phi_n(m) + B_n(m)/\eta)(x + 1)}{n \eta} \right).$$
Non-exact oracle inequalities for the penalized estimator

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

$$\| \max_{1 \leq i \leq n} \sup_{f \in m} \ell_f(Z_i) \|_{\psi_1} \leq \phi_n(m)$$

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 $(m, x) \in \mathcal{M} \times \mathbb{R}_+^*$,

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

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

$$\text{pen}^\ell(m) = \rho_n^\ell(m, x + x_m)$$

and the penalized estimator $\hat{f}_m$ associated with this penalty function.
Non-exact oracle inequalities for the penalized estimator

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

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

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

$$\rho^\ell_n(m, x) \geq \max \left( \lambda^*_\eta(m), c_0 \frac{(\phi_n(m) + B_n(m)/\eta)(x + 1)}{n\eta} \right).$$ 

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 $\text{pen}^\ell(m) = \rho^\ell_n(m, x + x_m)$ and the penalized estimator $\hat{f}_m$ associated with this penalty function. Then, with probability greater than $1 - c_2 e^{-x}$,

$$R(\hat{f}_m) \leq \frac{1 + \eta}{1 - \eta} \inf_{m \in \mathcal{M}} \left( \inf_{f \in m} P\ell_f + \text{pen}^\ell(m) \right).$$
Non-exact oracle inequalities for the penalized estimator

\[ \text{pen}^\ell(m) = \max \left( \lambda^*_\eta(m), c_0 \frac{(\phi_n(m) + B_n(m)/\eta)(x + x_m + 1)}{m\eta} \right) \sim \lambda^*_\eta(m) \]

where

\[ \mathbb{E}\|P_n - P\|_{\nu(\ell_m)\lambda^*_\eta(m)} \leq (\eta/4)\lambda^*_\eta(m). \]
Oracle inequality for the estimation problem

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

$$
\| \max_{1 \leq i \leq n} \sup_{f \in m} \mathcal{E}_f(Z_i) \|_{\psi_1} \leq \phi_n(m) \quad \text{and} \quad PE_f^2 \leq B_n(m)(PE_f)^\beta + B_n^2(m)/n.
$$
Oracle inequality for the estimation problem

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

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

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

$$
\rho_n^\mathcal{E}(m, x) \geq \max \left( \nu_{\eta}^*(m), c_2(B_n(m) + \phi_n(m)) \left( \frac{x + 1}{n\eta} \right)^{\frac{1}{2-\beta}} \right).
$$
Oracle inequality for the estimation problem

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

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

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

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

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 $\text{pen}^\mathcal{E}(m) = \rho_n^\mathcal{E}(m, x + x_m)$ and the penalized estimator $\hat{f}_m$ associated with this penalty function.
Oracle inequality for the estimation problem

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

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

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

$$\rho^\mathcal{E}_n(m, x) \geq \max \left( \nu^*_\eta(m), c_2(B_n(m) + \phi_n(m)) \left( \frac{x + 1}{mn} \right)^{\frac{1}{2-\beta}} \right).$$

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

$$\text{pen}^\mathcal{E}(m) = \rho^\mathcal{E}_n(m, x + x_m)$$

and the penalized estimator $\hat{f}_m$ associated with this penalty function. Then, with probability greater than $1 - c_3 e^{-x}$,

$$R(\hat{f}_m) - R(f^\star) \leq \frac{1 + \eta}{1 - \eta} \inf_{m \in \mathcal{M}} \left( \inf_{f \in m} (R(f) - R(f^\star)) + \text{pen}^\mathcal{E}(m) \right).$$
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

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

where

$$\mathbb{E}\|P_n - P\|_{\nu(\epsilon_m)\nu_\eta^*(m)} \leq (\eta/4)\nu_\eta^*(m).$$
Exact oracle inequality for the penalized estimator

Take $\mathcal{M} = (m_n)_{n \in \mathbb{N}}$ s.t. $m_0 \subset m_1 \subset m_2 \subset \cdots$. 
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 $\phi_n$ and $B_n$ such that for every $m \in \mathcal{M}$ and every $f \in m$,

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

where $\mathcal{L}_{m,f} = \ell_f - \ell_{f^*}$. 
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 \( \phi_n \) and \( B_n \) such that for every \( m \in \mathcal{M} \) and every \( f \in m \),

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

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

\[
\rho_{n}^\mathcal{C}(m, x) \geq \max \left( \mu_{1/2}^*(m), (B_n(m) + \phi_n(m)) \left( \frac{x + 1}{n} \right)^{\frac{1}{2 - \beta}} \right).
\]
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 $\phi_n$ and $B_n$ such that for every $m \in \mathcal{M}$ and every $f \in m$,

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

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

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

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

$$\text{pen}^\mathcal{L}(m) = (7/2)\rho_n^\mathcal{L}(m, x + x_m)$$

and the penalized estimator $\hat{f}_m$ 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 $\phi_n$ and $B_n$ such that for every $m \in \mathcal{M}$ and every $f \in m$,

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

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

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

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 $\text{pen}^\mathcal{L}(m) = (7/2)\rho_n^\mathcal{L}(m, x + x_m)$ and the penalized estimator $\hat{f}_m$ associated with this penalty function. Then, with probability greater than $1 - c_1 e^{-x}$,

$$
R(\hat{f}_m) \leq \inf_{m \in \mathcal{M}} \left( \inf_{f \in m} R(f) + (18/7)\text{pen}^\mathcal{L}(m) \right).
$$
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

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

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

where

$$\mathbb{E}\|P_n - P\|_{\mathcal{L}_m, \mu_{1/2}(m)} \leq \left( \frac{1}{8} \right) \mu_{1/2}(m)$$

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

Thanks!!