Interconnexions entre complexité, concentration et géométrie en théorie de l'apprentissage et applications aux problèmes en grandes dimensions

Guillaume Lecué

CNRS, Laboratoire d'analyse mathématiques appliquées, Université Paris-Est Marne-la-vallée

8 Décembre 2011

modéle d'entrées/sorties i.i.d. pour la perte quadratique

- $(X_1, Y_1), \ldots, (X_n, Y_n)$: n i.i.d.$\sim (X, Y)$ variable aléatoires à valeurs dans $\mathcal{X} \times \mathbb{R}$

modéle d'entrées/sorties i.i.d. pour la perte quadratique

- $(X_1, Y_1), \ldots, (X_n, Y_n)$: n i.i.d.$\sim (X, Y)$ variable aléatoires à valeurs dans $\mathcal{X} \times \mathbb{R}$

- $\ell : (f, (x, y)) \longmapsto \ell_f(x, y) = (y - f(x))^2$: fonction de perte quadratique de $f : \mathcal{X} \to \mathbb{R}$

modéle d'entrées/sorties i.i.d. pour la perte quadratique

- $(X_1, Y_1), \ldots, (X_n, Y_n)$: n i.i.d.$\sim (X, Y)$ variable aléatoires à valeurs dans $\mathcal{X} \times \mathbb{R}$

- $\ell : (f, (x, y)) \longmapsto \ell_f(x, y) = (y - f(x))^2$: fonction de perte quadratique de $f : \mathcal{X} \to \mathbb{R}$

- $R(f) = \mathbb{E}\ell_f(X, Y) = \mathbb{E}(Y - f(X))^2$: risque quadratique

modéle d'entrées/sorties i.i.d. pour la perte quadratique

- $(X_1, Y_1), \ldots, (X_n, Y_n)$: n i.i.d.$\sim (X, Y)$ variable aléatoires à valeurs dans $\mathcal{X} \times \mathbb{R}$
- $\ell : (f, (x, y)) \longmapsto \ell_f(x, y) = (y - f(x))^2$: fonction de perte quadratique de $f : \mathcal{X} \to \mathbb{R}$
- $R(f) = \mathbb{E}\ell_f(X, Y) = \mathbb{E}(Y - f(X))^2$: risque quadratique
- Le risque quadratique d'une statistique $\hat{f}_n$ est

$$R(\hat{f}_n) = \mathbb{E}\big[(Y - \hat{f}_n(X))^2 | \mathcal{D}\big]$$

où $\mathcal{D} := ((X_1, Y_1), \cdots, (X_n, Y_n))$.

▸ Modèle général

Trois problèmes d'agrégation

Les trois problémes d'agrégation (Nemirovski et Tsybakov)

Soit $F = \{f_1, \ldots, f_M\}$ une classe de fonctions $f_j : \mathcal{X} \to \mathbb{R}$ (dictionnaire).

Les trois problémes d'agrégation (Nemirovski et Tsybakov)

Soit $F = \{f_1, \ldots, f_M\}$ une classe de fonctions $f_j : \mathcal{X} \to \mathbb{R}$ (dictionnaire).

1. **Agrégation en Sélection de Modèles :** construire $\hat{f}_n$ tel que a.g.p.

$$R(\hat{f}_n) \leq \min_{f \in F} R(f) + \psi_n^{(MS)}(M)$$

Les trois problémes d'agrégation (Nemirovski et Tsybakov)

Soit $F = \{f_1, \ldots, f_M\}$ une classe de fonctions $f_j : \mathcal{X} \to \mathbb{R}$ (dictionnaire).

1. **Agrégation en Sélection de Modèles :** construire $\hat{f}_n$ tel que a.g.p.

$$R(\hat{f}_n) \leq \min_{f \in F} R(f) + \psi_n^{(MS)}(M)$$

2. **Agrégation Convexe :** construire $\hat{f}_n$ tel que a.g.p.

$$R(\hat{f}_n) \leq \min_{f \in \mathrm{Conv}(F)} R(f) + \psi_n^{(C)}(M)$$

Les trois problémes d'agrégation (Nemirovski et Tsybakov)

Soit $F = \{f_1, \ldots, f_M\}$ une classe de fonctions $f_j : \mathcal{X} \to \mathbb{R}$ (dictionnaire).

1. **Agrégation en Sélection de Modèles :** construire $\hat{f}_n$ tel que a.g.p.

$$R(\hat{f}_n) \leq \min_{f \in F} R(f) + \psi_n^{(MS)}(M)$$

2. **Agrégation Convexe :** construire $\hat{f}_n$ tel que a.g.p.

$$R(\hat{f}_n) \leq \min_{f \in \mathrm{Conv}(F)} R(f) + \psi_n^{(C)}(M)$$

3. **Agrégation Linéaire :** construire $\hat{f}_n$ tel que a.g.p.

$$R(\hat{f}_n) \leq \min_{f \in \mathrm{Span}(F)} R(f) + \psi_n^{(L)}(M)$$

Les vitesses optimales d'agrégation (Tsybakov)

Il existe deux constantes absolues $c_0, c_1 > 0$ telles que pour tout n (nombre d'observations) et tout M (taille du dictionnaire) :

Les vitesses optimales d'agrégation (Tsybakov)

Il existe deux constantes absolues $c_0, c_1 > 0$ telles que pour tout n (nombre d'observations) et tout M (taille du dictionnaire) :

1. **Agrégation en Sélection de Modèle :** il existe F tel que $|F| = M$ et pour tout $\hat{f}_n$ il existe (X, Y) tel que a.p. $\geq c_0$

$$R(\hat{f}_n) \geq \min_{f \in F} R(f) + c_1 \psi_n^{(MS)}(M); \qquad \psi_n^{(MS)}(M) = \frac{\log M}{n}$$

Les vitesses optimales d'agrégation (Tsybakov)

Il existe deux constantes absolues $c_0, c_1 > 0$ telles que pour tout n (nombre d'observations) et tout M (taille du dictionnaire) :

1. **Agrégation en Sélection de Modèle :** il existe F tel que $|F| = M$ et pour tout $\hat{f}_n$ il existe (X, Y) tel que a.p. $\geq c_0$

$$R(\hat{f}_n) \geq \min_{f \in F} R(f) + c_1 \psi_n^{(MS)}(M); \qquad \psi_n^{(MS)}(M) = \frac{\log M}{n}$$

2. **Agrégation Convexe :** il existe F tel que $|F| = M$ et pour tout $\hat{f}_n$ il existe (X, Y) tel que a.p. $\geq c_0$

$$R(\hat{f}_n) \geq \min_{f \in \mathrm{Conv}(F)} R(f) + c_1 \psi_n^{(C)}(M); \psi_n^{(C)}(M) = \begin{cases} M/n & M \leq \sqrt{n} \\ \sqrt{\frac{\log(eM/\sqrt{n})}{n}} & M \geq \sqrt{n} \end{cases}$$

Les vitesses optimales d'agrégation (Tsybakov)

Il existe deux constantes absolues $c_0, c_1 > 0$ telles que pour tout n (nombre d'observations) et tout M (taille du dictionnaire) :

1. **Agrégation en Sélection de Modèle :** il existe F tel que $|F| = M$ et pour tout $\hat{f}_n$ il existe (X, Y) tel que a.p. $\geq c_0$

$$R(\hat{f}_n) \geq \min_{f \in F} R(f) + c_1 \psi_n^{(MS)}(M); \qquad \psi_n^{(MS)}(M) = \frac{\log M}{n}$$

2. **Agrégation Convexe :** il existe F tel que $|F| = M$ et pour tout $\hat{f}_n$ il existe (X, Y) tel que a.p. $\geq c_0$

$$R(\hat{f}_n) \geq \min_{f \in \mathrm{Conv}(F)} R(f) + c_1 \psi_n^{(C)}(M); \psi_n^{(C)}(M) = \begin{cases} M/n & M \leq \sqrt{n} \\ \sqrt{\frac{\log(eM/\sqrt{n})}{n}} & M \geq \sqrt{n} \end{cases}$$

3. **Agrégation Linéaire :** il existe F tel que $|F| = M$ et pour tout $\hat{f}_n$ il existe (X, Y) tel que a.p. $\geq c_0$

$$R(\hat{f}_n) \geq \min_{f \in \mathrm{Span}(F)} R(f) + c_1 \psi_n^{(L)}(M); \qquad \psi_n^{(L)}(M) = \frac{M}{n}$$

Minimisation du risque empirique

Un candidat naturel pour les problèmes d'agrégation est la procédure de minimisation du risque empirique (MRE) definie par :

Minimisation du risque empirique

Un candidat naturel pour les problèmes d'agrégation est la procédure de minimisation du risque empirique (MRE) definie par :

1. **Agrégation en Sélection de Modèle :**

$$\hat{f}_n^{(MRE)} \in \underset{f \in F}{\text{argmin}}\, R_n(f) \text{ où } R_n(f) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2,$$

$R_n(\cdot)$ est appelé le risque empirique ; $(\mathbb{E}R_n(f) = R(f))$.

Minimisation du risque empirique

Un candidat naturel pour les problèmes d'agrégation est la procédure de minimisation du risque empirique (MRE) definie par :

1. **Agrégation en Sélection de Modèle :**

$$\hat{f}_n^{(MRE)} \in \operatorname*{argmin}_{f \in F} R_n(f) \text{ où } R_n(f) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2,$$

   $R_n(\cdot)$ est appelé le risque empirique ; $(\mathbb{E}R_n(f) = R(f))$.

2. **Agrégation Convexe :**

$$\hat{f}_n^{(MRE-C)} \in \operatorname*{argmin}_{f \in \mathrm{Conv}(F)} R_n(f).$$

Minimisation du risque empirique

Un candidat naturel pour les problèmes d'agrégation est la procédure de minimisation du risque empirique (MRE) definie par :

1. **Agrégation en Sélection de Modèle :**

$$\hat{f}_n^{(MRE)} \in \operatorname*{argmin}_{f \in F} R_n(f) \text{ où } R_n(f) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2,$$

   $R_n(\cdot)$ est appelé le risque empirique ; $(\mathbb{E}R_n(f) = R(f))$.

2. **Agrégation Convexe :**

$$\hat{f}_n^{(MRE-C)} \in \operatorname*{argmin}_{f \in \mathrm{Conv}(F)} R_n(f).$$

3. **Agrégation Linéaire :**

$$\hat{f}_n^{(MRE-L)} \in \operatorname*{argmin}_{f \in \mathrm{Span}(F)} R_n(f).$$

Minimisation du risque empirique

Un candidat naturel pour les problèmes d'agrégation est la procédure de minimisation du risque empirique (MRE) definie par :

1. **Agrégation en Sélection de Modèle :**

$$\hat{f}_n^{(MRE)} \in \operatorname*{argmin}_{f \in F} R_n(f) \text{ où } R_n(f) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2,$$

$R_n(\cdot)$ est appelé le risque empirique ; $(\mathbb{E}R_n(f) = R(f))$.

2. **Agrégation Convexe :**

$$\hat{f}_n^{(MRE-C)} \in \operatorname*{argmin}_{f \in \mathrm{Conv}(F)} R_n(f).$$

3. **Agrégation Linéaire :**

$$\hat{f}_n^{(MRE-L)} \in \operatorname*{argmin}_{f \in \mathrm{Span}(F)} R_n(f).$$

Question : le MRE est-il optimal pour les trois problèmes d'agrégation ?

MRE est optimal pour le problème d'agrégation Linéaire

MRE est optimal pour le problème d'agrégation Linéaire

Vladimir Koltchinksii : pour tout $n, M \geq 1$, tout F de taille M et tout couple (X, Y) tel que $\max_{f \in F} |f(X)|, |Y| \leq b$, pour tout $x > 0$, avec probabilité plus grande que $1 - 4\exp(-x)$,

MRE est optimal pour le problème d'agrégation Linéaire

Vladimir Koltchinksii : pour tout $n, M \geq 1$, tout F de taille M et tout couple (X, Y) tel que $\max_{f \in F} |f(X)|, |Y| \leq b$, pour tout $x > 0$, avec probabilité plus grande que $1 - 4\exp(-x)$,

$$R(\hat{f}_n^{(MRE-L)}) \leq \min_{f \in \mathrm{Span}(F)} R(f) + c_0 b^2 \max\left(\frac{M}{n}, \frac{x}{n}\right).$$

MRE est optimal pour le problème d'agrégation convexe

MRE est optimal pour le problème d'agrégation convexe

L. : pour tout $n, M \geq 1$, tout F de taille M et tout couple (X, Y) tel que $\max_{f \in F} |f(X)|, |Y| \leq b$, pour tout $x > 0$, avec probabilité plus grande que $1 - 4\exp(-x)$,

MRE est optimal pour le problème d'agrégation convexe

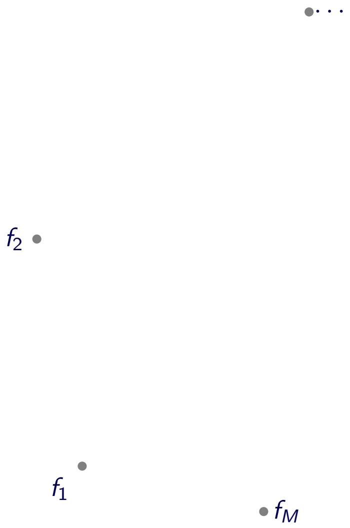
L. : pour tout $n, M \geq 1$, tout F de taille M et tout couple (X, Y) tel que $\max_{f \in F} |f(X)|, |Y| \leq b$, pour tout $x > 0$, avec probabilité plus grande que $1 - 4\exp(-x)$,

$$R(\hat{f}_n^{(MRE-C)}) \leq \min_{f \in \mathrm{Conv}(F)} R(f) + c_0 b^2 \max\left(\psi_n^{(C)}(M), \frac{x}{n}\right).$$

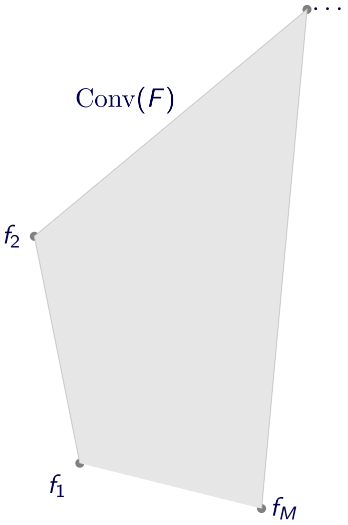
où

$$\psi_n^{(C)}(M) = \begin{cases} M/n & M \leq \sqrt{n} \\ \sqrt{\frac{\log(eM/\sqrt{n})}{n}} & M \geq \sqrt{n}. \end{cases}$$

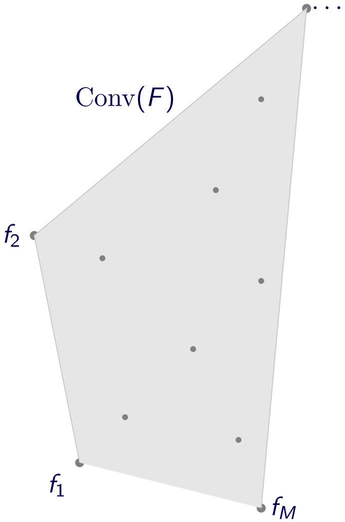
Brève preuve de l'optimalité du MRE pour l'agrégation convexe ($M \geq \sqrt{n}$)



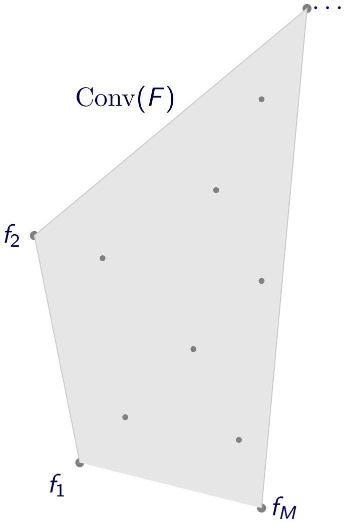
Brève preuve de l'optimalité du MRE pour l'agrégation convexe ($M \geq \sqrt{n}$)



Brève preuve de l'optimalité du MRE pour l'agrégation convexe ($M \geq \sqrt{n}$)



Brève preuve de l'optimalité du MRE pour l'agrégation convexe ($M \geq \sqrt{n}$)

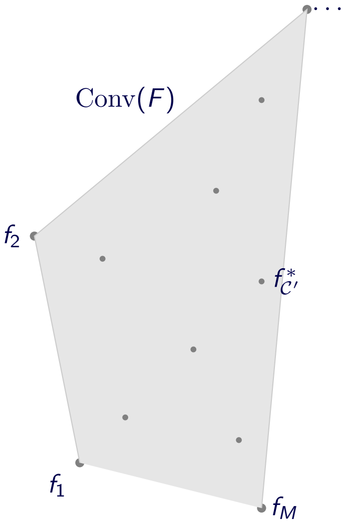


$$m = \left\lceil \sqrt{\frac{n}{\log(eM/\sqrt{n})}} \right\rceil$$

$$\mathcal{C}' = \left\{ \frac{1}{m} \sum_{j=1}^{m} h_j : h_j \in F \right\}$$

Carl-Maurey ; Nemirovskii ; Tsybakov

Brève preuve de l'optimalité du MRE pour l'agrégation convexe ($M \geq \sqrt{n}$)

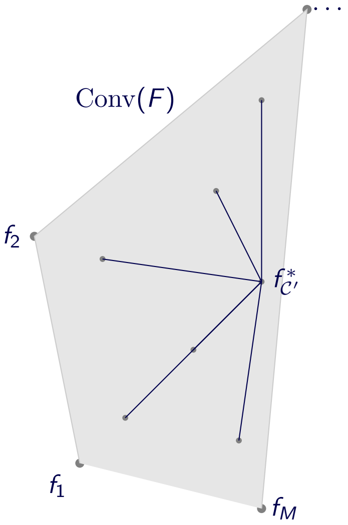


$$m = \left\lceil \sqrt{\frac{n}{\log(eM/\sqrt{n})}} \right\rceil$$

$$\mathcal{C}' = \left\{ \frac{1}{m} \sum_{j=1}^{m} h_j : h_j \in F \right\}$$

Carl-Maurey ; Nemirovskii ; Tsybakov

Brève preuve de l'optimalité du MRE pour l'agrégation convexe ($M \geq \sqrt{n}$)

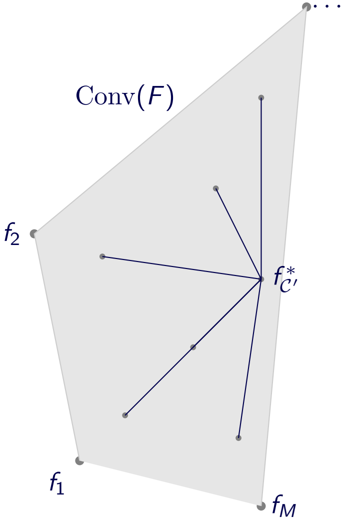


$$m = \left\lceil \sqrt{\frac{n}{\log(eM/\sqrt{n})}} \right\rceil$$

$$\mathcal{C}' = \left\{ \frac{1}{m} \sum_{j=1}^m h_j : h_j \in F \right\}$$

Carl-Maurey ; Nemirovskii ; Tsybakov

Brève preuve de l'optimalité du MRE pour l'agrégation convexe ($M \geq \sqrt{n}$)



$$m = \left\lceil \sqrt{\frac{n}{\log(eM/\sqrt{n})}} \right\rceil$$

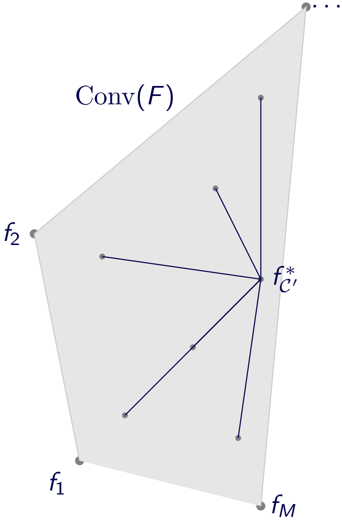
$$\mathcal{C}' = \left\{ \frac{1}{m} \sum_{j=1}^{m} h_j : h_j \in F \right\}$$

Carl-Maurey ; Nemirovskii ; Tsybakov

1) Isomorphie sur

$$\cup_{g \in \mathcal{C}'} [g, f^*_{\mathcal{C}'}]$$

Brève preuve de l'optimalité du MRE pour l'agrégation convexe ($M \geq \sqrt{n}$)



$$m = \left\lceil \sqrt{\frac{n}{\log(eM/\sqrt{n})}} \right\rceil$$

$$\mathcal{C}' = \left\{ \frac{1}{m} \sum_{j=1}^m h_j : h_j \in F \right\}$$

Carl-Maurey ; Nemirovskii ; Tsybakov

1) Isomorphie sur

$$\cup_{g \in \mathcal{C}'} [g, f_{\mathcal{C}'}^*]$$

2) Propriétés d'approximation de $\mathcal{C}'$

Sous-optimalité du MRE pour le problème d'agrégation MS

Sous-optimalité du MRE pour le problème d'agrégation MS

L. & Mendelson : $\forall n, M$, $\exists F$ de taille M et (X, Y) tel que a.p. $\geq c_0$

$$R(\hat{f}_n^{(MRE)}) \geq \min_{f \in F} R(f) + c\sqrt{\frac{\log M}{n}}$$

Sous-optimalité du MRE pour le problème d'agrégation MS

L. & Mendelson : $\forall n, M, \exists F$ de taille M et (X, Y) tel que a.p. $\geq c_0$

$$R(\hat{f}_n^{(MRE)}) \geq \min_{f \in F} R(f) + c\sqrt{\frac{\log M}{n}}$$

$\sqrt{\log M/n} >> (\log M)/n \Longrightarrow \hat{f}_n^{(MRE)}$ n'atteint pas $(\log M)/n$ en général.

Sous-optimalité du MRE pour le problème d'agrégation MS

L. & Mendelson : $\forall n, M$, $\exists F$ de taille M et (X, Y) tel que a.p. $\geq c_0$

$$R(\hat{f}_n^{(MRE)}) \geq \min_{f \in F} R(f) + c\sqrt{\frac{\log M}{n}}$$

$\sqrt{\log M/n} >> (\log M)/n \Longrightarrow \hat{f}_n^{(MRE)}$ n'atteint pas $(\log M)/n$ en général.

[Birgé, Massart], [Bartlett, Lee, Williamson], [Juditskii, Rigollet, Tsybakov].

Long Version

Brève preuve de la sous-optimalité du MRE pour l'agrégation MS quand $M = 2$

$\bullet f_2(X)$

$\bullet f_1(X)$

Brève preuve de la sous-optimalité du MRE pour l'agrégation MS quand $M = 2$

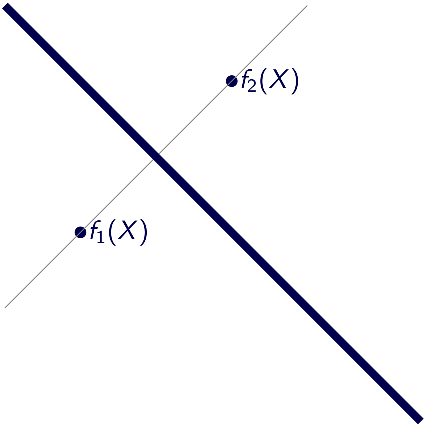
$\bullet f_2(X)$

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

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

$\bullet f_1(X)$

Brève preuve de la sous-optimalité du MRE pour l'agrégation MS quand $M = 2$

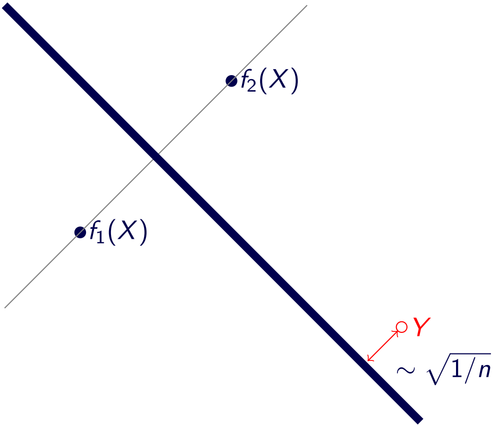


$$F = \{f_1, f_2\}$$

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

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

Brève preuve de la sous-optimalité du MRE pour l'agrégation MS quand $M = 2$

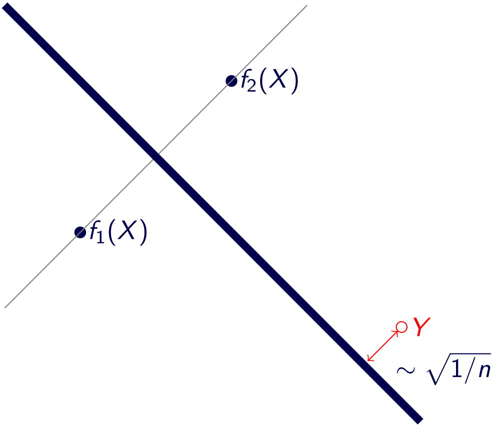


$$F = \{f_1, f_2\}$$

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

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

Brève preuve de la sous-optimalité du MRE pour l'agrégation MS quand $M = 2$



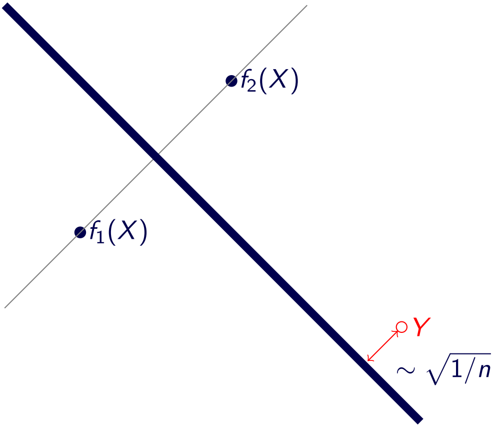
$F = \{f_1, f_2\}$

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

a.p. $\geq c_0$, $\hat{f}_n^{(MRE)} = f_1$

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

Brève preuve de la sous-optimalité du MRE pour l'agrégation MS quand $M = 2$



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

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

$$\text{a.p.} \geq c_0,\ \hat{f}_n^{(MRE)} = f_1$$

$$\text{terme résiduel} \sim 1/\sqrt{n}$$

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

MRE sur $\mathrm{Conv}(F)$ pour le problème d'agrégation MS

MRE sur $\mathrm{Conv}(F)$ pour le problème d'agrégation MS

Pascal Massart : Quelles sont les performances du MRE-C dans le contexte de l'agrégation MS ?

MRE sur $\mathrm{Conv}(F)$ pour le problème d'agrégation MS

Pascal Massart : Quelles sont les performances du MRE-C dans le contexte de l'agrégation MS ?

$$\hat{f}_n^{(MRE-C)} \in \underset{f \in \mathrm{Conv}(F)}{\operatorname{argmin}} R_n(f)$$

est-il vrai qu'avec grande probabilité

$$R(\hat{f}_n^{(MRE-C)}) \leq \min_{f \in F} R(f) + c_0 \frac{\log M}{n} \quad ?$$

MRE sur $\mathrm{Conv}(F)$ pour le problème d'agrégation MS

Pascal Massart : Quelles sont les performances du MRE-C dans le contexte de l'agrégation MS ?

$$\hat{f}_n^{(MRE-C)} \in \underset{f \in \mathrm{Conv}(F)}{\mathrm{argmin}} R_n(f)$$

est-il vrai qu'avec grande probabilité

$$R(\hat{f}_n^{(MRE-C)}) \leq \min_{f \in F} R(f) + c_0 \frac{\log M}{n} \quad ?$$

L. & Mendelson : $\forall n, M, \exists F$ de taille M et (X, Y) tel que, a.p. $\geq c_0$

$$R(\hat{f}_n^{(MRE-C)}) \geq \min_{f \in F} R(f) + c_0 \psi_n(M);$$

$$\psi_n(M) = \begin{cases} M/n & \text{if } M \leq \sqrt{n} \\ (n \log(eM/\sqrt{n}))^{-1/2} & \text{if } M \geq \sqrt{n} \end{cases} >> \frac{\log M}{n}.$$

Brève preuve de la sous-optimalitè de MRE-C pour l'agrégation MS - 1

$(\phi_i)_{i\in\mathbb{N}}$ suite de Rademacher i.i.d.

$$F := \{0, \pm\phi_1, \ldots, \pm\phi_M\} \text{ et la sortie } Y := \phi_{M+1}$$

Brève preuve de la sous-optimalitè de MRE-C pour l'agrégation MS - 1

$(\phi_i)_{i\in\mathbb{N}}$ suite de Rademacher i.i.d.

$$F := \{0, \pm\phi_1, \ldots, \pm\phi_M\} \text{ et la sortie } Y := \phi_{M+1}$$

1. pas de gain dans la terme d'approximation :

$$\min_{f\in F} R(f) = \min_{f\in \mathrm{Conv}(F)} R(f)$$

Brève preuve de la sous-optimalitè de MRE-C pour l'agrégation MS - 1

$(\phi_i)_{i\in\mathbb{N}}$ suite de Rademacher i.i.d.

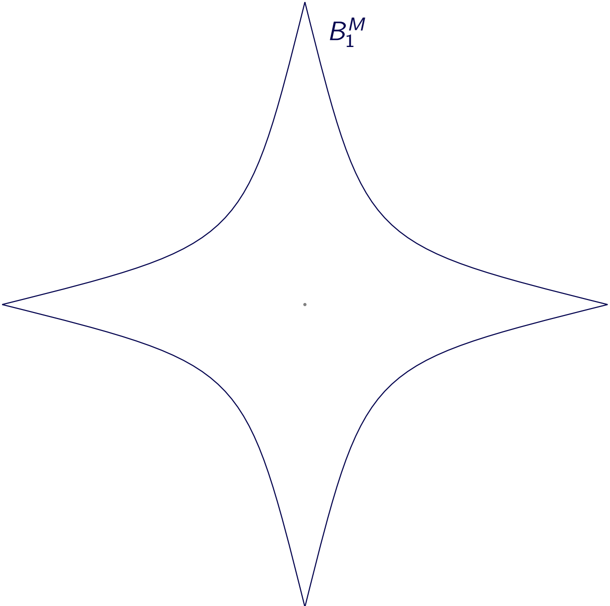
$$F := \{0, \pm\phi_1, \dots, \pm\phi_M\} \text{ et la sortie } Y := \phi_{M+1}$$

1. pas de gain dans la terme d'approximation :

$$\min_{f\in F} R(f) = \min_{f\in \mathrm{Conv}(F)} R(f)$$

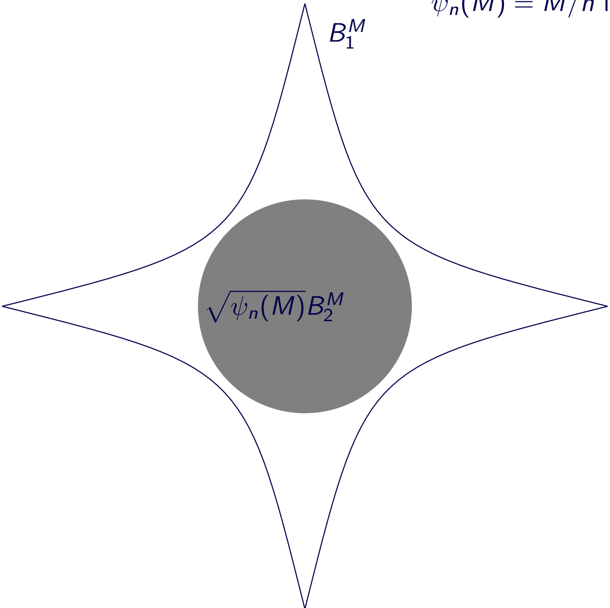
2. maximisation de la complexité de $\mathrm{Conv}(F)$

Brève preuve de la sous-optimalité de MRE-C pour l'agrégation MS - 2



Brève preuve de la sous-optimalitè de MRE-C pour l'agrégation MS - 2

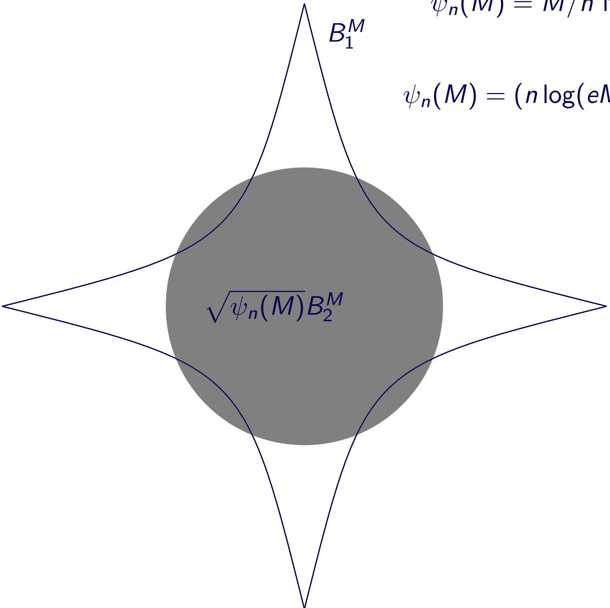
$\psi_n(M) = M/n$ if $M \leq \sqrt{n}$



Brève preuve de la sous-optimalitè de MRE-C pour l'agrégation MS - 2

$$\psi_n(M) = M/n \text{ if } M \leq \sqrt{n}$$

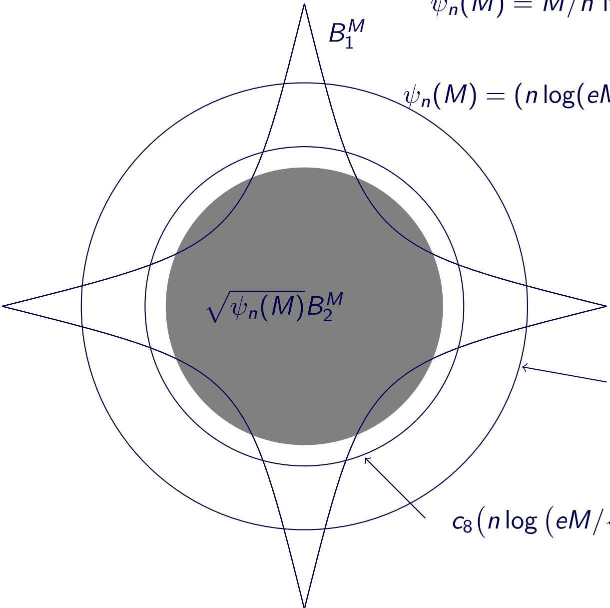
$$\psi_n(M) = (n\log(eM/\sqrt{n}))^{-1/2} \text{ if } M \geq \sqrt{n}$$



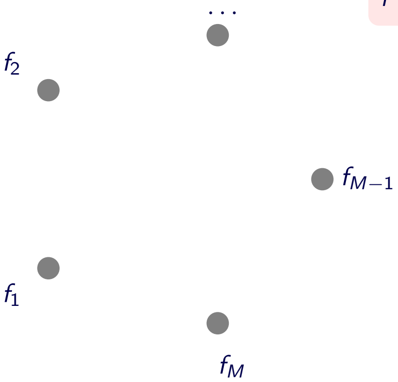
Brève preuve de la sous-optimalitè de MRE-C pour l'agrégation MS - 2

$$\psi_n(M) = M/n \text{ if } M \leq \sqrt{n}$$

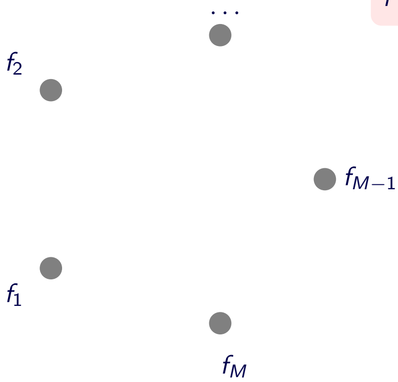
$$\psi_n(M) = (n \log(eM/\sqrt{n}))^{-1/2} \text{ if } M \geq \sqrt{n}$$



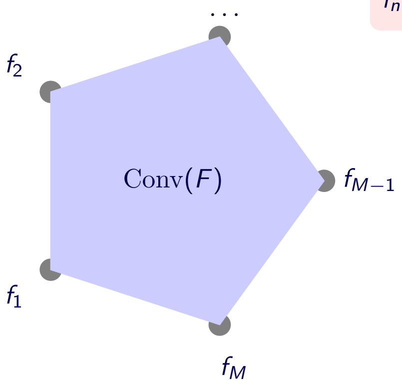
$$\hat{f}^{(MRE)} \in \operatorname{argmin}_{f \in F} R_n(f)$$



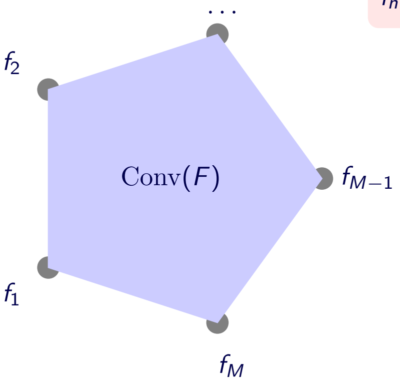
$$\hat{f}^{(MRE)} \in \operatorname{argmin}_{f \in F} R_n(f)$$



Sous-optimal pour des raisons géométriques.



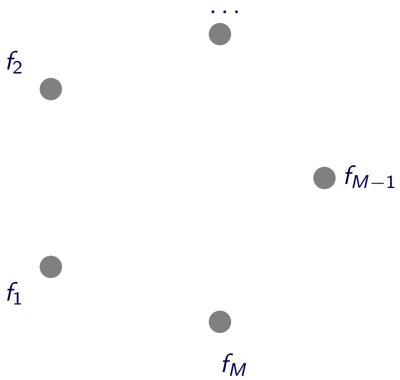
$$\hat{f}_n^{(MRE-C)} \in \mathrm{argmin}_{f \in \mathrm{Conv}(F)} R_n(f)$$



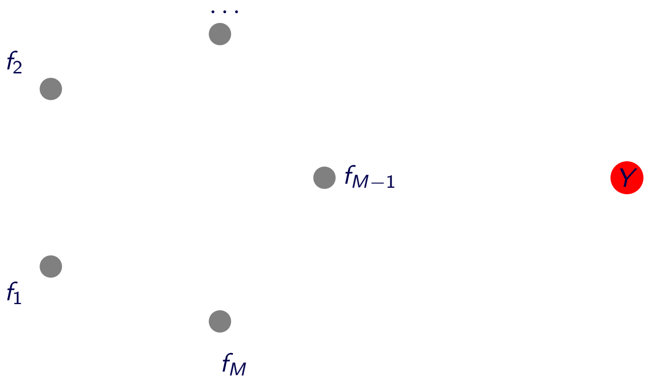
$$\hat{f}_n^{(MRE-C)} \in \mathrm{argmin}_{f \in \mathrm{Conv}(F)} R_n(f)$$

Sous-optimal pour des raisons de complexité

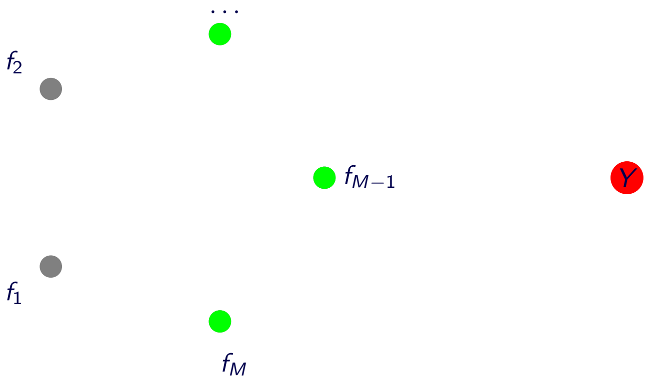
Construction d'une procédure optimale pour le problème d'agrégation MS



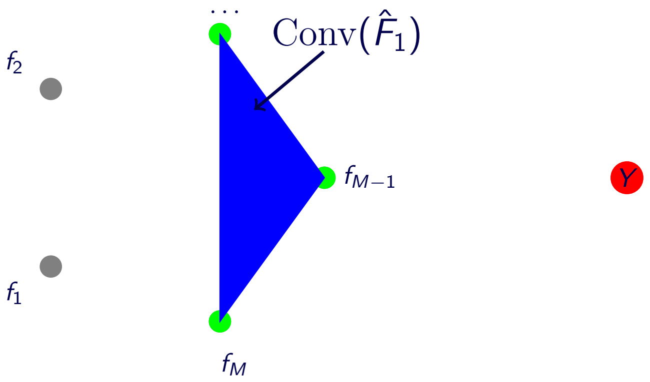
Construction d'une procédure optimale pour le problème d'agrégation MS



Construction d'une procédure optimale pour le problème d'agrégation MS

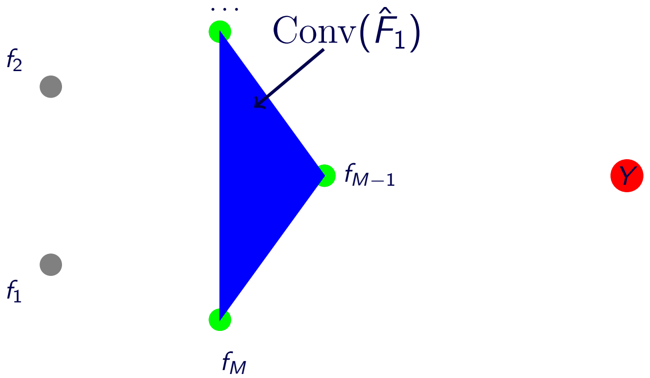


Construction d'une procédure optimale pour le problème d'agrégation MS



Construction d'une procédure optimale pour le problème d'agrégation MS

$\tilde{f} \in \mathrm{argmin}_{f \in \mathrm{Conv}(\hat{F}_1)} R_n(f)$



Découpe des données

$2n$ observations :
$D_{2n} = ((X_1, Y_1), \dots, (X_n, Y_n), (X_{n+1}, Y_{n+1}), \dots, (X_{2n}, Y_{2n}))$.

Découpe des données

$2n$ observations :
$D_{2n} = ((X_1, Y_1), \dots, (X_n, Y_n), (X_{n+1}, Y_{n+1}), \dots, (X_{2n}, Y_{2n}))$.

$D_1 = ((X_1, Y_1), \dots, (X_n, Y_n))$: construction de l'ensemble $\hat{F}_1$ des points approximativement empiriquement bons

Découpe des données

$2n$ observations :
$D_{2n} = ((X_1, Y_1), \dots, (X_n, Y_n), (X_{n+1}, Y_{n+1}), \dots, (X_{2n}, Y_{2n}))$.

$D_1 = ((X_1, Y_1), \dots, (X_n, Y_n))$: construction de l'ensemble $\hat{F}_1$ des points approximativement empiriquement bons

$D_2 = ((X_{n+1}, Y_{n+1}), \dots, (X_{2n}, Y_{2n}))$: construction du MRE sur $\mathrm{Conv}(\hat{F}_1)$

Construction de $\hat{F}_1$ (étape de pré-sélection)

Etape 1 : MRE sur F (construit à partir de D_1) :

$$\hat{f} \in \underset{f \in F}{\operatorname{argmin}} R_n(f) \qquad R_n(f) = \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2$$

Construction de $\hat{F}_1$ (étape de pré-sélection)

Etape 1 : MRE sur F (construit à partir de D_1) :

$$\hat{f} \in \underset{f \in F}{\operatorname{argmin}} R_n(f) \qquad R_n(f) = \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2$$

Etape 2 : ensemble des presques minimiseurs du risque empirique (sur D_1) :

$$\hat{F}_1 = \Big\{ f \in F : R_n(f) \le R_n(\hat{f}) + C_1 \max(\alpha \|\hat{f} - f\|_{L_2^n}, \alpha^2) \Big\},$$

pour $\alpha = \sqrt{(x + \log M)/n}$, où $x > 0$ est le niveau de confiance et

$$\|\hat{f} - f\|_{L_2^n}^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}(X_i) - f(X_i))^2$$

MRE sur $\mathrm{Conv}(\hat{F}_1)$

$$\mathrm{Conv}(\hat{F}_1) = \Big\{ \sum_{f \in \hat{F}_1} \lambda_f f : \lambda_f \geq 0 \text{ and } \sum_{f \in \hat{F}_1} \lambda_f = 1 \Big\},$$

MRE sur $\mathrm{Conv}(\hat{F}_1)$

$$\mathrm{Conv}(\hat{F}_1) = \Big\{ \sum_{f \in \hat{F}_1} \lambda_f f : \lambda_f \geq 0 \text{ and } \sum_{f \in \hat{F}_1} \lambda_f = 1 \Big\},$$

Construction du MRE sur $\mathrm{Conv}(F_1)$ pour un risque empirique construit sur D_2 :

$$\tilde{f} \in \underset{f \in \mathrm{Conv}(\hat{F}_1)}{\mathrm{argmin}} \frac{1}{n} \sum_{i=n+1}^{2n} (f(X_i) - Y_i)^2.$$

Théorème (L. & Mendelson)

Pour tout $n \geq 1$, *tout dictionaire* F *de taille* M, *tout couple* (X, Y) *tel que* $\max_{f \in F} |f(X)|, |Y| \leq b$ *p.s.*, $\forall x > 0$, *a.p.* $\geq 1 - 2\exp(-x)$,

$$R(\tilde{f}) \leq \min_{f \in F} R(f) + c_0 b^2 \max\left(\frac{\log M}{n}, \frac{x}{n}\right).$$

Théorème (L. & Mendelson)

Pour tout $n \geq 1$, tout dictionaire F de taille M, tout couple (X, Y) tel que $\max_{f \in F} |f(X)|, |Y| \leq b$ p.s., $\forall x > 0$, a.p. $\geq 1 - 2\exp(-x)$,

$$R(\tilde{f}) \leq \min_{f \in F} R(f) + c_0 b^2 \max\left(\frac{\log M}{n}, \frac{x}{n}\right).$$

Conclusion : $(\log M)/n$ est la vitesse optimale d'agrégation en déviation et $\tilde{f}$ est une procédure optimale d'agrégation.

Théorème (L. & Mendelson)

Pour tout $n \geq 1$, *tout dictionaire* F *de taille* M, *tout couple* (X, Y) *tel que* $\max_{f \in F} |f(X)|, |Y| \leq b$ *p.s.*, $\forall x > 0$, *a.p.* $\geq 1 - 2\exp(-x)$,

$$R(\tilde{f}) \leq \min_{f \in F} R(f) + c_0 b^2 \max\left(\frac{\log M}{n}, \frac{x}{n}\right).$$

Conclusion : $(\log M)/n$ est la vitesse optimale d'agrégation en déviation et $\tilde{f}$ est une procédure optimale d'agrégation.

Remarque : Cette procédure d'agrégation est "parcimonieuse", càd que les éléments non-pertinants de F sont affectés d'un poids nul.

Inégalités oracles pour le MRE, le MRE régularisé et le MRE pénalisé

Trois types d'inégalité oracle. Exemple en théorie de l'agrégation.

Le Minimiseur du Risque Empirique (MRE) sur un modèle fini F est

$$\hat{f}_n^{(MRE)} \in \operatorname*{argmin}_{f\in F} R_n(f) \text{ où } R_n(f) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2$$

Trois types d'inégalité oracle. Exemple en théorie de l'agrégation.

Le Minimiseur du Risque Empirique (MRE) sur un modèle fini F est

$$\hat{f}_n^{(MRE)} \in \underset{f \in F}{\operatorname{argmin}}\, R_n(f) \text{ où } R_n(f) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2$$

Théorème

Pour tout $x > 0$, avec probabilité plus grande que $1 - 4\exp(-x)$,

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

Trois types d'inégalité oracle. Exemple en théorie de l'agrégation.

Le Minimiseur du Risque Empirique (MRE) sur un modèle fini F est

$$\hat{f}_n^{(MRE)} \in \operatorname*{argmin}_{f \in F} R_n(f) \text{ où } R_n(f) = \frac{1}{n}\sum_{i=1}^n (Y_i - f(X_i))^2$$

Théorème

Pour tout $x > 0$, avec probabilité plus grande que $1 - 4\exp(-x)$,

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

et pour tout $\epsilon > 0$,

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

Trois types d'inégalité oracle. Exemple en théorie de l'agrégation.

Le Minimiseur du Risque Empirique (MRE) sur un modèle fini F est

$$\hat{f}_n^{(MRE)} \in \operatorname*{argmin}_{f \in F} R_n(f) \text{ où } R_n(f) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2$$

Théorème

Pour tout $x > 0$, avec probabilité plus grande que $1 - 4\exp(-x)$,

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

et pour tout $\epsilon > 0$,

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

et pour $f^*(X) = \mathbb{E}[Y|X]$

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

Trois types d'inégalité oracle. Exemple en théorie de l'agrégation.

Deux inégalités oracle avec deux résidues différents :

Trois types d'inégalité oracle. Exemple en théorie de l'agrégation.

Deux inégalités oracle avec deux résidues différents :

- un résidue à décroissance rapide pour l'inégalité oracle non-exacte :

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

Trois types d'inégalité oracle. Exemple en théorie de l'agrégation.

Deux inégalités oracle avec deux résidues différents :

- un résidue à décroissance rapide pour l'inégalité oracle non-exacte :

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

- un résidue à décroissance lente (optimal : il existe des bornes inférieures) pour l'inégalité oracle exacte :

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

Trois types d'inégalité oracle. Exemple en théorie de l'agrégation.

Deux inégalités oracle avec deux résidues différents :

- un résidue à décroissance rapide pour l'inégalité oracle non-exacte :

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

- un résidue à décroissance lente (optimal : il existe des bornes inférieures) pour l'inégalité oracle exacte :

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

Question : quelle est l'origine d'une telle différence entre ces deux types d'inégalité oracle ?

Trois types d'inégalité oracle. Exemple en théorie de l'agrégation.

Deux inégalités oracle avec deux résidues différents :

- un résidue à décroissance rapide pour l'inégalité oracle non-exacte :

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

- un résidue à décroissance lente (optimal : il existe des bornes inférieures) pour l'inégalité oracle exacte :

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

Question : quelle est l'origine d'une telle différence entre ces deux types d'inégalité oracle ?(raisons fondamentales ? géométrie - concentration - compléxité).

Inégalités oracle exactes et non-exactes

- classe des fonctions de perte :

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

Inégalités oracle exactes et non-exactes

- classe des fonctions de perte :

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

- classe des fonctions de pertes en excès :

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

(où $R(f_F^*) = \min_{f\in F} R(f) = \min_{f\in F} \mathbb{E}(Y - f(X))^2$).

Inégalités oracle exactes et non-exactes

- classe des fonctions de perte :

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

- classe des fonctions de pertes en excès :

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

(où $R(f_F^*) = \min_{f \in F} R(f) = \min_{f \in F} \mathbb{E}(Y - f(X))^2$).

Pour toute classe H de fonctions, *l'ensemble de H étoilé en 0* est

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

Inégalités oracle exactes et non-exactes

- classe des fonctions de perte :

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

- classe des fonctions de pertes en excès :

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

(où $R(f_F^*) = \min_{f \in F} R(f) = \min_{f \in F} \mathbb{E}(Y - f(X))^2$).

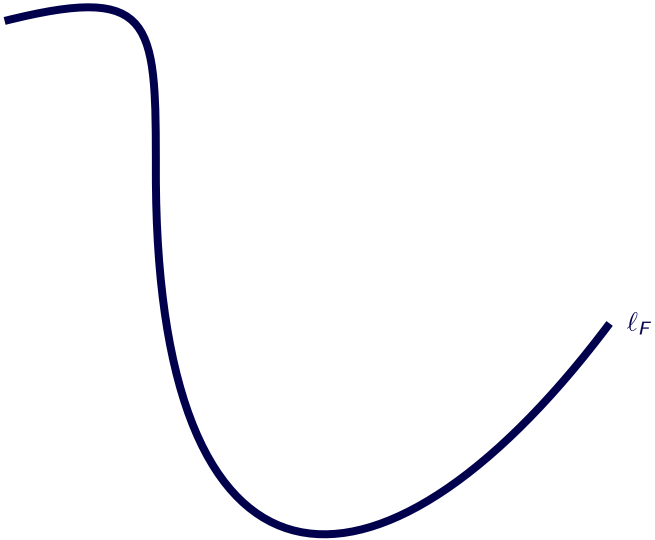
Pour toute classe H de fonctions, *l'ensemble de H étoilé en 0* est

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

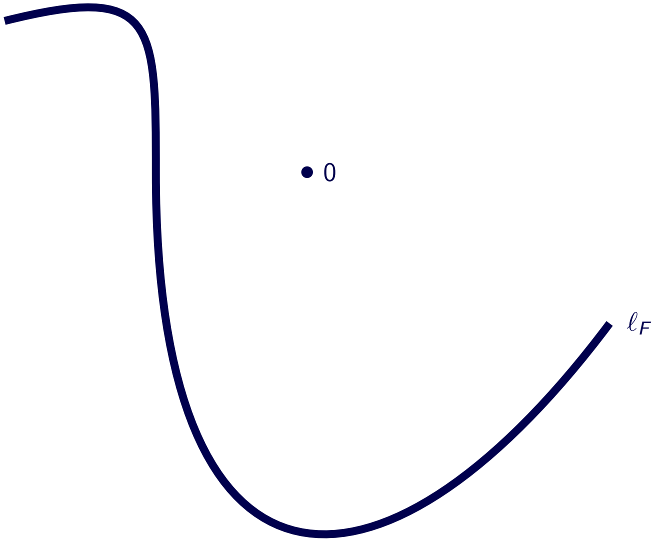
et son ensemble *localisé au niveau $\lambda > 0$* est

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

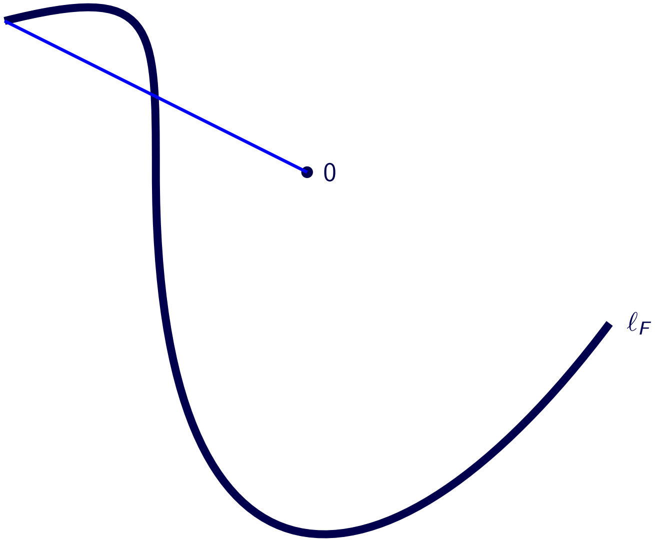
Inégalités oracle exactes et non-exactes



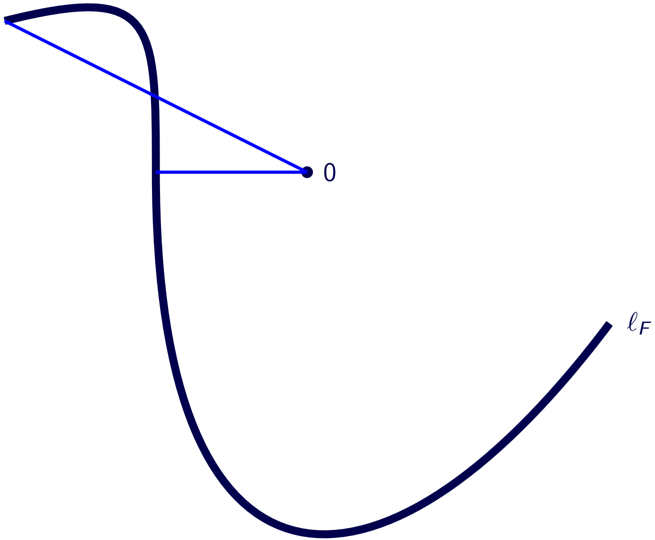
Inégalités oracle exactes et non-exactes



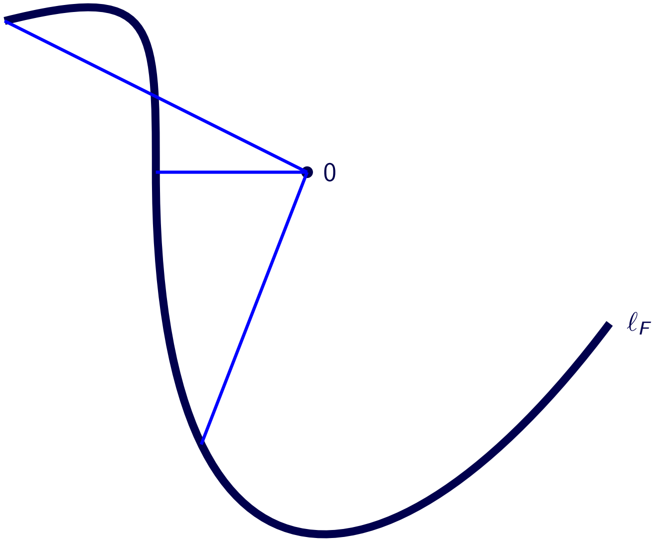
Inégalités oracle exactes et non-exactes



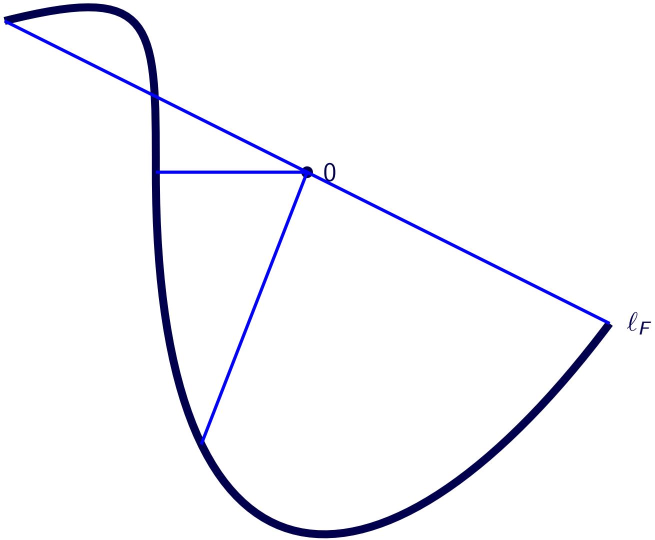
Inégalités oracle exactes et non-exactes



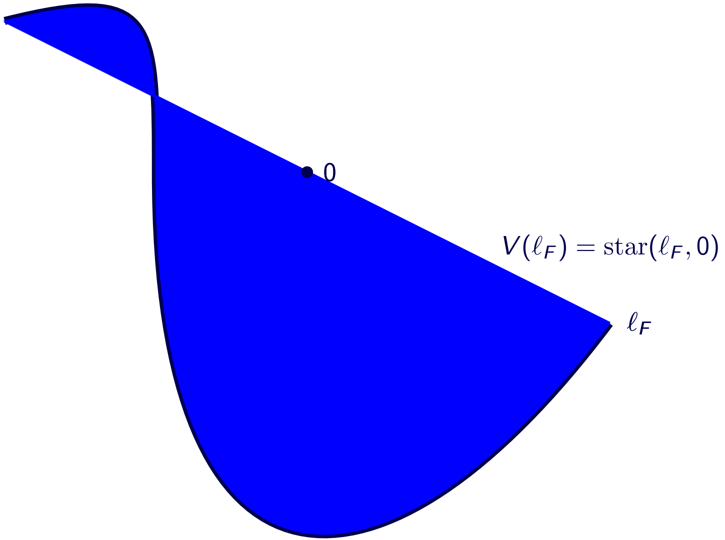
Inégalités oracle exactes et non-exactes



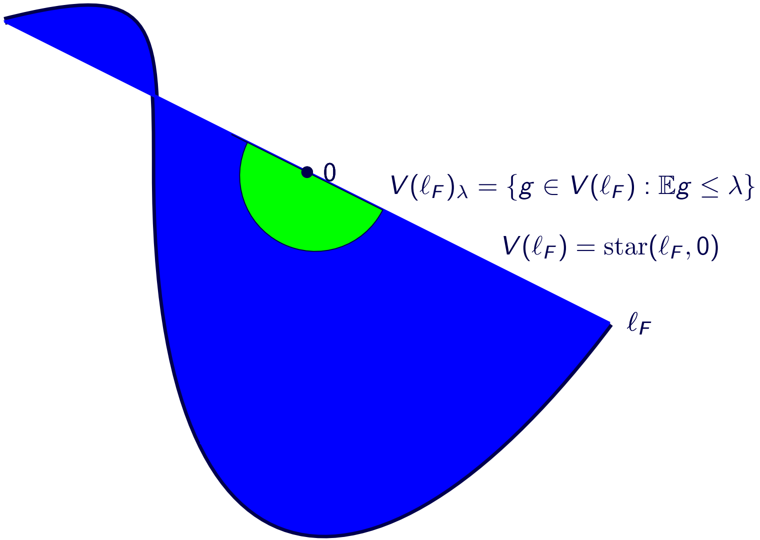
Inégalités oracle exactes et non-exactes



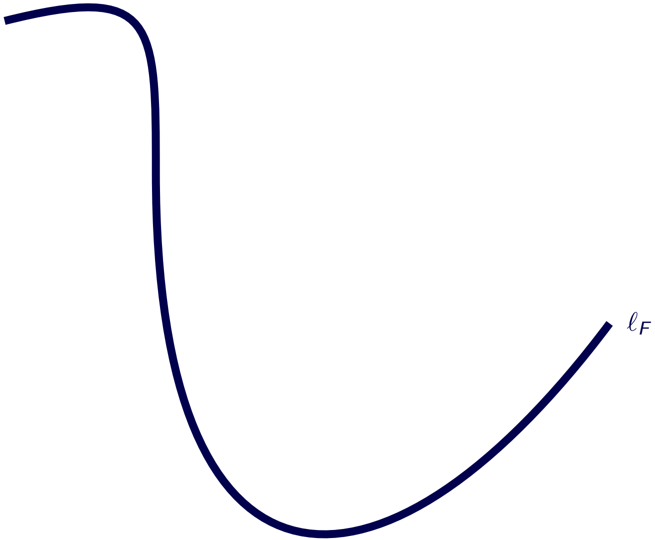
Inégalités oracle exactes et non-exactes



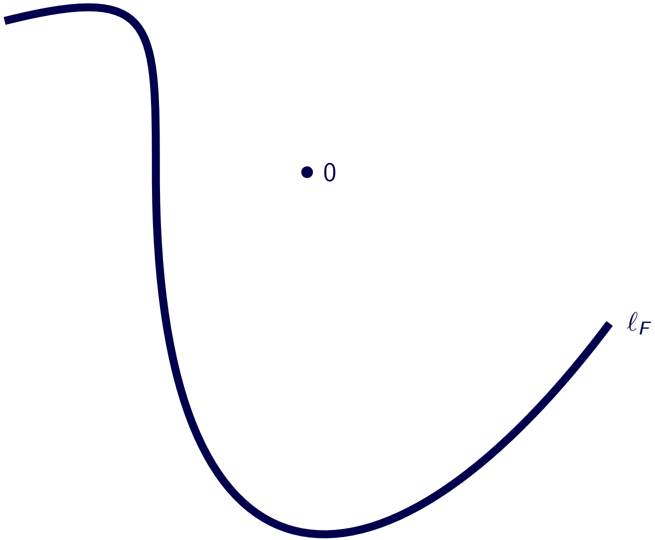
Inégalités oracle exactes et non-exactes



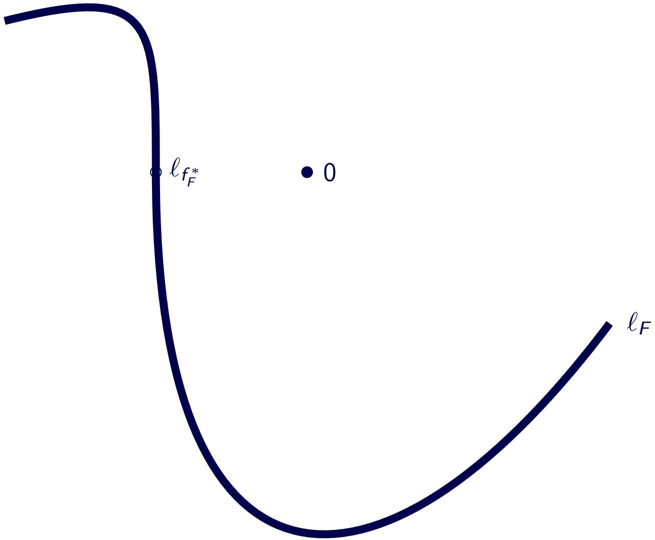
Inégalités oracle exactes et non-exactes



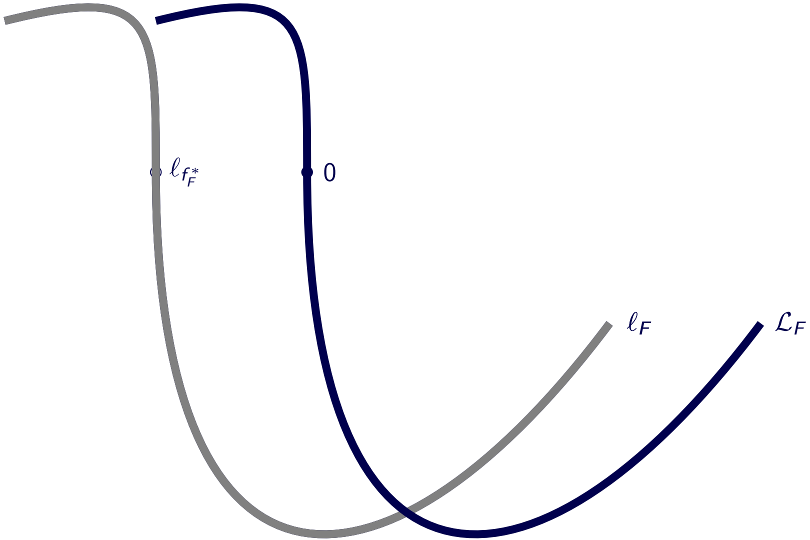
Inégalités oracle exactes et non-exactes



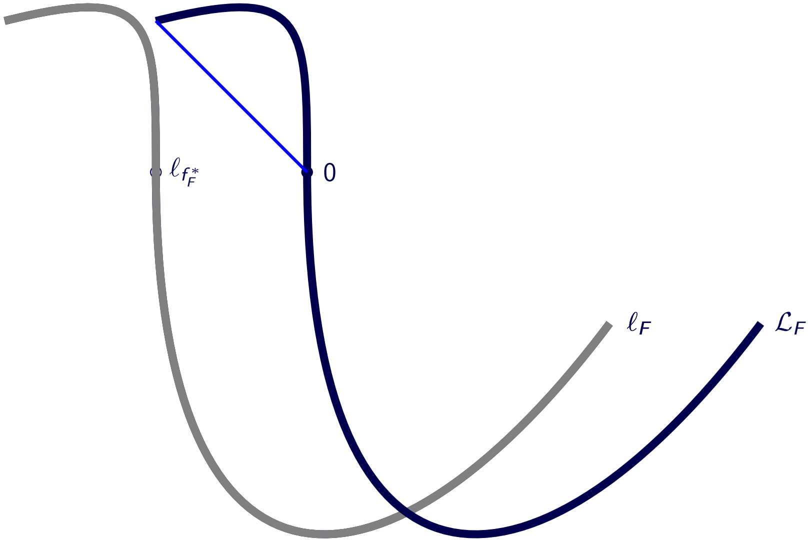
Inégalités oracle exactes et non-exactes



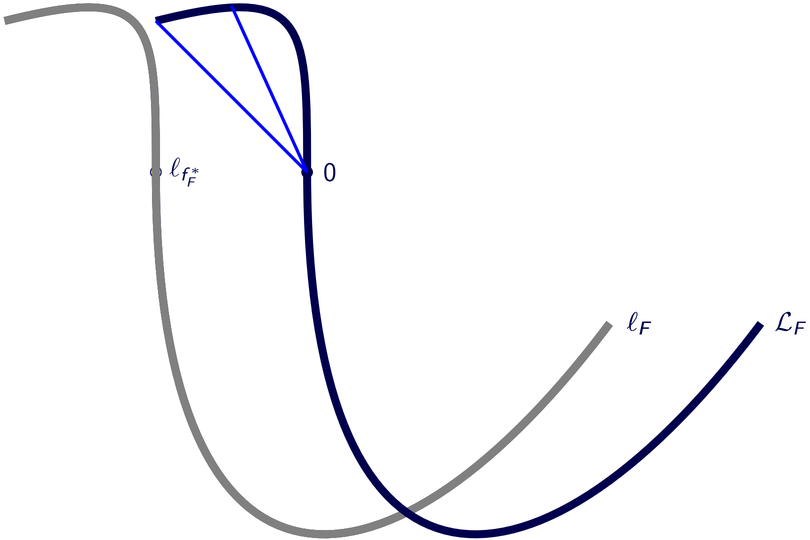
Inégalités oracle exactes et non-exactes



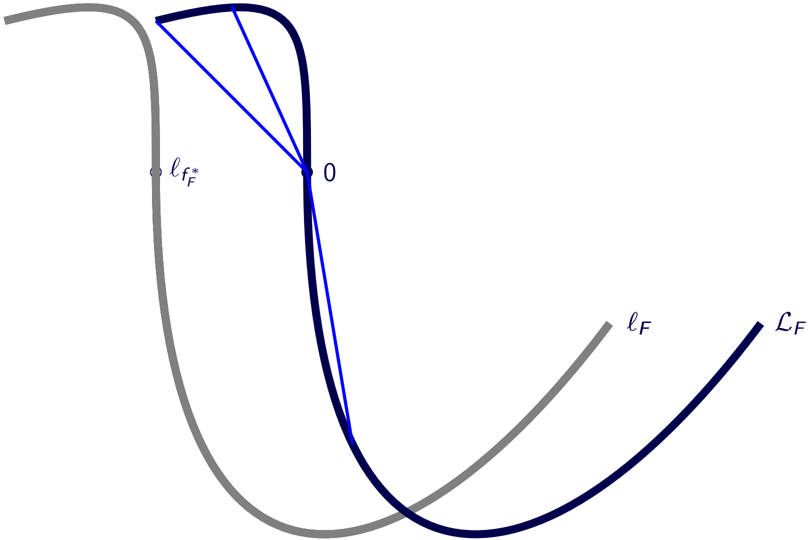
Inégalités oracle exactes et non-exactes



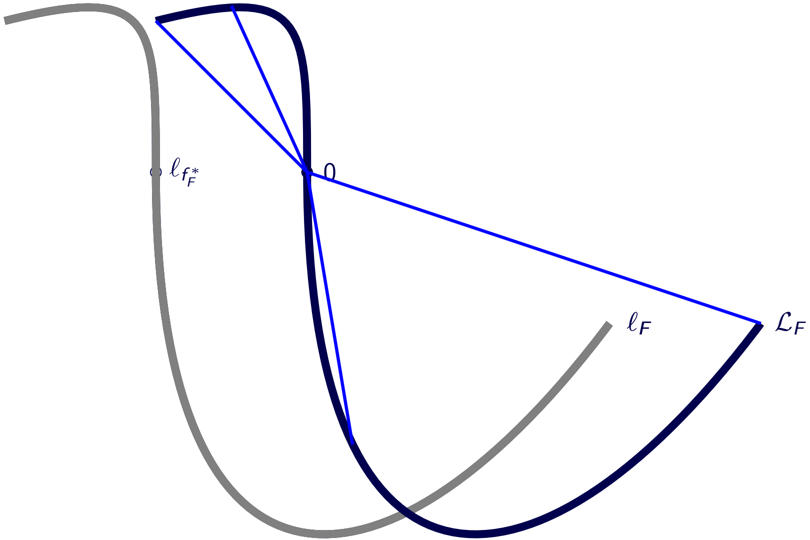
Inégalités oracle exactes et non-exactes



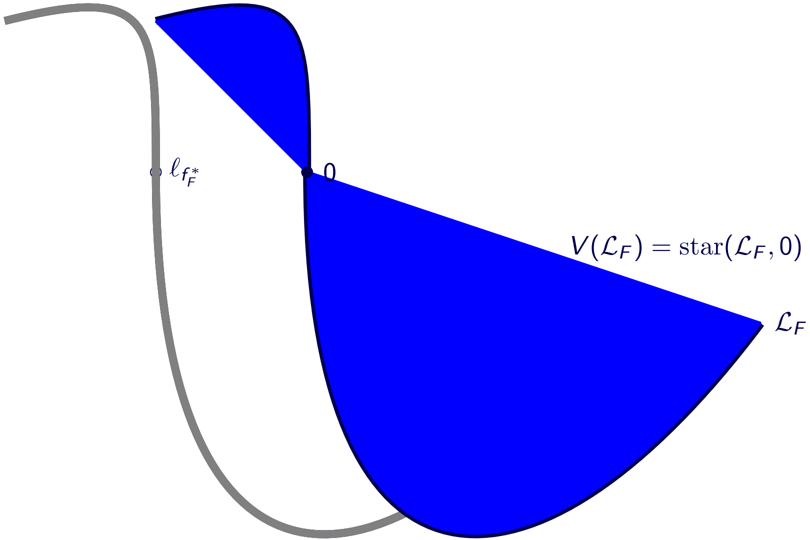
Inégalités oracle exactes et non-exactes



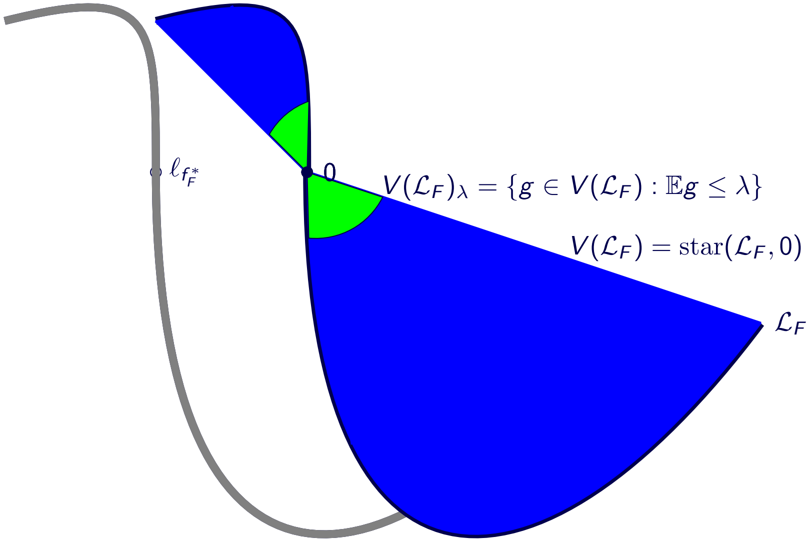
Inégalités oracle exactes et non-exactes



Inégalités oracle exactes et non-exactes

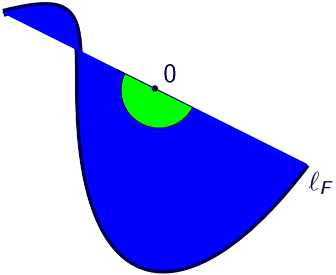


Inégalités oracle exactes et non-exactes

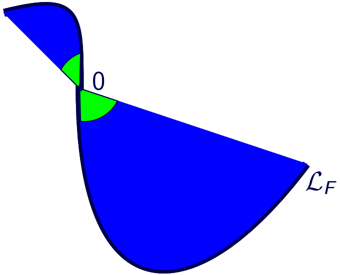


Inégalités oracle exactes et non-exactes

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

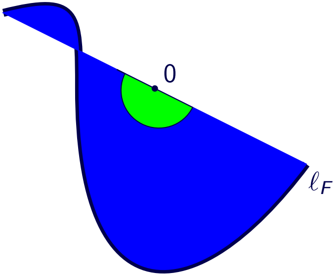


$$V(\mathcal{L}_F)_\lambda = \{g \in V(\mathcal{L}_F) : \mathbb{E}g \leq \lambda\}$$



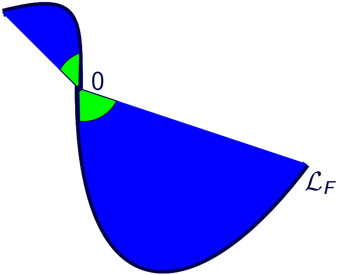
Inégalités oracle exactes et non-exactes

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



Inégalités oracle non-exactes

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



Inégalités oracle exactes

Inégalités oracle exactes et non-exactes

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

où

$$Ph := \mathbb{E}h(X, Y) \text{ and } P_n h := \frac{1}{n}\sum_{i=1}^{n} h(X_i, Y_i)$$

Inégalités oracle exactes et non-exactes

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

où

$$Ph := \mathbb{E}h(X, Y) \text{ and } P_n h := \frac{1}{n}\sum_{i=1}^{n} h(X_i, Y_i)$$

Deux points fixes caractérisent les inégalités oracle exactes et non-exactes :

Inégalités oracle exactes et non-exactes

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

où

$$Ph := \mathbb{E}h(X, Y) \text{ and } P_n h := \frac{1}{n} \sum_{i=1}^{n} h(X_i, Y_i)$$

Deux points fixes caractérisent les inégalités oracle exactes et non-exactes :

- pour les inégalités oracle exactes :

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

Inégalités oracle exactes et non-exactes

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

où

$$Ph := \mathbb{E}h(X, Y) \text{ and } P_n h := \frac{1}{n}\sum_{i=1}^{n} h(X_i, Y_i)$$

Deux points fixes caractérisent les inégalités oracle exactes et non-exactes :

- pour les inégalités oracle exactes :

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

- inégalités oracle non-exactes :

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

Inégalités oracle exactes

Théorème (Bartlett et Mendelson)

Soit F une classe de fonctions telle qu'il existe $B > 0$ satisfaisant $\forall f \in F$,

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

(où $\mathcal{L}_f = \ell_f - \ell_{f_F^*}$). Soit $\mu^* > 0$ tel que $\mathbb{E}\|P_n - P\|_{V(\mathcal{L}_F)_{\mu^*}} \leq \mu^*/8$.

Inégalités oracle exactes

Théorème (Bartlett et Mendelson)

Soit F une classe de fonctions telle qu'il existe $B > 0$ satisfaisant $\forall f \in F$,

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

(où $\mathcal{L}_f = \ell_f - \ell_{f_F^*}$). Soit $\mu^* > 0$ tel que $\mathbb{E}\|P_n - P\|_{V(\mathcal{L}_F)_{\mu^*}} \leq \mu^*/8$. Alors pour tout $x > 0$, avec probabilité plus grande que $1 - 4\exp(-x)$,

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

Inégalités oracle exactes

Théorème (Bartlett et Mendelson)

Soit F une classe de fonctions telle qu'il existe $B > 0$ satisfaisant $\forall f \in F$,

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

(où $\mathcal{L}_f = \ell_f - \ell_{f_F^*}$). Soit $\mu^* > 0$ tel que $\mathbb{E}\|P_n - P\|_{V(\mathcal{L}_F)_{\mu^*}} \leq \mu^*/8$. Alors pour tout $x > 0$, avec probabilité plus grande que $1 - 4\exp(-x)$,

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

où ρ_n est une fonction croissante telle que

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

Inégalités oracle non-exactes

Théorème (L. et Mendelson)

Soit F une classe de fonctions telle qu'il existe $B \geq 0$ satisfaisant pour tout $f \in F$,

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

Soit $0 < \epsilon < 1$ et $\lambda_\epsilon^* > 0$ tel que

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

Inégalités oracle non-exactes

Théorème (L. et Mendelson)

Soit F une classe de fonctions telle qu'il existe $B \geq 0$ satisfaisant pour tout $f \in F$,

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

Soit $0 < \epsilon < 1$ et $\lambda_\epsilon^* > 0$ tel que

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

Alors pour tout $x > 0$, avec probabilité plus grande que $1 - 8\exp(-x)$,

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

Inégalités oracle non-exactes

Théorème (L. et Mendelson)

Soit F une classe de fonctions telle qu'il existe $B \geq 0$ satisfaisant pour tout $f \in F$,

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

Soit $0 < \epsilon < 1$ et $\lambda_\epsilon^* > 0$ tel que

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

Alors pour tout $x > 0$, avec probabilité plus grande que $1 - 8\exp(-x)$,

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

où ρ_n est une fonction croissante telle que

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

La condition de Bernstein/Marge

1. Inégalités oracle exactes : $\forall f \in F, P\mathcal{L}_f^2 \leq BP\mathcal{L}_f$; (càd $P(\ell_f - \ell_{f_F^*})^2 \leq BP(\ell_f - \ell_{f_F^*})$).

La condition de Bernstein/Marge

1. Inégalités oracle exactes : $\forall f \in F, P\mathcal{L}_f^2 \leq BP\mathcal{L}_f$; (càd $P(\ell_f - \ell_{f_F^*})^2 \leq BP(\ell_f - \ell_{f_F^*}))$.
2. Inégalités oracle non-exactes : $\forall f \in F, P\ell_f^2 \leq BP\ell_f + B^2/n$.

La condition de Bernstein/Marge

1. Inégalités oracle exactes : $\forall f \in F, P\mathcal{L}_f^2 \leq BP\mathcal{L}_f$; (càd $P(\ell_f - \ell_{f_F^*})^2 \leq BP(\ell_f - \ell_{f_F^*}))$.
2. Inégalités oracle non-exactes : $\forall f \in F, P\ell_f^2 \leq BP\ell_f + B^2/n$.

Lemme

Pour toute fonction f telle que $\ell_f \geq 0$ p.s. et $\|\ell_f\|_{\psi_1} \leq D$ pour $D \geq 1$, on a pour tout n,

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

La condition de Bernstein/Marge

1. Inégalités oracle exactes : $\forall f \in F, P\mathcal{L}_f^2 \leq BP\mathcal{L}_f$; (càd $P(\ell_f - \ell_{f_F^*})^2 \leq BP(\ell_f - \ell_{f_F^*}))$.
2. Inégalités oracle non-exactes : $\forall f \in F, P\ell_f^2 \leq BP\ell_f + B^2/n$.

Lemme

Pour toute fonction f telle que $\ell_f \geq 0$ p.s. et $\|\ell_f\|_{\psi_1} \leq D$ pour $D \geq 1$, on a pour tout n,

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

Conclusion 1 : Pour les inégalités oracle non-exactes, la condition de Bernstein/Marge est presque trivialement satisfaite.

La condition de Bernstein/Marge

$\bullet f_2(X)$

$\bullet f_1(X)$

La condition de Bernstein/Marge

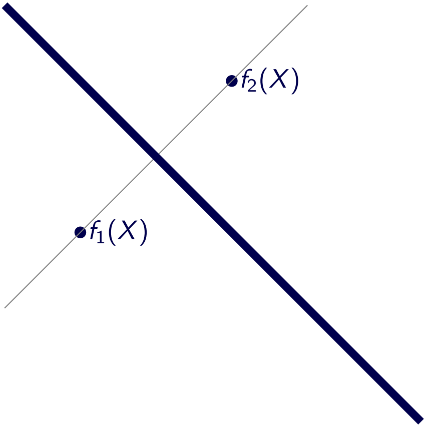
$\bullet f_2(X)$

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

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

$\bullet f_1(X)$

La condition de Bernstein/Marge

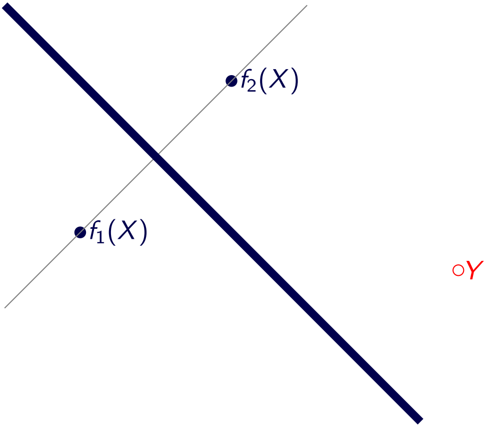


$$F = \{f_1, f_2\}$$

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

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

La condition de Bernstein/Marge

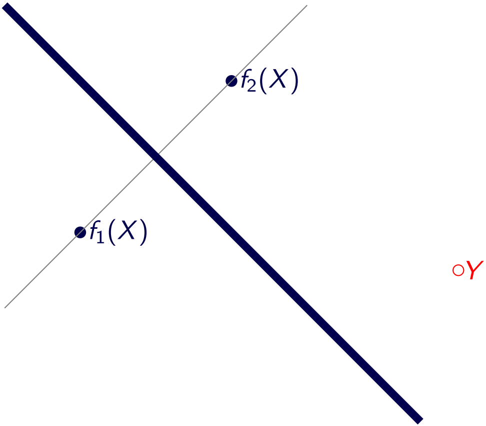


$$F = \{f_1, f_2\}$$

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

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

La condition de Bernstein/Marge



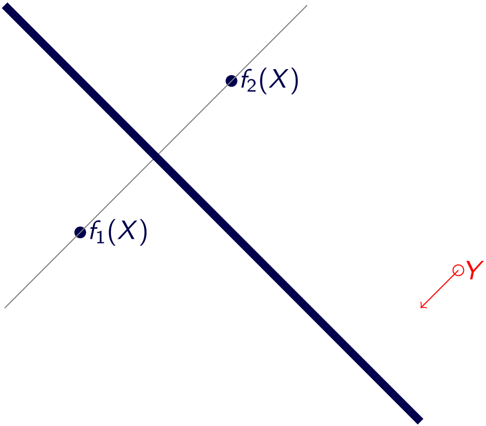
$$F = \{f_1, f_2\}$$

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

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

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

La condition de Bernstein/Marge



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

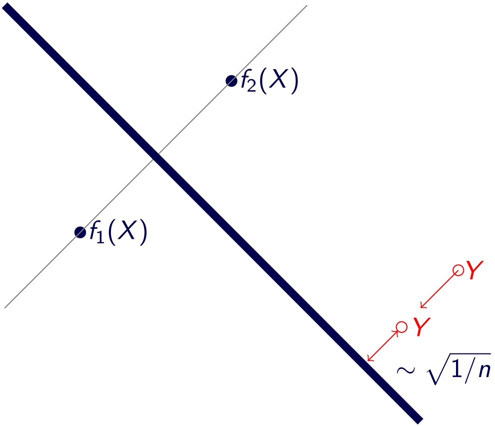
$$R(f) = \mathbb{E}(Y - f(X))^2$$

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

B croît

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

La condition de Bernstein/Marge



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

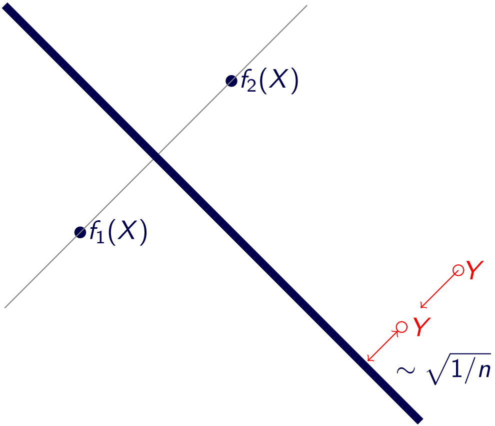
$$R(f) = \mathbb{E}(Y - f(X))^2$$

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

B croît

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

La condition de Bernstein/Marge



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

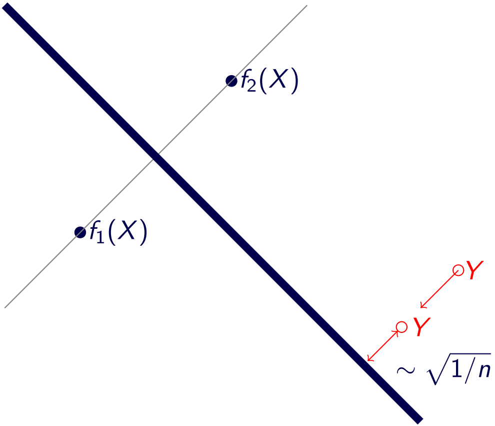
$$R(f) = \mathbb{E}(Y - f(X))^2$$

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

B croît $\qquad B \sim \sqrt{n}$

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

La condition de Bernstein/Marge



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

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

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

$$B \sim \sqrt{n}$$

résidue $\sim 1/\sqrt{n}$

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

La condition de Bernstein/Marge

Conclusion 2 : Pour les inégalités oracle exactes, la condition de Bernstein dépend de la géométrie du couple (F, Y).

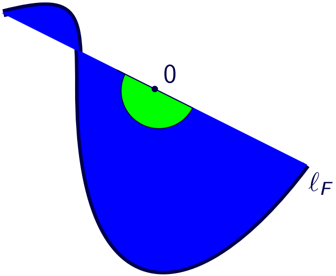
La condition de Bernstein/Marge

Conclusion 2 : Pour les inégalités oracle exactes, la condition de Bernstein dépend de la géométrie du couple (F, Y).

L'aspect géométrique explique la différence de résidue pour le problème d'agrégation MS : l'ensemble M_F des sorties Y ayant plus de deux oracles n'est jamais vide. On peut donc toujours trouver une sortie Y dans une "mauvaise position" telle que la constante de Bernstein est de l'ordre de $\sqrt{n}$ et donc un terme résiduel en $\sim \sqrt{1/n}$ pour le MRE.

Les termes de complexité : μ^* et λ_ϵ^*

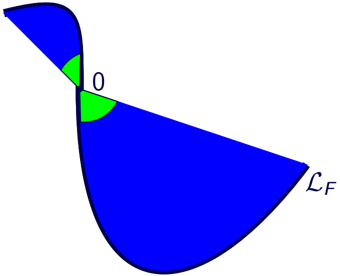
$$V(\ell_F)_\lambda = \{g \in V(\ell_F) : \mathbb{E}g \leq \lambda\}$$



Inégalités oracle non-exactes

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

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



Inégalités oracle exactes

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

Un exemple de calcul des points fixes μ^* et λ_ϵ^*

"Peeling argument" :

Un exemple de calcul des points fixes μ^* et λ^*_ϵ

"Peeling argument" : H une classe de fonctions telle que $Ph \geq 0, \forall h \in H$. Pour $H_\mu = \{h \in H : Ph \leq \mu\}$, on a

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

Un exemple de calcul des points fixes μ^* et λ^*_ϵ

"Peeling argument" : H une classe de fonctions telle que $Ph \geq 0, \forall h \in H$. Pour $H_\mu = \{h \in H : Ph \leq \mu\}$, on a

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

Plusieures manières de calculer $\mathbb{E}\|P - P_n\|_{H_\mu}$:

1. Symmétrisation + argument de contraction + Intégrale de Dudley ;

Un exemple de calcul des points fixes μ^* et λ_ϵ^*

"Peeling argument" : H une classe de fonctions telle que $Ph \geq 0, \forall h \in H$. Pour $H_\mu = \{h \in H : Ph \leq \mu\}$, on a

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

Plusieures manières de calculer $\mathbb{E}\|P - P_n\|_{H_\mu}$:

1. Symmétrisation + argument de contraction + Intégrale de Dudley ;
2. Méthode de chaining dans certains cas particuliers ;

Un exemple de calcul des points fixes μ^* et λ_ϵ^*

"Peeling argument" : H une classe de fonctions telle que $Ph \geq 0, \forall h \in H$. Pour $H_\mu = \{h \in H : Ph \leq \mu\}$, on a

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

Plusieures manières de calculer $\mathbb{E}\|P - P_n\|_{H_\mu}$:

1. Symmétrisation + argument de contraction + Intégrale de Dudley ;
2. Méthode de chaining dans certains cas particuliers ;
3. Calcul de complexités Gaussiennes ;

Un exemple de calcul des points fixes μ^* et λ_ϵ^*

"Peeling argument" : H une classe de fonctions telle que $Ph \geq 0, \forall h \in H$. Pour $H_\mu = \{h \in H : Ph \leq \mu\}$, on a

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

Plusieures manières de calculer $\mathbb{E}\|P - P_n\|_{H_\mu}$:

1. Symmétrisation + argument de contraction + Intégrale de Dudley ;
2. Méthode de chaining dans certains cas particuliers ;
3. Calcul de complexités Gaussiennes ;
4. Méthode ℓ_∞^n de M. Rudelson,...

Un exemple de calcul des points fixes μ^* et λ_ϵ^*

Calcul de λ_ϵ^* et μ^* dans le cas de la régression pour la perte quadratique :

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

et

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

Un exemple de calcul des points fixes μ^* et λ^*_ϵ

Calcul de λ^*_ϵ et μ^* dans le cas de la régression pour la perte quadratique :

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

et

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

Mesure de complexité de F :

Un exemple de calcul des points fixes μ^* et λ^*_ϵ

Calcul de λ^*_ϵ et μ^* dans le cas de la régression pour la perte quadratique :

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

et

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

Mesure de complexité de F :

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

Un exemple de calcul des points fixes μ^* et λ^*_ϵ

Calcul de λ^*_ϵ et μ^* dans le cas de la régression pour la perte quadratique :

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

et

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

Mesure de complexité de F :

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

Un exemple de calcul des points fixes μ^* et λ^*_ϵ

Calcul de λ^*_ϵ et μ^* dans le cas de la régression pour la perte quadratique :

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

et

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

Mesure de complexité de F :

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

où $\gamma_2(T,d) := \inf_{(T_s)} \sup_{t \in T} \sum_{s=0}^\infty 2^{-s/2} d(t, T_s)$ t.q. $|T_s| \leq 2^{2^s}$ et $\tilde{A} = A - A$.

Un exemple de calcul des points fixes μ^* et λ^*_ϵ

Calcul de λ^*_ϵ et μ^* dans le cas de la régression pour la perte quadratique :

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

et

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

Mesure de complexité de F :

$$P_\sigma F := \{(f(X_1), \cdots, f(X_n)) : f \in F\} \text{ et } U_n(F) := \mathbb{E}\gamma_2(\widetilde{P_\sigma F}, \ell^n_\infty)^2$$

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

Un exemple de calcul des points fixes μ^* et λ_ϵ^*

Calcul de λ_ϵ^* et μ^* dans le cas de la régression pour la perte quadratique :

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

et

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

Mesure de complexité de F :

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

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

Lemme

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

Un exemple de calcul des points fixes μ^* et λ_ϵ^*

Calcul de λ_ϵ^* et μ^* dans le cas de la régression pour la perte quadratique :

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

et

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

Mesure de complexité de F :

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

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

Lemme

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

Un exemple de calcul des points fixes μ^* et λ_ϵ^*

On combine

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

Un exemple de calcul des points fixes μ^* et λ_ϵ^*

On combine

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

Un exemple de calcul des points fixes μ^* et λ_ϵ^*

On combine

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

on obtient approximativement

Un exemple de calcul des points fixes μ^* et λ_ϵ^*

On combine

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

on obtient approximativement

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

Un exemple de calcul des points fixes μ^* et λ_ϵ^*

On combine

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

on obtient approximativement

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

Car $R^*=\inf_{f\in F} R(f)\neq 0$ en général. Donc λ_ϵ^* est en général le carré de μ^* (biensûr dans des cas particuliers, on peut obtenir des vitesses rapides pour le MRE pour des inégalités oracle exactes).

On peut donc expliquer la différence de résidue entre les inégalités oracle exacte et non-exactes par deux raisons :

1. la géométrie de (F, Y) est très importante pour les inégalités oracle exactes alors qu'elle n'a pas de rôle particulier pour les inégalités non-exactes : condition de Bernstein ;

On peut donc expliquer la différence de résidue entre les inégalités oracle exacte et non-exactes par deux raisons :

1. la géométrie de (F, Y) est très importante pour les inégalités oracle exactes alors qu'elle n'a pas de rôle particulier pour les inégalités non-exactes : condition de Bernstein ;
2. les complexités de $V(\mathcal{L}_F)_\lambda$ et $V(\ell_F)_\lambda$ peuvent être très différentes.

▸ Inégalités oracle pour les RERM

▸ Applications à la complétion de matrices

Perspectives

1. Comprendre pourquoi le MRE est optimale pour les problèmes d'agrégation L et C, mais sous-optimal pour l'agrégation MS.

Perspectives

1. Comprendre pourquoi le MRE est optimale pour les problèmes d'agrégation L et C, mais sous-optimal pour l'agrégation MS. Pour $\Lambda \subset \mathbb{R}^M$, on peut définir le problème de Λ-agrégation ($\Lambda = \mathbb{R}^M$: agrégation L ; $\Lambda = \{\lambda : \lambda_j \geq 0, \sum \lambda_j = 1\}$: agrégation C ; $\Lambda = \{e_1, \ldots, e_M\}$: agrégation MS).

Perspectives

1. Comprendre pourquoi le MRE est optimale pour les problèmes d'agrégation L et C, mais sous-optimal pour l'agrégation MS. Pour $\Lambda \subset \mathbb{R}^M$, on peut définir le problème de Λ-agrégation ($\Lambda = \mathbb{R}^M$: agrégation L ; $\Lambda = \{\lambda : \lambda_j \geq 0, \sum \lambda_j = 1\}$: agrégation C ; $\Lambda = \{e_1, \dots, e_M\}$: agrégation MS). Montrer que

   MRE est optimal pour Λ-agrégation pour tout $n \Longleftrightarrow \Lambda$ est convexe.

Perspectives

1. Comprendre pourquoi le MRE est optimale pour les problèmes d'agrégation L et C, mais sous-optimal pour l'agrégation MS. Pour $\Lambda \subset \mathbb{R}^M$, on peut définir le problème de Λ-agrégation ($\Lambda = \mathbb{R}^M$: agrégation L ; $\Lambda = \{\lambda : \lambda_j \geq 0, \sum \lambda_j = 1\}$: agrégation C ; $\Lambda = \{e_1, \ldots, e_M\}$: agrégation MS). Montrer que

   MRE est optimal pour Λ-agrégation pour tout $n \Longleftrightarrow \Lambda$ est convexe.
2. Comprendre quels sont les problèmes de Λ-agrégation pour lesquels la vitesses de convergence du MRE est la même pour les inégalités oracle exactes et non-exactes.

Merci pour votre attention !

Lower bound for MRE for non-convex models

Théorème (L. & Mendelson)

Let $F \subset B(L_\infty)$ which is μ-Donsker and assume that $\exists T_0 \in \tau$ s.t. $\mathcal{M}(T_0) \geq 2$.

Lower bound for MRE for non-convex models

Théorème (L. & Mendelson)

Let $F \subset B(L_\infty)$ which is μ-Donsker and assume that $\exists T_0 \in \tau$ s.t. $\mathcal{M}(T_0) \geq 2$. Set $Q = \{\ell_f - \ell_{f_F^*} : f \in F,\ \mathbb{E}(f(X) - T_0(X))^2 = \mathbb{E}(f_F^*(X) - T_0(X))^2\}$ and $H(Q) = \mathbb{E} \sup_{q \in Q} G_q$ ($(G_q)_{q \in Q}$ is the canonical Gaussian process associated with Q.)

Lower bound for MRE for non-convex models

Théorème (L. & Mendelson)

Let $F \subset B(L_\infty)$ which is μ-Donsker and assume that $\exists T_0 \in \tau$ s.t. $\mathcal{M}(T_0) \geq 2$. Set $Q = \{\ell_f - \ell_{f_F^*} : f \in F,\ \mathbb{E}(f(X) - T_0(X))^2 = \mathbb{E}(f_F^*(X) - T_0(X))^2\}$ and $H(Q) = \mathbb{E}\sup_{q \in Q} G_q$ ($(G_q)_{q\in Q}$ is the canonical Gaussian process associated with Q.) $\exists c_0, c_1, N(F)$ s.t. $\forall n \geq N(F)$, if $Y = T_{\lambda_n}(X)$ then w.p.g. c_0,

$$R(\hat{f}_n^{(MRE)}) \geq \inf_{f \in F} R(f) + C_2 \frac{H(Q)}{\sqrt{n}} \delta^2 \| T - f^* \|$$

where δ is s.t. $\forall n \geq N(F)$, $\mathrm{osc}_n(F, f^*, \delta) \leq C_2 H(Q)/\sqrt{n}$ and $\lambda_n = C_2 H(Q)/\sqrt{n}$.

Short version

Sketch of the proof of the suboptimality of the MRE for the MS aggregation problem

Aim

Lower bound for the excess risk of the Empirical risk minimization algorithm ($Y = T(X)$; any model F)

Sketch of the proof of the suboptimality of the MRE for the MS aggregation problem

Aim

Lower bound for the excess risk of the Empirical risk minimization algorithm ($Y = T(X)$; any model F)

$\exists T \in \tau$ (set of targets) such that w.p.g. c_0

$$\mathbb{E}\big[(\hat{f}_n^{(MRE)}(X) - T(X))^2 | \mathcal{D}\big] - \inf_{f \in F} \mathbb{E}(f(X) - T(X))^2 \geq r_n(F).$$

Sketch of the proof of the suboptimality of the MRE for the MS aggregation problem

Aim

Lower bound for the excess risk of the Empirical risk minimization algorithm ($Y = T(X)$; any model F)

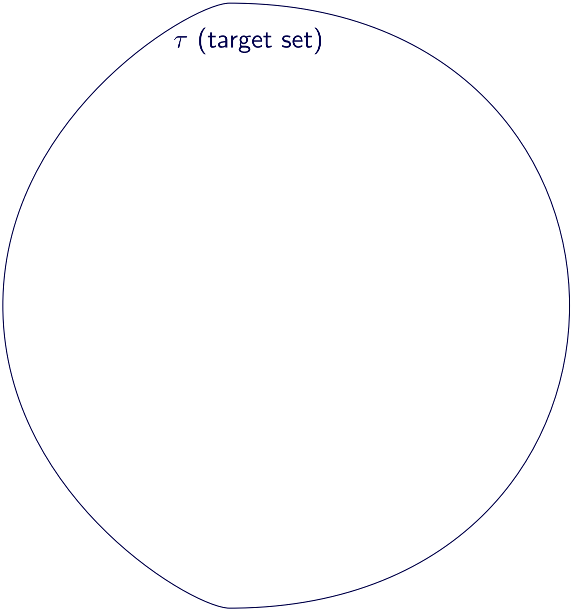
$\exists T \in \tau$ (set of targets) such that w.p.g. c_0

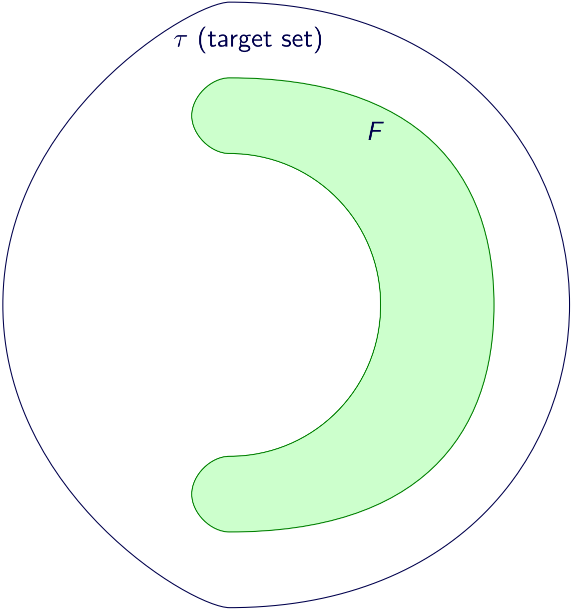
$$\mathbb{E}\big[(\hat{f}_n^{(MRE)}(X) - T(X))^2|\mathcal{D}\big] - \inf_{f\in F} \mathbb{E}(f(X) - T(X))^2 \geq r_n(F).$$

Assumption :

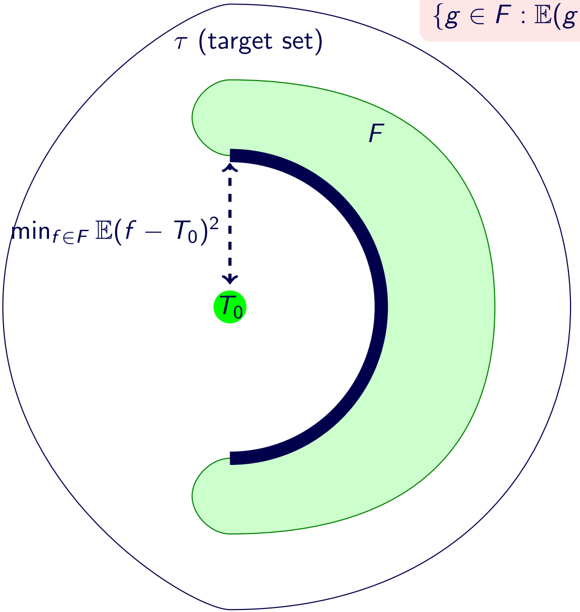
$\exists T_0 \in \tau$ s.t. $\mathrm{card}(\mathcal{M}(T_0)) \geq 2$ where

$$\mathcal{M}(T) = \{f \in F : \mathbb{E}(f(X) - T(X))^2 = \inf_{f\in F} \mathbb{E}(f(X) - T(X))^2\}$$


τ (target set)

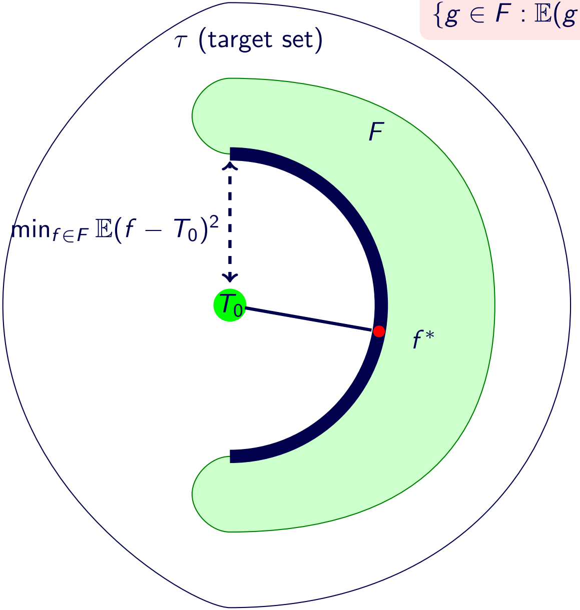


τ (target set)
F



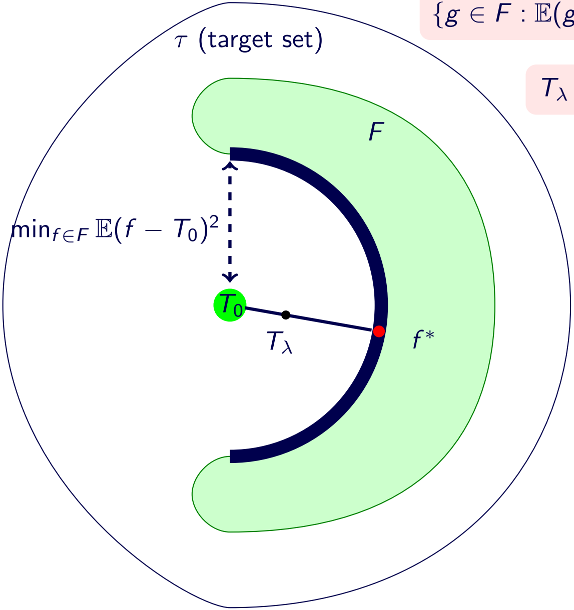
$\{g \in F : \mathbb{E}(g-T_0)^2 = \inf_F \mathbb{E}(f-T_0)^2\}$

$\{g \in F : \mathbb{E}(g - T_0)^2 = \inf_F \mathbb{E}(f - T_0)^2\}$



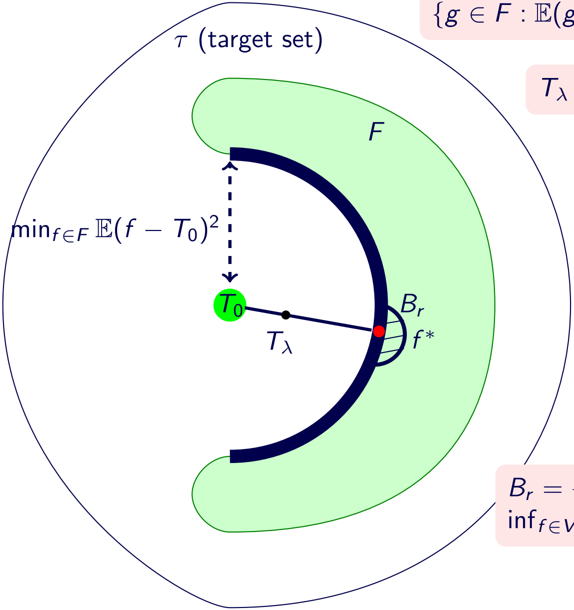
$$\{g \in F : \mathbb{E}(g - T_0)^2 = \inf_F \mathbb{E}(f - T_0)^2\}$$

$$T_\lambda = (1-\lambda)T_0 + \lambda f^*$$



$$\{g \in F : \mathbb{E}(g - T_0)^2 = \inf_F \mathbb{E}(f - T_0)^2\}$$

$$T_\lambda = (1-\lambda)T_0 + \lambda f^*$$



$$B_r = \{f : R(f) \le R(f^*) + r\}$$
$$\inf_{f\in V_m} R_n^\lambda(f) < \inf_{f\in B_r} R_n^\lambda(f)$$

Outline of the proof

The core of the proof is to find a set Q that can "compete" with $B_r = \{f \in F : \mathbb{E}\mathcal{L}_\lambda(f) \leq r\}$ in the sense that the empirical excess risk function

$$\mathcal{E}_n : f \in F \longmapsto \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_\lambda(f)(X_i) = P_n\mathcal{L}_\lambda(f)$$

will be more negative on Q than on it can possibly be on B_r ($\mathcal{L}_\lambda(f) := (f - T_\lambda)^2 - (f^* - T_\lambda)^2$).

Outline of the proof

The core of the proof is to find a set Q that can "compete" with $B_r = \{f \in F : \mathbb{E}\mathcal{L}_\lambda(f) \leq r\}$ in the sense that the empirical excess risk function

$$\mathcal{E}_n : f \in F \longmapsto \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_\lambda(f)(X_i) = P_n\mathcal{L}_\lambda(f)$$

will be more negative on Q than on it can possibly be on B_r ($\mathcal{L}_\lambda(f) := (f - T_\lambda)^2 - (f^* - T_\lambda)^2$).
Thus, the MRE $\hat{f}_\lambda \notin B_r$, and thus, with a certain probability,

$$\mathbb{E}\big[\mathcal{L}_\lambda(\hat{f}_\lambda)|D\big] > r.$$

Proof in two parts :

- $\mathcal{E}_n$ is likely to be very negative on $\{f \in F : \mathbb{E}\mathcal{L}(f) = \min_{f\in F} \mathbb{E}\mathcal{L}(f)\}$;
- find some r on which the oscillations of $\mathcal{E}_n$ in B_r are small.

Gaussian process end multidimensional CLT

Gaussian process : Let $Q \subset L_2(\mu)$.

Gaussian process end multidimensional CLT

Gaussian process : Let $Q \subset L_2(\mu)$. $(G_q)_{q\in Q}$ is a *Canonical Gaussian process* associated with Q when $\forall N \in \mathbb{N}, \forall q_1, \ldots, q_N \in Q$, $(G_{q_1}, \ldots, G_{q_N}) \sim \mathcal{N}_N(0, (\langle q_i, q_j \rangle)_{1\leq i,j\leq N})$.

Gaussian process end multidimensional CLT

Gaussian process : Let $Q \subset L_2(\mu)$. $(G_q)_{q\in Q}$ is a *Canonical Gaussian process* associated with Q when $\forall N \in \mathbb{N}, \forall q_1, \ldots, q_N \in Q$, $(G_{q_1}, \ldots, G_{q_N}) \sim \mathcal{N}_N(0, (\langle q_i, q_j \rangle)_{1\leq i,j\leq N})$.
A common measure of the "complexity" of Q is

$$H(Q) := \mathbb{E} \sup_{q\in Q} G_q$$

Gaussian process end multidimensional CLT

Gaussian process : Let $Q \subset L_2(\mu)$. $(G_q)_{q\in Q}$ is a *Canonical Gaussian process* associated with Q when $\forall N \in \mathbb{N}, \forall q_1, \ldots, q_N \in Q$, $(G_{q_1}, \ldots, G_{q_N}) \sim \mathcal{N}_N(0, (\langle q_i, q_j \rangle)_{1\leq i,j\leq N})$.
A common measure of the "complexity" of Q is

$$H(Q) := \mathbb{E} \sup_{q\in Q} G_q$$

Multidimensional CLT : $(V_i)_{i\in\mathbb{N}}$: sequence of d-dimensional i.i.d.r.v. with zero mean and finite L_2-norm.

Gaussian process end multidimensional CLT

Gaussian process : Let $Q \subset L_2(\mu)$. $(G_q)_{q \in Q}$ is a *Canonical Gaussian process* associated with Q when $\forall N \in \mathbb{N}, \forall q_1, \ldots, q_N \in Q$, $(G_{q_1}, \ldots, G_{q_N}) \sim \mathcal{N}_N(0, (\langle q_i, q_j \rangle)_{1 \leq i,j \leq N})$.
A common measure of the "complexity" of Q is

$$H(Q) := \mathbb{E} \sup_{q \in Q} G_q$$

Multidimensional CLT : $(V_i)_{i \in \mathbb{N}}$: sequence of d-dimensional i.i.d.r.v. with zero mean and finite L_2-norm.

$$\left| \mathbb{P}\Big(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_i \in A(t_1, \ldots, t_d)\Big) - \mathbb{P}(G \in A(t_1, \ldots, t_d)) \right| \longrightarrow 0 \text{ as } n \longrightarrow +\infty,$$

Gaussian process end multidimensional CLT

Gaussian process : Let $Q \subset L_2(\mu)$. $(G_q)_{q\in Q}$ is a *Canonical Gaussian process* associated with Q when $\forall N \in \mathbb{N}, \forall q_1, \ldots, q_N \in Q$, $(G_{q_1}, \ldots, G_{q_N}) \sim \mathcal{N}_N(0, (\langle q_i, q_j\rangle)_{1\le i,j\le N})$.
A common measure of the "complexity" of Q is

$$H(Q) := \mathbb{E} \sup_{q\in Q} G_q$$

Multidimensional CLT : $(V_i)_{i\in\mathbb{N}}$: sequence of d-dimensional i.i.d.r.v. with zero mean and finite L_2-norm.

$$\Big|\mathbb{P}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n V_i \in A(t_1,\ldots,t_d)\Big) - \mathbb{P}(G \in A(t_1,\ldots,t_d))\Big| \longrightarrow 0 \text{ as } n \longrightarrow +\infty,$$

where $A(t_1,\ldots,t_d) = \{v = (v_1,\ldots,v_d) \in \mathbb{R}^d : x_j \le t_j, \forall j\}$

Gaussian process end multidimensional CLT

Gaussian process : Let $Q \subset L_2(\mu)$. $(G_q)_{q\in Q}$ is a *Canonical Gaussian process* associated with Q when $\forall N \in \mathbb{N}, \forall q_1, \ldots, q_N \in Q$, $(G_{q_1}, \ldots, G_{q_N}) \sim \mathcal{N}_N(0, (\langle q_i, q_j \rangle)_{1\leq i,j\leq N})$.
A common measure of the "complexity" of Q is

$$H(Q) := \mathbb{E} \sup_{q\in Q} G_q$$

Multidimensional CLT : $(V_i)_{i\in\mathbb{N}}$: sequence of d-dimensional i.i.d.r.v. with zero mean and finite L_2-norm.

$$\left|\mathbb{P}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i \in A(t_1, \ldots, t_d)\Big) - \mathbb{P}(G \in A(t_1, \ldots, t_d))\right| \longrightarrow 0 \text{ as } n \longrightarrow +\infty,$$

where $A(t_1, \ldots, t_d) = \{v = (v_1, \ldots, v_d) \in \mathbb{R}^d : x_j \leq t_j, \forall j\}$ and G is a d-dimensional Gaussian process with zero mean and covariance matrix $(\mathbb{E}V^{(i)}V^{(j)})_{1\leq i,j\leq d}$.

Multivariate CLT outside B_r

Fix a finite set $Q' \subset Q := \{\mathcal{L}(f) : \mathbb{E}\mathcal{L}(f) = \min_{f\in F} \mathbb{E}\mathcal{L}(f)\}$ for which $H(Q') \geq H(Q)/2$ and $0 \in Q'$.

Multivariate CLT outside B_r

Fix a finite set $Q' \subset Q := \{\mathcal{L}(f) : \mathbb{E}\mathcal{L}(f) = \min_{f \in F} \mathbb{E}\mathcal{L}(f)\}$ for which $H(Q') \geq H(Q)/2$ and $0 \in Q'$.

Théorème

$\exists n_0 = n_0(Q')$ s.t. $\forall n \geq n_0$, with μ^n-probability at least c_1,

$$\inf_{\mathcal{L}(f) \in Q'} P_n \mathcal{L}_{\lambda_n}(f) \leq -c_2 \frac{H(Q)}{\sqrt{n}}$$

where $\lambda_n = c_3 H(Q)/\sqrt{n}$.

Uniform Central Limit Theorem

Recall that a bounded class of functions F is μ-Donsker if and only if for $\forall u > 0, \exists \delta > 0, \exists n_0$ s.t. $\forall n \geq n_0, \mathrm{osc}_n(F, \delta) \leq u$ where

$$\mathrm{osc}_n(F, \delta) = \mathbb{E} \sup_{\{f, h \in F : \|f - h\| \leq \delta\}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_i (f - h)(X_i) \right|,$$

where $g_1, \ldots, g_n$ are n i.i.d. standard Gaussian variables.

Uniform Central Limit Theorem

Recall that a bounded class of functions F is μ-Donsker if and only if for $\forall u > 0, \exists \delta > 0, \exists n_0$ s.t. $\forall n \geq n_0, \mathrm{osc}_n(F, \delta) \leq u$ where

$$\mathrm{osc}_n(F, \delta) = \mathbb{E} \sup_{\{f, h \in F : \|f - h\| \leq \delta\}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_i (f - h)(X_i) \right|,$$

where $g_1, \ldots, g_n$ are n i.i.d. standard Gaussian variables.

δ s.t. $\forall n \geq N(F)$, $\mathrm{osc}_n(F, f^*, \delta) \leq C_2 H(Q)/\sqrt{n}$.

UCLT around f^*

Now we are ready to control the oscillation of the empirical excess risk function uniformly over the set $B_r = \{f \in F : \mathbb{E}\mathcal{L}_\lambda \leq r\}$.

UCLT around f^*

Now we are ready to control the oscillation of the empirical excess risk function uniformly over the set $B_r = \{f \in F : \mathbb{E}\mathcal{L}_\lambda \leq r\}$.

Théorème

$\exists c_3$ s.t. $\forall n \geq n_1$, with μ^n-probability at least $1 - c_1/2$,

$$\inf_{\{f \in F : \mathbb{E}\mathcal{L}_{\lambda_n}(f) \leq r_n\}} P_n \mathcal{L}_{\lambda_n}(f) \geq -\frac{c_2 H(Q)}{2\sqrt{n}}$$

where

$$r_n = c_3 \frac{H(Q)}{\sqrt{n}} \delta^2 \| T - f^* \|^2$$

Exact and non-exact oracle inequalities for regularized MRE

▸ Oracle inequality for MRE

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

Regularized Empirical risk minimization - Part 1

A problem in learning theory is given by

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

Regularized Empirical risk minimization - Part 1

A problem in learning theory is given by

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

Regularized Empirical risk minimization - Part 1

A problem in learning theory is given by

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

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

Regularized Empirical risk minimization - Part 1

A problem in learning theory is given by

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

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

Example in regression : when we construct

$$\hat{f}_n^{MRE} \in \operatorname*{argmin}_{f \in F} \frac{1}{n} \sum_{i=1}^n (Y_i - f(X_i))^2,$$

Regularized Empirical risk minimization - Part 1

A problem in learning theory is given by

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

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

Example in regression : when we construct

$$\hat{f}_n^{MRE} \in \operatorname*{argmin}_{f \in F} \frac{1}{n} \sum_{i=1}^n (Y_i - f(X_i))^2,$$

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

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

Regularized Empirical risk minimization - Part 1

A problem in learning theory is given by

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

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

Example in regression : when we construct

$$\hat{f}_n^{MRE} \in \operatorname*{argmin}_{f \in F} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2,$$

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

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

(=> Oracle inequalities)

Regularized Empirical risk minimization - Part 1

A problem in learning theory is given by

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

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

Example in regression : when we construct

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

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

$$f_F^* \in \underset{f \in F}{\operatorname{argmin}} \, \mathbb{E}(Y - f(X))^2$$

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

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

Regularized Empirical risk minimization - Part 2

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

Regularized Empirical risk minimization - Part 2

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

Regularized Empirical risk minimization - Part 2

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

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

called a criterion. So that

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

Regularized Empirical risk minimization - Part 2

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

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

called a criterion. So that

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

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

Regularized Empirical risk minimization - Part 2

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

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

called a criterion. So that

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

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

Regularized Empirical risk minimization - Part 2

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

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

called a criterion. So that

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

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

Regularized Empirical risk minimization - Part 2

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

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

called a criterion. So that

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

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

Regularized Empirical risk minimization procedure - Part 3

Model :

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

Regularized Empirical risk minimization procedure - Part 3

Model :

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

Regularized Empirical risk minimization procedure - Part 3

Model :

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

Regularized Empirical risk minimization procedure - Part 3

Model :

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

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

Regularized Empirical risk minimization procedure - Part 3

Model :

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

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

$$\hat{f}_n^{RMRE} \in \underset{f \in \mathcal{F}}{\mathrm{argmin}} \big(R_n(f) + \mathrm{reg}(f)\big),$$

where $\mathrm{reg}(f) = \lambda \mathrm{crit}^\alpha(f)$;

Regularized Empirical risk minimization procedure - Part 3

Model :

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

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

$$\hat{f}_n^{RMRE} \in \operatorname*{argmin}_{f \in \mathcal{F}} \big(R_n(f) + \mathrm{reg}(f)\big),$$

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

Regularized Empirical risk minimization procedure - Part 3

Model :

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

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

$$\hat{f}_n^{RMRE} \in \operatorname*{argmin}_{f \in \mathcal{F}} \big(R_n(f) + \mathrm{reg}(f)\big),$$

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

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

Regularized Empirical risk minimization procedure - Part 3

Model :

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

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

$$\hat{f}_n^{RMRE} \in \operatorname*{argmin}_{f \in \mathcal{F}} \big(R_n(f) + \mathrm{reg}(f)\big),$$

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

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

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

Regularized Empirical risk minimization procedure - Part 3

Model :

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

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

$$\hat{f}_n^{RMRE} \in \operatorname*{argmin}_{f \in \mathcal{F}} \big(R_n(f) + \mathrm{reg}(f)\big),$$

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

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

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

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

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

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

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

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

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

For every $r \geq 0$:

- loss functions classes :

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

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

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

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

For every $r \geq 0$:

- loss functions classes :

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

- excess loss functions classes :

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

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

Theorem (L. and Mendelson)

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

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

Theorem (L. and Mendelson)

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

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

Theorem (L. and Mendelson)

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

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

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

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

Theorem (L. and Mendelson)

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

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

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

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

Let $x > 0$ and set

$$\hat{f}_n^{RMRE} \in \operatorname*{argmin}_{f \in \mathcal{F}} \left(R_n(f) + \frac{1}{1+\epsilon} \rho_n(\mathrm{crit}(f) + 1, x)\right).$$

Theorem (L. and Mendelson)

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

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

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

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

Let $x > 0$ and set

$$\hat{f}_n^{RMRE} \in \underset{f\in\mathcal{F}}{\operatorname{argmin}}\left(R_n(f) + \frac{1}{1+\epsilon}\rho_n(\mathrm{crit}(f)+1,x)\right).$$

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

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

Theorem (L. and Mendelson)

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

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

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

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

Let $x > 0$ and set

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

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

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

Theorem (Bartlett, Neeman and Mendelson)

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

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

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

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

Let $x>0$ *and set*

$$\hat{f}_n^{RMRE}\in \operatorname*{argmin}_{f\in\mathcal{F}}\Big(R_n(f)+\frac{1}{1+\epsilon}\rho_n(\mathrm{crit}(f)+1,x)\Big).$$

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

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

Conclusion on Exact and Non-exact oracle inequalities for RMRE

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

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

Conclusion on Exact and Non-exact oracle inequalities for RMRE

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

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

- loss functions classes :

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

Conclusion on Exact and Non-exact oracle inequalities for RMRE

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

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

- loss functions classes :

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

- excess loss functions classes :

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

Conclusion on Exact and Non-exact oracle inequalities for RMRE

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

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

- loss functions classes :

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

- excess loss functions classes :

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

1. RMRE with regularizing function $\mathrm{reg}(f) \gtrsim \lambda_\epsilon^*(\mathrm{crit}(f))$

Conclusion on Exact and Non-exact oracle inequalities for RMRE

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

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

- loss functions classes :

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

- excess loss functions classes :

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

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

Conclusion on Exact and Non-exact oracle inequalities for RMRE

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

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

- loss functions classes :

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

- excess loss functions classes :

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

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

Conclusion on Exact and Non-exact oracle inequalities for RMRE

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

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

- loss functions classes :

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

- excess loss functions classes :

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

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

Conclusion on Exact and Non-exact oracle inequalities for RMRE

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

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

- loss functions classes :

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

- excess loss functions classes :

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

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

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

Applications in matrix completion

Oracle inequality for MRE

Example in matrix completion

Model :

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

Example in matrix completion

Model :

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

Example in matrix completion

Model :

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

Notation :

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

Example in matrix completion

Model :

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

Notation :

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

Example in matrix completion

Model :

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

Notation :

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

Example in matrix completion

Model :

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

Notation :

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

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

Example in matrix completion

Model :

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

Notation :

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

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

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

Matrix Completion - Convexification

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

Matrix Completion - Convexification

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

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

Matrix Completion - Convexification

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

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

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

Matrix Completion - Convexification

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

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

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

bibliography :

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

Matrix Completion - Convexification

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

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

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

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

Matrix Completion - Application of the general result

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

Matrix Completion - Application of the general result

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

For non-exact oracle inequalities for RMRE :

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

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

Matrix Completion - Application of the general result

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

For non-exact oracle inequalities for RMRE :

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

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

For exact oracle inequalities for RMRE :

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

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

Computation of the fixed point

Computation of the fixed point

Lemma (L. and Mendelson)

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

Computation of the fixed point

Lemma (L. and Mendelson)

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

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

Computation of the fixed point

Lemma (L. and Mendelson)

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

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

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

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

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

Theorem (L. and Mendelson)

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

Theorem (L. and Mendelson)

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

Theorem (L. and Mendelson)

Assume that $\|Y\|_{\psi_q}, \| \|X\|_{S_2} \|_{\psi_q} \leq K(mT)$ *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 RMRE procedure*

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

Theorem (L. and Mendelson)

Assume that $\|Y\|_{\psi_q}, \|\|X\|_{S_2}\|_{\psi_q} \leq K(mT)$ *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 RMRE procedure*

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

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

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

where $\eta_\epsilon(n, mT, x) = c_1 K(mT)^q (\log n)^{(4q-2)/q}(x + \log n)$*.*

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 $\|X\|_{S_2}$.

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. 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. 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. We have fast rates $\sim \|A_0\|_{S_1}^2/n$.

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

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

Matrix Completion - Part 5

Remarks :

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

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

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

Matrix Completion - Part 5

Remarks :

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

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

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

Matrix Completion - Part 5

Remarks :

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

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

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

Matrix Completion - Part 5

Remarks :

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

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

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

Matrix Completion - Part 5

Remarks :

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

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

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

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

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

Theorem (Gaïffas and L.)

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

Theorem (Gaïffas and L.)

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

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

Theorem (Gaïffas and L.)

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

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

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

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

General model

Model-Notation-Aim in learning theory

Model :

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

Model-Notation-Aim in learning theory

Model :

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

Model-Notation-Aim in learning theory

Model :

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

Notation :

- ℓ_f : loss function of a function f

Model-Notation-Aim in learning theory

Model :

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

Notation :

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

Model-Notation-Aim in learning theory

Model :

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

Notation :

- ℓ_f : loss function of a function f
- $R(f) = \mathbb{E}\ell_f(Z)$: risk of a function f
- A statistic is a function $\hat{f}_n$ of the data $(Z_1, \cdots, Z_n)$ and its risk is

$$R(\hat{f}_n) = \mathbb{E}\big[\ell_{\hat{f}_n}(Z) | Z_1, \cdots, Z_n\big].$$

▸ Model input/output - square loss