Regularization methods in learning theory

Guillaume Lecué

CREST, ENSAE

Journées MAS - Rouen - August 2022

joint works with G. Chinot, M. Lerasle, S. Mendelson.

Definitions and Aims

1. **data:** $(X_i, Y_i)_{i=1}^N \overset{i.i.d.}{\sim} (X, Y) \in \mathcal{X} \times \mathbb{R}$.
2. **functions class:** $F \subset L_2(\mathcal{X}, \mu)$ where $X \sim \mu$
3. **oracle:** $f^* \in \operatorname{argmin}_{f \in F} \mathbb{E}(Y - f(X))^2$
4. **regularization function:** $\|\cdot\| : \operatorname{span}(F) \to \mathbb{R}$ (any norm)

Statistical goal: Estimate f^* in $L_2(\mu)$ knowing that f^* has some structure "related to $\|\cdot\|$" with the **Regularized Empirical risk minimization (RERM):**

$$\hat{f} \in \operatorname*{argmin}_{f \in F} \left(\frac{1}{N} \sum_{i=1}^{N} (Y_i - f(X_i))^2 + \lambda \|f\| \right) \quad \lambda : \text{ regularization parameter}$$

Aim of the talk: Find the key properties of F and $\|\cdot\|$ which drives the statistical performances of $\hat{f}$.
Aim of the talk: Only two parameters drive everything:

1. the size of the subdifferential of $\|\cdot\|$ around f^*
2. the local Gaussian mean widths of the family of sub-models $f^* + \rho B := \{f \in F : \|f - f^*\| \leq \rho\}, \rho > 0$.

Examples

LASSO:

1. $F = \{\langle \cdot, t \rangle : t \in \mathbb{R}^d\}$
2. $\|\langle \cdot, t \rangle\| = \|t\|_1$

SLOPE:

1. $F = \{\langle \cdot, t \rangle : t \in \mathbb{R}^d\}$
2. $\|\langle \cdot, t \rangle\| = \|t\|_{SLOPE} = \sum_{j=1}^{d} \sqrt{\log(ed/j)} t_j^\sharp$ where $t_1^\sharp \geq \cdots \geq t_d^\sharp$

Nuclear norm regularization:

1. $F = \{\langle \cdot, A \rangle : A \in \mathbb{R}^{m \times T}\}$
2. $\|\langle \cdot, A \rangle\| = \|A\|_{S_1} = \sum \sigma_j(A)$

SVM:

1. $F = RKHS$
2. $\|f\| = \|f\|_{RKHS}$

Subdifferential of $\|\cdot\|$ at f^*

Size of the subdifferential of $\|\cdot\|$

The subdifferential of $\|\cdot\|$ in f:

$$\partial \|\cdot\| (f) := \left\{ g \in \operatorname{span}(F) : \|f + h\| \geq \|f\| + \langle g, h \rangle, \forall h \in \operatorname{span}(F) \right\}$$

We have

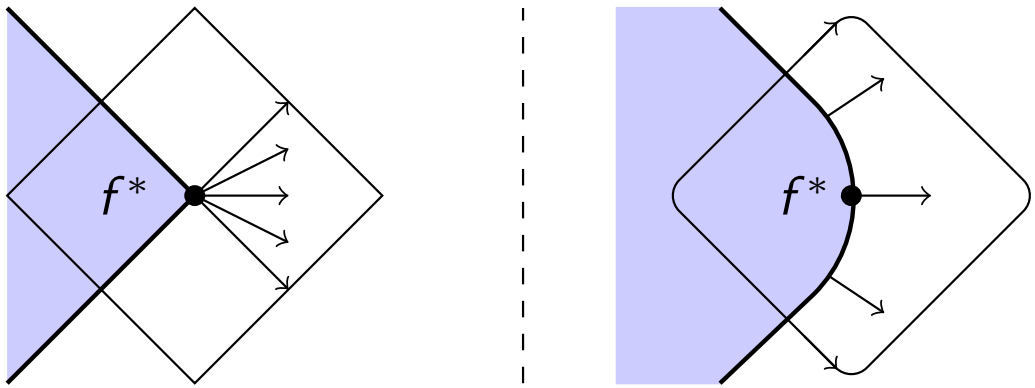
$$\partial \|\cdot\| (f) = \begin{cases} \{g \in S^* : \langle g, f \rangle = \|f\|\} & \text{if } f \neq 0 \\ B^* & \text{if } f = 0 \end{cases}$$

where for $\|f\|^* = \sup_{\|g\| \leq 1} \langle f, g \rangle$,

$$S^* = \{f : \|f\|^* = 1\} \text{ and } B^* = \{f : \|f\|^* \leq 1\}$$

Size of the subdifferential of $\|\cdot\|$

Subdifferential are large sets at points where $\|\cdot\|$ is not differentiable



Examples:

$$\partial \|\cdot\|_1 \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \left\{ \begin{pmatrix} 1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} : |x_i| \leq 1 \right\}; \qquad \partial \|\cdot\|_1 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \left\{ \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \right\}$$

Size of the subdifferential of $\|\cdot\|$

LASSO:

$$\partial \|\cdot\|_1 (t) \text{ is large when } t \text{ is sparse}$$

SLOPE:

$$\partial \|\cdot\|_{SLOPE} (t) \text{ is large when } t \text{ is sparse}$$

Nuclear norm:

$$\partial \|\cdot\|_{S_1} (A) \text{ is large when } A \text{ is low rank}$$

SVM:

$$\partial \|\cdot\|_{RKHS} (f) \text{ is never large except when } f = 0$$

$\ell_1, SLOPE, S_1$ are sparsity inducing norms: they promotes some "hidden" low dimensional structure

RKHS: no sparsity inducing power but some smoothing property.

Local Gaussian mean width of
$f^* + \rho B := \{f \in F : \|f^* - f\| \leq \rho\}, \rho > 0$

Definition

Let $H \subset L_2(\mu)$ and denote by $(G_h)_{h\in H}$ the canonical Gaussian process over H. The **Gaussian mean width of** H is

$$\ell^*(H) = \mathbb{E} \sup_{h\in H} G_h$$

Ex.: $H = \{\langle \cdot, t\rangle : t \in B_1^d\}$ then for $h = \langle \cdot, t\rangle$,

$$G_h = G_t = \sum_{j=1}^{d} g_j t_j = \langle G, t\rangle$$

where $g_1, \dots, g_d \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, $G = (g_1, \dots, g_d)^\top$ and

$$\ell^*(H) = \ell^*(B_1^d) = \mathbb{E} \sup_{t\in B_1^d} \langle G, t\rangle = \mathbb{E}\, \|G\|_\infty \sim \sqrt{\log(ed)}$$

1. $H = \{\langle \cdot, t \rangle : t \in B_{SLOPE}\}$ where

$$B_{SLOPE} = \left\{t \in \mathbb{R}^d : \|t\|_{SLOPE} \leq 1\right\} \text{ and } \|t\|_{SLOPE} = \sum_{j=1}^{d} \sqrt{\log(ed/j)}\, t_j^{\sharp},$$

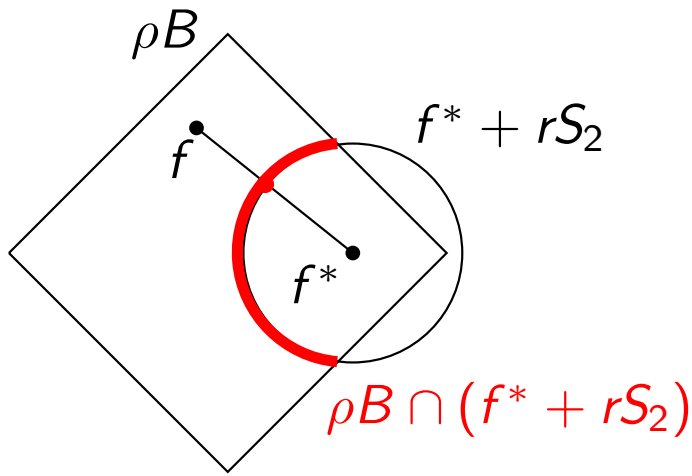
then for $h = \langle \cdot, t \rangle$, $G_h = G_t = \langle G, t \rangle$ and $\ell^*(H) = \ell^*(B_{SLOPE}) \sim 1$.

2. $H = \{\langle \cdot, A \rangle : A \in B_{S_1}\}$ then for $h = \langle \cdot, A \rangle$, $G_h = G_A = \langle G, A \rangle$ where G is a standard Gaussian matrix in $\mathbb{R}^{m \times T}$ and so

$$\ell^*(H) = \ell^*(B_{S_1}) = \mathbb{E} \sup_{\|A\|_{S_1} \leq 1} \langle G, A \rangle = \mathbb{E}\, \|G\|_{S_\infty} \sim \sqrt{m + T}$$

The complexity of $\rho B := \{f \in F : \|f\| \leq \rho\}, \rho > 0$

Statictical complexities are local: $\ell^*(\rho B \cap (f^* + rS_2))$



Definition Two complexity parameters of ρB

$$r_Q(\rho) = \inf\left\{r > 0 : \ell^*(\rho B \cap (rS_2 + f^*)) \leq r\sqrt{N}\right\}$$

$$r_M(\rho) = \inf\left\{r > 0 : \sigma\ell^*(\rho B \cap (rS_2 + f^*)) \leq r^2\sqrt{N}\right\}$$

$$r(\rho) = \max(r_Q(\rho), r_M(\rho))$$

The complexity of $\rho B := \{f \in F : \|f\| \leq \rho\}, \rho > 0$

Theorem (L. & Mendelson)

In the sub-Gaussian regression model $Y = f^*(X) + \sigma g$ where:

1. X is a sub-Gaussian vector and g is sub-gaussian,

then $r(\rho)^2$ is the rate of convergence achieved by the ERM over ρB: $\operatorname{argmin}_{f \in \rho B} \sum (Y_i - f(X_i))^2$.

Example: $\rho B = \{\langle \cdot, t \rangle : t \in \rho B_1^d\}$ and $\mathbb{E}\langle X, t \rangle^2 = \|t\|_2^2$,

$$(r_M(\rho)^2 \sim \begin{cases} \rho\sigma\sqrt{\frac{\log d}{N}} & \text{if } \rho^2 N \leq \sigma^2 \log d, \\ \rho\sigma\sqrt{\frac{1}{N}\log\left(\frac{ed^2\sigma^2}{\rho^2 N}\right)} & \text{if } \sigma^2 \log d \leq \rho^2 N \leq \sigma^2 d^2, \\ \frac{\sigma^2 d}{N} & \text{if } \rho^2 N \geq \sigma^2 d. \end{cases}$$

$$(r_Q(\rho))^2 \sim \begin{cases} \frac{\rho^2}{N}\log\left(\frac{ed}{N}\right) & \text{if } N \leq c_1 d, \\ 0 & \text{if } N > c_2 d. \end{cases}$$

Statistical properties of RERM

Assumptions : sub-gaussian framework

$$\|f(X)\|_{\psi_2} := \inf\left\{c > 0 : \mathbb{E}\exp\left(\frac{f^2(X)}{c^2}\right) \leq e\right\}$$

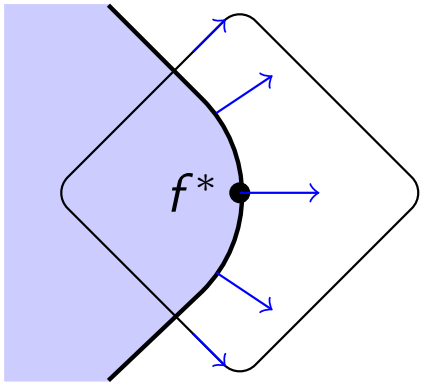
$$\|f(X)\|_{\psi_2} \leq L\|f(X)\|_{L_2} \Leftrightarrow \mathbb{P}[|f(X)| \geq tL\|f(X)\|_{L_2}] \leq 2\exp(-t^2/2), \forall t \geq 1.$$

Example: $f(X) = \langle X, t\rangle$ where $X = (x_1, \ldots, x_d)^\top$ and $x_1, \ldots, x_d \overset{i.i.d.}{\sim} x$ and $x \in L_{\psi_2}$.

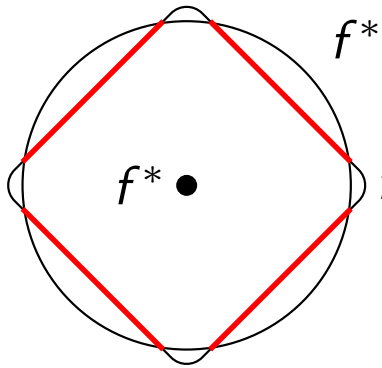
Assumption (subgaussian framework)

- A1 F is convex
- A2 $\|f(X)\|_{\psi_2} \leq L\|f(X)\|_{L_2}$ for all $f \in F$,
- A3 $\sigma := \|Y - f^*(X)\|_{\psi_2} < \infty$

The sparsity equation: a deterministic property



$$\Gamma_{f^*}(\rho) := \cup \{\partial \|\cdot\|(f) : f \in f^* + \rho B\}$$



$$H_\rho := F \cap (f^* + (\rho S \cap r(\rho)B_2))$$

$$\Delta(\rho) := \inf_{f \in H_\rho} \sup_{g \in \Gamma_{f^*}(\rho)} \langle g, f - f^* \rangle$$

Sparsity equation:

ρ^* is such that $\Delta(\rho^*) \geq 4\rho^*/5$

The sparsity equation: examples

LASSO: If there exists $v \in t^* + (\rho/20)B_1^d$ such that $\|v\|_0 = s$ then $\Delta(\rho) \geq 4\rho/5$ when

$$\sqrt{s} \lesssim \frac{\rho}{r(\rho)}$$

and so one can take

$$\rho^* \sim \sigma s \sqrt{\frac{\log(ed)}{N}}$$

SLOPE: If there exists $v \in t^* + (\rho/20)B_{SLOPE}$ such that $\|v\|_0 = s$ then $\Delta(\rho) \geq 4\rho/5$ when

$$\sum_{j=1}^{s} \sqrt{\frac{\log(ed/j)}{j}} \lesssim \frac{\rho}{r(\rho)}$$

and so one can take

$$\rho^* \sim \sigma s \frac{\log(ed)}{\sqrt{N}}$$

The sparsity equation: examples

Trace norm: If there exists $V \in A^* + (\rho/20)B_{S_1}$ such that $\mathrm{rank}(V) = r$ then $\Delta(\rho) \geq 4\rho/5$ when

$$\sqrt{s} \lesssim \frac{\rho}{r(\rho)}$$

and so one can take

$$\rho^* \sim \sigma r \sqrt{\frac{m+T}{N}}$$

RKHS norm: If $\rho^* > \|f^*\|$ then $0 \in f^* + \rho^* B$ and since

$$\partial \|\cdot\| (0) = B^*$$

then $\Delta(\rho^*) = \rho^*$ and the **sparsity equation is satisfied** for $\rho^* \sim \|f^*\|_{RKHS}$ (this works for any norm).

Main result in the sub-gaussian case (L.& Mendelson)

We assume that

A1 F is convex

A2 $\|f(X)\|_{\psi_2} \leq L \|f(X)\|_{L_2}$ for all $f \in F$,

A3 $\sigma := \|Y - f^*(X)\|_{\psi_2} < \infty$

Let ρ^* satisfy the sparsity equation $\Delta(\rho^*) \geq 4\rho^*/5$ then for the regularization parameter

$$\lambda = c_0 \frac{r(\rho^*)^2}{\rho^*}$$

the RERM $\hat{f}$ satisfies

$$\left\|\hat{f} - f^*\right\| \leq \rho^* \text{ and } \left\|\hat{f} - f^*\right\|_{L_2} \leq r(\rho^*)$$

with probability larger than

$$1 - 2\exp(-c_1 N \min(r_M(\rho^*)^2/\sigma^2), 1))$$

Examples

Applications: **LASSO**

Assume that:

1. $\mathbb{E}\langle X, t\rangle^2 = \|t\|_2^2$, $\forall t \in \mathbb{R}^d$
2. $\left\|\langle X, t\rangle\right\|_{\psi_2} \leq L\|t\|_2$, $\forall t \in \mathbb{R}^d$
3. $\left\|Y - \langle X, t^*\rangle\right\|_{\psi_2} = \sigma < \infty$ where $t^* \in \operatorname{argmin}_{t\in\mathbb{R}^d} \mathbb{E}(Y - \langle X, t\rangle)^2$
4. $\exists v \in t^* + (\rho^*/20)B_1^d$ such that $\|v\|_0 = s$ for $\rho^* \sim \sigma s\sqrt{\log(ed)/N}$
5. $N \gtrsim s\log(ed/s)$

then with probability larger than

$$1 - 2\exp\left(-c_0 s\sqrt{N\log(ed)}/\sigma\right)$$

for any $1 \leq p \leq 2$

$$\left\|\hat{t} - t^*\right\|_p \lesssim \sigma s^{1/p}\sqrt{\frac{\log(ed)}{N}}$$

where

$$\hat{t} \in \operatorname*{argmin}_{t\in\mathbb{R}^d}\left(\frac{1}{N}\sum_{i=1}^{N}(Y_i - \langle X_i, t\rangle)^2 + c_0\sigma\sqrt{\frac{\log(ed)}{N}}\|t\|_1\right)$$

Applications: **SLOPE**

Assume that:

1. $\mathbb{E}\langle X, t\rangle^2 = \|t\|_2^2$, $\forall t \in \mathbb{R}^d$
2. $\left\|\langle X, t\rangle\right\|_{\psi_2} \leq L\|t\|_2$, $\forall t \in \mathbb{R}^d$
3. $\left\|Y - \langle X, t^*\rangle\right\|_{\psi_2} = \sigma < \infty$ where $t^* \in \operatorname{argmin}_{t\in\mathbb{R}^d} \mathbb{E}(Y - \langle X, t\rangle)^2$
4. $\exists v \in t^* + (\rho^*/20)B_{SLOPE}$ s.t. $\|v\|_0 = s$ for $\rho^* \sim \sigma s \log(ed)/\sqrt{N}$
5. $N \gtrsim s\log(ed/s)$

then with probability larger than $1 - 2\exp\left(-c_0 s\sqrt{N}\log(ed)/\sigma\right)$

$$\left\|\hat{t} - t^*\right\|_{SLOPE} \lesssim \sigma s\frac{\log(ed/s)}{\sqrt{N}}, \quad \left\|\hat{t} - t^*\right\|_2 \lesssim \sigma\sqrt{\frac{s}{N}\log\left(\frac{ed}{s}\right)}$$

where

$$\hat{t} \in \operatorname*{argmin}_{t\in\mathbb{R}^d}\left(\frac{1}{N}\sum_{i=1}^{N}(Y_i - \langle X_i, t\rangle)^2 + \frac{c_0\sigma}{\sqrt{N}}\|t\|_{SLOPE}\right)$$

Applications: **Trace norm regularization**

Assume that:

1. $\mathbb{E}\langle X, A\rangle^2 = \|A\|_{S_2}^2$, $\forall A \in \mathbb{R}^{m\times T}$
2. $\left\|\langle X, A\rangle\right\|_{\psi_2} \le L\|A\|_{S_2}$, $\forall A \in \mathbb{R}^{m\times T}$
3. $\left\|Y - \langle X, A^*\rangle\right\|_{\psi_2} = \sigma$ where $A^* \in \operatorname{argmin}_{A\in\mathbb{R}^{m\times T}} \mathbb{E}(Y - \langle X, A\rangle)^2$
4. $\exists V \in A^* + (\rho^*/20)B_{S_1}$ s.t. $\operatorname{rank}(V) = r$ for $\rho^* \sim \sigma r\sqrt{(m+T)/N}$
5. $N \gtrsim r(m+T)$

then with probability larger than $1 - 2\exp\left(-c_0 s\sqrt{N(m+T)}/\sigma\right)$, for all $1 \le p \le 2$,

$$\left\|\hat{A} - A^*\right\|_{S_p} \lesssim \sigma r^{1/p}\sqrt{\frac{m+T}{N}},$$

where

$$\hat{A} \in \operatorname*{argmin}_{A\in\mathbb{R}^{m\times T}}\left(\frac{1}{N}\sum_{i=1}^{N}(Y_i - \langle X_i, A\rangle)^2 + c_0\sigma\sqrt{\frac{m+T}{N}}\|A\|_{S_1}\right)$$

Applications: **SVM** (1/2)

Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a (s.d.) kernel s.t. $\|K\|_{L_2} < \infty$ and denote by *RKHS* the associated RKHS. Denote by $(\lambda_j)_j$ the eigenvalues of $T_k : f \to \int K(\cdot, y) f(y) d\mu(y)$.

$$r_Q(\rho) = \inf \left\{ r > 0 : \left(\sum_j (\rho^2 \lambda_j) \wedge r^2 \right)^{1/2} \leq r\sqrt{N} \right\}$$

$$r_M(\rho) = \inf \left\{ r > 0 : \sigma \left(\sum_j (\rho^2 \lambda_j) \wedge r^2 \right)^{1/2} \leq r^2\sqrt{N} \right\}$$

$$r(\rho) = \max(r_Q(\rho), r_M(\rho))$$

$$\lambda_0 \sim \sigma \sup_{\rho > 0} \frac{\left(\sum_j (\rho^2 \lambda_j) \wedge r^2(\rho) \right)^{1/2}}{\rho\sqrt{N}}.$$

Applications: **SVM** (2/2)

Assume that:

1. $\|f(X)\|_{\psi_2} \leq L \|f(X)\|_{L_2}$, $\forall f \in RKHS$
2. $\|Y - f^*(X)\|_{\psi_2} = \sigma$ where $f^* \in \operatorname{argmin}_{f \in RKHS} \mathbb{E}(Y - f(X))^2$

then for $\rho^* \sim \|f^*\|_{RKHS}$, with probability larger than $1 - 2\exp\left(-c_0 N \min(r_M(\rho^*)^2/\sigma, 1)\right)$,

$$\left\|\hat{f} - f^*\right\|_{L_2} \lesssim r(\rho^*)$$

where

$$\hat{f} \in \operatorname*{argmin}_{f \in RKHS} \left(\frac{1}{N} \sum_{i=1}^{N} (Y_i - f(X_i))^2 + \lambda_0 \|f\| \right)$$

Key steps in the proof:

1) Homogenity argument

2) Quadratic / multiplier / regularization decomposition of the excess regularized risk

Definitions and strategy

The regularized loss:

$$\ell_f(x, y) + \lambda \|f\| = (y - f(x))^2 + \lambda \|f\|$$

The excess regularized loss:

$$\mathcal{L}_f^\lambda = \left(\ell_f + \lambda \|f\|\right) - \left(\ell_{f^*} + \lambda \|f^*\|\right)$$

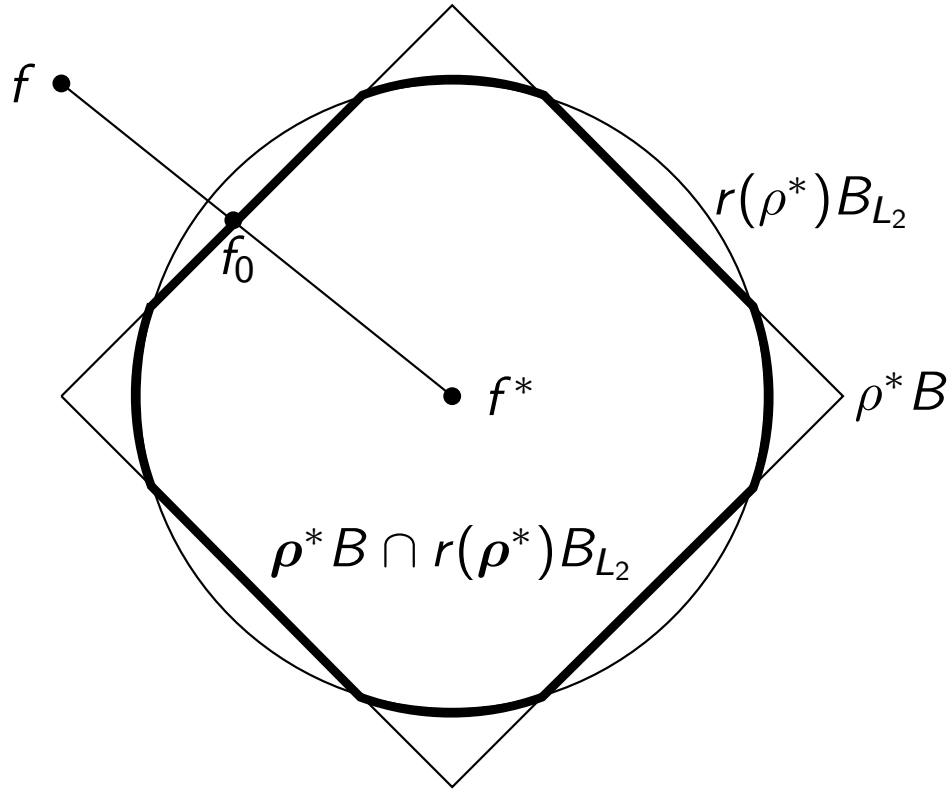
Strategy: show that for all $f \in F$

$$\|f - f^*\| > \rho^* \text{ or } \|f - f^*\|_{L_2} > r(\rho^*) \text{ implies } P_N \mathcal{L}_f^\lambda > 0.$$

This proves the result since the RERM $\hat{f}$ satisfies $P_N \mathcal{L}_{\hat{f}}^\lambda \leq 0$ where

$$P_N g = \frac{1}{N} \sum_{i=1}^{N} g(X_i, Y_i).$$

The homogenity argument



Let $f \in F$ be such that $\left\|\hat{f} - f^*\right\| > \rho^*$ or $\left\|\hat{f} - f^*\right\|_{L_2} > r(\rho^*)$ then there exists f_0 in the border of $\boldsymbol{\rho}^* B \cap r(\boldsymbol{\rho}^*)B_{L_2}$ such that $P_N \mathcal{L}_f^\lambda \geq \alpha P_N \mathcal{L}_{f_0}^\lambda$ for some $0 \leq \alpha \leq 1$.

CCL. show that $P_N \mathcal{L}_f^\lambda > 0$ for all f in the border of $\boldsymbol{\rho}^* B \cap r(\boldsymbol{\rho}^*)B_{L_2}$.

The quadratic / multiplier / regularization decomposition

Let $f \in F$ be in the border of $\boldsymbol{\rho}^* B \cap r(\boldsymbol{\rho}^*) B_{L_2}$. We want to prove that

$$P_N \mathcal{L}_f^\lambda > 0.$$

We decompose the excess regularized risk as

$$\begin{aligned}
P_N \mathcal{L}_f^\lambda &= P_N(\ell_f - \ell_{f^*}) + \lambda\big(\|f\| - \|f^*\|\big) \\
&= \frac{1}{N}\sum_{i=1}^N (Y_i - f(X_i))^2 - (Y_i - f^*(X_i))^2 + \lambda\big(\|f\| - \|f^*\|\big) \\
&= \frac{1}{N}\sum_{i=1}^N (f(X_i) - f^*(X_i))^2 \rightsquigarrow \text{quadratic process (Q)} \\
&+ \frac{2}{N}\sum_{i=1}^N (Y_i - f^*(X_i))(f^*(X_i) - f(X_i)) \rightsquigarrow \text{multiplier process (M)} \\
&+ \lambda\big(\|f\| - \|f^*\|\big) \rightsquigarrow \text{regularization term (R)}
\end{aligned}$$

3 tasks

- Lower bound on the quadratic process: w.h.p. for all $f \in f^* + \rho^* B$ such that $\|f - f^*\|_{L_2} \geq r_Q(\rho^*)$

$$\frac{1}{N}\sum_{i=1}^{N}(f(X_i) - f^*(X_i))^2 \geq \frac{1}{2}\|f - f^*\|_{L_2}^2 .$$

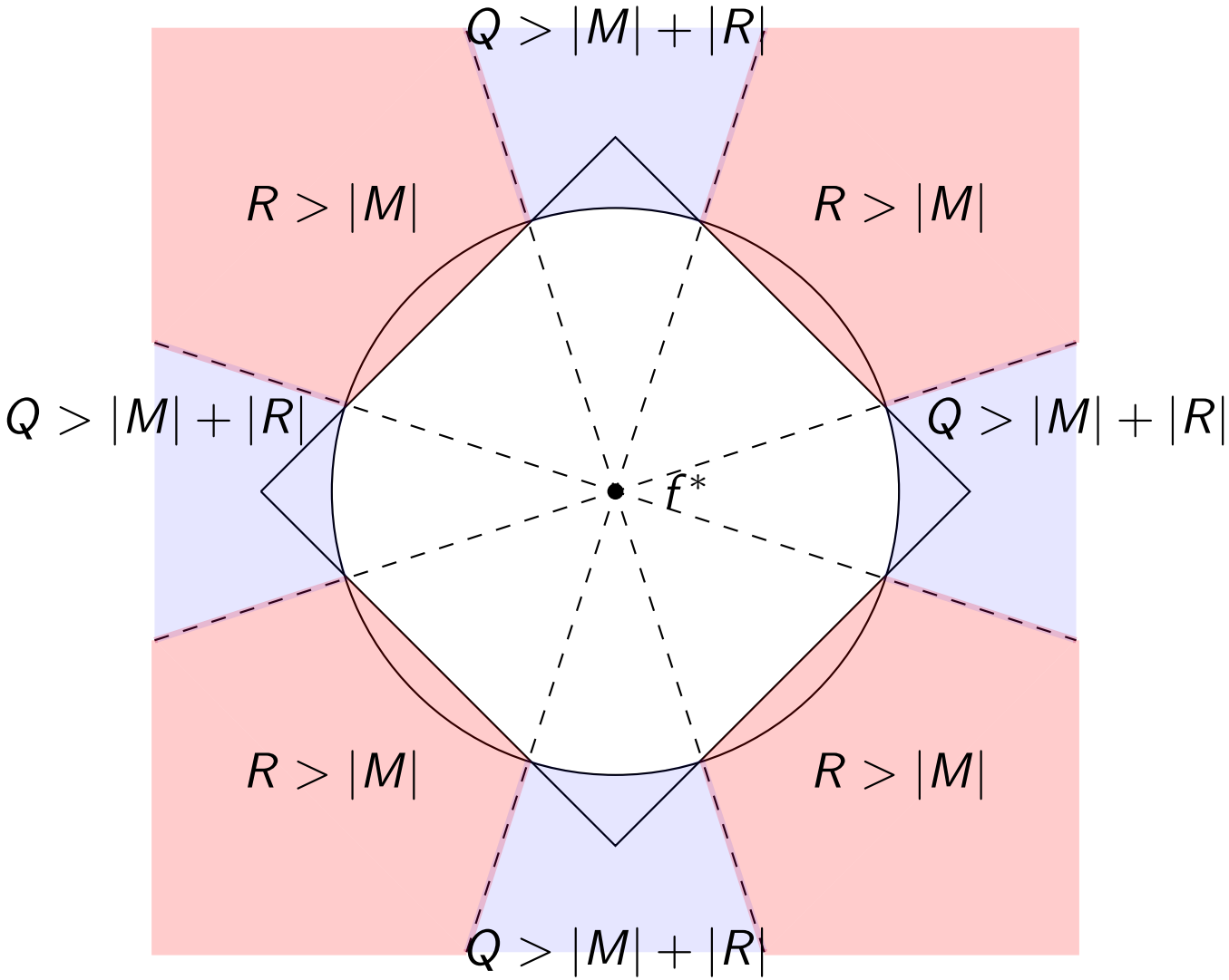
- Upper bounds on the multiplier process: w.h.p. for all $f \in f^* + \rho^* B$,

$$\left|\frac{1}{N}\sum_{i=1}^{N}(Y_i - f^*(X_i))(f(X_i) - f^*(X_i))\right| \lesssim \max\left((r_M(\rho^*))^2, \|f - f^*\|_{L_2}^2\right)$$

- lower on the regularization term via the sub-differential: for all f such that $\|f - f^*\| = \rho^*$ and $\|f - f^*\|_{L_2} \leq r(\rho^*)$,

$$\lambda\left(\|f\| - \|f^*\|\right) \geq c_0 \frac{(r(\rho^*))^2}{\rho^*} \sup_{g \in \partial\|\cdot\|(f^*)} \langle g, f - f^* \rangle \gtrsim (r(\rho^*))^2$$

Geometry of the Q/M/R decomposition



Quadratic loss function

- G. Lecué and S. Mendelson. Learning Subgaussian classes : Upper and minimax bounds
- G. Lecué and S. Mendelson. Regularization and the small-ball method II: complexity dependent error rates. JMLR 2017
- G. Lecué and S. Mendelson. Regularization and the small-ball method I: sparse recovery. AOS 2018

Lipschitz and linear loss functions

- P. Alquier, V. Cottet and G. Lecué. Estimation bounds and sharp oracle inequalities of regularized procedures with Lipschitz loss functions. AOS 2019.
- S. Chrétien, M. Cucuringu, G. Lecué and L. Neirac Learning with Semi-Definite Programming: new statistical bounds based on fixed point analysis and excess risk curvature. JMLR 2021.

MOM estimators

- G. Lecué and M. Lerasle. Robust Machine Learning by median of means: theory and practice. AOS 2020.
- G. Chinot, G. Lecué and M. Lerasle. Statistical Learning with Lipschitz and convex loss functions. PTRF 2020.
- G. Chinot, G. Lecué and M. Lerasle Robust high dimensional learning for Lipschitz and convex losses. JMLR 2020.

Thanks for your attention

Minoration of the quadratic process via the small ball method

Definition (Mendelson, Mendelson & Koltchinski)

The functions class F satisfies the small ball assumption when there exists $\kappa > 0$ and $\epsilon > 0$ such that for every $f, g \in F$,

$$\mathbb{P}\left[|f(X) - h(X)| \geq \kappa \|f - h\|_{L_2}\right] \geq \epsilon$$

Theorem (Mendelson)

If F satisfies the small ball assumption then w.p. $1 - 2\exp(-N\epsilon^2/2)$,

$$\frac{1}{N}\sum_{i=1}^{N}(f(X_i) - f^*(X_i))^2 \geq \frac{\kappa^2\epsilon}{16}\|f - f^*\|_{L_2}^2$$

for every $f \in F$ such that $\|f - f^*\|_{L_2} \geq r_Q$ for

$$r_Q = \inf\left(r > 0 : \mathbb{E}\sup_{f \in F \cap (f^* + rB_2)}\left|\frac{1}{N}\sum_{i=1}^{N}\epsilon_i(f - f^*)(X_i)\right| \leq \frac{\kappa\epsilon}{32}r\right)$$

LASSO under weak moment assumptions

Denote $X = (x_j)_{j=1}^d$ the random design. Assume that

1. $\mathbb{E}\langle X, t\rangle^2 = \|t\|_2^2$ for every $t \in \mathbb{R}^d$,
2. $\|x_j\|_{L_p} \leq \kappa_0 \sqrt{p}\, \|x_j\|_{L_2}$ for $p \sim \log d$
3. $\mathbb{P}\big[|\langle X, t\rangle| \geq \kappa \|t\|_2\big] \geq \epsilon$ for all $t \in \mathbb{R}^d$
4. $\sigma_q := \big\|Y - \langle X, t^*\rangle\big\|_{L_q}$ for some $q > 2$

Then, with probability larger than

$$1 - \frac{c_0 \log^q N}{N^{q/2-1}},$$

$$\big\|\hat{t} - t^*\big\|_2^2 \lesssim \min\left\{\frac{\sigma_q^2 \|t^*\|_0 \log(ed)}{N}, \sigma_q \|t^*\|_1 \sqrt{\frac{\log(ed)}{N}}\right\}$$

where

$$\hat{t} \in \operatorname*{argmin}_{t \in \mathbb{R}^d}\left(\frac{1}{N}\sum_{i=1}^N (Y_i - \langle X_i, t\rangle)^2 + c_0 \sigma_q \sqrt{\frac{\log(ed)}{N}}\, \|t\|_1\right)$$

Conclusion

We recover the classical rates of convergence of the LASSO for the same regularization parameter

1. under weak moments assumptions
2. under the small ball property
3. no statistical model
4. the sparsity of t^* appears because the subdifferential of $\|\cdot\|_1$ is large at sparse vectors

Rates of convergence of ERM

Some complexity measures of a set $\mathcal{F} \subset L_2(\mu)$?

1. Gaussian mean width

$$\mathbb{E}\,\|G\|_{\mathcal{F}} = \mathbb{E} \sup_{f \in \mathcal{F}} G_f$$

where $(G_f)_{f \in \mathcal{F}}$ is the canonical Gaussian process indexed by $\mathcal{F} \subset L_2(\mu)$.
(Ex.: $\mathcal{F} = \{\langle \cdot, t \rangle : t \in T\}$, $T \subset \mathbb{R}^d$ then $G_t = \langle G, t \rangle$)

2. covering - entropy

$$N(\mathcal{F}, \epsilon D)$$

is the minimal number of translated of the ball ϵD needed to cover $\mathcal{F}$ (D is the unit ball of $L_2(\mu)$).

3. Gelfand k-width : $c_k(\mathcal{F}) = \inf_{L: L_2(\mu) \mapsto \mathbb{R}^k} \operatorname{diam}(\mathcal{F} \cap \ker L, L_2(\mu))$.

4. statistical complexity: minimax rate of convergence over $\mathcal{F}$.

How are they related?

$$\sup_{\epsilon>0} \epsilon\sqrt{\log N(\mathcal{F}, \epsilon D)} \quad \underset{\uparrow}{\lesssim} \quad \mathbb{E}\,\|G\|_{\mathcal{F}} \quad \underset{\uparrow}{\lesssim} \quad \int \sqrt{\log N(\mathcal{F}, \epsilon D)}\, d\epsilon$$

Sudakov *Dudley*

$$\sup_{\epsilon>0} \epsilon\sqrt{\log N(\mathcal{F}, \epsilon D)} \quad \underset{\uparrow}{\lesssim} \quad \sup_{k\in\mathbb{N}} \sqrt{k} c_k(\mathcal{F}) \quad \underset{\uparrow}{\lesssim} \quad \mathbb{E}\,\|G\|_{\mathcal{F}}$$

Carl *Pajor / Tomczak – Jaegermann*

($\mathcal{F}$ convex body) ($\mathcal{F}$ star-shapped in 0)

ex.: $\mathcal{F} = \{\langle \cdot, t\rangle : t \in B_1^d\}$, $X \sim \mu$ is isotropic (i.e. $\mathbb{E}\langle X, t\rangle^2 = \|t\|_{\ell_2^d}^2$) then Sudakov, Carl and [P./T.-J.] are sharp $= \sqrt{\log d}$ but Dudley is not sharp $= (\log d)^{3/2}$.

Learning theory framework and ERM

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

model: $\mathcal{F} \subset L_2(\mathcal{X}, \mu)$ (where $X \sim \mu$),

estimator: Emprical risk minimization (ERM) :

$$\hat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} R_N(f) \text{ where } R_N(f) = \frac{1}{N} \sum_{i=1}^{N} \big(Y_i - f(X_i)\big)^2.$$

results : Fix $0 < \delta < 1$. With probability greater than $1 - \delta$,

$$R(\hat{f}) \leq \inf_{f \in \mathcal{F}} R(f) + r_N(\mathcal{F}, \delta)$$

where $R(f) = \mathbb{E}(Y - f(X))^2$, $R(\hat{f}) = \mathbb{E}\big[(Y - \hat{f}(X))^2 | (X_i, Y_i)\big]$.

questions :

a) How large is $r_N(\mathcal{F}, \delta)$? (complexity of $\mathcal{F}$, value of δ,...)

b) Can we do better than ERM ? (minimax results - depending on δ, the complexity structure of $\mathcal{F}$,...)

Assumptions: sub-gaussian and convex class framework

1. $\mathcal{F}$ is L-sub-Gaussian: $\forall f, g \in \mathcal{F} \cup \{0\}$,

$$\|f - g\|_{\psi_2(\mu)} \leq L \|f - g\|_{L_2(\mu)}$$

$(\|f\|_{\psi_2(\mu)} = \inf \left(c > 0 : \mathbb{E} \exp(f^2(X)/c^2) \leq 2\right) \sim \sup_{p \geq 1} \frac{\|f\|_{L_p}}{\sqrt{p}})$.

2. $\|Y - f^*_{\mathcal{F}}(X)\|_{\psi_2} \leq \sigma$ (noise level) where $f^*_{\mathcal{F}} \in \operatorname{argmin}_{f \in \mathcal{F}} R(f)$

3. $\mathcal{F}$ is convex

Theorem [L.& Mendelson]: sharp oracle inequality for ERM in Sub-Gaussian framework

Let D be the unit ball in $L_2(\mu)$ and define

$$s_N^* = \inf\left\{0 < s \le d_{\mathcal{F}}(L_2) : \mathbb{E}\|G\|_{(\mathcal{F}-f_{\mathcal{F}}^*)\cap sD} \le (c_0/\sigma)s^2\sqrt{N}\right\},$$

$$r_N^* = \inf\left\{0 < r \le d_{\mathcal{F}}(L_2) : \mathbb{E}\|G\|_{(\mathcal{F}-f_{\mathcal{F}}^*)\cap rD} \le c_1 r\sqrt{N}\right\}.$$

1. If $\sigma \ge c_3 r_N^*$ then with probability at least $1 - 4\exp(-c_4 N\sigma^{-2}(s_N^*)^2)$,

$$R(\hat{f}) \le \inf_{f\in\mathcal{F}} R(f) + (s_N^*)^2,$$

2. If $\sigma \le c_3 r_N^*$ then with probability at least $1 - 4\exp(-c_4 N)$,

$$R(\hat{f}) \le \inf_{f\in\mathcal{F}} R(f) + (r_N^*)^2.$$

Classical fixed points - two main streams

<1996 Fixed points were associated to Dudley entropy integrals: [van de Geer, AOS90, AOS93, EP in M-estimation] or [Birgé, Massart PTRF93]: $\text{residue} = (t_N^*)^2$

$$t_N^* = \inf\left(s > 0 : \int_{c_0 s^2}^{c_1 s} \sqrt{\log N\big((\mathcal{F} - f_{\mathcal{F}}^*) \cap sD, \epsilon D\big)}\, d\epsilon \leq (c_2/\sigma) s^2 \sqrt{N}\right).$$

>1996 Fixed points were associated to the expected supremum of the empirical process (indexed by localized classes) or weighted, symmetrized version, ...: [Massart, Saint Flour 2003] [Koltchinksii, Saint Flour 2008], [Bartlett, Mendelson, PTRF06], [Blanchard, Bousquet, Massart]:

$$\text{residue} = \inf\left\{ s > 0 : \mathbb{E} \sup_{\{f \in \mathcal{F} : P\mathcal{L}_f \leq s\}} |(P - P_N)\mathcal{L}_f| \leq c_0 s \right\}.$$

2 regimes for the noise - 2 statistical complexities - 2 empirical processes

1. If $\sigma \geq c_3 r_N^*$ then $\text{residue} = (s_N^*)^2$
2. If $\sigma \leq c_3 r_N^*$ then $\text{residue} = (r_N^*)^2$

We want to be as good as $f_{\mathcal{F}}^*$ using observations $(X_i, Y_i)_{i=1}^N$. There are two different sources of statistical complexity:

- the projection $P : f \in L_2(\mu) \mapsto \big(f(X_i)\big)_{i=1}^N$ is a source of complexity because we want procedures having good "generalization" capabilities (being good even outside of the data sample). Main source of complexity when $\sigma \lesssim r_N^*$.
- the noise $\|Y - f_{\mathcal{F}}^*(X)\|_{\psi_2} = \sigma$ is a source of complexity because *it is a noise !*: the values $f_{\mathcal{F}}^*(X_i)$ are hidden by the "noise" $Y_i - f_{\mathcal{F}}^*(X_i)$. Main source of complexity when $\sigma \gtrsim r_N^*$.

Decomposition of the excess loss function:

$$\begin{aligned}\mathcal{L}_f(x,y) = (\ell_f - \ell_{f_{\mathcal{F}}^*})(x,y) &= \big(y - f(x)\big)^2 - \big(y - f_{\mathcal{F}}^*(x)\big)^2 \\ &= (f - f_{\mathcal{F}}^*)^2(x) + 2(y - f_{\mathcal{F}}^*(x))(f_{\mathcal{F}}^* - f)(x)\end{aligned}$$

2 regimes for the noise - 2 statistical complexities - 2 empirical processes

1. The quadratic process $((P - P_N)(f - f_{\mathcal{F}}^*)^2)_{f \in \mathcal{F} \cap (f_{\mathcal{F}}^* + rD)}$. [Mendelson-Pajor-Tomczak]: w.h.p.

$$\sup_{h \in \mathcal{H}} \left| \frac{1}{N} \sum_{i=1}^{N} h^2(X_i) - \mathbb{E}h^2(X) \right| \lesssim \left(d_{\psi_2}(\mathcal{H}) \frac{\mathbb{E}\|G\|_{\mathcal{H}}}{\sqrt{N}} + \frac{(\mathbb{E}\|G\|_{\mathcal{H}})^2}{N} \right).$$

   This measures the statistical complexity coming from the projection via r_N^*.

2. The linear process $((P - P_N)(y - f_{\mathcal{F}}^*)(f_{\mathcal{F}}^* - f))_{f \in \mathcal{F} \cap (f_{\mathcal{F}}^* + rD)}$. [Mendelson]: w.h.p.

$$\sup_{h \in \mathcal{H}} \left| \frac{1}{N} \sum_{i=1}^{N} \xi_i h(X_i) - \mathbb{E}\xi h(X) \right| \lesssim \|\xi\|_{\psi_2} \frac{\mathbb{E}\|G\|_{\mathcal{H}}}{\sqrt{N}}.$$

   This measures the statistical complexity coming from the noise $(= \|\xi\|_{\psi_2} = \|Y - f_{\mathcal{F}}^*(X)\|_{\psi_2} = \sigma)$ via s_N^*.

Can we do better? (better rates for ERM? - better procedures than ERM?)

Consider the gaussian regression model:

$$Y = f^*(X) + W \text{ where } W \sim \mathcal{N}(0, \sigma^2) \text{ indep. of } X \text{ and } f^* \in \mathcal{F}$$

(note that $f^* = f^*_{\mathcal{F}} \in \mathcal{F}$).

If $\sigma \gtrsim r^*_N$ then with probability at least $1 - 4\exp(-c_4 N\sigma^{-2}(s^*_N)^2)$, $R(\hat{f}_{ERM}) \leq \inf_{f\in\mathcal{F}} R(f) + (s^*_N)^2(s^*_N)^2$.

Theorem (L.& Mendelson - minimax lower bound)

If $\tilde{f}_N$ is a procedure such that, for every $f^* \in \mathcal{F}$, with probability at least $1 - 4\exp(-c_4 N\sigma^{-2}(s^*_N)^2)$, $R(\tilde{f}_N) \leq \inf_{f\in\mathcal{F}} R(f) +$ residue, then necessarily residue $\gtrsim (s^*_N)^2(s^*_N)^2$.

ERM is minimax in the Gaussian regression model over sub-Gaussian and convex models (for this confidence bound and noise level $\sigma \gtrsim r^*_N$).

ERM is minimax for high confidence but not for constant confidence

Corollary

In the Gaussian regression model with respect to a sub-Gaussian and convex model for the confidence $1 - 4\exp(-c_4 N\sigma^{-2}(s_N^*)^2)$ *and for a noise level* $\sigma \gtrsim r_N^*$*, ERM is minimax.*

Theorem (Birgé and Massart, PTRF93)

In the Gaussian regression model over 1*-dimensional* α*-Hölderian spaces,*

1. *the ERM is minimax in* **expectation** *when* $\alpha > 1/2$*,*
2. *the ERM is not minimax* **in the constant regime** *when* $\alpha < 1/2$*: it satisfies a* $n^{-\alpha/2}$ *lower bound with constant probability (the minimax rate being* $n^{-\alpha/(2\alpha+1)}$*).*

Remark: Note that Gaussian mean widths of localized sets of $\alpha-$Hölderian classe for $\alpha < 1/2$ are infinite.

1. This is a "high confidence minimax bound". Classical minimax bounds are given in expectation or with constant probability. We used the *Gaussian shift theorem*.
2. For this "high confidence" minimax bound, two points in $\mathcal{F}$ are enough. We did not use the complexity (or richness) of $\mathcal{F}$.
3. What is hard (from a statistical point of view) in learning with high confidence bound is *the high confidence bound* (not the complexity or shape of $\mathcal{F}$).
4. What happens, if we want to learn with constant probability in the Gaussian regression model? Construct $\tilde{f}_N$ such that, for every $f^* \in \mathcal{F}$, with probability greater than $3/4$,

$$\left\| \tilde{f}_N - f^* \right\|_2^2 = R(\tilde{f}_N) - \inf_{f \in \mathcal{F}} R(f) \leq \text{residue}.$$

This is where the complexity of $\mathcal{F}$ comes into the game.

1. what complexity ? (are we going to recover the Gaussian complexity of localized sets obtained in the upper bound ?) No - Sudakov - Gelfand widths
2. can we use the Gaussian shift theorem in this case? Yes
3. are we going to recover the classical minimax results in this regime? Yes [Le Cam - Birgé - Tsybakov - Yang/Barron - Fano - Assouad]

Minimax result for constant confidence - large noise

Theorem (L. & Mendelson)

If $\tilde{f}_N$ is a procedure in the Gaussian regression model $Y = f^*(X) + W$ ($W \sim \mathcal{N}(0, \sigma^2)$ ind. of X) such that for every $f^* \in \mathcal{F}$, with probability greater than $3/4$, $R(\tilde{f}_N) \leq \inf_{f \in \mathcal{F}} R(f) + \text{residue}$ then necessarily

$$\text{residue} \gtrsim (q_N^*)^2$$

where

$$q_N^* = \inf\left\{ s > 0 : s\sqrt{\log N\big((\mathcal{F} - f^*) \cap 2sD, sD\big)} \leq (c_0/\sigma) s^2 \sqrt{N} \right\}.$$

Sudakov inequality (for the localized set $(\mathcal{F} - f^*) \cap 2sD$):

$$\sup_{0 < \epsilon < 2s} \epsilon \sqrt{\log N\big((\mathcal{F} - f^*) \cap 2sD, \epsilon D\big)} \lesssim \mathbb{E} \left\| G \right\|_{(\mathcal{F} - f^*) \cap 2sD}.$$

"Sudakov complexity" of the localized set $(\mathcal{F} - f^*) \cap 2sD$ at level s:

$$s\sqrt{\log N\big((\mathcal{F} - f^*) \cap 2sD, sD\big)}$$

The same result follows from

1. Yang and Barron,"Information-theoretic determination of minimax rates of convergence". AOS 1999; via Fano inequality.
2. Tsybakov, "Introduction to non-parametric estimation". Springer 2009 (cf. Theorem 2.5); via second Pinsker inequality and the Kullback-Leiber divergence between two Gaussian measures.

Here, via the Gaussian shift theorem; i.e. the Gaussian isoperimetry (cf. [Li& Kuelbs]).
This minimax result is to be compared with the result of the upper bound in the large noise regime ($\sigma \gtrsim r_N^*$).

1. minimax lower bound $= (q_N^*)^2$ where

$$q_N^* = \inf_{s>0} \left\{ s\sqrt{\log N\big((\mathcal{F}-f^*)\cap 2sD, sD\big)} \leq (c_0/\sigma)s^2\sqrt{N} \right\}$$

2. upper bound for ERM $= (s_N^*)^2$ where

$$s_N^* = \inf_{0<s\leq d_{\mathcal{F}}(L_2)} \left\{ \mathbb{E}\|G\|_{(\mathcal{F}-f^*)\cap 2sD} \leq (c_0/\sigma)s^2\sqrt{N} \right\}$$

conclusion for large noise

$$\sigma \gtrsim r_N^* = \inf_r \left\{ \mathbb{E}\|G\|_{(\mathcal{F}-f^*)\cap rD} \le c_1 r\sqrt{N} \right\}$$

w.p.g. $1 - 4\exp(-c_0 N\sigma^{-2}(s_N^*)^2)$, $R(\hat{f}_{ERM}) \le \inf_{f\in\mathcal{F}} R(f) + (s_N^*)^2$, where

$$s_N^* = \inf\left\{0 < s \le d_{\mathcal{F}}(L_2) : \mathbb{E}\|G\|_{(\mathcal{F}-f^*)\cap sD} \le (c_0/\sigma)s^2\sqrt{N}\right\}.$$

In the Gaussian regression model, if a procedure satisfies a sharp oracle inequality

- with the same confidence then $\text{residue} \ge (s_N^*)^2$;
- with constant probability then $\text{residue} \ge (q_N^*)^2$ where

$$q_N^* = \inf\left\{s > 0 : s\sqrt{\log N((\mathcal{F}-f^*)\cap 2sD, sD)} \le (c_0/\sigma)s^2\sqrt{N}\right\}.$$

If "Sudakov is sharp at the level q_N^*":

$$q_N^*\sqrt{\log N((\mathcal{F}-f^*)\cap 2q_N^*D, q_N^*D)} \sim \mathbb{E}\,\|G\|_{(\mathcal{F}-f^*)\cap 2q_N^*D}$$

then upper and lower bounds match and therefore ERM is minimax in the Gaussian model for any subgaussian and convex model for both exponentially large and constant confidences.

conclusion for small noise $\sigma \lesssim r_N^*$

w.p.g. $1 - 4\exp(-c_4 N)$, $R(\hat{f}_{ERM}) \leq \inf_{f \in \mathcal{F}} R(f) + (r_N^*)^2$, where

$$r_N^* = \inf\left\{0 < r \leq d_{\mathcal{F}}(L_2) : \mathbb{E}\|G\|_{(\mathcal{F}-f^*)\cap rD} \leq c_1 r\sqrt{N}\right\}.$$

In the Gaussian regression model, if a procedure satisfies (for any $f^* \in \mathcal{F}$) a sharp oracle inequality w.p.g. $1/2$ then
$\text{residue} \gtrsim \inf_{f^* \in \mathcal{F}} \left(c_N(\mathcal{F} - f^*)\right)^2$.
If "Pajor/Tomczak-Jaegermann is sharp at level N" (for some $f_0^* \in \mathcal{F}$):

$$\sqrt{N} c_N((\mathcal{F} - f_0^*) \cap r_N^* D) \sim \mathbb{E}\|G\|_{(\mathcal{F}-f_0^*)\cap r_N^* D}$$

then upper and lower bounds match and therefore ERM is minimax in the Gaussian model for any subgaussian and convex model for both exponentially large and constant confidences.

An example of application - ERM over the unit ball of the MAX-norm

data: $(X_i, Y_i)_{i=1}^N$ i.i.d. $\in \mathbb{R}^{p\times q} \times \mathbb{R}$,

model: $\mathcal{F} = \{\langle \cdot, A\rangle : \|A\|_{max} \le R\}$,
$\|A\|_{max} = \min_{A=UV^\top} \|U\|_{2\mapsto\infty} \|V\|_{2\mapsto\infty}$.

estimator: Emprical risk minimization (ERM) :

$$\hat{A} \in \operatorname*{argmin}_{\|A\|_{max}\le R} \frac{1}{N}\sum_{i=1}^N \left(Y_i - \langle X_i, A\rangle\right)^2.$$

Assumptions:

- X is isotropic ($\mathbb{E}\langle A, X\rangle^2 = (pq)^{-1}\|A\|_F^2$) and subgaussian ($\left\|\langle X, A\rangle\right\|_{\psi_2} \le L(pq)^{-1}\|A\|_F$),
- $A^*_{max} \in \operatorname{argmin}_{\|A\|_{max}\le R} \mathbb{E}(Y - \langle X, A\rangle)^2$ and $\left\|Y - \langle X, A^*_{max}\rangle\right\|_{\psi_2} \le \sigma$.

Gaussian mean width: $\operatorname{conv}\mathcal{X}_\pm \subset \mathcal{B}_{max} \subset K_G \operatorname{conv}\mathcal{X}_\pm$ where $\mathcal{X}_\pm = \{uv^\top : u \in \{\pm 1\}^p, v \in \{\pm 1\}^q\}$. Let $\mathfrak{G} = (\mathcal{N}(0, (pq)^{-1})_{ij}) \in \mathbb{R}^{p\times q}$

$$\mathbb{E}\|G\|_{(\mathcal{F}-f^*)\cap sD} = \mathbb{E}\sup_{\|A\|_{max}\le R; \|A\|_F \le s\sqrt{pq}} \langle \mathfrak{G}, A\rangle$$

$$\le K_G R \mathbb{E}\sup_{A\in\mathcal{X}_\pm} \langle \mathfrak{G}, A\rangle \le K_G R \max_{A\in\mathcal{X}_\pm} \frac{\|A\|_F}{\sqrt{pq}} \sqrt{\log|\mathcal{X}_\pm|} \le K_G R\sqrt{p+q}.$$

Therefore, $(s_N^*)^2 \sim \sigma R\sqrt{(p+q)/N}$ and $(r_N^*)^2 \sim R(p+q)/N$.
If $\sigma \gtrsim R\sqrt{(p+q)/N}$ then w.p.g. $1-2\exp(-c_1\sqrt{N(p+q)}/(\sigma R))$,

$$\mathbb{E}(Y - \langle \hat{A}, X \rangle)^2 \leq \inf_{A \in R\mathcal{B}_{max}} \mathbb{E}(Y - \langle A, X \rangle)^2 + c_2 \sigma R \sqrt{\frac{p+q}{N}}.$$

If $\sigma \lesssim R\sqrt{(p+q)/N}$, then w.p.g. $1-2\exp(-c_1 N)$,

$$\mathbb{E}(Y - \langle \hat{A}, X \rangle)^2 \leq \inf_{A \in R\mathcal{B}_{max}} \mathbb{E}(Y - \langle A, X \rangle)^2 + c_2 R \frac{p+q}{N}.$$

In the Gaussian linear model ($Y = \langle A^*, X \rangle + W$), we obtain a minimax bound (for constant and exponentially large confidence)

$$R\sqrt{\frac{p+q}{N}}$$

via the entropy estimate of [Cai&Wenxin, 2013].
Therefore, ERM is minmax over the MAX-norm model in the Gaussian linear model. A similar result was obtained in [Cai&Wenxin, 2013] for the ERM over $R\mathcal{B}_{max} \cap (\alpha B_\infty^{pq})$.

Take home message on the minimaximality of ERM

For the model $Y = f^*(X) + W$ where $W \sim \mathcal{N}(0, \sigma^2)$ is ind. of X and $f^* \in \mathcal{F}$ — a sub-gaussian and convex class, we have

1. in the large noise regime $\sigma \gtrsim r_N^*$,

complexity	constant confidence	large confidence
ERM (sharp)	Gaussian $((s_N^*)^2)$	Gaussian $((s_N^*)^2)$
minimax (achievable)	Sudakov $((q_N^*)^2)$	Gaussian $((s_N^*)^2)$

2. in the small noise regime $\sigma \lesssim r_N^*$,

complexity	constant confidence or large confidence
ERM	Gaussian $((r_N^*)^2)$ (sharp ?)
minimax	Gelfand (c_N^2) (achievable ?)

A word about the convexity assumption; example in aggregation

$F = \{f_1, \ldots, f_M\}$; $\hat{f} \in \operatorname{argmin}_{f \in F} R_n(f)$.

1. with probability larger than $1/2$,

$$R(\hat{f}) \geq \min_{f \in F} R(f) + c_0 \sqrt{\frac{\log |F|}{n}}.$$

2. There exists a procedure $\tilde{f}$ such that w.h.p.

$$R(\hat{f}) \leq \min_{f \in F} R(f) + c_0 \frac{\log |F|}{n}.$$

Convexity of the model is very important for ERM (more than the complexity structure).

Examples of L-sub-Gaussian classes

$\mathcal{F}$ is a set of linear functional and X is distributed like:

1. $X = (X^1, \cdots, X^d)$ where $X^1, \cdots, X^d$ are independent L-sub-gaussian variables (i.e. $\left\|X^i\right\|_{\psi_2} \leq L \left\|X^i\right\|_2$).
2. the uniform measure on $d^{1/p} B_p^d$.
3. $X = (X^1, \cdots, X^d)$ unconditional, supported in RB_∞^d and $\mathbb{E}(X^i)^2 \geq c > 0$.
4. $X \in \mathcal{M}_{p,q}$ uniformly distributed over $\{\pm E_{ij} : 1 \leq i \leq p, 1 \leq j \leq q\}$ (where (E_{ij}) is the canonical basis of $\mathcal{M}_{p,q}$) is a sub-gaussian design.
5. $X \in \mathcal{M}_{p,q}$ uniformly distributed over $\{E_{ij} : 1 \leq i \leq p, 1 \leq j \leq q\}$ (matrix completion design) and $\mathcal{B} \subset \mathcal{M}_{p,q}$ such that $|A_{ij}| \leq R, \forall i, j, A \in \mathcal{B}$. Then $\{\langle \cdot, A \rangle : A \in \mathcal{B}\}$ is L-sub-gaussian for X.

Other examples of Gaussian mean widths

1. If $p \geq 2$ then $\ell_*(B_p^d \cap sB_2^d) = \ell_*(sB_2^d) = s\sqrt{d}$.
2. If $p < 2$ then set $1 = 1/p + 1/q$ and put $1/r = 1/2 - 1/q$. For any $d^{-1/r} < s \leq 1$,

$$\ell_*(B_p^d \cap sB_2^d) \sim \begin{cases} \sqrt{q} d^{1/q} & \text{if } 2 < q < \log(2d) \text{ and } s^{-1} \leq c_1^{q/r} d^{1/r} \\ \sqrt{\log\left(2ds^2\right)} & \text{if } q \geq \log(2d) \end{cases}$$

and if $2 < q < \log(2d)$ and $s^{-1} > c_1^{q/r} d^{1/r}$ then

$$s\sqrt{d} \lesssim \ell_*(B_p^d \cap sB_2^d) \lesssim c_1^{-q/r} s\sqrt{d}.$$

Furthermore, if $s \leq d^{-1/r}$, then $\ell_*(B_p^d \cap sB_2^d) = \ell_*(sB_2^d) = s\sqrt{d}$.

Examples of Gelfand width [Foucart, Pajor, Rauhut, Ullrich]

$0 < p \leq 1$, $p < q \leq 2$, $N < d$,

$$c_N(B_p^d, \ell_q^d) \sim \min\left(1, \frac{\log(ed/N)}{N}\right)^{\frac{1}{p}-\frac{1}{q}}$$

and for $B_{p,\infty} = \{x \in \mathbb{R}^d : x_j^* \leq j^{-1/p}\}$, when $p < 1$

$$c_N(B_{p,\infty}^d, \ell_q^d) \sim \min\left(1, \frac{\log(ed/N)}{N}\right)^{\frac{1}{p}-\frac{1}{q}}$$

2 regimes for the noise - 2 statistical complexities - 2 empirical processes

The proof of the upper bound follows from the strategy developped in [Bartlett-Mendelson, PTRF 06]: the isomoprhic method.
If on an event, one has for every $f \in \mathcal{F}$, s.t. $P\mathcal{L}_f \geq \lambda^*$,

$$\frac{1}{2} P\mathcal{L}_f \leq P_N \mathcal{L}_f \leq \frac{3}{2} P\mathcal{L}_f \tag{1}$$

then, on the same event, $R(\hat{f}_{ERM}) \leq \inf_{f \in \mathcal{F}} R(f) + \lambda^*$ (because $P_N \mathcal{L}_{\hat{f}} \leq 0$).
To prove (1), it is enough to prove that

$$\sup_{\{f : P\mathcal{L}_f \geq \lambda^*\}} \left| \frac{1}{N} \sum_{i=1}^{N} \frac{\mathcal{L}_f(X_i, Y_i)}{P\mathcal{L}_f} - 1 \right| \leq \frac{1}{2}.$$

Why do we work at this confidence bound?

Classical results in the bounded case are written like (cf. Koltchinksii or Massart): $\forall t \geq c_0$, with probability greater than $1 - 4\exp(-c_1 t)$,

$$R(\hat{f}_{ERM}) \leq \inf_{f \in \mathcal{F}} R(f) + \|\mathcal{F}\|_\infty \max\left((s_N^*)^2, \frac{t}{N}\right).$$

The trade-off is obtained for $t = N(s_N^*)^2$.

1. below $t \leq N(s_N^*)^2$ the probability estimate is damaged (the residue is still $(s_N^*)^2$).
2. above $t \geq N(s_N^*)^2$, the residue is damaged.

Sharp estimates on the risk of ERM

For the problem of estimation of the mean of a gaussian vector:

$$Y \sim \mathcal{N}(\mu, I_d)$$

where $\mu \in T$ and T is a convex subset of $\mathbb{R}^d$. The ERM is

$$\hat{t} \in \operatorname*{argmin}_{t \in T} \|Y - t\|_2 , \hat{t} = P_T Y$$

fixed point equation:

$$t_\mu \in \operatorname*{argmax}_{r>0} \left(\mathbb{E} \|G\|_{T \cap (\mu + rB_2^d)} - r^2/2 \right)$$

Theorem (Chatterjee)

For every $x > 0$,

$$\mathbb{P}\Big(\big| \|\hat{\mu} - \mu\| - t_\mu \big| \geq x t_\mu \Big) \leq 3 \exp\left(\frac{-x^4 t_\mu^2}{32(1+x)^2} \right).$$

Beyond convex class

The upper bound results are true under the following assumptions:

1. $\mathcal{F}$ is B-Bernstein: $\forall f \in \mathcal{F}$,

$$\|f - f_{\mathcal{F}}^*\|_{L_2(\mu)}^2 \leq BP\mathcal{L}_f = B\big(R(f) - R(f_{\mathcal{F}}^*)\big).$$

2. $\mathcal{F} - \mathcal{F}$ is star-shaped around 0 ($[f - g, 0] \subset \mathcal{F} - \mathcal{F}, \forall f, g \in \mathcal{F}$).

But for locally compact classes, the Bernstein condition holds for all $f_{\mathcal{F}}^*$ iff the class $\mathcal{F}$ is convex.

Complexity is important but geometry is even more important

Theorem

Let $X \sim \mu$. Let $\mathcal{F} \subset L_2(\mu)$ be locally compact. The following are equivalent:

i) *for any real valued random variable $Y \in L_2$,*
$\exists f^*_{\mathcal{F}} \in \operatorname{argmin}_{f \in \mathcal{F}} \mathbb{E}(Y - f(X))^2$ *and for every $f \in \mathcal{F}$,*

$$\mathbb{E}\big(f(X) - f^*_{\mathcal{F}}(X)\big)^2 \leq \mathbb{E}\big((Y - f(X))^2 - (Y - f^*_{\mathcal{F}}(X))^2\big). \quad (2)$$

ii) *$\mathcal{F}$ is non-empty and convex.*

For non-convex model, ERM cannot do better than $1/\sqrt{N}$.

$\Longrightarrow$ the shape of the model really matters in Learning theory.