Learning sub-Gaussian classes : Upper and minimax bounds

Guillaume Lecué

CNRS, centre de mathématiques appliquées, Ecole Polytechnique.

Postdam - Conference on Structural Inference in Statistics - 2013

joint work with Shahar Mendelson

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

$$\text{data}: \ (X_i, Y_i)_{i=1}^N \text{ i.i.d.} \sim (X, Y) \in \mathcal{X} \times \mathbb{R},$$
$$\text{model}: \ \mathcal{F} \subset L_2(\mathcal{X}, \mu) \text{ (where } X \sim \mu),$$

$$\begin{aligned} \text{data} &: (X_i, Y_i)_{i=1}^N \text{ i.i.d.} \sim (X, Y) \in \mathcal{X} \times \mathbb{R}, \\ \text{model} &: \mathcal{F} \subset L_2(\mathcal{X}, \mu) \text{ (where } X \sim \mu), \\ \text{estimator} &: \text{Emprical risk minimization (ERM) :} \end{aligned}$$

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

$$\begin{aligned} \text{data}: &\ (X_i, Y_i)_{i=1}^N \text{ i.i.d.} \sim (X, Y) \in \mathcal{X} \times \mathbb{R}, \\ \text{model}: &\ \mathcal{F} \subset L_2(\mathcal{X}, \mu) \text{ (where } X \sim \mu), \\ \text{estimator}: &\ \text{Emprical risk minimization (ERM)}: \end{aligned}$$

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

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

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) :

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

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 $r_N(\mathcal{F}, \delta)$?

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) :

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

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,...)

$$\begin{aligned} \text{data} &: (X_i, Y_i)_{i=1}^N \text{ i.i.d.} \sim (X, Y) \in \mathcal{X} \times \mathbb{R}, \\ \text{model} &: \mathcal{F} \subset L_2(\mathcal{X}, \mu) \text{ (where } X \sim \mu), \\ \text{estimator} &: \text{Emprical risk minimization (ERM) :} \end{aligned}$$

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

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 ?

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) :

$$\hat{f} \in \underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^N \left(Y_i - f(X_i)\right)^2.$$

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

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 $\mathcal{F}$,...)

How do we measure the complexity 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 $\mathcal{F} \subset L_2(\mu)$.

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 $\mathcal{F} \subset L_2(\mu)$.

2. $N(\mathcal{F}, \epsilon D)$ is the minimal number of balls ϵD (D is the unit ball of $L_2(\mu)$) needed to cover $\mathcal{F}$: covering - entropy

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 $\mathcal{F} \subset L_2(\mu)$.

2. $N(\mathcal{F}, \epsilon D)$ is the minimal number of balls ϵD (D is the unit ball of $L_2(\mu)$) needed to cover $\mathcal{F}$: covering - entropy

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

$$\sup_{\epsilon>0} \epsilon \log^{1/2} N(\mathcal{F}, \epsilon D) \quad \underset{\substack{\uparrow \\ \textit{Sudakov}}}{\lesssim} \quad \mathbb{E}\|G\|_{\mathcal{F}} \quad \underset{\substack{\uparrow \\ \textit{Dudley}}}{\lesssim} \quad \int \log^{1/2} N(\mathcal{F}, \epsilon D) d\epsilon$$

$$\sup_{\epsilon>0} \epsilon \log^{1/2} N(\mathcal{F}, \epsilon D) \quad \underset{\substack{\uparrow \\ \textit{Sudakov}}}{\lesssim} \quad \mathbb{E}\|G\|_{\mathcal{F}} \quad \underset{\substack{\uparrow \\ \textit{Dudley}}}{\lesssim} \quad \int \log^{1/2} N(\mathcal{F}, \epsilon D) d\epsilon$$

$$\sup_{\epsilon>0} \epsilon \log^{1/2} N(\mathcal{F}, \epsilon D) \quad \underset{\substack{\uparrow \\ \textit{Carl} \\ (\mathcal{F} \text{ convex body})}}{\lesssim} \quad \sup_{k\in\mathbb{N}} \sqrt{k} c_k(\mathcal{F}) \quad \underset{\substack{\uparrow \\ \textit{Pajor / Tomczak - Jaegermann} \\ (\mathcal{F} \text{ star-shapped in } 0)}}{\lesssim} \quad \mathbb{E}\|G\|_{\mathcal{F}}$$

$$\sup_{\epsilon>0} \epsilon \log^{1/2} N(\mathcal{F}, \epsilon D) \underset{\substack{\uparrow \\ \textit{Sudakov}}}{\lesssim} \mathbb{E}\|G\|_{\mathcal{F}} \underset{\substack{\uparrow \\ \textit{Dudley}}}{\lesssim} \int \log^{1/2} N(\mathcal{F}, \epsilon D) d\epsilon$$

$$\sup_{\epsilon>0} \epsilon \log^{1/2} N(\mathcal{F}, \epsilon D) \underset{\substack{\uparrow \\ \textit{Carl} \\ (\mathcal{F} \text{ convex body})}}{\lesssim} \sup_{k\in\mathbb{N}} \sqrt{k} c_k(\mathcal{F}) \underset{\substack{\uparrow \\ \textit{Pajor/Tomczak} - \textit{Jaegermann} \\ (\mathcal{F} \text{ star-shapped in } 0)}}{\lesssim} \mathbb{E}\|G\|_{\mathcal{F}}$$

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 $= (\log d)^{3/2}$.

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

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

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

Assumptions : sub-gaussian framework

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

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

   $(\|f\|_{\psi_2(\mu)} = \inf\left(c > 0 : \mathbb{E}\exp(f^2(X)/c^2) \leq 2\right))$.
2. $\|Y - f^*_{\mathcal{F}}(X)\|_{\psi_2} \leq \sigma$ (noise level) where $f^*_{\mathcal{F}} \in \text{argmin}_{f\in\mathcal{F}} R(f)$

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

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

   $(\|f\|_{\psi_2(\mu)} = \inf\big(c > 0 : \mathbb{E}\exp(f^2(X)/c^2) \leq 2\big))$.
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 B-Bernstein : $\forall f \in \mathcal{F}$,

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

1. $\mathcal{F}$ is L-sub-Gaussian : $\forall f, g \in \mathcal{F} \cup \{0\}$,
$$\|f-g\|_{\psi_2(\mu)} \leq L\|f-g\|_{L_2(\mu)}$$
$(\|f\|_{\psi_2(\mu)} = \inf\big(c>0 : \mathbb{E}\exp(f^2(X)/c^2) \leq 2\big))$.
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 B-Bernstein : $\forall f \in \mathcal{F}$,
$$\|f - f^*_{\mathcal{F}}\|^2_{L_2(\mu)} \leq BP\mathcal{L}_f = B\big(R(f) - R(f^*_{\mathcal{F}})\big).$$
4. $\mathcal{F} - \mathcal{F}$ is star-shaped around 0 ($[f-g, 0] \subset \mathcal{F} - \mathcal{F}, \forall f, g \in \mathcal{F}$).

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

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

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

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

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

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

where, for D the unit ball in $L_2(\mu)$,

$$s_N^* = \inf\left\{0 < s \leq d_{\mathcal{F}}(L_2) : \mathbb{E}\|G\|_{sD\cap(\mathcal{F}-\mathcal{F})} \leq (c_0/\sigma)s^2\sqrt{N}\right\},$$
$$r_N^* = \inf\left\{0 < r \leq d_{\mathcal{F}}(L_2) : \mathbb{E}\|G\|_{rD\cap(\mathcal{F}-\mathcal{F})} \leq c_1 r\sqrt{N}\right\}.$$

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

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

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

where, for D the unit ball in $L_2(\mu)$,

$$s_N^* = \inf\left\{0 < s \leq d_{\mathcal{F}}(L_2) : \mathbb{E}\|G\|_{sD\cap(\mathcal{F}-\mathcal{F})} \leq (c_0/\sigma)s^2\sqrt{N}\right\},$$
$$r_N^* = \inf\left\{0 < r \leq d_{\mathcal{F}}(L_2) : \mathbb{E}\|G\|_{rD\cap(\mathcal{F}-\mathcal{F})} \leq c_1 r\sqrt{N}\right\}.$$

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

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

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

$$\sigma^* = \inf\left(s > 0 : \int_{c_0 s^2}^{c_1 s} \log^{1/2} N((\mathcal{F}-\mathcal{F})\cap(2sD), \epsilon D) d\epsilon \leq c_2 s^2 \sqrt{N}\right).$$

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

$$\sigma^* = \inf\left(s > 0 : \int_{c_0 s^2}^{c_1 s} \log^{1/2} N((\mathcal{F}-\mathcal{F})\cap(2sD), \epsilon D)d\epsilon \le c_2 s^2\sqrt{N}\right).$$

$>$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] :

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

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

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

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

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

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

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

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

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

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 $P_\tau : f \in L_2(\mu) \mapsto \big(f(X_i)\big)_1^N$ is a source of complexity because we want procedures having good "generalization" capabilities (being good even outside of the data sample).

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

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

where $r_N^* = \inf \left\{ 0 < r \leq d_{\mathcal{F}}(L_2) : \mathbb{E}\|G\|_{rD\cap(\mathcal{F}-\mathcal{F})} \leq c_1 r\sqrt{N} \right\}$.
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 $P_\tau : f \in L_2(\mu) \mapsto \big(f(X_i)\big)_1^N$ is a source of complexity because we want procedures having good "generalization" capabilities (being good even outside of the data sample).
- 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" $Y_i - f_{\mathcal{F}}^*(X_i)$.

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

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

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 $P_\tau : f \in L_2(\mu) \mapsto \big(f(X_i)\big)_1^N$ 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" $Y_i - f_{\mathcal{F}}^*(X_i)$.

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

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

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 $P_\tau : f \in L_2(\mu) \mapsto \big(f(X_i)\big)_1^N$ 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 $\sigma \gtrsim r_N^*$.

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

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

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 $P_\tau : f \in L_2(\mu) \mapsto \big(f(X_i)\big)_1^N$ 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 :

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2. The linear process $((P - P_N)(y - f^*_{\mathcal{F}})(f^*_{\mathcal{F}} - f))_{f \in \mathcal{F} \cap rD}$. [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 rate for ERM ? - better procedure than ERM ?)

Can we do better ? (better rate for ERM ? - better procedure than ERM ?)

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

Can we do better ? (better rate for ERM ? - better procedure than ERM ?)

The gaussian regression model : $Y = f^*(X) + W$ where $W \sim \mathcal{N}(0, \sigma^2)$ indep. of X and $f^* = f^*_{\mathcal{F}} \in \mathcal{F}$. If $\sigma \gtrsim r_N^*$ then with probability at least

$1 - 4\exp(-c_4 N \sigma^{-2} (s_N^*)^2)$, $\qquad R(\hat{f}_{ERM}) \leq \inf_{f \in \mathcal{F}} R(f) + (s_N^*)^2$.

Can we do better? (better rate for ERM? - better procedure than ERM?)

The gaussian regression model : $Y = f^*(X) + W$ where $W \sim \mathcal{N}(0, \sigma^2)$ indep. of X and $f^* = f^*_{\mathcal{F}} \in \mathcal{F}$. If $\sigma \gtrsim r^*_N$ then with probability at least

$$1 - 4\exp(-c_4 N \sigma^{-2}(s^*_N)^2), \qquad R(\hat{f}_{ERM}) \leq \inf_{f \in \mathcal{F}} R(f) + (s^*_N)^2.$$

Theorem (L.& Mendelson)

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

Can we do better? (better rate for ERM? - better procedure than ERM?)

The gaussian regression model : $Y = f^*(X) + W$ where $W \sim \mathcal{N}(0, \sigma^2)$ indep. of X and $f^* = f^*_{\mathcal{F}} \in \mathcal{F}$. If $\sigma \gtrsim r^*_N$ then with probability at least

$$1 - 4\exp(-c_4 N \sigma^{-2}(s^*_N)^2), \qquad R(\hat{f}_{ERM}) \leq \inf_{f \in \mathcal{F}} R(f) + (s^*_N)^2.$$

Theorem (L.& Mendelson)

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

ERM is minimax in the Gaussian regression model over sub-Gaussian 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 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.*

ERM is minimax for high confidence but not for constant confidence

Corollary

In the Gaussian regression model with respect to a sub-Gaussian 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* $\alpha > 1/2$*,*

ERM is minimax for high confidence but not for constant confidence

Corollary

In the Gaussian regression model with respect to a sub-Gaussian 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* $n^{-\alpha/(2\alpha+1)}$*).*

ERM is minimax for high confidence but not for constant confidence

Corollary

In the Gaussian regression model with respect to a sub-Gaussian 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 $n^{-\alpha/(2\alpha+1)}$).*

Comparison : The two results cannot be compared because the Gaussian mean widths of localized sets of the $\alpha-$Hölderian classe for $\alpha < 1/2$ are infinite.

ERM is minimax for high confidence but not for constant confidence

Corollary

In the Gaussian regression model with respect to a sub-Gaussian 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 $n^{-\alpha/(2\alpha+1)}$).*

Comparison : The two results cannot be compared because the Gaussian mean widths of localized sets of the $\alpha-$Hölderian classe for $\alpha < 1/2$ are infinite. **Nevertheless**, we believe that there should be two regimes for the minimaximality of ERM.

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*.

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 $\mathcal{F}$.

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 $\mathcal{F}$).

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 ?

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 $3/4$,

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

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 $3/4$,

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

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

1. 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 $3/4$,

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

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 ?)

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 $3/4$,

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

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 ?)
2. can we use the Gaussian shift theorem in this case ?

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 $3/4$,

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

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 ?)
2. can we use the Gaussian shift theorem in this case ?
3. are we going to recover the classical minimax results in this regime ?

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,

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

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 ?
3. are we going to recover the classical minimax results in this regime ?

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 $3/4$,

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

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 ?

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,

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

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 [Birgé - Tsybakov - Yang/Barron]

Minimax result for constant confidence - large noise

Theorem (L. & Mendelson)

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

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

where

$$q_N^* = \inf \left\{ s > 0 : s \log^{1/2} N((\mathcal{F} - \mathcal{F}) \cap 2sD, sD) \leq (c_0/\sigma) s^2 \sqrt{N} \right\}.$$

Theorem (L. & Mendelson)

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

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

where

$$q_N^* = \inf\left\{s > 0 : s \log^{1/2} N((\mathcal{F} - \mathcal{F}) \cap 2sD, sD) \leq (c_0/\sigma) s^2 \sqrt{N}\right\}.$$

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

$$\sup_{0 < \epsilon < 2s} \epsilon \log^{1/2} N((\mathcal{F} - \mathcal{F}) \cap 2sD, \epsilon D) \lesssim \mathbb{E} \| G \|_{(\mathcal{F} - \mathcal{F}) \cap 2sD}.$$

Theorem (L. & Mendelson)

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

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

where

$$q_N^* = \inf\left\{s > 0 : s \log^{1/2} N((\mathcal{F} - \mathcal{F}) \cap 2sD, sD) \leq (c_0/\sigma) s^2 \sqrt{N}\right\}.$$

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

$$\sup_{0<\epsilon<2s} \epsilon \log^{1/2} N((\mathcal{F} - \mathcal{F}) \cap 2sD, \epsilon D) \lesssim \mathbb{E}\|G\|_{(\mathcal{F}-\mathcal{F})\cap 2sD}.$$

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

$$s \log^{1/2} N((\mathcal{F} - \mathcal{F}) \cap 2sD, sD)$$

The same result follows from

1. Yang and Barron,"Information-theoretic determination of minimax rates of convergence". AOS 1999 ; via Fano inequality.

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.

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]).

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 :

$$q_N^* = \inf_{s>0} \left\{ s \log^{1/2} N((\mathcal{F} - \mathcal{F}) \cap 2sD, sD) \leq (c_0/\sigma) s^2 \sqrt{N} \right\}$$

2. upper bound for ERM :

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

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 :

$$q_N^* = \inf_{s>0} \left\{ s \log^{1/2} N((\mathcal{F} - \mathcal{F}) \cap 2sD, sD) \leq (c_0/\sigma) s^2 \sqrt{N} \right\}$$

2. upper bound for ERM :

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

For the small noise regime, we obtain the following minimax bound.

Minimax result for constant confidence - small noise

Theorem (L. & Mendelson)

Denote $f^*(X) = \mathbb{E}[Y|X]$. *For every procedure* $\tilde{f}_N$,

$$\sup_{f^* \in \mathcal{F}} \mathbb{P}_{f^*}\left[R(\tilde{f}_N) - \inf_{f \in \mathcal{F}} R(f) \geq \frac{\mathcal{D}(f^*, \tau)^2}{4}\right] \geq \frac{1}{2}$$

where $\mathcal{D}(f^*, \tau) = \mathrm{diam}\big(\{h \in \mathcal{F} : (h(X_i))_1^N = (f^*(X_i))_1^N\}, L_2(\mu)\big)$.

Minimax result for constant confidence - small noise

Theorem (L. & Mendelson)

Denote $f^*(X) = \mathbb{E}[Y|X]$. *For every procedure* $\tilde{f}_N$,

$$\sup_{f^* \in \mathcal{F}} \mathbb{P}_{f^*}\left[R(\tilde{f}_N) - \inf_{f \in \mathcal{F}} R(f) \geq \frac{\mathcal{D}(f^*, \tau)^2}{4}\right] \geq \frac{1}{2}$$

where $\mathcal{D}(f^*, \tau) = \mathrm{diam}\big(\{h \in \mathcal{F} : (h(X_i))_1^N = (f^*(X_i))_1^N\}, L_2(\mu)\big)$.

$$\begin{aligned}\mathcal{D}(f^*, \tau) &= \mathrm{diam}((\mathcal{F} - f^*) \cap \ker P_\tau, L_2) \text{ where } P_\tau : f \in L_2(\mu) \mapsto (f(X_i))_1^N \in \mathbb{R}^N \\ &\geq c_N(\mathcal{F} - f^*) \text{ Gelfand } N - \text{width.}\end{aligned}$$

Theorem (L. & Mendelson)

Denote $f^*(X) = \mathbb{E}[Y|X]$. For every procedure $\tilde{f}_N$,

$$\sup_{f^* \in \mathcal{F}} \mathbb{P}_{f^*}\left[R(\tilde{f}_N) - \inf_{f \in \mathcal{F}} R(f) \geq \frac{\mathcal{D}(f^*, \tau)^2}{4}\right] \geq \frac{1}{2}$$

where $\mathcal{D}(f^*, \tau) = \mathrm{diam}\big(\{h \in \mathcal{F} : (h(X_i))_1^N = (f^*(X_i))_1^N\}, L_2(\mu)\big)$.

$$\begin{aligned} \mathcal{D}(f^*, \tau) &= \mathrm{diam}((\mathcal{F} - f^*) \cap \ker P_\tau, L_2) \text{ where } P_\tau : f \in L_2(\mu) \mapsto (f(X_i))_1^N \in \mathbb{R}^N \\ &\geq c_N(\mathcal{F} - f^*) \text{ Gelfand } N-\text{width.} \end{aligned}$$

Therefore, the square of the minimal Gelfand N-width

$\sup_{f^* \in \mathcal{F}} \big(c_N(\mathcal{F} - f^*)\big)^2$ is a minimax lower bound in the regression model.

Minimax result for constant confidence - small noise

Theorem (L. & Mendelson)

Denote $f^*(X) = \mathbb{E}[Y|X]$. For every procedure $\tilde{f}_N$,

$$\sup_{f^* \in \mathcal{F}} \mathbb{P}_{f^*}\left[R(\tilde{f}_N) - \inf_{f \in \mathcal{F}} R(f) \geq \frac{\mathcal{D}(f^*, \tau)^2}{4}\right] \geq \frac{1}{2}$$

where $\mathcal{D}(f^*, \tau) = \mathrm{diam}\big(\{h \in \mathcal{F} : (h(X_i))_1^N = (f^*(X_i))_1^N\}, L_2(\mu)\big)$.

$$\begin{aligned}\mathcal{D}(f^*, \tau) &= \mathrm{diam}((\mathcal{F} - f^*) \cap \ker P_\tau, L_2) \text{ where } P_\tau : f \in L_2(\mu) \mapsto (f(X_i))_1^N \in \mathbb{R}^N \\ &\geq c_N(\mathcal{F} - f^*) \text{ Gelfand } N-\text{width.}\end{aligned}$$

Therefore, the square of the minimal Gelfand N-width

$\sup_{f^* \in \mathcal{F}} \big(c_N(\mathcal{F} - f^*)\big)^2$ is a minimax lower bound in the regression model.

Similar lower bounds have been obtained in Compressed Sensing by [Donoho, IEEE06] or [Cohen, Dahmen, DeVore, JAMS09].

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

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

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

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

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

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

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

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

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

- with the same confidence then $\mathrm{residue} \ge (s_N^*)^2$;

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

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

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

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

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

$$q_N^* = \inf \left\{ s > 0 : s \log^{1/2} N((\mathcal{F} - \mathcal{F}) \cap 2sD, sD) \leq (c_0/\sigma)s^2\sqrt{N} \right\}.$$

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

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

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

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

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

$$q_N^* = \inf\left\{s > 0 : s\log^{1/2} N((\mathcal{F}-\mathcal{F})\cap 2sD, sD) \leq (c_0/\sigma)s^2\sqrt{N}\right\}.$$

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

$$q_N^* \log^{1/2} N((\mathcal{F}-\mathcal{F})\cap 2q_N^* D, q_N^* D) \sim \mathbb{E}\|G\|_{(\mathcal{F}-\mathcal{F})\cap 2q_N^* D}$$

then upper and lower bounds match and therefore ERM is minimax in the Gaussian model for any subgaussian 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

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

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

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

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 $\mathrm{residue} \gtrsim \sup_{f^*\in\mathcal{F}} \left(c_N(\mathcal{F}-f^*)\right)^2$.

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

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

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

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
$\mathrm{residue} \gtrsim \sup_{f^* \in \mathcal{F}} \left(c_N(\mathcal{F} - f^*)\right)^2$.
If "Pajor/Tomczak-Jaegermann is sharp at level N" (for some $f_0^* \in \mathcal{F}$) :

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

then upper and bounds match and therefore ERM is minimax in the Gaussian model for any subgaussian 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. $\in \mathbb{R}^{p\times q} \times \mathbb{R}$,

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

$$\begin{aligned}
\text{data} &: (X_i, Y_i)_{i=1}^N \text{ i.i.d.} \in \mathbb{R}^{p\times q} \times \mathbb{R},\\
\text{model} &: \mathcal{F} = \{\langle \cdot, A\rangle : \|A\|_{max} \leq R\},\\
&\quad \|A\|_{max} = \min_{A=UV^\top} \|U\|_{2\mapsto\infty}\|V\|_{2\mapsto\infty}.
\end{aligned}$$

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 : $\mathcal{F} = \{\langle \cdot, A\rangle : \|A\|_{max} \leq R\}$,

$\|A\|_{max} = \min_{A=UV^\top} \|U\|_{2\mapsto\infty}\|V\|_{2\mapsto\infty}$.

estimator : Emprical risk minimization (ERM) :

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

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) :

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

Assumptions :

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

data : $(X_i, Y_i)_{i=1}^N$ i.i.d. $\in \mathbb{R}^{p\times q} \times \mathbb{R}$,
model : $\mathcal{F} = \{\langle \cdot, A\rangle : \|A\|_{max} \leq R\}$,
$\|A\|_{max} = \min_{A=UV^\top} \|U\|_{2\mapsto\infty}\|V\|_{2\mapsto\infty}$.
estimator : Emprical risk minimization (ERM) :

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

Assumptions :

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

Gaussian mean width : $\mathrm{conv}\mathcal{X}_\pm \subset \mathcal{B}_{max} \subset K_G \mathrm{conv}\mathcal{X}_\pm$ where $\mathcal{X}_\pm = \{uv^\top : u \in \{\pm 1\}^p, v \in \{\pm 1\}^q\}$.

data : $(X_i, Y_i)_{i=1}^N$ i.i.d. $\in \mathbb{R}^{p\times q} \times \mathbb{R}$,
model : $\mathcal{F} = \{\langle \cdot, A\rangle : \|A\|_{max} \leq R\}$,
$\|A\|_{max} = \min_{A=UV^\top} \|U\|_{2\mapsto\infty}\|V\|_{2\mapsto\infty}$.
estimator : Emprical risk minimization (ERM) :

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

Assumptions :

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

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

data : $(X_i, Y_i)_{i=1}^N$ i.i.d. $\in \mathbb{R}^{p\times q} \times \mathbb{R}$,
model : $\mathcal{F} = \{\langle \cdot, A\rangle : \|A\|_{max} \leq R\}$,
$\|A\|_{max} = \min_{A=UV^\top} \|U\|_{2\mapsto\infty}\|V\|_{2\mapsto\infty}$.
estimator : Emprical risk minimization (ERM) :

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

Assumptions :

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

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

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

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

Therefore, $(s_N^*)^2 \sim \sigma R \sqrt{(p+q)/N}$ and $(r_N^*)^2 \sim R(p+q)/N$.

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

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

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

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

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

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

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, $(s_N^*)^2 \sim \sigma R\sqrt{(p+q)/N}$ 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.

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

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

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

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

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 $R\mathcal{B}_{max} \cap (\alpha B_\infty^{pq})$.

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 $f \in \mathcal{F}$,*

$$\mathbb{E}\big(f(X) - f^*_{\mathcal{F}}(X)\big)^2 \leq \mathbb{E}\big((Y - f(X))^2 - (Y - f^*_{\mathcal{F}}(X))^2\big). \quad (1)$$

ii) *$\mathcal{F}$ is non-empty and 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 $f \in \mathcal{F}$,*

$$\mathbb{E}\big(f(X) - f^*_{\mathcal{F}}(X)\big)^2 \leq \mathbb{E}\big((Y - f(X))^2 - (Y - f^*_{\mathcal{F}}(X))^2\big). \quad (1)$$

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

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

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 $Y \in L_2$,* $\exists f^*_{\mathcal{F}} \in \mathrm{argmin}_{f\in\mathcal{F}} \mathbb{E}(Y - f(X))^2$ *and for every $f \in \mathcal{F}$,*

$$\mathbb{E}\big(f(X) - f^*_{\mathcal{F}}(X)\big)^2 \leq \mathbb{E}\big((Y - f(X))^2 - (Y - f^*_{\mathcal{F}}(X))^2\big). \quad (1)$$

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

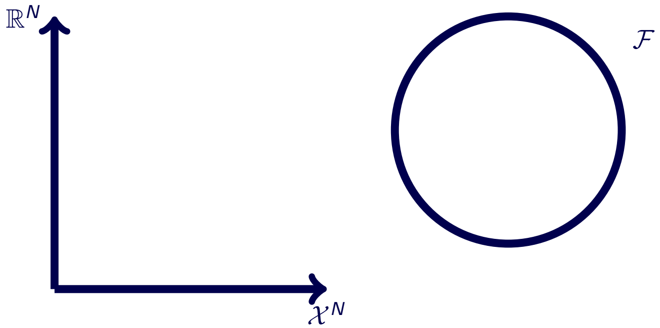
For non-convex model, ERM cannot do better than $1/\sqrt{N}$.

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

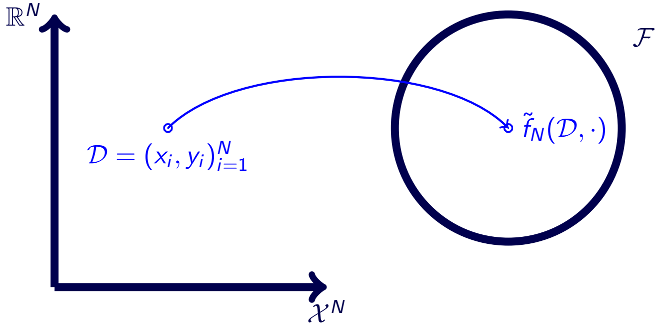
$\tilde{f}_N$ a procedure such that, for every $f^* \in \mathcal{F}$, with probability greater than $1-\delta$, $R(\tilde{f}_N) \leq \inf_{f \in \mathcal{F}} R(f) + \mathrm{residue}$ i.e. $\|\tilde{f}_N - f^*\|_{L_2(\mu)}^2 \leq \mathrm{res.}$

minimax results for high confidence bounds in $Y = f^*(X) + W$ - sketch of proof

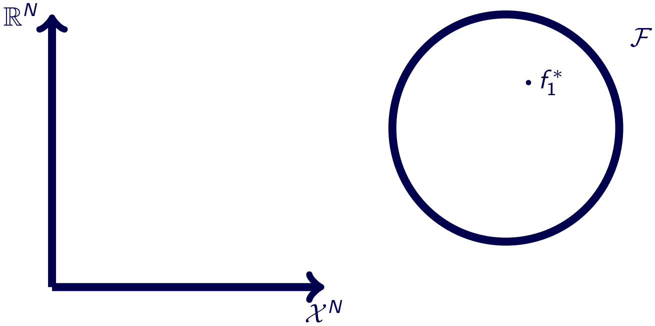
$\tilde{f}_N$ a procedure such that, for every $f^* \in \mathcal{F}$, with probability greater than $1-\delta$, $R(\tilde{f}_N) \le \inf_{f\in\mathcal{F}} R(f) + \mathrm{residue}$ i.e. $\|\tilde{f}_N - f^*\|_{L_2(\mu)}^2 \le \mathrm{res.}$



$\tilde{f}_N$ a procedure such that, for every $f^* \in \mathcal{F}$, with probability greater than $1-\delta$, $R(\tilde{f}_N) \le \inf_{f\in\mathcal{F}} R(f) + \mathrm{residue}$ i.e. $\|\tilde{f}_N - f^*\|_{L_2(\mu)}^2 \le \mathrm{res.}$

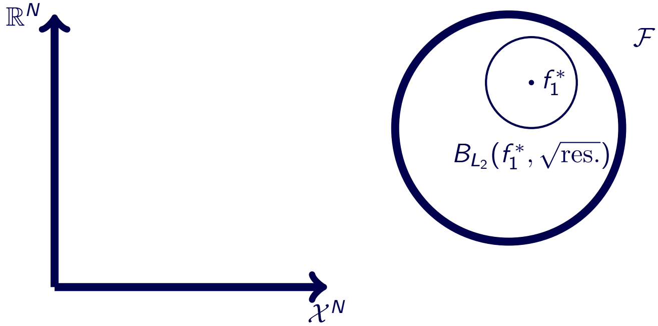


$\tilde{f}_N$ a procedure such that, for every $f^* \in \mathcal{F}$, with probability greater than $1-\delta$, $R(\tilde{f}_N) \leq \inf_{f\in\mathcal{F}} R(f) + \mathrm{residue}$ i.e. $\|\tilde{f}_N - f^*\|_{L_2(\mu)}^2 \leq \mathrm{res.}$

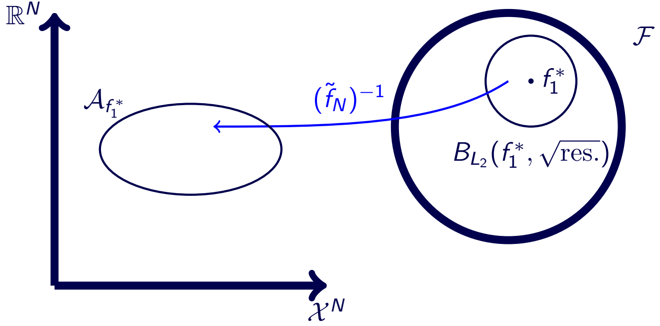


minimax results for high confidence bounds in $Y = f^*(X) + W$ - sketch of proof

$\tilde{f}_N$ a procedure such that, for every $f^* \in \mathcal{F}$, with probability greater than $1-\delta$, $R(\tilde{f}_N) \leq \inf_{f\in\mathcal{F}} R(f) + \mathrm{residue}$ i.e. $\|\tilde{f}_N - f^*\|^2_{L_2(\mu)} \leq \mathrm{res.}$

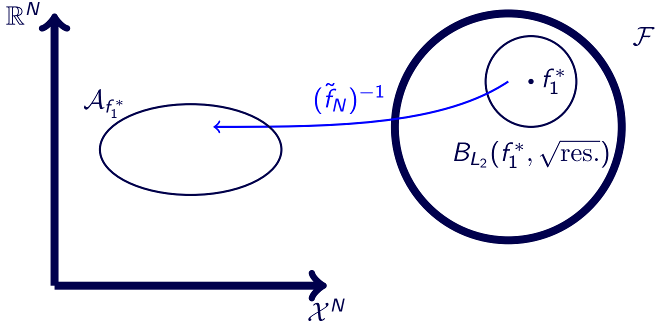


$\tilde{f}_N$ a procedure such that, for every $f^* \in \mathcal{F}$, with probability greater than $1-\delta$, $R(\tilde{f}_N) \leq \inf_{f\in\mathcal{F}} R(f) + \mathrm{residue}$ i.e. $\|\tilde{f}_N - f^*\|^2_{L_2(\mu)} \leq \mathrm{res.}$



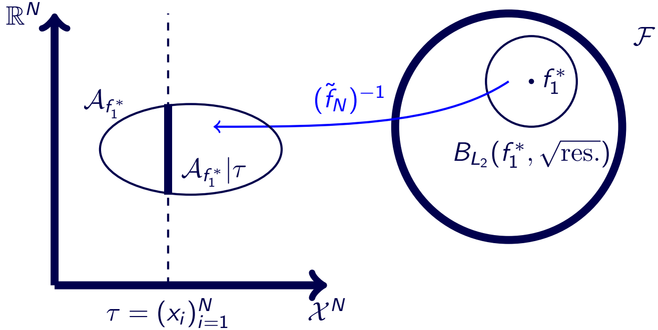
$\mathcal{A}_{f_1^*} = (\tilde{f}_N)^{-1}(B_{L_2}(f_1^*, \sqrt{\mathrm{res.}}))$

$\tilde{f}_N$ a procedure such that, for every $f^* \in \mathcal{F}$, with probability greater than $1-\delta$, $R(\tilde{f}_N) \leq \inf_{f \in \mathcal{F}} R(f) + \mathrm{residue}$ i.e. $\|\tilde{f}_N - f^*\|_{L_2(\mu)}^2 \leq \mathrm{res.}$



$\mathcal{A}_{f_1^*} = (\tilde{f}_N)^{-1}(B_{L_2}(f_1^*, \sqrt{\mathrm{res.}}))$
$(\nu_{f_1^*} \otimes \mu^N)(\mathcal{A}_{f_1^*}) \geq 1 - \delta \qquad \nu_{f_1^*} \sim \mathcal{N}((f_1^*(x_i))_1^N, \sigma^2 I_N)),\ X \sim \mu$

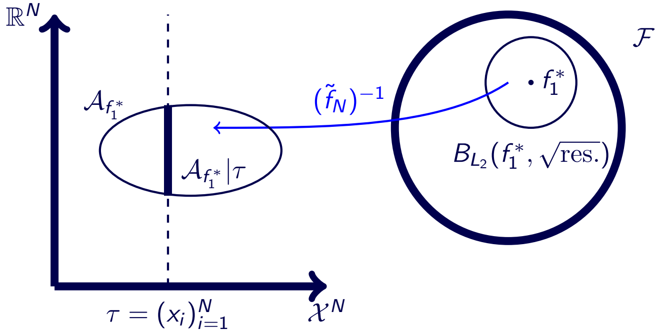
$\tilde{f}_N$ a procedure such that, for every $f^* \in \mathcal{F}$, with probability greater than $1-\delta$, $R(\tilde{f}_N) \leq \inf_{f\in\mathcal{F}} R(f) + \mathrm{residue}$ i.e. $\|\tilde{f}_N - f^*\|_{L_2(\mu)}^2 \leq \mathrm{res.}$



$\mathcal{A}_{f_1^*} = (\tilde{f}_N)^{-1}(B_{L_2}(f_1^*, \sqrt{\mathrm{res.}}))$
$(\nu_{f_1^*} \otimes \mu^N)(\mathcal{A}_{f_1^*}) \geq 1-\delta$

$\mathcal{A}_{f_1^*}|\tau \subset \mathbb{R}^N$: fiber of $\mathcal{A}_{f_1^*}$

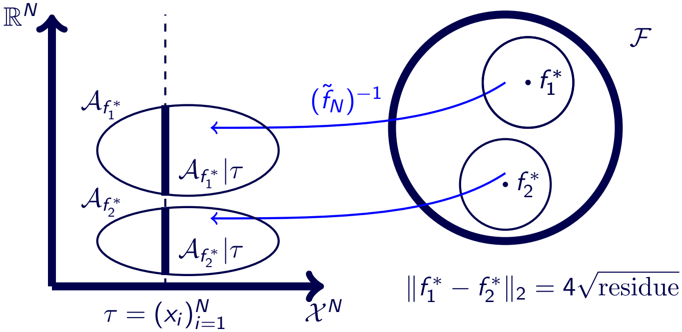
$\tilde{f}_N$ a procedure such that, for every $f^* \in \mathcal{F}$, with probability greater than $1-\delta$, $R(\tilde{f}_N) \leq \inf_{f\in\mathcal{F}} R(f) + \mathrm{residue}$ i.e. $\|\tilde{f}_N - f^*\|^2_{L_2(\mu)} \leq \mathrm{res.}$



$\mathcal{A}_{f_1^*} = (\tilde{f}_N)^{-1}(B_{L_2}(f_1^*, \sqrt{\mathrm{res.}}))$ $\qquad$ $\mathcal{A}_{f_1^*}|\tau \subset \mathbb{R}^N$: fiber of $\mathcal{A}_{f_1^*}$

$(\nu_{f_1^*} \otimes \mu^N)(\mathcal{A}_{f_1^*}) \geq 1-\delta \Longrightarrow \mu^N\big(\tau : \nu_{f_1^*,\tau}(\mathcal{A}_{f_1^*}|\tau) \geq 1-\sqrt{\delta}\big) \geq 1-\sqrt{\delta}$.

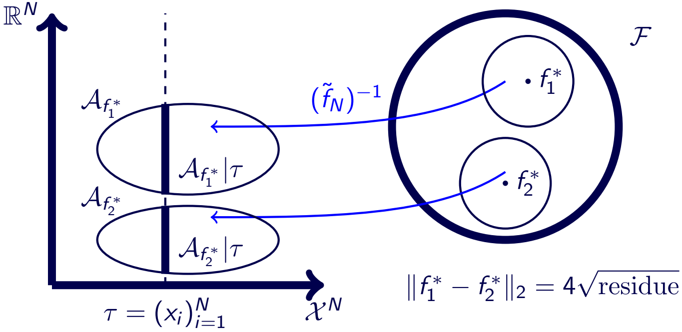
$\tilde{f}_N$ a procedure such that, for every $f^* \in \mathcal{F}$, with probability greater than $1-\delta$, $R(\tilde{f}_N) \leq \inf_{f\in\mathcal{F}} R(f) + \mathrm{residue}$ i.e. $\|\tilde{f}_N - f^*\|^2_{L_2(\mu)} \leq \mathrm{res.}$



$\mathcal{A}_{f_1^*} = (\tilde{f}_N)^{-1}(B_{L_2}(f_1^*, \sqrt{\mathrm{res.}}))$ $\qquad$ $\mathcal{A}_{f_1^*}|\tau \subset \mathbb{R}^N$: fiber of $\mathcal{A}_{f_1^*}$

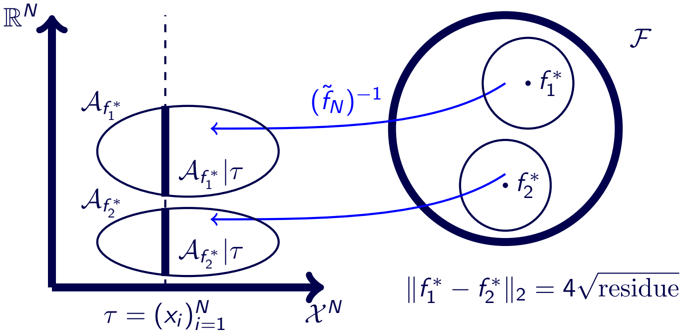
$(\nu_{f_1^*} \otimes \mu^N)(\mathcal{A}_{f_1^*}) \geq 1-\delta \Longrightarrow \mu^N\big(\tau : \nu_{f_1^*,\tau}(\mathcal{A}_{f_1^*}|\tau) \geq 1-\sqrt{\delta}\big) \geq 1-\sqrt{\delta}.$

$\tilde{f}_N$ a procedure such that, for every $f^* \in \mathcal{F}$, with probability greater than $1 - \delta$, $R(\tilde{f}_N) \le \inf_{f \in \mathcal{F}} R(f) + \mathrm{residue}$ i.e. $\|\tilde{f}_N - f^*\|^2_{L_2(\mu)} \le \mathrm{res.}$



$\mathcal{A}_{f_1^*}|\tau, \mathcal{A}_{f_2^*}|\tau \subset \mathbb{R}^N$ disjoint. $\nu_{f_1^*,\tau}(\mathcal{A}_{f_1^*}|\tau), \nu_{f_2^*,\tau}(\mathcal{A}_{f_2^*}|\tau) \ge 1 - \sqrt{\delta}$ for many τ's.

$\tilde{f}_N$ a procedure such that, for every $f^* \in \mathcal{F}$, with probability greater than $1-\delta$, $R(\tilde{f}_N) \leq \inf_{f \in \mathcal{F}} R(f) + \mathrm{residue}$ i.e. $\|\tilde{f}_N - f^*\|_{L_2(\mu)}^2 \leq \mathrm{res.}$



$\mathcal{A}_{f_1^*}|\tau, \mathcal{A}_{f_2^*}|\tau \subset \mathbb{R}^N$ disjoint. $\nu_{f_1^*,\tau}(\mathcal{A}_{f_1^*}|\tau), \nu_{f_2^*,\tau}(\mathcal{A}_{f_2^*}|\tau) \geq 1 - \sqrt{\delta}$ for many τ's.$\nu_{f_1^*,\tau} \sim \mathcal{N}\big((f_1^*(x_i))_1^N, \sigma^2 I_N\big)$ and $\nu_{f_2^*,\tau} \sim \mathcal{N}\big((f_2^*(x_i))_1^N, \sigma^2 I_N\big)$.

1. $\mathcal{A}_{f_1^*}|\tau, \mathcal{A}_{f_2^*}|\tau \subset \mathbb{R}^N$ disjoint.
2. $\nu_{f_1^*,\tau}(\mathcal{A}_{f_1^*}|\tau), \nu_{f_2^*,\tau}(\mathcal{A}_{f_2^*}|\tau) \geq 1 - \sqrt{\delta}$ for many τ's.
3. $\nu_{f_1^*,\tau} \sim \mathcal{N}\big((f_1^*(x_i))_1^N, \sigma^2 I_N\big)$ and $\nu_{f_2^*,\tau} \sim \mathcal{N}\big((f_2^*(x_i))_1^N, \sigma^2 I_N\big)$.

1. $\mathcal{A}_{f_1^*}|\tau, \mathcal{A}_{f_2^*}|\tau \subset \mathbb{R}^N$ disjoint.
2. $\nu_{f_1^*,\tau}(\mathcal{A}_{f_1^*}|\tau), \nu_{f_2^*,\tau}(\mathcal{A}_{f_2^*}|\tau) \geq 1 - \sqrt{\delta}$ for many τ's.
3. $\nu_{f_1^*,\tau} \sim \mathcal{N}\big((f_1^*(x_i))_1^N, \sigma^2 I_N\big)$ and $\nu_{f_2^*,\tau} \sim \mathcal{N}\big((f_2^*(x_i))_1^N, \sigma^2 I_N\big)$.

$\Rightarrow$ $\nu_{f_1^*,\tau}(\mathcal{A}_{f_2^*}|\tau) < \sqrt{\delta}$ and $\nu_{f_2^*,\tau}(\mathcal{A}_{f_2^*}|\tau) \geq 1 - \sqrt{\delta}$ for many τ's.

minimax results for high confidence bounds in $Y = f^*(X) + W$ - sketch of proof

1. $\mathcal{A}_{f_1^*}|\tau, \mathcal{A}_{f_2^*}|\tau \subset \mathbb{R}^N$ disjoint.
2. $\nu_{f_1^*,\tau}(\mathcal{A}_{f_1^*}|\tau), \nu_{f_2^*,\tau}(\mathcal{A}_{f_2^*}|\tau) \geq 1 - \sqrt{\delta}$ for many τ's.
3. $\nu_{f_1^*,\tau} \sim \mathcal{N}\big((f_1^*(x_i))_1^N, \sigma^2 I_N\big)$ and $\nu_{f_2^*,\tau} \sim \mathcal{N}\big((f_2^*(x_i))_1^N, \sigma^2 I_N\big)$.

$\Rightarrow$ $\nu_{f_1^*,\tau}(\mathcal{A}_{f_2^*}|\tau) < \sqrt{\delta}$ and $\nu_{f_2^*,\tau}(\mathcal{A}_{f_2^*}|\tau) \geq 1 - \sqrt{\delta}$ for many τ's.

4. $\|(f_1^*(x_i))_1^N - (f_2^*(x_i))_1^N\|_{L_2^N} \lesssim \|f_1^* - f_2^*\|_2 = 4\sqrt{\mathrm{residue}}$.

1. $\mathcal{A}_{f_1^*}|\tau, \mathcal{A}_{f_2^*}|\tau \subset \mathbb{R}^N$ disjoint.
2. $\nu_{f_1^*,\tau}(\mathcal{A}_{f_1^*}|\tau), \nu_{f_2^*,\tau}(\mathcal{A}_{f_2^*}|\tau) \geq 1 - \sqrt{\delta}$ for many τ's.
3. $\nu_{f_1^*,\tau} \sim \mathcal{N}\big((f_1^*(x_i))_1^N, \sigma^2 I_N\big)$ and $\nu_{f_2^*,\tau} \sim \mathcal{N}\big((f_2^*(x_i))_1^N, \sigma^2 I_N\big)$.

$\Rightarrow$ $\nu_{f_1^*,\tau}(\mathcal{A}_{f_2^*}|\tau) < \sqrt{\delta}$ and $\nu_{f_2^*,\tau}(\mathcal{A}_{f_2^*}|\tau) \geq 1 - \sqrt{\delta}$ for many τ's.

4. $\|(f_1^*(x_i))_1^N - (f_2^*(x_i))_1^N\|_{L_2^N} \lesssim \|f_1^* - f_2^*\|_2 = 4\sqrt{\mathrm{residue}}$.

$\Rightarrow$ This forces the residue to be large.

1. $\mathcal{A}_{f_1^*}|\tau, \mathcal{A}_{f_2^*}|\tau \subset \mathbb{R}^N$ disjoint.
2. $\nu_{f_1^*,\tau}(\mathcal{A}_{f_1^*}|\tau), \nu_{f_2^*,\tau}(\mathcal{A}_{f_2^*}|\tau) \geq 1 - \sqrt{\delta}$ for many τ's.
3. $\nu_{f_1^*,\tau} \sim \mathcal{N}((f_1^*(x_i))_1^N, \sigma^2 I_N)$ and $\nu_{f_2^*,\tau} \sim \mathcal{N}((f_2^*(x_i))_1^N, \sigma^2 I_N)$.

$\Rightarrow$ $\nu_{f_1^*,\tau}(\mathcal{A}_{f_2^*}|\tau) < \sqrt{\delta}$ and $\nu_{f_2^*,\tau}(\mathcal{A}_{f_2^*}|\tau) \geq 1 - \sqrt{\delta}$ for many τ's.

4. $\|(f_1^*(x_i))_1^N - (f_2^*(x_i))_1^N\|_{L_2^N} \lesssim \|f_1^* - f_2^*\|_2 = 4\sqrt{\mathrm{residue}}.$

$\Rightarrow$ This forces the residue to be large.

Theorem ("Gaussian shift theorem". Li & Kuelbs, 98)

Let $\nu \sim \mathcal{N}(0, I_N)$. Let $H_+ = \{x \in \mathbb{R}^N : \langle x, w\rangle \geq b\}$ for some $w \in \mathbb{R}^N, b \in \mathbb{R}$. Let $B \subset \mathbb{R}^N$ such that $\nu(H_+) = \nu(B)$. Then,

$$\nu(w + B) \geq \nu(w + H_+).$$

1. $\mathcal{A}_{f_1^*}|\tau, \mathcal{A}_{f_2^*}|\tau \subset \mathbb{R}^N$ disjoint.
2. $\nu_{f_1^*,\tau}(\mathcal{A}_{f_1^*}|\tau), \nu_{f_2^*,\tau}(\mathcal{A}_{f_2^*}|\tau) \geq 1 - \sqrt{\delta}$ for many τ's.
3. $\nu_{f_1^*,\tau} \sim \mathcal{N}((f_1^*(x_i))_1^N, \sigma^2 I_N)$ and $\nu_{f_2^*,\tau} \sim \mathcal{N}((f_2^*(x_i))_1^N, \sigma^2 I_N)$.

$\Rightarrow$ $\nu_{f_1^*,\tau}(\mathcal{A}_{f_2^*}|\tau) < \sqrt{\delta}$ and $\nu_{f_2^*,\tau}(\mathcal{A}_{f_2^*}|\tau) \geq 1 - \sqrt{\delta}$ for many τ's.

4. $\|(f_1^*(x_i))_1^N - (f_2^*(x_i))_1^N\|_{L_2^N} \lesssim \|f_1^* - f_2^*\|_2 = 4\sqrt{\text{residue}}$.

$\Rightarrow$ This forces the residue to be large.

Theorem ("Gaussian shift theorem". Li & Kuelbs, 98)

Let $\nu \sim \mathcal{N}(0, I_N)$. Let $H_+ = \{x \in \mathbb{R}^N : \langle x, w \rangle \geq b\}$ for some $w \in \mathbb{R}^N, b \in \mathbb{R}$. Let $B \subset \mathbb{R}^N$ such that $\nu(H_+) = \nu(B)$. Then,

$$\nu(w + B) \geq \nu(w + H_+).$$

If $\nu_u \sim \mathcal{N}(u, \sigma^2 I_N)$ and $\nu_v \sim \mathcal{N}(v, \sigma^2 I_N)$ then

$$\nu_u(A) \geq 1 - \Phi\left(\Phi^{-1}(1 - \nu_v(A)) + \|u - v\|_{\ell_2^N}/\sigma\right)$$

for $\Phi(t) = \mathbb{P}[\mathcal{N}(0,1) \leq t]$.

For many τ's :

$$\sqrt{\delta} > \nu_{f_1^*,\tau}(\mathcal{A}_{f_2^*}|\tau)$$

For many τ's :

$$\begin{aligned}\sqrt{\delta} &> \nu_{f_1^*,\tau}(\mathcal{A}_{f_2^*}|\tau) \\ &\geq 1 - \Phi\big(\Phi^{-1}(1 - \nu_{f_2^*,\tau}(\mathcal{A}_{f_2^*}|\tau)) + \|(f_1^*(x_i))_1^N - (f_2^*(x_i))_1^N\|_{L_2^N}/\sigma\big)\end{aligned}$$

For many τ's :

$$
\begin{aligned}
\sqrt{\delta} &> \nu_{f_1^*,\tau}(\mathcal{A}_{f_2^*}|\tau) \\
&\geq 1 - \Phi\big(\Phi^{-1}(1 - \nu_{f_2^*,\tau}(\mathcal{A}_{f_2^*}|\tau)) + \|(f_1^*(x_i))_1^N - (f_2^*(x_i))_1^N\|_{L_2^N}/\sigma\big) \\
&\geq 1 - \Phi\big(\Phi^{-1}(\sqrt{\delta}) + c_0\sqrt{N}\|f_1^* - f_2^*\|_2/\sigma\big) = (\star)
\end{aligned}
$$

For many τ's :

$$\begin{aligned}\sqrt{\delta} &> \nu_{f_1^*,\tau}(\mathcal{A}_{f_2^*}|\tau)\\ &\geq 1 - \Phi\big(\Phi^{-1}(1-\nu_{f_2^*,\tau}(\mathcal{A}_{f_2^*}|\tau)) + \|(f_1^*(x_i))_1^N - (f_2^*(x_i))_1^N\|_{L_2^N}/\sigma\big)\\ &\geq 1 - \Phi\big(\Phi^{-1}(\sqrt{\delta}) + c_0\sqrt{N}\|f_1^* - f_2^*\|_2/\sigma\big) = (\star)\end{aligned}$$

If $c_0\sqrt{N}\|f_1^* - f_2^*\|_2/\sigma \leq |\Phi^{-1}(\sqrt{\delta})|$ then $(\star) \geq 1/2$ which is impossible if $\delta < 1/4$ ($\Phi(\sqrt{\delta}) < 0$). Therefore,

$$16 \times \mathrm{residue} = \|f_1^* - f_2^*\|_2^2 \gtrsim \sigma^2 \frac{(\Phi^{-1}(\sqrt{\delta}))^2}{N} \gtrsim \sigma^2 \frac{\log(1/\delta)}{N}.$$

For many τ's :

$$\begin{aligned}\sqrt{\delta} &> \nu_{f_1^*,\tau}(\mathcal{A}_{f_2^*}|\tau)\\ &\geq 1 - \Phi\Big(\Phi^{-1}(1-\nu_{f_2^*,\tau}(\mathcal{A}_{f_2^*}|\tau)) + \|(f_1^*(x_i))_1^N - (f_2^*(x_i))_1^N\|_{L_2^N}/\sigma\Big)\\ &\geq 1 - \Phi\Big(\Phi^{-1}(\sqrt{\delta}) + c_0\sqrt{N}\|f_1^* - f_2^*\|_2/\sigma\Big) = (\star)\end{aligned}$$

If $c_0\sqrt{N}\|f_1^* - f_2^*\|_2/\sigma \leq |\Phi^{-1}(\sqrt{\delta})|$ then $(\star) \geq 1/2$ which is impossible if $\delta < 1/4$ ($\Phi(\sqrt{\delta}) < 0$). Therefore,

$$16 \times \mathrm{residue} = \|f_1^* - f_2^*\|_2^2 \gtrsim \sigma^2 \frac{(\Phi^{-1}(\sqrt{\delta}))^2}{N} \gtrsim \sigma^2 \frac{\log(1/\delta)}{N}.$$

The result follows for $\delta = 4\exp(-c_4 N\sigma^{-2}(s_N^*)^2)$. ■

Thanks for your attention

Examples of L-sub-Gaussian classes

1. $X = (X^1, \cdots, X^d)$ where $X^1, \cdots, X^d$ are independent L-sub-gaussian variables (i.e. $\|X^i\|_{\psi_2} \leq L\|X^i\|_2$).

1. $X = (X^1, \cdots, X^d)$ where $X^1, \cdots, X^d$ are independent L-sub-gaussian variables (i.e. $\|X^i\|_{\psi_2} \leq L\|X^i\|_2$).
2. the uniform measure on $d^{1/p}B_p^d$.

1. $X = (X^1, \cdots, X^d)$ where $X^1, \cdots, X^d$ are independent L-sub-gaussian variables (i.e. $\|X^i\|_{\psi_2} \leq L\|X^i\|_2$).
2. the uniform measure on $d^{1/p} B_p^d$.
3. $X = (X^1, \cdots, X^d)$ unconditional, supported in RB_∞^d and $\mathbb{E}(X^i)^2 \geq c > 0$.

1. $X = (X^1, \cdots, X^d)$ where $X^1, \cdots, X^d$ are independent L-sub-gaussian variables (i.e. $\|X^i\|_{\psi_2} \leq L\|X^i\|_2$).
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.

1. $X = (X^1, \cdots, X^d)$ where $X^1, \cdots, X^d$ are independent L-sub-gaussian variables (i.e. $\|X^i\|_{\psi_2} \leq L\|X^i\|_2$).
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.

Why do we work at this confidence bound ?

We do have : $\forall t \geq c_0$, with probability greater than $1 - 4\exp(-c_0 t^2 N (s_N^*/\sigma)^2)$,

$$R(\hat{f}_{ERM}) \leq \inf_{f \in \mathcal{F}} R(f) + t^3 (s_N^*)^2.$$

We do have : $\forall t \geq c_0$, with probability greater than $1 - 4\exp(-c_0 t^2 N (s_N^*/\sigma)^2)$,

$$R(\hat{f}_{ERM}) \leq \inf_{f \in \mathcal{F}} R(f) + t^3 (s_N^*)^2.$$

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

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

We do have : $\forall t \geq c_0$, with probability greater than $1 - 4\exp(-c_0 t^2 N(s_N^*/\sigma)^2)$,

$$R(\hat{f}_{ERM}) \leq \inf_{f \in \mathcal{F}} R(f) + t^3 (s_N^*)^2.$$

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

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

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.

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 $d^{-1/r} < s \leq 1$,

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

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

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

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 $d^{-1/r} < s \leq 1$,

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

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 $\ell_*(B_p^d \cap sB_2^d) = \ell_*(sB_2^d) = s\sqrt{d}$.