PhD thesis presentation

Statistical Learning Theory: A PAC-Bayesian Approach

Jean-Yves Audibert

Université Pierre et Marie Curie

PhD advisor : Olivier Catoni

Introduction

Statistical Learning Theory

$\Rightarrow$ how to make predictions about the future based on past experiences

1. Aggregated estimators in L_2 regression
2. A better variance control in classification
   - PAC-Bayesian complexities
   - Compression schemes complexities
3. Classification under Tsybakov's type assumptions

Setup (1/2)

- Training set: $Z_1^N = \left\{ Z_i \triangleq (X_i, Y_i) : i = 1, \ldots, N \right\}$,
  $X_i \in \mathcal{X}, Y_i \in \mathcal{Y}$, $\mathcal{Z} \triangleq \mathcal{X} \times \mathcal{Y}$, (X_i, Y_i) i.i.d. $\sim \mathbb{P}$
- Prediction function : a mapping from $\mathcal{X}$ to $\mathcal{Y}$.
  $$\mathcal{F}(\mathcal{X}, \mathcal{Y}) \triangleq \left\{ \text{prediction functions} \right\}$$
- Loss function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$
- Expected and empirical risks :
  $$R(f) \triangleq \mathbb{P} L[Y, f(X)] \triangleq \mathbb{E}_{\mathbb{P}(dX,dY)} L[Y, f(X)]$$
  $$r(f) \triangleq \bar{\mathbb{P}} L[Y, f(X)] \triangleq \frac{1}{N} \sum_{i=1}^{N} L[Y_i, f(X_i)]$$

Target : Using the data Z_1^N, find a prediction function with the smallest generalization error R.

Setup (2/2)

- In general there is no estimator $\hat{f} : \mathcal{Z}^N \to \mathcal{F}(X;Y)$ s.t.

$$\lim_{N\to+\infty} \sup_{\mathbb{P}\in\mathcal{M}_+^1(\mathcal{Z})} \left\{ \mathbb{P}^{\otimes N} R(\hat{f}_{Z_1^N}) - \inf_{f\in\mathcal{F}(\mathcal{X},\mathcal{Y})} R(f) \right\} = 0.$$

$$\Rightarrow \text{model } \mathcal{F} \subset \mathcal{F}(\mathcal{X},\mathcal{Y})$$

- $\tilde{f} \in \operatorname*{argmin}_{f\in\mathcal{F}} R(f)$
- Classification : $|\mathcal{Y}| < +\infty$ and $L(y,y') \triangleq \mathbb{1}_{y\neq y'}$
- Regression : $\mathcal{Y} = \mathbb{R}$ and $L(y,y') \triangleq (y-y')^2$

Aggregated estimator in L_2 regression (1/4)

- Initial set of function $\mathcal{F}_\Theta \triangleq \left\{ f_\theta : \theta \in \Theta \right\}$
- Mixture model : $\mathcal{F} \triangleq \left\{ \mathbb{E}_{\rho(d\theta)} f_\theta : \rho \in \mathcal{M}_+^1(\Theta) \right\}$
  - Interest in mixtures comes from theoretical and practical results
- $f^* \triangleq \mathbb{E}_{\mathbb{P}}(Y/X = \cdot) \in \operatorname{argmin}_{\mathcal{F}(\mathcal{X},\mathcal{Y})} R$
- Assumptions:
  1. $\forall f, g \in \mathcal{F}_\Theta \cup \{f^*\}, \forall x \in \mathcal{X},$
     $$|f(x) - g(x)| \leq B$$
  2. $\exists \alpha, M > 0$ s.t. $\forall x \in \mathcal{X},$
     $$\mathbb{E}_{\mathbb{P}(dY)} \exp(\alpha |Y - f^*(X)| / X = x) \leq M$$
  - Assumptions satisfied in binary classification

Aggregated estimator in L_2 regression (2/4)

Targets :

- obtain an empirical bound of the efficiency of any mixture
- study the properties of the mixture minimizing the empirical bound

Empirical bound : $\exists\, C_1, C_2 > 0$ depending only on the constants B, α and M s.t. $\forall \epsilon > 0$ and $\forall 0 < \lambda < C_1$, with $\mathbb{P}^{\otimes N}$-probability at least $1 - 2\epsilon$, $\forall \rho \in \mathcal{M}_+^1(\Theta)$,

$$(B_\lambda) \qquad R(\mathbb{E}_{\rho(d\theta)} f_\theta) - R(\tilde{f}) \leq (1+\lambda)\left[r(\mathbb{E}_{\rho(d\theta)} f_\theta) - r(\tilde{f})\right] + 2\lambda \mathbb{E}_{\bar{\mathbb{P}}} \mathrm{Var}_{\rho(d\theta)} f_\theta + \frac{C_2}{N} \frac{K(\rho,\pi) + \log(\epsilon^{-1})}{\lambda}.$$

Aggregated estimator in L_2 regression (3/4)

Estimator

- cut the training set into two pieces
- $\Lambda \triangleq$ geometric grid of $\left[\frac{C}{\sqrt{N}}; C\right]$
- $\forall \lambda \in \Lambda$, define $\hat{\rho}_\lambda$ as the minimizer of bound (B_λ) associated with the first half
- Choose $\hat{\lambda}$ as the ERM over the second half of the training set

Theorem. Let $\mathcal{C} \triangleq \frac{K(\tilde{\rho},\pi)+\log\log N}{N}$ and $\tilde{\rho} \in \mathcal{M}_+^1(\Theta)$ s.t. $R(\mathbb{E}_{\tilde{\rho}(d\theta)} f_\theta) = \min_{\mathcal{F}} R$. For the previous estimator:

$$\mathbb{P}^{\otimes N} R(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) - \min_{\mathcal{F}} R \leq C\left(\sqrt{\mathcal{C}\, \mathbb{E}_{\mathbb{P}} \mathrm{Var}_{\tilde{\rho}(d\theta)} f_\theta} \vee \mathcal{C}\right).$$

Aggregated estimator in L_2 regression (4/4)

Corollary. If $|\Theta| < +\infty$, then

$$\mathbb{P}^{\otimes N} R(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) - R(\tilde{f}) \leq \begin{cases} C \dfrac{\log |\Theta|}{N} & \text{when } \tilde{f} \in \mathcal{F}_\Theta \\ C \sqrt{\dfrac{\log |\Theta|}{N}} & \text{in any case} \end{cases}$$

Aggregated estimator in L_2 regression (4/4)

Corollary. If $|\Theta| < +\infty$, then

$$\mathbb{P}^{\otimes N} R(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) - R(\tilde{f}) \leq \begin{cases} C \dfrac{\log |\Theta|}{N} & \text{when } \tilde{f} \in \mathcal{F}_\Theta \\ C \sqrt{\dfrac{\log |\Theta|}{N}} & \text{in any case} \end{cases}$$

- $\dfrac{\log |\Theta|}{N}$ is the optimal convergence rate in model selection

Aggregated estimator in L_2 regression (4/4)

Corollary. If $|\Theta| < +\infty$, then

$$\mathbb{P}^{\otimes N} R(\mathbb{E}_{\hat{\rho}(d\theta)} f_\theta) - R(\tilde{f}) \leq \begin{cases} C \dfrac{\log|\Theta|}{N} & \text{when } \tilde{f} \in \mathcal{F}_\Theta \\ C\sqrt{\dfrac{\log|\Theta|}{N}} & \text{in any case} \end{cases}$$

- $\dfrac{\log|\Theta|}{N}$ is the optimal convergence rate in model selection
- $\sqrt{\dfrac{\log|\Theta|}{N}}$ is the optimal convergence rate for convex combination when $|\Theta| > \sqrt{N}$

Application to binary classification

Setting

- $\mathcal{X} = \mathbb{R}^d$
- $\mathcal{Y} = \{0; 1\}$, $f^*(X) = \mathbb{P}(Y = 1/X)$, plug-in : $\mathbb{1}_{\hat{f} \geq \frac{1}{2}}$
- $\mathcal{F}_\Theta \triangleq \{0_\mathbb{R}\} \cup \{1_\mathbb{R}\} \bigcup_{\substack{j \in \{1,\ldots,d\} \\ \tau \in \mathbb{R}}} \{\mathbb{1}_{x_j \geq \tau}\} \bigcup_{\substack{j' \in \{1,\ldots,d\} \\ \tau' \in \mathbb{R}}} \{\mathbb{1}_{x'_j < \tau'}\}$,
- π : smooth prior distribution
- Labels generated from Breiman's generators

Application to binary classification

Setting

- $\mathcal{X} = \mathbb{R}^d$
- $\mathcal{Y} = \{0; 1\}$, $f^*(X) = \mathbb{P}(Y = 1/X)$, plug-in : $\mathbb{1}_{\hat{f} \geq \frac{1}{2}}$
- $\mathcal{F}_\Theta \triangleq \{0_\mathbb{R}\} \cup \{1_\mathbb{R}\} \bigcup_{\substack{j \in \{1,\ldots,d\} \\ \tau \in \mathbb{R}}} \{\mathbb{1}_{x_j \geq \tau}\} \bigcup_{\substack{j' \in \{1,\ldots,d\} \\ \tau' \in \mathbb{R}}} \{\mathbb{1}_{x'_j < \tau'}\}$,
- π : smooth prior distribution
- Labels generated from Breiman's generators

Results

A better variance control in classification

- Classification : $|\mathcal{Y}| < +\infty$ and $L(y, y') \triangleq \mathbb{1}_{y \neq y'}$

Transductive setting : we are given the training set Z_1^N and N points to classify $X_{N+1}, \ldots, X_{2N}$.
Target : predict unknown labels $Y_{N+1}, \ldots, Y_{2N}$

$$
\begin{cases}
\bar{\mathbb{P}} & \triangleq & \frac{1}{N} \sum_{i=1}^{N} \delta_{(X_i, Y_i)} \\
\bar{\mathbb{P}}' & \triangleq & \frac{1}{N} \sum_{i=N+1}^{2N} \delta_{(X_i, Y_i)} \\
\bar{\bar{\mathbb{P}}} & \triangleq & \frac{1}{2N} \sum_{i=1}^{2N} \delta_{(X_i, Y_i)} \\
r(f) & \triangleq & \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{Y_i \neq f(X_i)} = \bar{\mathbb{P}}[Y \neq f(X)] \\
r'(f) & \triangleq & \frac{1}{N} \sum_{i=N+1}^{2N} \mathbb{1}_{Y_i \neq f(X_i)} = \bar{\mathbb{P}}'[Y \neq f(X)] \\
\bar{\bar{\mathbb{P}}}_{f_1, f_2} & \triangleq & \bar{\mathbb{P}}[f_1(X) \neq f_2(X)]
\end{cases}
$$

Relative PAC-Bayesian bounds

Definitions. • A function Q on $\mathcal{Z}^{2N}$ is said to be exchangeable iff for any permutation σ, $Q_{Z_{\sigma(1)},\ldots,Z_{\sigma(2N)}} = Q_{Z_1,\ldots,Z_{2N}}$. • $\pi_h \triangleq \frac{\exp(h)}{\pi \exp(h)} \cdot \pi$

Relative PAC-Bayesian bounds

Definitions. • A function Q on $\mathcal{Z}^{2N}$ is said to be exchangeable iff for any permutation σ, $Q_{Z_{\sigma(1)},\ldots,Z_{\sigma(2N)}} = Q_{Z_1,\ldots,Z_{2N}}$. • $\pi_h \triangleq \frac{\exp(h)}{\pi \exp(h)} \cdot \pi$

Theorem. Let π_1 and π_2 be exchangeable prior distributions. Define $\mathcal{K}_{1,2} \triangleq K(\rho_1, \pi_1) + K(\rho_2, \pi_2) + \log(\epsilon^{-1})$. For any $\epsilon > 0$, $\lambda > 0$, with $\mathbb{P}^{\otimes 2N}$-probability at least $1 - \epsilon$, for any $\rho_1, \rho_2 \in \mathcal{M}_+^1(\mathcal{F})$,

$$\rho_2 r' - \rho_1 r' + \rho_1 r - \rho_2 r \leq \frac{2\lambda}{N}(\rho_1 \otimes \rho_2)\overline{\overline{\mathbb{P}}}_{\cdot,\cdot} + \frac{\mathcal{K}_{1,2}}{\lambda}.$$

Relative PAC-Bayesian bounds

Definitions. • A function Q on $\mathcal{Z}^{2N}$ is said to be exchangeable iff for any permutation σ, $Q_{Z_{\sigma(1)},\ldots,Z_{\sigma(2N)}} = Q_{Z_1,\ldots,Z_{2N}}$. • $\pi_h \triangleq \frac{\exp(h)}{\pi\exp(h)} \cdot \pi$

Theorem. Let π_1 and π_2 be exchangeable prior distributions. Define $\mathcal{K}_{1,2} \triangleq K(\rho_1,\pi_1) + K(\rho_2,\pi_2) + \log(\epsilon^{-1})$. For any $\epsilon > 0$, $\lambda > 0$, with $\mathbb{P}^{\otimes 2N}$-probability at least $1-\epsilon$, for any $\rho_1, \rho_2 \in \mathcal{M}_+^1(\mathcal{F})$,

$$\rho_2 r' - \rho_1 r' + \rho_1 r - \rho_2 r \leq \tfrac{2\lambda}{N}(\rho_1 \otimes \rho_2)\bar{\bar{\mathbb{P}}}_{\cdot,\cdot} + \frac{\mathcal{K}_{1,2}}{\lambda}.$$

Theorem. For any $\xi \in]0;1[$ and $\lambda, \lambda_1, \lambda_2 > 0$, define

$$\mathcal{K}_{1,2}^{\mathrm{loc}} \triangleq K\big(\rho_1, (\pi_1)_{-\lambda_1 r}\big) + K\big(\rho_2, (\pi_2)_{-\lambda_2 r}\big) + \log(\pi_1)_{-\lambda_1 r} \exp\left(\tfrac{\lambda_1^2}{2\xi N}\rho_1 \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\right)$$
$$+ \log(\pi_2)_{-\lambda_2 r} \exp\left(\tfrac{\lambda_2^2}{2\xi N}\rho_2 \bar{\bar{\mathbb{P}}}_{\cdot,\cdot}\right) + (1+\xi)\log(\epsilon^{-1}).$$

With $\mathbb{P}^{\otimes 2N}$-probability at least $1-\epsilon$, for any $\rho_1, \rho_2 \in \mathcal{M}_+^1(\mathcal{F})$,

$$\rho_2 r' - \rho_1 r' + \rho_1 r - \rho_2 r \leq \tfrac{2\lambda}{N}(\rho_1 \otimes \rho_2)\bar{\bar{\mathbb{P}}}_{\cdot,\cdot} + \frac{\mathcal{K}_{1,2}^{\mathrm{loc}}}{(1-\xi)\lambda}$$

Application to VC theory (1/3)

- $\mathbb{X} \triangleq X_1^{2N}$
- $\mathcal{A}(\mathbb{X}) \triangleq \left\{ \left\{ f \in \mathcal{F} : \forall 1 \leq i \leq N, f(X_i) = \sigma_i \right\}; \sigma_1^{2N} \in \{0;1\}^{2N} \right\}$
- $N(\mathbb{X}) \triangleq \left| \mathcal{A}(\mathbb{X}) \right| = \left| \left\{ [f(X_k)]_{k=1}^{2N} : f \in \mathcal{F} \right\} \right|$
- $\pi_{\mathcal{U}(\mathbb{X})}$: exchangeable distribution uniform on $\mathcal{A}(\mathbb{X})$ to the extent that $\pi_{\mathcal{U}(\mathbb{X})}(A) = \frac{1}{N(\mathbb{X})}$ for any $A \in \mathcal{A}(\mathbb{X})$.

Theorem. With $\mathbb{P}^{\otimes 2N}$-probability at least $1-\epsilon$, for any $f_1, f_2 \in \mathcal{F}$,

$$r'(f_2) - r'(f_1) \leq r(f_2) - r(f_1) + \sqrt{\frac{8 \bar{\bar{\mathbb{P}}}_{f_1,f_2} \left[2 \log N(\mathbb{X}) + \log(\epsilon^{-1}) \right]}{N}}.$$

In particular, introducing $\tilde{f}' \triangleq \operatorname{argmin}_{\mathcal{F}} r'$, we obtain

$$r'(\hat{f}_{\mathrm{ERM}}) - r'(\tilde{f}') \leq r(\hat{f}_{\mathrm{ERM}}) - r(\tilde{f}') + \sqrt{\frac{8 \bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},\tilde{f}'} [2 \log N(\mathbb{X}) + \log(\epsilon^{-1})]}{N}}.$$

Application to VC theory (2/3)

Localized theorem. For any $\lambda \geq 0$, define

$$\mathcal{C}_\lambda(f) \triangleq \log \sum_{A \in \mathcal{A}(\mathbb{X})} \exp\Big\{ -\lambda\big[(r+r')_A - (r+r')(f)\big]\Big\}.$$

Let $\mathcal{C}(f,g) \triangleq \min_{\lambda \geq 0}\big\{\mathcal{C}_\lambda(f) + \mathcal{C}_\lambda(g)\big\}$. For any $\epsilon > 0$, with $\mathbb{P}^{\otimes 2N}$-probability at least $1-\epsilon$,

$$r'(\hat{f}_{\mathrm{ERM}}) - r'(\tilde{f}') \leq r(\hat{f}_{\mathrm{ERM}}) - r(\tilde{f}') + \sqrt{\frac{8\bar{\mathbb{P}}_{\hat{f}_{\mathrm{ERM}},\tilde{f}'}[\mathcal{C}(\hat{f}_{\mathrm{ERM}},\tilde{f}') + \log(\epsilon^{-1})]}{N}}.$$

Application to VC theory (2/3)

Localized theorem. For any $\lambda \geq 0$, define

$$\mathcal{C}_\lambda(f) \triangleq \log \sum_{A \in \mathcal{A}(\mathbb{X})} \exp\Big\{ -\lambda\big[(r+r')_A - (r+r')(f)\big]\Big\}.$$

Let $\mathcal{C}(f,g) \triangleq \min_{\lambda \geq 0}\big\{\mathcal{C}_\lambda(f) + \mathcal{C}_\lambda(g)\big\}$. For any $\epsilon > 0$, with $\mathbb{P}^{\otimes 2N}$-probability at least $1-\epsilon$,

$$r'(\hat{f}_{\text{ERM}}) - r'(\tilde{f}') \leq r(\hat{f}_{\text{ERM}}) - r(\tilde{f}') + \sqrt{\frac{8\bar{\mathbb{P}}_{\hat{f}_{\text{ERM}},\tilde{f}'}[\mathcal{C}(\hat{f}_{\text{ERM}},\tilde{f}') + \log(\epsilon^{-1})]}{N}}.$$

Illustration of localization efficiency by a toy example.

- $\mathcal{X} = [0;1]$, $\mathcal{F} = \{\mathbb{1}_{[\theta;1]}; \theta \in [0;1]\}$
- $Y = \mathbb{1}_{X \geq \tilde{\theta}}$ for some $\tilde{\theta} \in [0;1]$ and $\mathbb{P}(dX)$ absolutely continuous wrt Lebesgue measure.
- $\rightsquigarrow$ Non localized inequality gives $r'(\hat{f}_{\text{ERM}}) \leq \frac{8\log(2N+1)+4\log(\epsilon^{-1})}{N}$
- $\rightsquigarrow$ Localized inequality gives $r'(\hat{f}_{\text{ERM}}) \leq \frac{37+5\log(\epsilon^{-1})}{N}$

Application to VC theory (3/3)

Empirical VC-bound taking into account the variance term

- $$\bar{\mathcal{F}} \triangleq \left\{ f \in \mathcal{F} : r(f) \leq r(\hat{f}_{\mathrm{ERM}}) + \sqrt{\frac{8 \bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}}, f} [2 \log N(\mathbb{X}) + \log(\epsilon^{-1})]}{N}} \right\}.$$

Theorem. For any $\epsilon > 0$, with $\mathbb{P}^{\otimes 2N}$-probability at least $1 - \epsilon$,

$$r'(\hat{f}_{\mathrm{ERM}}) - r'(\tilde{f}') \leq \sup_{f \in \bar{\mathcal{F}}} \left\{ r(\hat{f}_{\mathrm{ERM}}) - r(f) + \sqrt{\frac{8 \bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}}, f} [2 \log N(\mathbb{X}) + \log(\epsilon^{-1})]}{N}} \right\}$$

To simplify, we can weaken the previous inequality into

$$r'(\hat{f}_{\mathrm{ERM}}) - r'(\tilde{f}') \leq \sqrt{\frac{8 \sup_{\bar{\mathcal{F}}} \bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}}, \cdot} [2 \log N(\mathbb{X}) + \log(\epsilon^{-1})]}{N}}.$$

Another way of controlling the variance term

Reminder

- $$\rho_2 r' - \rho_1 r' + \rho_1 r - \rho_2 r \leq \frac{2\lambda}{N}(\rho_1 \otimes \rho_2)\bar{\bar{\mathbb{P}}}_{.,.} + \frac{\mathcal{K}_{1,2}}{\lambda}$$
- Target : use the bounds to design efficient estimators

Basic approach : consider $(\rho_2, \pi_2, \rho_1, \pi_1) = \left(\rho, \pi, \delta_{\tilde{f}}, \delta_{\tilde{f}}\right)$.

$\rightsquigarrow$ $\rho r' - r'(\tilde{f}) \leq \rho r - r(\tilde{f}) + \frac{2\lambda}{N}\rho\bar{\bar{\mathbb{P}}}_{.,\tilde{f}} + \frac{K(\rho,\pi)+\log(\epsilon^{-1})}{\lambda}$

Another way of controlling the variance term

Reminder

- $$\rho_2 r' - \rho_1 r' + \rho_1 r - \rho_2 r \leq \frac{2\lambda}{N}(\rho_1 \otimes \rho_2)\bar{\bar{\mathbb{P}}}_{.,.} + \frac{\mathcal{K}_{1,2}}{\lambda}$$
- Target : use the bounds to design efficient estimators

Basic approach : consider $(\rho_2, \pi_2, \rho_1, \pi_1) = \left(\rho, \pi, \delta_{\tilde{f}}, \delta_{\tilde{f}}\right)$.

$$\rightsquigarrow \rho r' - r'(\tilde{f}) \leq \rho r - r(\tilde{f}) + \frac{2\lambda}{N}\rho\bar{\bar{\mathbb{P}}}_{.,\tilde{f}} + \frac{K(\rho,\pi)+\log(\epsilon^{-1})}{\lambda}$$

Main problem : control the variance term

Another way of controlling the variance term

Reminder

- $$\rho_2 r' - \rho_1 r' + \rho_1 r - \rho_2 r \leq \frac{2\lambda}{N}(\rho_1 \otimes \rho_2)\bar{\bar{\mathbb{P}}}_{.,.} + \frac{\mathcal{K}_{1,2}}{\lambda}$$
- Target : use the bounds to design efficient estimators

Basic approach : consider $(\rho_2, \pi_2, \rho_1, \pi_1) = \left(\rho, \pi, \delta_{\tilde{f}}, \delta_{\tilde{f}}\right)$.

$\rightsquigarrow \rho r' - r'(\tilde{f}) \leq \rho r - r(\tilde{f}) + \frac{2\lambda}{N}\rho\bar{\bar{\mathbb{P}}}_{.,\tilde{f}} + \frac{K(\rho,\pi)+\log(\epsilon^{-1})}{\lambda}$

Main problem : control the variance term

solution : use iteratively the bounds through comparisons between observable estimators

Non localized estimator

Theorem. Let $L \triangleq \log\left[\log(eN)\epsilon^{-1}\right]$ and

$$S(\rho', \rho'') \triangleq \min_{\lambda \in [\sqrt{N}; N]} \left\{ \tfrac{2\lambda}{N}\left(\rho' \otimes \rho''\right)\bar{\bar{\mathbb{P}}}_{.,.} + \sqrt{e}\,\tfrac{K(\rho',\pi)+K(\rho'',\pi)+L}{\lambda} \right\}.$$

With $\mathbb{P}^{\otimes N}$-proba at least $1-\epsilon$, $\forall \rho', \rho'' \in \mathcal{M}_+^1(\mathcal{F})$,

$$\rho'' r' - \rho' r' \leq \rho'' r - \rho' r + S(\rho', \rho'')$$

Algorithm. Let $\rho_0 = \pi$. For any $k \geq 1$, define ρ_k as the distribution with the smallest complexity $K(\rho_k, \pi)$ such that $\rho_k r - \rho_{k-1} r + S(\rho_{k-1}, \rho_k) \leq 0$. Classify using a function drawn according to the last posterior distribution ρ_K.

Non localized estimator

Theorem. Let

$$\mathbb{G}(\lambda) \triangleq -\tfrac{1}{\lambda}\log \pi \exp\big(-\lambda r'\big) + \tfrac{1}{2\lambda}\log \pi_{-\lambda r'} \exp\left(\tfrac{72\sqrt{e}\lambda^2}{N}\pi_{-\lambda r'}\bar{\bar{\mathbb{P}}}_{.,.}\right) + \tfrac{L}{2\lambda}.$$

With $\mathbb{P}^{\otimes 2N}$-probability at least $1-\epsilon$, for any $k \in \{1,\dots,K\}$,

- $\rho_k r - \rho_{k-1} r + S(\rho_k, \rho_{k-1}) = 0$, $\rho_k r < \rho_{k-1} r$ and $\rho_k r' \leq \rho_{k-1} r'$,
- $K(\rho_k, \pi) \geq K(\rho_{k-1}, \pi)$,
- $\rho_K r' \leq \min_{\frac{\sqrt{N}}{6\sqrt{e}} \leq \lambda \leq \frac{N}{6\sqrt{e}}} \mathbb{G}(\lambda)$.

Optimality of the estimator

Tsybakov's type assumptions:

- there exists $C' > 0$ and $0 < q < 1$ such that the covering entropy of the model $\mathcal{F}$ for the distance $\mathbb{P}_{.,.}$ satisfies for any $u > 0$, $H(u, \mathcal{F}, \mathbb{P}_{.,.}) \leq C' u^{-q}$,
- there exist $c'', C'' > 0$ and $\kappa \geq 1$ such that for any function $f \in \mathcal{F}$,

$$c'' \left[R(f) - R(\tilde{f}) \right]^{\frac{1}{\kappa}} \leq \mathbb{P}_{f,\tilde{f}} \leq C'' \left[R(f) - R(\tilde{f}) \right]^{\frac{1}{\kappa}},$$

$\Rightarrow$ with $\mathbb{P}^{\otimes 2N}$-probability at least $1 - \epsilon$,

$$\mathbb{G}(\lambda) \leq r'(\tilde{f}) + \log(e\epsilon^{-1}) \mathrm{O}\left(N^{-\frac{\kappa}{2\kappa - 1 + q}} \right)$$

provided that $\lambda = N^{\frac{\kappa}{2\kappa - 1 + q}} (\in [\sqrt{N}; N])$ and π is appropriately chosen.

Localized estimator

Use of localized inequalities leads to an improved estimator

Theorem. Let $\Lambda \triangleq \left\{\lambda_j \triangleq \sqrt{N} e^{\frac{j}{2}}; 0 \leq j \leq \log N\right\}$ and $\mathbb{G}_{\mathrm{loc}}(j) \triangleq$

$$\pi_{-\lambda_{j-1} r'} r' + \frac{\sup_{0 \leq i \leq j} \left\{ \log \pi_{-\lambda_i r'} \otimes \pi_{-\lambda_i r'} \exp\left(\frac{C\lambda_i^2}{N} \bar{\mathbb{P}}'_{\cdot,\cdot}\right)\right\}}{\lambda_j} + C \frac{\log[\log(eN)\epsilon^{-1}]}{\lambda_j}$$

for an appropriate constant $C > 0$. For any $\epsilon > 0$, for the localized estimator, with $\mathbb{P}^{\otimes 2N}$-probability at least $1 - \epsilon$,

$$\hat{\rho}_{\mathrm{loc}}\, r' \leq \min_{1 \leq j \leq \log N} \mathbb{G}_{\mathrm{loc}}(j).$$

$\rightsquigarrow$ improvement in the first term of the guarantee since $\pi_{-\lambda r'} r' \leq -\frac{1}{\lambda} \log \pi \exp\left(-\lambda r'\right)$

$\rightsquigarrow$ efficiency of Gibbs classif since $\hat{\rho}_{\mathrm{loc}} \in \left\{\pi_{-\lambda r}, \lambda \in \Lambda\right\}$

Compression schemes (1/4)

- family of algorithms: $\hat{F} : \cup_{n=0}^{+\infty} \mathcal{Z}^n \times \Theta \times \mathcal{X} \to \mathcal{Y}$.
  For any $\theta \in \Theta$, $\hat{F}_\theta$ is an algorithm to the extent that, with any training set Z_1^N, it associates a prediction function $\hat{F}_{Z_1^N,\theta} : \mathcal{X} \to \mathcal{Y}$.
- for any $h \in \mathbb{N}^*$, $\mathcal{I}_h \triangleq \{1,\dots,N\}^h$. Any $I \in \mathcal{I}_h$ can be written as $I = \{i_1,\dots,i_h\}$. Define $I^c \triangleq \{1,\dots,N\} - \{i_1,\dots,i_h\}$ and $Z_I \triangleq \left(Z_{i_1},\dots,Z_{i_h}\right)$. The law of the random variable Z_I will be denoted $\mathbb{P}^I$. For any $J \subset \{1,\dots,N\}$, let $\bar{\mathbb{P}}^J \triangleq \frac{1}{|J|}\sum_{i\in J}\delta_{Z_i}$.
  $\forall I, I_1, I_2$ in $\mathcal{I} \triangleq \underset{2\le h\le N-1}{\cup} \mathcal{I}_h$ and $\theta, \theta_1, \theta_2$ in Θ, introduce

$$\begin{cases} R(I,\theta) \triangleq \mathbb{P}[Y \neq \hat{F}_{Z_I,\theta}(X)] \quad r(I,\theta) \triangleq \bar{\mathbb{P}}^{I^c}[Y \neq \hat{F}_{Z_I,\theta}(X)] \\ \mathbb{P}(I_1,\theta_1,I_2,\theta_2) \triangleq \mathbb{P}[\hat{F}_{Z_{I_1},\theta_1}(X) \neq \hat{F}_{Z_{I_2},\theta_2}(X)] \\ \bar{\mathbb{P}}(I_1,\theta_1,I_2,\theta_2) \triangleq \bar{\mathbb{P}}^{(I_1\cup I_2)^c}[\hat{F}_{Z_{I_1},\theta_1}(X) \neq \hat{F}_{Z_{I_2},\theta_2}(X)] \end{cases}$$

Compression schemes (2/4)

- Let $\pi : \cup_{n=0}^{+\infty} \mathcal{Z}^n \to \mathcal{M}_+^1(\Theta)$ associate a prior distribution on the set Θ with any training sample Z_I.
- For any $\theta \in \Theta$ and any $I \in \mathcal{I}_h$, the complexity of the estimator $\hat{F}_{Z_I,\theta}$ is defined as $\mathcal{C}(I,\theta) \triangleq \log \pi_{Z_I}^{-1}(\theta) + h \log\left(\frac{N}{\alpha}\right)$. To shorten the formulae, introduce $C_{1,2} \triangleq \frac{\mathcal{C}(I_1,\theta_1)+\mathcal{C}(I_2,\theta_2)+\log[(1-\alpha)^{-2}\alpha^4\epsilon^{-1}]}{|(I_1 \cup I_2)^c|}$.
- For any $(I_1,\theta_1,I_2,\theta_2) \in \mathcal{I} \times \Theta \times \mathcal{I} \times \Theta$, define

$$S(I_1,\theta_1,I_2,\theta_2) \triangleq \sqrt{2C_{1,2}\bar{\mathbb{P}}(I_1,\theta_1,I_2,\theta_2) + C_{1,2}^2} + \frac{4C_{1,2}}{3}.$$

Compression schemes (3/4)

Algorithm. Let $I_0 \in \mathcal{I}_2$ and $\theta_0 \in \operatorname{argmax}_{\theta\in\Theta} \pi_{Z_{I_0}}(\theta)$. For any $k \geq 1$, define $I_k \in \bigcup_{2\leq h\leq N-1} \mathcal{I}_h$ and $\theta_k \in \Theta$ such that

$$(I_k, \theta_k) \in \operatorname*{argmin}_{(I,\theta): r(I,\theta) - r(I_{k-1},\theta_{k-1}) + S(I,\theta,I_{k-1},\theta_{k-1}) \leq 0} \mathcal{C}(I,\theta).$$

Classify using the function $\hat{F}_{Z_{I_K},\theta_K}$ where (I_K, θ_K) is the compression set and algorithm obtained at the last iteration.

$$\Downarrow$$

- Regularize any initial overfitting algorithm $\hat{f}$
- Way to choose the similarity measure on the input data, and in particular to choose the kernel of an algorithm
- take into account the variance term

Compression schemes (4/4)

- For any $(I,\theta) \in \mathcal{I} \times \Theta$,
  $k(I,\theta) \triangleq \max\left\{0 \leq k \leq K; \mathcal{C}(I_k,\theta_k) \leq \mathcal{C}(I,\theta)\right\}$.

Theorem. With $\mathbb{P}^{\otimes N}$-proba at least $1-2\epsilon$, for any $k \in \{1,\dots,K\}$, we have

- $r(I_k,\theta_k) < r(I_{k-1},\theta_{k-1})$ and $R(I_k,\theta_k) \leq R(I_{k-1},\theta_{k-1})$,
- $\mathcal{C}(I_k,\theta_k) \geq \mathcal{C}(I_{k-1},\theta_{k-1})$,
- $R(I_K,\theta_K) \leq \inf_{(I,\theta)\in\mathcal{I}\times\Theta}\left\{R(I,\theta) + 2S(I_{k(I,\theta)},\theta_{k(I,\theta)},I,\theta)\right\}$,
- $$R(I_K,\theta_K) \leq \inf_{\substack{(I,\theta)\in\mathcal{I}\times\Theta \\ \xi\geq 0}} \sup_{\substack{(I',\theta')\in\mathcal{I}\times\Theta: \\ \mathcal{C}(I',\theta')\leq\mathcal{C}(I,\theta)}} \Big\{ (1+\xi)R(I,\theta) - \xi R(I',\theta') + 2(1+\xi)S(I',\theta',I,\theta)\Big\}.$$

Proof of the compression schemes guarantee

Lemma. With $\mathbb{P}^{\otimes N}$-proba at least $1-2\epsilon$, $\forall I', I'' \in \mathcal{I}$ and $\theta', \theta'' \in \Theta$,

$$R(I'',\theta'') - R(I',\theta') \leq r(I'',\theta'') - r(I',\theta') + S(I',\theta',I'',\theta'').$$

- By definition of (I_k,θ_k), we get $R(I_k,\theta_k) \leq R(I_{k-1},\theta_{k-1})$.
- For any $(I,\theta) \in \mathcal{I}\times\Theta$, we have

$$\begin{aligned} R(I_K,\theta_K) &\leq R\Big(I_{k(I,\theta)},\theta_{I_{k(I,\theta)}}\Big) \\ &\leq R(I,\theta) + r\Big(I_{k(I,\theta)},\theta_{I_{k(I,\theta)}}\Big) - r(I,\theta) \\ &\qquad + S\Big(I_{k(I,\theta)},\theta_{I_{k(I,\theta)}},I,\theta\Big) \\ &\leq R(I,\theta) + 2S\Big(I_{k(I,\theta)},\theta_{I_{k(I,\theta)}},I,\theta\Big) \end{aligned}$$

Complexity and margin assumptions

- $(u, \mathcal{F}, \mathbb{P}_{.,.})$-covering entropy :

$$H(u, \mathcal{F}, \mathbb{P}_{.,.}) \triangleq \min \left\{ \log |\mathcal{N}| : \mathcal{N} \subset \mathcal{F}(\mathcal{X}, \mathcal{Y}) \text{ s.t. } \mathcal{F} \subset \mathcal{N} + \mathcal{B}_{\mathbb{P}}(u) \right\}$$

- $(u, \mathcal{F}, \mathbb{P}_{.,.})$-bracketing entropy :

$$H_{[]}(u, \mathcal{F}, \mathbb{P}_{.,.}) \triangleq \min \Big\{ \log |\mathcal{N}| : \mathcal{N} \subset \mathcal{F}(\mathcal{X}, \mathcal{Y}) \text{ s.t. } \forall f \in \mathcal{F}, \exists f', f'' \in \mathcal{N} \text{ satisfying } \mathbb{P}_{f',f''} \leq u \text{ and } f' \leq f \leq f'' \Big\}$$

- $$h_q(u) \triangleq \begin{cases} \log(eu^{-1}) & \text{when} \quad q = 0 \\ u^{-q} & \text{when} \quad q > 0 \end{cases} .$$

Complexity assumptions

- *(CA1)* : $\exists C' > 0$ s.t. the covering entropy of the model $\mathcal{F}$ for the distance $\mathbb{P}_{.,.}$ satisfies for any $u > 0$, $H(u, \mathcal{F}, \mathbb{P}_{.,.}) \leq C' h_q(u)$.
- *(CA2)* : $\exists C' > 0$ s.t. the bracketing entropy of the model $\mathcal{F}$ for the distance $\mathbb{P}_{.,.}$ satisfies for any $u > 0$, $H^{[]}(u, \mathcal{F}, \mathbb{P}_{.,.}) \leq C' h_q(u)$.
- *(CA3)* : $\exists C' > 0$ and $\pi \in \mathcal{M}_+^1(\mathcal{F})$ s.t. for any $t > 0$, for any $f' \in \mathcal{F}$, we have $\pi\left(\mathbb{P}_{.,f'} \leq t\right) \geq \exp\left[-C' h_q(t)\right]$.
- Let $t, C' > 0$. A probability distribution π is said to satisfy (t, C')-*(CA3)* when $\pi\left(\mathbb{P}_{.,\tilde{f}} \leq t\right) \geq \exp\left[-C' h_q(t)\right]$.

$\rightsquigarrow$ *(CA2)* $\Rightarrow$ *(CA1)* $\Leftrightarrow$ *(CA3)*

Margin assumptions

$\alpha \in \mathbb{R}_+ \cup \{+\infty\}$, $\kappa \in [1; +\infty]$, $\eta \triangleq \mathbb{E}_{\mathbb{P}}(Y|X = \cdot)$, $\Delta R \triangleq R - R(\tilde{f})$

- *(MA1)* : $\mathcal{Y} = \{0; 1\}$ and there exists $C'' > 0$ s.t. for any $t > 0$, $\mathbb{P}\big(0 < |\eta(X) - 1/2| \leq t\big) \leq C'' t^{\alpha}$.
- *(MA2)* : there exists $C'' > 0$ such that for any function $f \in \mathcal{F}$, $\mathbb{P}_{f,\tilde{f}} \leq C'' \big[\Delta R(f)\big]^{\frac{1}{\kappa}}$.
- *(MA3)* : there exist $c'', C'' > 0$ such that for any function $f \in \mathcal{F}$, $c'' \big[\Delta R(f)\big]^{\frac{1}{\kappa}} \leq \mathbb{P}_{f,\tilde{f}} \leq C'' \big[\Delta R(f)\big]^{\frac{1}{\kappa}}$.
- *(MA4)* : there exist $c'', C'' > 0$ such that $\mathbb{P}_{\cdot,\tilde{f}} \leq C'' \big[\Delta R\big]^{\frac{1}{\kappa}}$, and for any $t > 0$, $\pi(\Delta R \leq t) \geq c'' \pi\big(\mathbb{P}_{\cdot,\tilde{f}} \leq C'' t^{\frac{1}{\kappa}}\big)$

  $\rightsquigarrow$ *(MA3)* $\Rightarrow$ *(MA4)* $\Rightarrow$ *(MA2)*

ERM on nets (1/2)

Theorem. Assume *(MA2)* and *(CA1)*. When *(MA3)* holds, we define

$$(v_N, a_N) \triangleq \begin{cases} \left(\left[\frac{\log N}{N} \right]^{\frac{\kappa}{2\kappa-1}}, \exp\left\{ -\check{C}_1 (\log N)^{\frac{\kappa}{4\kappa-2}} N^{\frac{\kappa-1}{4\kappa-2}} \right\} \right) & \text{for } q = 0 \\ \left(N^{-\frac{\kappa}{2\kappa-1+q}}, \check{C}_1 N^{-\frac{(\kappa-1)\mathbb{1}_{q\leq 1}+q}{q(2\kappa-1+q)}} \right) & \text{for } q > 0 \end{cases}$$

and $b_N \triangleq \check{C}_2 (v_N)^{\frac{1}{\kappa}}$. When *(MA3)* does not hold, we define $(v_N, a_N) \triangleq$

$$\begin{cases} \left(\left[\frac{\log(eN^{1/\kappa})}{N} \right]^{\frac{\kappa}{2\kappa-1}}, \exp\left\{ -\check{C}_1 (\log[eN^{1/\kappa}])^{\frac{\kappa}{4\kappa-2}} N^{\frac{\kappa-1}{4\kappa-2}} \right\} \right) & \text{for } q = 0 \\ \left(N^{-\frac{\kappa}{2\kappa-1+q}}, \check{C}_1 N^{-\frac{\kappa-1+q}{q(2\kappa-1+q)}} \right) & \text{for } 0 < q < 1 \\ \left((\log N) N^{-\frac{1}{2}}, \check{C}_1 (\log N)^{-\frac{1}{2}} N^{-\frac{1}{2}} \right) & \text{for } q = 1 \\ \left(N^{-\frac{1}{1+q}}, \check{C}_1 N^{-\frac{1}{1+q}} \right) & \text{for } q > 1 \end{cases}$$

and $b_N \triangleq \check{C}_2 v_N$.

ERM on nets (2/2)

For any classifier minimizing the empirical risk among a u_N-covering net $\mathcal{N}_{u_N}$ such that

$$a_N \leq u_N \leq b_N \tag{1}$$

and

$$\log \left| \mathcal{N}_{u_N} \right| \leq \check{C}_3 h_q(u_N) \tag{2}$$

for some positive constants $\check{C}_i, i = 1, \ldots, 3$, we have

$$\mathbb{P}^{\otimes N} \left[R(\hat{f}) - R(\tilde{f}) \right] \leq C v_N$$

for some constant $C > 0$ $\big($depending on C'', $\check{C}_i, i = 1, \ldots, 3$ *[*and also on c'' under Assumption *(MA3)]*$\big)$.

Bracketing entropy conditions (1/2)

Theorem. Let us define

$$w_N \triangleq \begin{cases} \left[\frac{\log(eN^{1/\kappa})}{N}\right]^{\frac{\kappa}{2\kappa-1}} & \text{under Assumptions } \textit{(MA2)+(CA2)} \text{ for } q=0 \\ N^{-\frac{\kappa}{2\kappa-1+q}} & \text{under Assumptions } \textit{(MA2)+(CA2)} \text{ for } 0<q< \\ (\log N)N^{-\frac{1}{2}} & \text{under Assumptions } \textit{(MA2)+(CA2)} \text{ for } q=1 \\ N^{-\frac{1}{1+q}} & \text{under Assumptions } \textit{(MA2)+(CA2)} \text{ for } q>1 \end{cases}$$

For any classifier $\hat{f}_{\mathrm{ERM},\mathcal{N}}$ minimizing the empirical risk in a $u_N \triangleq \check{C}_1 w_N$-covering net $\mathcal{N}$ for some positive constant $\check{C}_1$, we have $\mathbb{P}^{\otimes N}\left[R(\hat{f}_{\mathrm{ERM},\mathcal{N}}) - R(\tilde{f})\right] \leq \check{C} w_N$ for some constant $\check{C} > 0$ $\big($depending on C', C'' and $\check{C}_1\big)$.

Bracketing entropy conditions (2/2)

Theorem. Let $\lambda_N \geq \check{C}_1 \frac{h_q(w_N)}{w_N}$ and π be a probability distribution satisfying $(\check{C}_2 w_N, \check{C}_3)$-*(CA3)* for some positive constants $\check{C}_i, i = 1, \ldots, 3$. Then we have

$$\mathbb{P}^{\otimes N}\left[\pi_{-\lambda_N r} R - R(\tilde{f})\right] \leq \breve{C} w_N$$

for some constant $\breve{C} > 0$ (depending on C'', $\check{C}_i, i = 1, \ldots, 3$).

Empirical covering nets

Theorem. Let $\breve{C}$ be positive constant and define

$$(\alpha_q, \beta_q) = \begin{cases} \left(\frac{1}{N}, \frac{\log N}{N}\right) & \text{when } q = 0 \\ \left(\exp\left\{-N^{\frac{q}{1+q}}\right\}, N^{-\frac{1}{1+q}}\right) & \text{when } q > 0 \end{cases}.$$

With $\mathbb{P}^{\otimes N}$-probability at least $1 - (\alpha_q)^{\breve{C}}$, there exists $\breve{C}_1, \breve{C}_2, \breve{C}_3, \breve{C}_4 > 0$ such that for any $u \geq \breve{C}_1 \beta_q$,

- a $(u, \mathcal{F}, \mathbb{P}_{.,.})$-covering net is a $(\breve{C}_3 u, \mathcal{F}, \bar{\mathbb{P}}_{.,.})$-covering net,
- a $(u, \mathcal{F}, \bar{\mathbb{P}}_{.,.})$-covering net is a $(\breve{C}_2 u, \mathcal{F}, \mathbb{P}_{.,.})$-covering net,
- $H(u, \mathcal{F}, \bar{\mathbb{P}}_{.,.}) \leq \breve{C}_4 h_q(u)$.

Chaining and VC theory (1/2)

- VC-dimension of the set $\mathcal{F}$

$$V \triangleq \max\left\{|A| : A \in \mathcal{X}^{2N} \text{ such that } |\{A \cap f^{-1}(1) : f \in \mathcal{F}\}| = 2^{|A|}\right\}$$

Theorem. For any $\epsilon > 0$, with $\mathbb{P}^{\otimes 2N}$-probability at least $1-\epsilon$, we have

$$R(\hat{f}_{\mathrm{ERM}}) \leq \inf_{f \in \mathcal{F}} \left\{ r'(f) + 47\sqrt{\frac{(V+1)\bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},f}}{N}\log\left(\frac{8e}{\bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},f}}\right)} + 34\sqrt{\frac{\bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},f}\log(\epsilon^{-1})}{N}} \right\}$$

$$\rightsquigarrow \mathbb{P}^{\otimes N} R(\hat{f}_{\mathrm{ERM}}) - R(\tilde{f})$$

$$\leq 47\sqrt{\frac{(V+1)\mathbb{P}^{\otimes 2N}\bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},\tilde{f}}}{N}\log\left(\frac{8e}{\mathbb{P}^{\otimes 2N}\bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},\tilde{f}}}\right)} + 34\sqrt{\frac{\mathbb{P}^{\otimes 2N}\bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},\tilde{f}}}{N}}.$$

Chaining and VC theory (2/2)

Theorem. Under assumption *(MA2)*, for any set $\mathcal{F}$ of VC-dimension V, the ERM-classifier satisfies

$$\mathbb{P}^{\otimes N} R(\hat{f}_{\mathrm{ERM}}) - R(\tilde{f}) \leq \breve{C} \begin{cases} \left(\frac{V}{N} \log N\right)^{\frac{\kappa}{2\kappa-1}} & \text{when } 1 \leq \kappa < +\infty \\ \sqrt{\frac{V}{N}} & \text{when } \kappa = +\infty \end{cases} .$$

Conclusion

- Relative bounds used iteratively allows a **better variance control**
- PAC-Bayesian and compression schemes complexities lead to **new algorithms** which have nice theoretical properties
- Future work might look deeper at their **practical efficiency**