PAC-Bayesian bounds

Jean-Yves Audibert[1,2]

1. Imagine - Université Paris Est,
2. Willow - CNRS/ENS/INRIA

July 2009

Outline

Supervised learning

- Training data = n input-output pairs :

$$Z_1 = (X_1, Y_1), \ldots, Z_n = (X_n, Y_n)$$

- A new input X comes.
- Goal: predict the corresponding output Y.
- Probabilistic assumption (batch setting):

$$Z = (X, Y), Z_1, \ldots, Z_n \qquad \text{i.i.d.}$$

from some unknown distribution P

Measuring the quality of prediction

- $\ell(y, y')$ = loss incurred for predicting y' while the true output is y
- Typical losses are:
  - the least square loss for real outputs

$$\ell(y, y') = (y - y')^2$$

  - the classification loss for discrete outputs

$$\ell(y, y') = \mathbf{1}_{y \neq y'}$$

- Prediction function: $f : \mathcal{X} \to \mathcal{Y}$
- Risk: $R(f) = \mathbb{E}\, \ell[Y, f(X)]$

Statistical learning theory (SLT)

- Achievable goal for an estimator $\hat{f}$: predict as well as the best function in a set of prediction functions $\mathcal{F}$ (provided that $\mathcal{F}$ is not too large)
- Central goal of SLT: study $R(\hat{f})$ (whatever $\hat{f}$ is)
- Prominent tool of SLT: probabilistic analysis of the supremum

$$\sup_{f \in \mathcal{F}} \left| R(f) - r(f) \right|$$

with

$$r(f) = \frac{1}{n} \sum_{i=1}^{n} \ell[Y_i, f(X_i)].$$

Kullback-Leibler (KL) divergence

$$K(\rho, \pi) = \begin{cases} \mathbb{E}_{\rho(df)} \log(\frac{\rho}{\pi}(f)) & \text{if } \rho \ll \pi \\ +\infty & \text{otherwise} \end{cases}$$

1. If $\rho \ll \pi$, then we have $K(\rho, \pi) = \mathbb{E}_{\pi(df)} \chi\big(\frac{\rho}{\pi}(f)\big)$ with $\chi : u \mapsto u \log(u) + 1 - u$ convex and nonnegative
2. $K(\rho, \pi) \geq 0$
3. $K(\rho, \pi) = 0 \Leftrightarrow \rho = \pi$
4. If $\mathcal{F}$ is finite and π is the uniform distribution on $\mathcal{F}$, let $H(\rho) = -\sum_{f \in \mathcal{F}} \rho(f) \log \rho(f)$, then

$$K(\rho, \pi) = \log(|\mathcal{F}|) - H(\rho) \leq \log |\mathcal{F}|.$$

Legendre transform of the KL divergence

Let $h : \mathcal{F} \to \mathbb{R}$ s.t. $\mathbb{E}_{\pi(df)} e^{h(f)} < +\infty$. Define

$$\pi_h(df) = \frac{e^{h(f)}}{\mathbb{E}_{\pi(df')} e^{h(f')}} \cdot \pi(df)$$

1. $K(\rho, \pi_h) = K(\rho, \pi) - \mathbb{E}_{\rho(df)} h(f) + \log \mathbb{E}_{\pi(df)} e^{h(f)}$
2. $\sup_\rho \left\{\mathbb{E}_{\rho(df)} h(f) - K(\rho, \pi)\right\} = \log \mathbb{E}_{\pi(df)} e^{h(f)}$
3. $\text{argmax}_\rho \left\{\mathbb{E}_{\rho(df)} h(f) - K(\rho, \pi)\right\} = \pi_h$
4. $\lambda \mapsto K(\pi_{\lambda h}, \pi)$ is nondecreasing on $[0, +\infty)$.

PAC-Bayesian analysis

- Study $\mathbb{E}_{\hat{\rho}(df)} R(f)$ for any distribution $\hat{\rho}$ on $\mathcal{F}$ depending on the training data
  $\longrightarrow$ similar to the study of $R(\hat{f})$ (whatever $\hat{f}$ is)
- Uses a (prior) distribution to evaluate the complexity of the data-dependent (or posterior) distribution
  $\longrightarrow$ different from VC bounds where the complexity is a global quantity characterizing the model $\mathcal{F}$
- The bound holds for any prior and posterior
  $\longrightarrow$ different from the usual Bayesian approach

McAllester's bound (1998,1999)

We assume $0 \le \ell(y, y') \le 1$ for any y, y'.

For any distribution π on $\mathcal{F}$, with probability at least $1 - \epsilon$, for any distribution ρ on $\mathcal{F}$

$$\left|\mathbb{E}_{\rho(df)} R(f) - \mathbb{E}_{\rho(df)} r(f)\right| \le \sqrt{\frac{K(\rho, \pi) + \log(4n\epsilon^{-1})}{2n - 1}}$$

Equivalently (measurability problems set aside), for any data-dependent (posterior) distribution $\hat{\rho}$, with probability at least $1 - \epsilon$,

$$\left|\mathbb{E}_{\hat{\rho}(df)} R(f) - \mathbb{E}_{\hat{\rho}(df)} r(f)\right| \le \sqrt{\frac{K(\hat{\rho}, \pi) + \log(4n\epsilon^{-1})}{2n - 1}}$$

Seeger's proof (slightly revisited)

The PAC lemma

Let V be a real-valued random variable s.t. $\mathbb{E}e^V \leq 1$, then with probability at least $1-\epsilon$, we have

$$V \leq \log(\epsilon^{-1}).$$

- McAllester's bound:

$$V = \sup_\rho \left\{(2n-1)\left[\mathbb{E}_{\rho(df)}R(f) - \mathbb{E}_{\rho(df)}r(f)\right]^2 - K(\rho,\pi) - \log(4n)\right\} \leq \log(\epsilon^{-1}).$$

- First step: Jensen's ineq. + Legendre transform of KL

$$V \leq \sup_\rho \left\{(2n-1)\mathbb{E}_{\rho(df)}[R(f)-r(f)]^2 - K(\rho,\pi) - \log(4n)\right\}$$

$$= -\log(4n) + \log \mathbb{E}_{\pi(df)} e^{(2n-1)[R(f)-r(f)]^2}$$

Seeger's proof (second step)

$$
\begin{aligned}
\mathbb{E}e^V &\le \frac{1}{4n}\mathbb{E}\mathbb{E}_{\pi(df)} e^{(2n-1)[R(f)-r(f)]^2} \\
&= \frac{1}{4n}\mathbb{E}_{\pi(df)}\Big(1+\mathbb{E}\big\{e^{(2n-1)[R(f)-r(f)]^2}-1\big\}\Big) \\
&= \frac{1}{4n}\mathbb{E}_{\pi(df)}\Big(1+\int_0^{+\infty}\mathbb{P}(e^{(2n-1)[R(f)-r(f)]^2}-1>t)dt\Big) \\
&= \frac{1}{4n}\mathbb{E}_{\pi(df)}\Big(1+\int_0^{+\infty}\mathbb{P}(|R(f)-r(f)|>\sqrt{\frac{\log(t+1)}{2n-1}})dt\Big) \\
&\le \frac{1}{4n}\mathbb{E}_{\pi(df)}\Big(1+\int_0^{+\infty}2e^{-2n\frac{\log(t+1)}{2n-1}}dt\Big) \\
&= \frac{1}{4n}\mathbb{E}_{\pi(df)}\Big(1+2\int_1^{+\infty}(t+1)^{-\frac{2n}{2n-1}}dt\Big) \\
&= \frac{4n-1}{4n} \le 1
\end{aligned}
$$

Minimizing McAllester's bound and Gibbs estimator

Let $B(\rho) = \mathbb{E}_{\rho(df)} r(f) + \sqrt{\frac{K(\rho,\pi)+\log(4n\epsilon^{-1})}{2n-1}}$.
McAllester's bound implies: for any distribution ρ

$$\mathbb{E}_{\rho(df)} R(f) \leq B(\rho).$$

Theorem

There exists $\hat{\lambda} \in [\lambda_1, \lambda_2]$ s.t. $B(\pi_{-\hat{\lambda}r}) = \min_\rho B(\rho)$ with $\lambda_1 = \sqrt{4(2n-1)\log(4n\epsilon^{-1})}$ and $\lambda_2 = 2\lambda_1 + 4(2n-1)$.
Besides, we have

1. $\hat{\lambda} = \sqrt{4(2n-1)[K(\pi_{-\hat{\lambda}r}, \pi) + \log(4n\epsilon^{-1})]}$
2. $\hat{\lambda} \in \operatorname{argmin}_{\lambda>0} \left\{ -\frac{1}{\lambda} \log \mathbb{E}_{\pi(df)} e^{-\lambda r(f)} + \frac{\lambda}{4(2n-1)} + \frac{\log(4n\epsilon^{-1})}{\lambda} \right\}$

Seeger's bound for classification (2002)

slightly revisited

- $K(p||q) = K\big(Be(p), Be(q)\big) = p\log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right)$

Theorem

With probability at least $1-\epsilon$, for any distribution ρ on $\mathcal{F}$,

$$K\big(\mathbb{E}_{\rho(df)}r(f)||\mathbb{E}_{\rho(df)}R(f)\big) \leq \frac{K(\rho,\pi) + \log(2\sqrt{n}\epsilon^{-1})}{n}$$

This time, it suffices to prove

$$V = \sup_\rho \left\{ nK(\mathbb{E}_{\rho(df)} r(f)) || \mathbb{E}_{\rho(df)} R(f)) - K(\rho, \pi) - \log(2\sqrt{n}) \right\} \leq \log(\epsilon^{-1}).$$

We have

$$\begin{aligned}
\mathbb{E} e^V &\leq \mathbb{E} e^{\sup_\rho \left\{ n\mathbb{E}_{\rho(df)} K(r(f)||R(f)) - K(\rho,\pi) - \log(2\sqrt{n}) \right\}} \\
&= \frac{1}{2\sqrt{n}} \mathbb{E}\mathbb{E}_{\pi(df)} e^{nK(r(f)),R(f))} \\
&= \frac{1}{2\sqrt{n}} \mathbb{E}_{\pi(df)} \sum_{k=0}^{n} \mathbb{P}(nr(f) = k) \Big(\frac{k}{nR(f)} \Big)^k \Big(\frac{n-k}{n[1-R(f)]} \Big)^{n-k} \\
&= \frac{1}{2\sqrt{n}} \mathbb{E}_{\pi(df)} \sum_{k=0}^{n} \binom{n}{k} \Big(\frac{k}{n} \Big)^k \Big(\frac{n-k}{n} \Big)^{n-k} \\
&\leq 1,
\end{aligned}$$

where the last inequality is obtained from computations using Stirling's approximation.

McAllester's bound vs Seeger's bound

- $\left|\mathbb{E}_{\rho(df)} R(f) - \mathbb{E}_{\rho(df)} r(f)\right| \leq \sqrt{\frac{K(\rho,\pi)+\log(4n\epsilon^{-1})}{2n-1}}$ (1)
- $K\left(\mathbb{E}_{\rho(df)} r(f) || \mathbb{E}_{\rho(df)} R(f)\right) \leq \frac{K(\rho,\pi)+\log(2\sqrt{n}\epsilon^{-1})}{n}$ (2)
- (2) $\Rightarrow$ (1) up to constant since from Pinsker's inequality:

$$\left|\mathbb{E}_{\rho(df)} R(f) - \mathbb{E}_{\rho(df)} r(f)\right| \leq \sqrt{K\left(\mathbb{E}_{\rho(df)} r(f) || \mathbb{E}_{\rho(df)} R(f)\right)}.$$

- (2) $\gg$ (1) when $\mathbb{E}_{\rho(df)} R(f)$ is close to 0 since (2) implies

$$\left|\mathbb{E}_{\rho(df)} R(f) - \mathbb{E}_{\rho(df)} r(f)\right| \leq \sqrt{\frac{2\mathbb{E}_{\rho(df)} r(f)[1 - \mathbb{E}_{\rho(df)} r(f)]\mathcal{K}}{n}} + \frac{4\mathcal{K}}{3n}$$

with

$$\mathcal{K} = K(\rho,\pi) + \log(2\sqrt{n}\epsilon^{-1}).$$

Catoni's old bound for classification (2002)

- Let $\Psi(\lambda) = \frac{e^t - 1 - t}{t^2} \underset{t \to 0}{\longrightarrow} \frac{1}{2}$.

Theorem

For $\lambda > 0$, with proba. at least $1 - \epsilon$, for any distribution ρ on $\mathcal{F}$,

$$\mathbb{E}_{\rho(df)} R(f) \leq \frac{\mathbb{E}_{\rho(df)} r(f)}{1 - \frac{\lambda}{n}\Psi(\frac{\lambda}{n})} + \frac{K(\rho, \pi) + \log(\epsilon^{-1})}{\lambda[1 - \frac{\lambda}{n}\Psi(\frac{\lambda}{n})]}$$

Since typical values of λ are in $[C\sqrt{n}; Cn]$, we roughly have

$$\mathbb{E}_{\rho(df)} R(f) \lesssim \mathbb{E}_{\rho(df)} r(f) + \frac{\lambda}{2n}\mathbb{E}_{\rho(df)} r(f) + \frac{K(\rho, \pi) + \log(\epsilon^{-1})}{\lambda}$$

$$\underset{\text{choice of } \lambda}{\approx} \mathbb{E}_{\rho(df)} r(f) + \sqrt{2\mathbb{E}_{\rho(df)} r(f) \frac{K(\rho, \pi) + \log(\epsilon^{-1})}{n}}$$

Audibert's bound (2004)

- Let $\Psi(\lambda) = \frac{e^t - 1 - t}{t^2} \underset{t \to 0}{\longrightarrow} \frac{1}{2}$.

Theorem

For $\lambda > 0$, with proba. at least $1 - \epsilon$, for any distribution ρ on $\mathcal{F}$,

$$\mathbb{E}_{\rho(df)} R(f) \leq \mathbb{E}_{\rho(df)} r(f) + \frac{\lambda}{n} \Psi\left(\frac{\lambda}{n}\right) \mathbb{E}_{\rho(df)} \mathbf{Var}_Z \ell(Y, f(X)) + \frac{K(\rho, \pi) + \log(\epsilon^{-1})}{\lambda}.$$

Zhang's bound (2005)

Theorem

For $\lambda > 0$, with proba. at least $1 - \epsilon$, for any distribution ρ on $\mathcal{F}$,

$$-\frac{n}{\lambda}\mathbb{E}_{\rho(df)} \log \mathbb{E}_Z e^{-\frac{\lambda}{n}\ell(Y,f(X))} \leq \mathbb{E}_{\rho(df)} r(f) + \frac{K(\rho,\pi) + \log(\epsilon^{-1})}{\lambda}.$$

Since we have

$$-\frac{1}{t}\log \mathbb{E}_Z e^{-t\ell(Y,f(X))} = R(f) - \frac{t}{2}\mathbf{Var}_Z \ell(Y, f(X)) + O(t^2),$$

we have

$$\text{l.h.s.} \approx \mathbb{E}_{\rho(df)} R(f) - \frac{\lambda}{2n}\mathbb{E}_{\rho(df)}\mathbf{Var}_Z \ell(Y, f(X))$$

Comparison of the bounds in classification

- Zhang and Audibert:

$$\mathbb{E}_{\rho(df)}R(f) \lessapprox \mathbb{E}_{\rho(df)}r(f)+\sqrt{2\mathbb{E}_{\rho(df)}\left(R(f)[1-R(f)]\right)\frac{K(\rho,\pi)+\log(\epsilon^{-1})}{n}}$$

- Catoni:

$$\mathbb{E}_{\rho(df)}R(f) \lessapprox \mathbb{E}_{\rho(df)}r(f)+\sqrt{2\mathbb{E}_{\rho(df)}R(f)\frac{K(\rho,\pi)+\log(\epsilon^{-1})}{n}}$$

- Seeger:

$$\mathbb{E}_{\rho(df)}R(f) \le \mathbb{E}_{\rho(df)}r(f)+\sqrt{\frac{2\mathbb{E}_{\rho(df)}R(f)[1-\mathbb{E}_{\rho(df)}R(f)]\mathcal{K}}{n}}+\frac{2\mathcal{K}}{3n}$$

with $\mathcal{K}=K(\rho,\pi)+\log(2\sqrt{n}\epsilon^{-1})$. Besides, we have

$$\mathbb{E}_{\rho(df)}R(f)[1-\mathbb{E}_{\rho(df)}R(f)] \ge \mathbb{E}_{\rho(df)}R(f)[1-R(f)]$$

$\Rightarrow$ similar PAC-Bayes bounds

Explicit Laplace transform in classification

- Instead of using

$$\log \mathbb{E} e^{-\frac{\lambda}{n}\ell(Y,f(X))} \leq -\frac{\lambda}{n} R(f) + \frac{\lambda^2}{n^2} \Psi\left(\frac{\lambda}{n}\right) R(f),$$

use

$$\begin{aligned} \log \mathbb{E} e^{-\frac{\lambda}{n}\ell(Y,f(X))} &= \log\left(1 - R(f)(1 - e^{-\frac{\lambda}{n}})\right) \\ &= -\frac{\lambda}{n} \Phi_{\frac{\lambda}{n}}\left(R(f)\right). \end{aligned}$$

with

$$\Phi_a(p) = -a^{-1} \log[1 - (1 - e^{-a})p] = p - \frac{a}{2} p(1-p) + O(a^2)$$

- Zhang's bound can be used to obtain exactly the same basic bound as Theorem 1.2.6.

Concentration of the risk w.r.t. the posterior distribution

- All PAC-Bayes bounds can be stated as: for any posterior distribution $\hat{\rho}$, with probability at least $1-\epsilon$ w.r.t. to the joint probability $P^{\otimes n}\hat{\rho}$ of the training set and the randomized prediction function $\hat{f} \sim \hat{\rho}$,

$$R(\hat{f}) \leq r(\hat{f}) + \text{ terms with } \log\left(\frac{\hat{\rho}}{\pi}(\hat{f})\right) \text{ instead of } K(\hat{\rho}, \pi)$$

- For instance, Seeger's bound becomes:

$$K\big(r(\hat{f})||R(\hat{f})\big) \leq \frac{\log(\frac{\hat{\rho}}{\pi}(\hat{f})) + \log(2\sqrt{n}\epsilon^{-1})}{n}$$

Unbiased empirical bounds

- Known problem of PAC bounds: pessimistic constants
- Proposed solution: find an empirical quantity $B(\rho)$ s.t.

$$\mathbb{E}[\mathbb{E}_{\rho(df)} R(f)] \leq \mathbb{E}[B(\rho)],$$

and choose the estimator or the parameters by minimizing $B(\rho)$.

Relative bounds

- Main idea: the difference of empirical risks of two close prediction functions has much smaller variations around its mean than the empirical risk of one of these functions.
- Typically, we start with

$$\mathbb{E}_{\rho_1(df)} R(f) - \mathbb{E}_{\rho_2(df)} R(f) \leq \mathbb{E}_{\rho_1(df)} r(f) - \mathbb{E}_{\rho_2(df)} r(f) + \frac{\lambda}{n}\Psi\left(\frac{2\lambda}{n}\right)\mathbb{E}_{\rho_1(df)}\mathbb{E}_{\rho_2(df)}\mathbf{Var}_Z\left[\ell(Y, f_1(X)) - \ell(Y, f_2(X))\right] + \frac{K(\rho_1, \pi_1) + K(\rho_2, \pi_2) + \log(\epsilon^{-1})}{\lambda}.$$

instead of

$$\mathbb{E}_{\rho(df)} R(f) \leq \mathbb{E}_{\rho(df)} r(f) + \frac{\lambda}{n}\Psi\left(\frac{\lambda}{n}\right)\mathbb{E}_{\rho(df)}\mathbf{Var}_Z \ell(Y, f(X)) + \frac{K(\rho, \pi) + \log(\epsilon^{-1})}{\lambda}.$$

1. Fast rates under margin assumptions
   - in classification
     - Mammen and Tsybakov's assumption: for a reference prediction function $\tilde{f} \in \mathcal{G}$, for any $f \in \mathcal{G}$,

$$\mathbb{P}[f(X) \neq \tilde{f}(X)] \leq C[R(f) - R(\tilde{f})]^{1/\kappa}$$

     - Catoni's margin functions:

$$\varphi(t) = \sup_{f \in \mathcal{F}} \left\{ \mathbb{E}\left|\mathbf{1}_{Y \neq f(X)} - \mathbf{1}_{Y \neq \tilde{f}(X)}\right| - t[R(f) - R(\tilde{f})] \right\}$$

$$\bar{\varphi}(t) = \sup_{f \in \mathcal{F}} \left\{ \bar{\mathbb{E}}\left|\mathbf{1}_{Y \neq f(X)} - \mathbf{1}_{Y \neq \tilde{f}(X)}\right| - t[r(f) - r(\tilde{f})] \right\}$$

   - in least squares regression, under reasonable assumptions,

$$\mathbf{Var}_Z\left[\ell(Y, f(X)) - \ell(Y, \tilde{f}(X))\right] \leq c[R(f) - R(\tilde{f})]$$

2. Algorithm design by successive improvement

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 $Y_{N+1}, \ldots, Y_{2N}$

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

Another way of controlling the variance term

Reminder

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

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

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

Main problem : control the variance term

Non localized estimator

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

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

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

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

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 $\mathbb{P}_{\cdot,\cdot}$ satisfies for any $u > 0$, $H(u, \mathcal{F}, \mathbb{P}_{\cdot,\cdot}) \leq C' u^{-q}$,
- there exist $c'', C'' > 0$ and $\kappa \geq 1$ such that for any function $f \in \mathcal{F}$,

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

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

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

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

PAC-Bayesian localization

- For a given $\hat{\rho}$, the prior minimizing the expected value of the bound for $\hat{\rho}$ is

$$\pi = \text{argmin}_{\pi'} \mathbb{E} K(\hat{\rho}, \pi') = \mathbb{E}[\hat{\rho}]$$

since $\mathbb{E} K(\hat{\rho}, \pi) = \mathbb{E} K(\hat{\rho}, \mathbb{E}[\hat{\rho}]) + K(\mathbb{E}[\hat{\rho}], \pi)$.

- Problem: $\mathbb{E}[\hat{\rho}]$ is not observable
- First solution (Catoni, 2003): apply basic PAC bound to $\pi_{-\beta R}$, expand $K(\hat{\rho}, \pi_{-\beta R})$ and develop additional empirical bounds to control the non observable terms
  - Zhang (2005) uses $\pi_{\alpha \log \mathbb{E}_Z e^{-\lambda \ell(Y, f(X))}}$.
  - Ambroladze, P.-H. and S.-T. (2006) localizes by cutting the training set into two parts
  - Catoni (2007) uses $\pi_{-\beta \Phi_{-\frac{\beta}{n}}[R(f)]}$.
  - Alquier (2007,2008) also uses $\pi_{-\beta R}$ but for general unbounded losses (regression, density estimation)

Properties of PAC-Bayesian localization

- Advantages
  - allow to replace $K(\rho, \pi)$ with $K(\rho, \pi_{-\lambda r})$
  - gain of logarithmic factor in parametric convergence rates
- Disadvantages = increase of the constant factors
- Open question = useful to build linear classifiers ? (Herbrich, Graepel, 2001; Langford, Shawe-Taylor, 2002; Germain, Lacasse, Laviolette, Marchand, 2009)

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 $X_{N+1},\dots,X_{2N}$.
Target : predict unknown labels $Y_{N+1},\dots,Y_{2N}$

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

Relative PAC-Bayesian bounds

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

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 $\rho_1, \rho_2 \in \mathcal{M}_+^1(\mathcal{F})$,

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

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

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

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

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

Application to VC theory (1/3)

- $\mathbb{X} \triangleq X_1^{2N}$
- $\mathcal{A}(\mathbb{X}) \triangleq \Big\{ \big\{ f \in \mathcal{F} : \forall 1 \le i \le N, f(X_i) = \sigma_i \big\} ; \sigma_1^{2N} \in \{0;1\}^{2N} \Big\}$
- $N(\mathbb{X}) \triangleq \big|\mathcal{A}(\mathbb{X})\big| = \Big| \big\{ [f(X_k)]_{k=1}^{2N} : f \in \mathcal{F} \big\} \Big|$
- $\pi_{\mathcal{U}(\mathbb{X})}$: 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 $f_1, f_2 \in \mathcal{F}$,

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

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

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

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 $1-\epsilon$,

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

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 $1-\epsilon$,

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

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.
- $\leadsto$ Non localized inequality gives $r'(\hat{f}_{\mathrm{ERM}}) \leq \frac{8 \log(2N+1) + 4 \log(\epsilon^{-1})}{N}$
- $\leadsto$ Localized inequality gives $r'(\hat{f}_{\mathrm{ERM}}) \leq \frac{37 + 5 \log(\epsilon^{-1})}{N}$

Target

- $\ell(y, y') = (y - y')^2$
- $R(f) = \mathbb{E}(Y - f(X))^2$
- $\varphi_1, \ldots, \varphi_d$ functions from $\mathcal{X}$ to $\mathbb{R}$

$$X \longrightarrow \begin{pmatrix} \varphi_1(X) \\ \vdots \\ \varphi_d(X) \end{pmatrix} = \varphi(X)$$

- $\Theta \subset \mathbb{R}^d$ closed convex
- $\mathcal{F} = \left\{ f_\theta = \sum_{j=1}^d \theta_j \varphi_j ; \theta = (\theta_1, \ldots, \theta_d) \in \Theta \right\}$
- Goal: predict as well as $f^* \in \text{argmin}_{f \in \mathcal{F}} R(f)$ (which is possibly different from $f^{(\text{reg})} : x \mapsto \mathbb{E}(Y|X = x)$)

Decomposition of the risk

- Gram matrix: $Q = \mathbb{E}[\varphi(X)\varphi^T(X)]$
- The risk is a quadratic form with matrix Q:

$$\begin{aligned} R(f_\theta) &= \mathbb{E}(Y - \theta^T\varphi(X))^2 \\ &= \mathbb{E}Y^2 - 2\theta^T\mathbb{E}[\varphi(X)Y] + \theta^T Q\theta \end{aligned}$$

Motivations

- Better understanding of the parametric linear least squares regression
- Central task for nonparametric regression with linear approximation space
- Two-stage model selection

Ordinary least squares and empirical risk minimization

- Linear aggregation: $\mathcal{F} = \mathcal{F}_{\text{lin}} = \text{span}\{\varphi_1, \ldots, \varphi_d\}$ and $f^*_{\text{lin}} = f^*$
- Let $\hat{f}^{(\text{ols})} \in \text{argmin}_{f \in \mathcal{F}_{\text{lin}}} \frac{1}{n} \sum_{i=1}^n [Y_i - f(X_i)]^2$.
- $\mathbb{E} R(\hat{f}^{(\text{ols})}) - R(f^*_{\text{lin}}) = \mathbb{E}\left[\hat{f}^{(\text{ols})}(X) - f^*_{\text{lin}}(X)\right]^2$.
- if $\sup_{x \in \mathcal{X}} \text{Var}\,(Y | X = x) = \sigma^2 < +\infty$ and $f^{(\text{reg})} = f^*_{\text{lin}}$, we have

$$\mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^n \left[\hat{f}^{(\text{ols})}(X_i) - f^*_{\text{lin}}(X_i)\right]^2\right\} \le \sigma^2 \frac{d}{n}.$$

- It does not imply a $\frac{d}{n}$ upper bound on $\mathbb{E} R(\hat{f}^{(\text{ols})}) - R(f^*_{\text{lin}})$.

Theorem (Györfi, Kohler, Krzyżak, Walk, 2004)

If $\sup_{x \in \mathcal{X}} \mathbf{V}ar(Y|X=x) = \sigma^2 < +\infty$ *and*

$$\|f^{(\text{reg})}\|_\infty = \sup_{x \in \mathcal{X}} |f^{(\text{reg})}(x)| \leq H$$

for some $H > 0$*, then the truncated estimator* $\hat{f}_H^{(\text{ols})} = (\hat{f}^{(\text{ols})} \wedge H) \vee -H$ *satisfies*

$$\mathbb{E}R(\hat{f}_H^{(\text{ols})}) - R(f^{(\text{reg})}) \leq 8[R(f^*_{\text{lin}}) - R(f^{(\text{reg})})] + \kappa \frac{(\sigma^2 \vee H^2) d \log n}{n}$$

for some numerical constant κ*.*

Theorem (Birgé, Massart, 1998)

Assume that for any f_1, f_2 in $\mathcal{F}$, $\|f_1 - f_2\|_\infty \leq H$ and $\exists f_0 \in \mathcal{F}$ satisfying

$$\text{for any } x \in \mathcal{X}, \quad \mathbb{E}\Big\{\exp\Big[A^{-1}\big|Y - f_0(X)\big|\Big] \,\Big|\, X = x\Big\} \leq M,$$

for some positive constants A and M. Let

$$\tilde{B} = \inf_{\phi_1,\ldots,\phi_d} \sup_{\theta \in \mathbb{R}^d - \{0\}} \frac{\|\sum_{j=1}^d \theta_j \phi_j\|_\infty^2}{\|\theta\|_\infty^2}$$

where the infimum is taken w.r.t. all possible orthonormal basis of $\mathcal{F}$ for $\langle f_1, f_2\rangle = \mathbb{E} f_1(X) f_2(X)$. Then, with probability at least $1-\epsilon$:

$$R(\hat{f}^{(\mathrm{erm})}) - R(f^*) \leq \kappa(A^2 + H^2)\frac{d\log[2 + (\tilde{B}/n) \wedge (n/d)] + \log(\epsilon^{-1})}{n},$$

where κ is a positive constant depending only on M.

Projection estimator

Theorem (Tsybakov, 2003)

Let $\phi_1, \dots, \phi_d$ be an o.n.b. of $\mathcal{F}_{\text{lin}}$ for $\langle f_1, f_2 \rangle = \mathbb{E} f_1(X) f_2(X)$. The projection estimator on this basis is $\hat{f}^{(\text{proj})} = \sum_{j=1}^{d} \hat{\theta}_j^{(\text{proj})} \phi_j$, with

$$\hat{\theta}^{(\text{proj})} = \frac{1}{n} \sum_{i=1}^{n} Y_i \phi_j(X_i).$$

If

$$\sup_{x \in \mathcal{X}} \mathbf{V}ar(Y|X = x) = \sigma^2 < +\infty$$

and

$$\|f^{(\text{reg})}\|_\infty = \sup_{x \in \mathcal{X}} |f^{(\text{reg})}(x)| \leq H < +\infty,$$

then we have

$$\mathbb{E} R(\hat{f}^{(\text{proj})}) - R(f^*_{\text{lin}}) \leq (\sigma^2 + H^2) \frac{d}{n}.$$

Conclusion of the survey

- $R(\hat{f}^{(\mathrm{erm})}) - R(f^*) = \mathrm{O}\big(\frac{d\log(2+n/d)+\log(\epsilon^{-1})}{n}\big)$ for L_∞-bounded $\mathcal{F}$ and exponential moments
- There is no simple d/n which does not require strong assumptions
- Degraded convergence rate when Q is ill-conditioned ?

Theorem

Assume $\mathbb{E}\left[\|\varphi(X)\|^4\right] < +\infty$ *and* $\sup \mathbb{E}\{[Y - \tilde{f}(X)]^2 | X\} \leq \sigma^2$. *For any* $\epsilon > 0$, *there is* n_ϵ *s.t. for any* $n \geq n_\epsilon$, *with proba. at least* $1 - \epsilon$,

$$R(\hat{f}^{(\mathrm{erm})}) \leq R(f^*_{\mathrm{lin}}) + \sigma^2 \frac{30d + 1000 \log(3\epsilon^{-1})}{n}$$

- Θ bounded
- π uniform distribution on $\mathcal{F}$
- $\lambda > 0$
- $W_i(f, f') = \frac{\lambda}{n}\{[Y_i - f(X_i)]^2 - [Y_i - f'(X_i)]^2\}$
- $\hat{\mathcal{E}}(f) = \log \mathbb{E}_{\pi(df')} \frac{1}{\prod_{i=1}^n [1 - W_i(f,f') + \frac{1}{2} W_i(f,f')^2]}$
- We consider the "posterior" distribution $\hat{\pi} = \pi_{-\hat{\mathcal{E}}(f)}$
- for $\frac{\lambda}{n}$ small enough, $1 - W_i(f, f') + \frac{1}{2} W_i(f, f')^2$ is close to $e^{-W_i(f,f')}$, and consequently

$$\hat{\mathcal{E}}(f) \approx \lambda r(f) + \log \mathbb{E}_{\pi(df')} e^{-\lambda r(f')},$$

and

$$\hat{\pi} \approx \pi_{-\lambda r}$$

Theorem

Assume $\sup_{f_1,f_2\in\mathcal{F}} \|f_1 - f_2\|_\infty \leq H$ *and, for some* $\sigma > 0$,

$$\sup_{x\in\mathcal{X}} \mathbb{E}\{[Y - f^*(X)]^2 | X = x\} \leq \sigma^2 < +\infty.$$

Let $\lambda = \frac{n}{3(2\sigma+H)^2}$ *and* $\hat{f}$ *be a prediction function drawn from the distribution* $\hat{\pi}$.
Then for any $\epsilon > 0$, *with probability at least* $1 - \epsilon$, *we have*

$$R(\hat{f}) - R(f^*) \leq 17(2\sigma + H)^2 \frac{d + \log(2\epsilon^{-1})}{n}.$$