PAC-Bayesian bounds and aggregation

Jean-Yves Audibert[1,2]

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

March 2010

Outline

1. Context
2. The different PAC-Bayes bounds
3. The three aggregation problems
   - Model selection type aggregation
   - Convex aggregation in high dimension
   - Linear aggregation
   - High-dimensional input and sparsity

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: $\ell(y, y') = (y - y')^2$
  - the classification loss for discrete outputs: $\ell(y, y') = \mathbf{1}_{y \neq y'}$
- Prediction function: $f : \mathcal{X} \to \mathcal{Y}$
- Risk: $R(f) = \mathbb{E}\,\ell[Y, f(X)]$
- Empirical risk: $r(f) = \frac{1}{n}\sum_{i=1}^{n} \ell[Y_i, f(X_i)]$

Kullback-Leibler (KL) divergence

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

1. If $\rho \ll \pi$, then we have $K(\rho,\pi) = \mathbb{E}_{\pi(df)} \chi\left(\frac{\rho}{\pi}(f)\right)$ 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. $\operatorname{argmax}_\rho \left\{ \mathbb{E}_{\rho(df)} h(f) - K(\rho, \pi) \right\} = \pi_h$
4. $\lambda \mapsto K(\pi_{\lambda h}, \pi)$ is nondecreasing on $[0, +\infty)$.

PAC-Bayesian analysis

- PAC-Bayesian approach: for any distribution ρ on $\mathcal{F}$,

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

where the bound $B(\rho)$ relies on the use at some point of

$$\sup_{\rho} \left\{ \mathbb{E}_{\rho(df)} d\big(R(f), r(f)\big) - K(\rho, \pi) \right\} = \log \mathbb{E}_{\pi(df)} e^{d(R(f), r(f))}$$

- Traditional SLT: for any $f \in \mathcal{F}$, $R(f) \leq \tilde{B}(f)$
- Dissimilarity between the approaches because of the KL term
- Uses a (prior) distribution to evaluate the complexity of the posterior distribution
- The bound holds for any prior and posterior
$\longrightarrow$ different from the usual Bayesian approach

McAllester's bound (1998,1999)

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

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

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

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

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

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)

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

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

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

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

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

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

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

Audibert's bound (2004)

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

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

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

we have

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

Catoni's bound (2007)

- 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

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

with

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

$\Rightarrow$ tighter constants and variance appearing implicitly

Comparison of the bounds in classification

- Zhang, A., Catoni (2007):

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

- Catoni (2002):

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

- Seeger:

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

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

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

$\Rightarrow$ similar PAC-Bayes bounds

Least square regression setting

- $R(g) = \mathbb{E}[Y - g(X)]^2$.
- Bounded noise setting: $Y \in [-1, 1]$
- $g_1, \ldots, g_d : \mathcal{X} \to \mathcal{Y}$, with $\|g_1\|_\infty, \ldots, \|g_d\|_\infty \leq 1$

$$g^*_{\mathbf{MS}} \in \operatorname*{argmin}_{g \in \{g_1, \ldots, g_d\}} R(g),$$

$$g^*_{\mathbf{C}} \in \operatorname*{argmin}_{g \in \{\sum_{j=1}^d \theta_j g_j ; \theta_1 \geq 0, \ldots, \theta_d \geq 0, \sum_{j=1}^d \theta_j = 1\}} R(g),$$

$$g^*_{\mathbf{L}} \in \operatorname*{argmin}_{g \in \{\sum_{j=1}^d \theta_j g_j ; \theta_1 \in \mathbb{R}, \ldots, \theta_d \in \mathbb{R}\}} R(g).$$

Optimal rates of aggregation

There exist $\hat{g}_{\mathbf{MS}}$, $\hat{g}_{\mathbf{C}}$ and $\hat{g}_{\mathbf{L}}$ such that

$$\mathbb{E}R(\hat{g}_{\mathbf{MS}}) - R(g^*_{\mathbf{MS}}) \leq C \min\left(\frac{\log d}{n}, 1\right),$$

$$\mathbb{E}R(\hat{g}_{\mathbf{C}}) - R(g^*_{\mathbf{C}}) \leq C \min\left(\sqrt{\frac{\log(1 + d/\sqrt{n})}{n}}, \frac{d}{n}, 1\right),$$

$$\mathbb{E}R(\hat{g}_{\mathbf{L}}) - R(g^*_{\mathbf{L}}) \leq C \min\left(\frac{d}{n}, 1\right),$$

where $\hat{g}_L$ requires the knowledge of the input distribution.

Optimal rates of aggregation (Tsybakov, 2003)

- $\sigma > 0$
- $\mathcal{P}_\sigma$ = set of proba. on $\mathcal{X} \times \mathbb{R}$ such that $Y = g(X) + \xi$, with $\|g\|_\infty \leq 1$, and $\xi \sim \mathcal{N}(0, \sigma^2)$
- For appropriate choices of $g_1, \ldots, g_d$:

$$\inf_{\hat{g}} \sup_{P \in \mathcal{P}_\sigma} \{\mathbb{E}R(\hat{g}) - R(g^*_{\mathbf{MS}})\} \geq C \min\left(\frac{\log d}{n}, 1\right),$$

$$\inf_{\hat{g}} \sup_{P \in \mathcal{P}_\sigma} \{\mathbb{E}R(\hat{g}) - R(g^*_{\mathbf{C}})\} \geq C \min\left(\sqrt{\frac{\log(1 + d/\sqrt{n})}{n}}, \frac{d}{n}, 1\right),$$

$$\inf_{\hat{g}} \sup_{P \in \mathcal{P}_\sigma} \{\mathbb{E}R(\hat{g}) - R(g^*_{\mathbf{L}})\} \geq C \min\left(\frac{d}{n}, 1\right).$$

Unusual properties

$$g^*_{\mathbf{MS}} \in \operatorname*{argmin}_{g \in \{g_1,\dots,g_d\}} R(g)$$

- To be "optimal", we need to choose $\hat{g}$ outside the model $\mathcal{G}$.
- Up to recently, the only known optimal algorithm is the progressive mixture rule
- The proof is neither based on bounds on the supremum of empirical processes nor on the PAC-Bayesian analysis

Progressive mixture rule (Catoni, 1999; Yang, 2000)

- π uniform distribution on the finite set $\{g_1, \ldots, g_d\}$
- $\lambda > 0$
- $\Sigma_i(g) = \sum_{k=1}^{i} [Y_k - g(X_k)]^2$: cumulative loss on the first i data points
- The progressive mixture rule: $\hat{g}_{\text{PM}} = \frac{1}{n+1} \sum_{i=0}^{n} \mathbb{E}_{g \sim \pi_{-\lambda\Sigma_i}} g$, i.e.,

$$\hat{g}_{\text{PM}}(x) = \frac{1}{n+1} \sum_{i=0}^{n} \frac{\sum_{j=1}^{d} g_j(x) e^{-\lambda \Sigma_i(g_j)}}{\sum_{j=1}^{d} e^{-\lambda \Sigma_i(g_j)}}.$$

- Theoretical guarantee:

$$\mathbb{E} R(\hat{g}_{\text{PM}}) - R(g^*_{\mathbf{MS}}) \leq \frac{8 \log d}{n+1}$$

Progressive indirect mixture rules (A., 2009)

- $\lambda > 0$
- For any $i \in \{0, \ldots, n\}$, let $\hat{h}_i$ be a prediction function s.t.

$$\forall X, Y \qquad [Y - \hat{h}_i(X)]^2 \leq -\frac{1}{\lambda} \log \mathbb{E}_{g \sim \pi_{-\lambda\Sigma_i}} e^{-\lambda[Y-g(X)]^2} \qquad (1)$$

- Progressive indirect mixture rule: $\hat{g}_\lambda = \frac{1}{n+1} \sum_{i=0}^{n} \hat{h}_i$.
- $\hat{h}_i = \mathbb{E}_{g \sim \pi_{-\lambda\Sigma_i}} g$ satisfies (1) for $\lambda \leq 1/8$.
- $\hat{h}_i$ exists even for $\lambda = 1/2$, and then

$$\mathbb{E}R(\hat{g}_{1/2}) - R(g^*_{\mathbf{MS}}) \leq \frac{2 \log d}{n+1}$$

Excess risk deviations abnormally high

- $\mathbb{E}R(\hat{g}_\lambda) - R(g^*_{\mathbf{MS}}) = \mathrm{O}\left(\frac{1}{n}\right) \not\Rightarrow R(\hat{g}) - R(g^*_{\mathbf{MS}}) = \mathrm{O}\left(\frac{1}{n}\right)$ w.h.p.
- $g_1 = 1$ and $g_2 = -1$
- For any $\lambda > 0$ and any training set size n large enough, there exist $\epsilon > 0$ and a distribution generating the data for which with probability larger than ϵ, we have

$$R(\hat{g}_\lambda) - R(g^*_{\mathbf{MS}}) \geq c\sqrt{\frac{\log(e\epsilon^{-1})}{n}}$$

Getting round the previous limitation (A., 2007)

- $r(g) = \frac{1}{n}\sum_{i=1}^{n}[Y_i - g(X_i)]^2$.
- $\hat{g}_{\text{ERM}} \in \underset{g\in\{g_1,\dots,g_d\}}{\operatorname{argmin}} \; r(g)$.
- $[g, g'] = \left\{\alpha g + (1-\alpha)g' : \alpha \in [0,1]\right\}$.
- The empirical star estimator is

$$\hat{g} \in \underset{g\in[\hat{g}_{\text{ERM}},g_1]\cup\dots\cup[\hat{g}_{\text{ERM}},g_d]}{\operatorname{argmin}} \; r(g).$$

- Theoretical guarantee: with probability at least $1-\epsilon$,

$$R(\hat{g}) - R(g^*_{\mathbf{MS}}) \le \frac{600\log(d\epsilon^{-1})}{n}.$$

See also Lecué and Mendelson (2009)

Different approaches

$$g_{\mathbf{C}}^* \in \underset{g \in \{\sum_{j=1}^d \theta_j g_j; \theta_1 \geq 0, \ldots, \theta_d \geq 0, \sum_{j=1}^d \theta_j = 1\}}{\operatorname{argmin}} R(g)$$

$$\sqrt{n} \ll d \ll e^n$$

- Apply the previous progressive mixture rule on an appropriate grid (Tsybakov, 2003)
- Use the exponentiated gradient algorithm (Kivinen and Warmuth, 1997; Cesa-Bianchi, 1999)
- Use a stochastic version of the mirror descent algorithm (Juditsky, Nazin, Tsybakov, Vayatis, 2005)

Results in expectation, based on a sequential procedure

A PAC-Bayesian approach (A., 2004)

$$\mathbb{E}[Y - \mathbb{E}_{g \sim \rho} g(X)]^2 = \mathbb{E}_{(g',g'') \sim \rho \otimes \rho} \mathbb{E}[Y - g'(X)][Y - g''(X)]$$

- Apply the PAC-Bayesian analysis for distributions on the product space $\{g_1, \dots, g_d\} \times \{g_1, \dots, g_d\}$
- PAC-Bayes bound: with probability at least $1 - \epsilon$,

$$R(\mathbb{E}_{g \sim \hat{\rho}} g) - R(g^*_{\mathbf{C}}) \leq \min_{\lambda \in [0, C_1]} \Big\{ (1 + \lambda) \big[r(\mathbb{E}_{g \sim \hat{\rho}} g) - r(g^*_{\mathbf{C}}) \big] + \frac{2\lambda}{n} \sum_{i=1}^{n} \mathrm{Var}_{g \sim \hat{\rho}} g(X_i) + C_2 \frac{1}{n} \frac{K(\hat{\rho}, \pi) + \log(2 \log(2n) \epsilon^{-1})}{\lambda} \Big\}.$$

The minimizer of the PAC-Bayes bound

- π = uniform distribution on $\{g_1, \ldots, g_d\}$
- $\hat{\rho}_{\mathbf{C}}$ = distribution minimizing the upper bound
- $g^*_{\mathbf{C}} = \mathbb{E}_{g \sim \rho^*_{\mathbf{C}}} g$.
- Theoretical guarantee: with probability at least $1 - \epsilon$,

$$R(\mathbb{E}_{g \sim \hat{\rho}_{\mathbf{C}}} g) - R(g^*_{\mathbf{C}}) \leq C \sqrt{\frac{\log(d \log(2n)\epsilon^{-1})}{n} \mathbb{E}\mathrm{Var}_{g \sim \rho^*_{\mathbf{C}}} g(X)} + C \frac{\log(d \log(2n)\epsilon^{-1})}{n},$$

- Excess risk at most of order $\sqrt{\frac{\log(d \log(2n))}{n}}$
- If $\rho^*_{\mathbf{C}}$ is a Dirac, excess risk at most of order $\frac{\log(d \log(2n))}{n}$

$$g_{\mathbf{L}}^* \in \underset{g \in \{\sum_{j=1}^d \theta_j g_j ; \theta_1 \in \mathbb{R}, \dots, \theta_d \in \mathbb{R}\}}{\operatorname{argmin}} R(g).$$

- Linear aggregation = linear least squares regression
- Assume that we know that $g_{\mathbf{L}}^* \in \mathcal{G}$, where $\mathcal{G}$ is L_∞ bounded
- There is no simple d/n bound which does not require strong assumptions if we care about logarithmic factors

$$R(\hat{g}_{\text{ERM}}) - R(g^*) \leq C \frac{d \log(2 + n/d) + \log(\epsilon^{-1})}{n}.$$

(Birgé and Massart, 1998)

A PAC-Bayesian approach with Gaussian prior (A. and Catoni, 2009)

- π uniform distribution on $\mathcal{G}$
- For an appropriate $\lambda > 0$, with probability at least $1 - \epsilon$,

$$R(\mathbb{E}_{g\sim\pi_{-\lambda r}} g) - R(g^*) \leq C\,\frac{d + \log(2\epsilon^{-1})}{n},$$

- Shrinking effect of $\pi_{-\lambda r}$ when compared to $\hat{g}_{\text{ERM}}$.

$$n \ll d \ll e^n$$

- predicting as $g^*_{\mathbf{C}}$ = achievable : $\sqrt{\frac{\log d}{n}}$
- predicting as $g^*_{\mathbf{L}}$ = not achievable : $\frac{d}{n}$

$$g^* \in \underset{g \in \{\sum_{j=1}^d \theta_j g_j; \theta_1 \in \mathbb{R}, \dots, \theta_d \in \mathbb{R}, \sum_{j=1}^d \mathbf{1}_{\theta_j \neq 0} \leq s\}}{\operatorname{argmin}} R(g).$$

- g^* achievable by Lasso (L_1 regularization) under strong assumptions on the correlations of $g_1(X), \dots, g_d(X)$

A model selection approach

- $\mathcal{L}_1 = \{Z_1, \ldots, Z_{n/2}\}$, and $\mathcal{L}_2 = \{Z_{n/2+1}, \ldots, Z_n\}$
- For any $I \subset \{1, \ldots, d\}$ of size s, let $\hat{g}_I$ be the Gibbs estimator for linear aggregation of $(g_j)_{j \in I}$ trained on $\mathcal{L}_1$
- Let $\hat{g}$ be the empirical star estimator trained on $\mathcal{L}_2$ and associated with the $\binom{d}{s}$ functions $\hat{g}_I$

$$R(\hat{g}) - R(g^*) \leq C \frac{s \log(d/s) + \log(2\epsilon^{-1})}{n}$$