Une approche PAC-bayésienne de la régression en grande dimension

Benjamin Guedj

Avant : UPMC & Telecom ParisTech
Bientôt : INRIA Lille - Nord Europe

En collaboration avec
Pierre Alquier, Gérard Biau,
Éric Moulines et Sylvain Robbiano

Statistical framework

From a training sample

$$\mathcal{D}_n = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$$

of i.i.d. replications of a r.v.

$$(\mathbf{X}, Y) \in \mathbb{R}^d \times \mathcal{Y} \qquad\qquad \text{with } \mathcal{Y} = \mathbb{R} \text{ or } \mathcal{Y} = \{\pm 1\}$$

learn the relationship between Y and $\mathbf{X}$.

Goal

Estimation of a transform Φ of the regression function

$$\mathbf{x} \mapsto \Phi\left(\mathbb{E}[Y|\mathbf{X} = \mathbf{x}]\right)$$

Highlights

- High-dimensional setting: $d \gg n$.
- Sparsity-based perspective, carrying no assumptions on the design.
- ***Modus operandi***: PAC-Bayesian theory.
  Main references: Shawe-Taylor and Williamson (1997), McAllester (1999), Catoni (2004, 2007), Audibert (2004, 2010), Alquier (2006, 2008), Dalalyan and Tsybakov (2008, 2012)...
- **Implementation**: MCMC algorithm favoring local moves of the Markov chain.
  Main references: Carlin and Chib (1995), Leung and Barron (2006), Hans et al. (2007), Petralias (2010), Petralias and Dellaportas (2012)...

Aggregation approach

From a known dictionary $\mathbb{D} = \{\phi_1, \phi_2, \ldots, \phi_M\}$, aggregated estimators are of the form $f_\theta = \theta^\top \mathbb{D} = \sum_{k=1}^M \theta_k \phi_k$ where:

- $\theta \in \{e_1, \ldots, e_M\}$ (selectors),
- $\theta \in \Lambda^M = \{\lambda \in \mathbb{R}_+^M \colon \sum_{k=1}^M \lambda_k = 1\}$ (convex aggregation),
- $\theta \in \mathbb{R}^M$ (linear aggregation),
- ...

Aggregation approach

From a known dictionary $\mathbb{D} = \{\phi_1, \phi_2, \ldots, \phi_M\}$, aggregated estimators are of the form $f_\theta = \theta^\top \mathbb{D} = \sum_{k=1}^M \theta_k \phi_k$ where:

- $\theta \in \{e_1, \ldots, e_M\}$ (selectors),
- $\theta \in \Lambda^M = \{\lambda \in \mathbb{R}_+^M \colon \sum_{k=1}^M \lambda_k = 1\}$ (convex aggregation),
- $\theta \in \mathbb{R}^M$ (linear aggregation),
- ...

$$\left\{ f_\theta = \sum_{j=1}^{d} \sum_{k=1}^{m_j} \theta_{jk} \phi_k \ , \quad \theta \in \Theta = \mathbb{R}^{\sum_{j=1}^d m_j}, \quad |f_\theta|_\infty \leq C \right\},$$

where $\mathbf{m} = (m_1, \ldots, m_d) \in \{0, \ldots, M\}^d$ is a model.

For some loss function ℓ, risk and empirical risk of an estimator f_θ

$$R(f_\theta) = \mathbb{E}\ell(Y, f_\theta(\mathbf{X})), \quad R_n(f_\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f_\theta(\mathbf{X}_i)).$$

PAC-Bayesian estimators

- Set a prior probability measure π on Θ, promoting sparsity.
- Constrained optimization problem:

$$\arg\min_{\rho} \left\{ \int_{\Theta} R_n(f_\theta)\rho(\mathrm{d}\theta) + \frac{\lambda}{n}\mathcal{KL}(\rho,\pi) \right\},$$

with the Kullback-Leibler divergence

$$\mathcal{KL}(\rho,\pi) = \int \log\left[\frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta)\right]\rho(\mathrm{d}\theta).$$

PAC-Bayesian estimators

- Set a prior probability measure π on Θ, promoting sparsity.
- Constrained optimization problem:

$$\arg\min_{\rho} \left\{ \int_{\Theta} R_n(f_\theta)\rho(\mathrm{d}\theta) + \frac{\lambda}{n}\mathcal{KL}(\rho, \pi) \right\},$$

with the Kullback-Leibler divergence

$$\mathcal{KL}(\rho, \pi) = \int \log\left[\frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta)\right]\rho(\mathrm{d}\theta).$$

- Unique solution: Gibbs posterior distribution

$$\hat{\rho}_\lambda(\mathrm{d}\theta) \propto \exp[-\lambda R_n(f_\theta)]\pi(\mathrm{d}\theta).$$

PAC-Bayesian estimators

- Two estimators in this talk:

$$\hat{\theta} \sim \hat{\rho}_\lambda \quad \text{(Randomized estimator)},$$

$$\bar{\theta} = \int_\Theta \theta \hat{\rho}_\lambda(\mathrm{d}\theta) = \mathbb{E}_{\hat{\rho}_\lambda}\theta \quad \text{(Posterior mean)}.$$

- PAC-Bayesian theory is a great tool to produce estimators with nearly minimax optimal properties.
- PAC-Bayesian bounds depend on the KL divergence and hold for any prior π.

PAC-Bayesian estimators

- Two estimators in this talk:

$$\hat{\theta} \sim \hat{\rho}_\lambda \quad \text{(Randomized estimator)},$$

$$\bar{\theta} = \int_\Theta \theta \hat{\rho}_\lambda(\mathrm{d}\theta) = \mathbb{E}_{\hat{\rho}_\lambda}\theta \quad \text{(Posterior mean)}.$$

- PAC-Bayesian theory is a great tool to produce estimators with nearly minimax optimal properties.
- PAC-Bayesian bounds depend on the KL divergence and hold for any prior π.

Take-home message

PAC-Bayesian theory adapts nicely to high-dimensional problems when coupled with a sparsity-inducing prior.

Sparsity-inducing prior

$$\pi(\theta) \propto \sum_{\mathbf{m}} \binom{d}{|\mathbf{m}|_0}^{-1} \beta^{\sum_{j=1}^{d} m_j} \, \mathrm{Unif}_{\mathcal{B}_{\mathbf{m}}}(\theta),$$

where $\beta \in (0,1)$ and

$$\mathcal{B}_{\mathbf{m}} = \left\{ \theta, \quad \sum_{j=1}^{d} \sum_{k=1}^{m_j} |\theta_{jk}| \leq C \right\}.$$

This prior distribution gives larger mass to sparse parameters.

Goal

Obtain oracle inequalities on the excess risk of the PAC-Bayesian estimators $f_{\hat{\theta}}$ and $f_{\bar{\theta}}$: For any $\varepsilon \in (0,1)$,

$$\mathbb{P}\left[R(f_{\hat{\theta}}) - R^\star \leq \mathrm{K}_\lambda \inf_{\theta} \left\{ R(f_\theta) - R^\star + \Delta_{n,d,M,\varepsilon}(\theta) \right\} \right] \geq 1 - \varepsilon.$$

Regression I

- $\mathcal{Y} = \mathbb{R}$, model $Y = \psi^\star(\mathbf{X}) + W$.
- Assumption: $|\psi^\star|_\infty \leq C$.

Theorem (G. and Alquier, 2013)

For any $\varepsilon \in (0,1)$, *any* $0 < \lambda < n/(4\sigma^2 + 4C^2)$, *with probability at least* $1 - \varepsilon$,

$$\left.\begin{array}{l} R(f_{\hat{\theta}}) - R(\psi^\star) \\ R(f_{\bar{\theta}}) - R(\psi^\star) \end{array}\right\} \leq \mathrm{K}_\lambda \times \inf_{\mathbf{m}} \inf_{\theta \in \mathcal{B}_{\mathbf{m}}} \left\{ R(f_\theta) - R(\psi^\star) \right.$$

$$\left. + |\mathbf{m}|_0 \frac{\log(d/|\mathbf{m}|_0)}{n} + \frac{\log(n)}{n} \sum_{j=1}^{d} m_j + \frac{\log(2/\varepsilon)}{n} \right\},$$

where $\mathrm{K}_\lambda \xrightarrow[\lambda \to 0]{} 1$ and $\mathrm{K}_\lambda \xrightarrow[\lambda \to n/(4\sigma^2 + C^2)]{} +\infty$.

Regression II

- Let $\phi_1, \phi_2, \dots$ refer to the trigonometric basis and assume that $\psi^\star = \sum_{j \in S^\star} \psi_j^\star$, where

$$\psi_j^\star \in \mathcal{W}(r_j, \ell_j) = \left\{ f \in \mathrm{L}^2([-1,1]) \colon f = \sum_{k=1}^{\infty} \theta_k \phi_k \text{ and } \sum_{i=1}^{\infty} i^{2r_j} \theta_i^2 \leq \ell_j \right\}.$$

Theorem (G. and Alquier, 2013)

For any real $\varepsilon \in (0,1)$, *any* $0 < \lambda < n/(4\sigma^2 + 4C^2)$, *with probability at least* $1 - \varepsilon$,

$$\left.\begin{array}{l} R(f_{\hat{\theta}}) - R(\psi^\star) \\ R(f_{\bar{\theta}}) - R(\psi^\star) \end{array}\right\} \leq$$

$$\mathrm{K}_\lambda \times \left\{ \sum_{j \in S^\star} \ell_j^{\frac{1}{2r_j+1}} \left(\frac{\log(n)}{2nr_j} \right)^{\frac{2r_j}{2r_j+1}} + \frac{|S^\star| \log(d/|S^\star|)}{n} + \frac{\log(2/\varepsilon)}{n} \right\}.$$

Logistic regression I

$\mathcal{Y} = \{\pm 1\}$, model

$$\log \frac{\mathbb{P}(Y = 1 | \mathbf{X} = \mathbf{x})}{1 - \mathbb{P}(Y = 1 | \mathbf{X} = \mathbf{x})} = \nu(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d.$$

Logistic loss function:

$$\ell \colon (Y, f_\theta(\mathbf{X})) \mapsto \log\left[1 + \exp(-Y f_\theta(\mathbf{X}))\right].$$

Simplified framework where $\mathbf{m} = (m_1, \dots, m_d) \in \{0, M\}^d$.

Theorem (G., 2013)

For any $\varepsilon \in (0,1)$,

$$\mathbb{P}\left[R(f_{\bar{\theta}}) \leq \mathrm{K}_\lambda \inf_\rho \left\{ \int R(f_\theta)\rho(\mathrm{d}\theta) + \frac{\mathcal{KL}(\rho, \pi)}{n} + \frac{\log(2/\varepsilon)}{n} \right\}\right] \geq 1 - \varepsilon.$$

Logistic regression II

Theorem (G., 2013)

For any $\varepsilon \in (0,1)$*, with probability at least* $1-\varepsilon$*,*

$$R(f_{\bar{\theta}}) \leq \mathrm{K}_\lambda \inf_{\mathbf{m}\in\mathcal{M}} \inf_{\theta\in\mathcal{B}_{\mathbf{m}}} \left\{ R(f_\theta) + \frac{|\mathbf{m}|_0}{n}\left[M\log\left(\frac{n}{M|\mathbf{m}|_0}\right) + \log\left(\frac{de}{|\mathbf{m}|_0}\right) + \log(1/\beta)\right] + \frac{\log(2/\varepsilon)}{n} \right\},$$

where $\mathrm{K}_\lambda \xrightarrow[\lambda\to 0]{} 1$*.*

Binary ranking I

- $\mathcal{Y} = \{\pm 1\}$, model $\eta \colon \mathbf{x} \mapsto \mathbb{P}\{Y = 1 | \mathbf{X} = \mathbf{x}\}$.
- Ranking consists in ordering $\mathbb{R}^d$ such that the order of labels is preserved.
- Goal: construct a so-called scoring function $s : \mathbb{R}^d \to \mathbb{R}$, such that for any pair $(\mathbf{x}, \mathbf{x}') \in \mathbb{R}^d \times \mathbb{R}^d$, $s(\mathbf{x}) \leq s(\mathbf{x}') \Leftrightarrow \eta(\mathbf{x}) \leq \eta(\mathbf{x}')$.
- Ranking risk:

$$R(s) \stackrel{def}{=} \mathbb{P}\left\{(s(\mathbf{X}) - s(\mathbf{X}')) \cdot (Y' - Y) < 0\right\},$$

and empirical counterpart

$$R_n(s) \stackrel{def}{=} \frac{1}{n(n-1)} \sum_{i \neq j} \mathbb{1}\{(Y_i - Y_j)(s(\mathbf{X}_i) - s(\mathbf{X}_j)) < 0\}.$$

Binary ranking II

- Set of scoring functions:

$$\mathcal{S}_\Theta = \left\{ s_\theta \colon \mathbf{x} \mapsto \sum_{j=1}^{d} \sum_{k=1}^{M} \theta_{jk}\phi_k(x_j), \quad \theta \in \mathbb{R}^{dM} \right\}.$$

  Simplified framework where $\mathbf{m} = (m_1, \ldots, m_d) \in \{0, M\}^d$.

- (Empirical) Excess risk s :

$$\mathcal{E}(s) \stackrel{def}{=} R(s) - R^\star, \quad \mathcal{E}_n(s) \stackrel{def}{=} R_n(s) - R_n(\eta).$$

- PAC-Bayesian estimator $\hat{s} = s_{\hat{\theta}}$ where $\hat{\theta} \sim \hat{\rho}_\lambda$.

Binary ranking III

Condition (C)

For any $\lambda > 0$*, and any scoring function* s*,*

$$\mathbb{E} \exp \left[\lambda \left(\mathcal{E}_n(s) - \mathcal{E}(s) \right) \right] \leq \exp(\psi),$$

where ψ *may depend on* n *and* λ*.*

Theorem

Under **C***, for any* $\varepsilon \in (0,1)$*,*

$$\mathbb{P} \left[\mathcal{E}(\hat{s}) \leq \inf_{\rho} \left\{ \mathcal{E}(s) + \frac{2\psi + 2\log(2/\varepsilon) + 2\mathcal{KL}(\rho, \pi)}{\lambda} \right\} \right] \geq 1 - \varepsilon,$$

where $s \sim \rho$*.*

Binary ranking IV

Corollary

For any distribution of $(\mathbf{X}, Y)$, **C** *holds for* $\psi = \lambda^2/4n$.
With $\lambda = \sqrt{n}$, *for any* $\varepsilon \in (0,1)$,

$$\mathbb{P}\left[\mathcal{E}(\hat{s}) \leq \inf_{\rho} \left\{\mathcal{E}(s) + \frac{1/2 + 2\log(2/\varepsilon) + 2\mathcal{KL}(\rho,\pi)}{\sqrt{n}}\right\}\right] \geq 1 - \varepsilon.$$

Corollary

Using the sparsity-inducing prior π, *with*

$$\lambda = c\sqrt{n|\mathbf{m}|_0 \log(d)},$$

for any $\varepsilon \in (0,1)$, *with probability at least* $1 - \varepsilon$,

$$\mathcal{E}(\hat{s}) \leq \inf_{\mathbf{m}} \inf_{\theta \in \mathcal{B}_{\mathbf{m}}} \left\{\mathcal{E}(s_\theta) + c' \frac{\sqrt{\log(2/\varepsilon) + |\mathbf{m}|_0 \left(\log(1/\beta) + \log(d)\right)}}{\sqrt{n}}\right\}.$$

Binary ranking V

Condition (MA(α))

The distribution of $(\mathbf{X}, Y)$ satisfies a margin condition **MA**(α) *of parameter $\alpha \in (0,1)$ if there exists $C < \infty$ such that for any scoring function s,*

$$\mathbb{P}\left[(s(\mathbf{X}) - s(\mathbf{X}'))(\eta(\mathbf{X}) - \eta(\mathbf{X}')) < 0\right] \leq C(R(s) - R^\star)^{\frac{\alpha}{1+\alpha}}.$$

Lemma

Let s be a scoring function, and

$$T = \mathbb{1}_{\{(s(\mathbf{X})-s(\mathbf{X}'))(Y-Y')<0\}} - \mathbb{1}_{\{(\eta(\mathbf{X})-\eta(\mathbf{X}'))(Y-Y')<0\}}.$$

Under the condition **MA**(α)*,*

$$\mathrm{Var}(T) \leq \mathcal{C}(R(s) - R^\star)^{\frac{\alpha}{1+\alpha}}.$$

Binary ranking VI

Corollary

Under **MA**(α)*, condition* **C** *holds for* $\psi = \frac{n}{2}\mathrm{Var}(T)\phi\left(\frac{2\lambda}{n}\right)$*, with* $\phi\colon t \mapsto e^t - t - 1$*. With* $\lambda = C_1^{-1} n^{\frac{1+\alpha}{2+\alpha}}$*, for any* $\varepsilon \in (0,1)$*, with probability at least* $1-\varepsilon$ *:*

$$\mathcal{E}(\hat{s}) \leq \inf_{\rho} \left\{ 2\mathcal{E}(s) + C_1 n^{-\frac{1+\alpha}{2+\alpha}} \left[\log(2/\varepsilon) + \mathcal{KL}(\rho, \pi)\right] \right\}$$

where C_1 *depends on* α *and* $\mathcal{C}$*.*

Binary ranking VII

Proposition

With the sparsity-inducing prior π*, with* $\lambda = C_1 \log(d)^{\frac{1}{2+\alpha}} n^{\frac{1+\alpha}{2+\alpha}}$*, for any* $\varepsilon \in (0,1)$*, with probability at least* $1-\varepsilon$ *:*

$$\mathcal{E}(\hat{s}) \leq \inf_{\mathbf{m}} \inf_{\theta \in \mathcal{B}_{\mathbf{m}}(t)} \left\{ 2\mathcal{E}(s_\theta) + C_2 n^{-\frac{1+\alpha}{2+\alpha}} K^{\frac{1+\alpha}{2+\alpha}} \right\},$$

where C_1 *and* C_2 *depend on* $\mathcal{C}$ *and* α*, and*

$$K = \log(2/\varepsilon) + |\mathbf{m}|_0 \left[\log(1/\beta) + \log(d)\right].$$

A challenging problem

- Goal: Sample a chain with stationary distribution $\hat{\rho}_\lambda$.
- The sample space is very high-dimensional, and its structure is non standard.
- Existing PAC-Bayesian implementations:
  - RJMCMC for the Single-Index model (Alquier and Biau, 2013).
  - Langevin Monte-Carlo for fixed design regression (Dalalyan and Tsybakov, 2012).
  - ...

A subspace Carlin & Chib-like approach

- Metropolized version of the Carlin & Chib algorithm (originally introduced by Petralias and Dellaportas (2012) for Bayesian model selection).
- Key idea: Introduce pseudopriors and define a neighborhood relationship on the models space.

[1] Least-squares fit, maximum likelihood estimator, ...

A subspace Carlin & Chib-like approach

- Metropolized version of the Carlin & Chib algorithm (originally introduced by Petralias and Dellaportas (2012) for Bayesian model selection).

- Key idea: Introduce pseudopriors and define a neighborhood relationship on the models space.

- For any model $\mathbf{m}$, define its neighborhood $\mathbb{V} = \{\mathbb{V}^+, \mathbb{V}^-\}$:
  - $\mathbb{V}^+$: All models with the regressors from $\mathbf{m}$ plus one.
  - $\mathbb{V}^-$: All models with the regressors from $\mathbf{m}$ but one.

- For any model $\mathbf{m}$, pseudoprior defined as Gaussian with mean equal to some default estimator[1] in model $\mathbf{m}$ and covariance matrix $\Sigma = \sigma^2 \mathcal{I}$, σ^2 being a parameter.

[1] Least-squares fit, maximum likelihood estimator, ...

Algorithm (R package `pacbpred`)

At iteration $t = 1, \ldots, T$:

1. Pick a move: Add, delete a covariate, or stay in the current model.

Algorithm (R package pacbpred)

At iteration $t = 1, \ldots, T$:

1. Pick a move: Add, delete a covariate, or stay in the current model.
2. For each of the neighbors models, draw a candidate estimator from the Gaussian pseudoprior (whose density is denoted by φ).

Algorithm (R package pacbpred)

At iteration $t = 1, \ldots, T$:

1. Pick a move: Add, delete a covariate, or stay in the current model.
2. For each of the neighbors models, draw a candidate estimator from the Gaussian pseudoprior (whose density is denoted by φ).
3. Pick the model j and candidate parameter θ_j with probability

$$\frac{\hat{\rho}_\lambda(\theta_j)/\varphi(\theta_j)}{\sum_k \hat{\rho}_\lambda(\theta_k)/\varphi(\theta_k)}.$$

Algorithm (R package pacbpred)

At iteration $t = 1, \ldots, T$:

1. Pick a move: Add, delete a covariate, or stay in the current model.
2. For each of the neighbors models, draw a candidate estimator from the Gaussian pseudoprior (whose density is denoted by φ).
3. Pick the model j and candidate parameter θ_j with probability

$$\frac{\hat{\rho}_\lambda(\theta_j)/\varphi(\theta_j)}{\sum_k \hat{\rho}_\lambda(\theta_k)/\varphi(\theta_k)}.$$

4. The Metropolis-Hastings acceptance ratio is

$$\alpha = \min\left(1, \frac{\hat{\rho}_\lambda(\theta_j)\varphi(\theta^{t-1})}{\hat{\rho}_\lambda(\theta^{t-1})\varphi(\theta_j)}\right).$$

Highlights

Take-home message

- Nearly minimax optimal estimators in a variety of high-dimensional models.
- Oracle risk bounds in probability under little or no assumption.
- Competitive implementation via MCMC, enforcing sparse models.

References

- G. and Alquier (2013), *PAC-Bayesian Estimation and Prediction in Sparse Additive Models*. Electronic Journal of Statistics.
- G. (2012), R package *pacbpred*, version 0.92.2.
- G. (2013), *Agrégation d'estimateurs et de classificateurs : théorie et méthodes*. Ph.D. thesis, UPMC.
- G. and Robbiano (2014), *Une approche PAC-bayésienne d'un problème de ranking binaire en grande dimension. 46èmes Journées de Statistique de la SFdS, Rennes*.

Key result

Lemma (Catoni, 2004)

Let $(A, \mathcal{A})$ *be a measurable space. For any probability measure* μ *on* $(A, \mathcal{A})$ *and any measurable function* $h : A \to \mathbb{R}$ *such that* $\int (\exp \circ h) \mathrm{d}\mu < \infty$,

$$\log \int (\exp \circ h) \mathrm{d}\mu = \sup_{m \in \mathcal{M}^1_\mu(A,\mathcal{A})} \left\{ \int h \mathrm{d}m - \mathcal{KL}(m, \mu) \right\},$$

with the convention $\infty - \infty = -\infty$. *Further, if* h *is upper-bounded on the support of* μ, *the supremum with respect to* m *in the right-hand term is reached for the Gibbs distribution* g *defined by*

$$\frac{\mathrm{d}g}{\mathrm{d}\mu}(a) = \frac{\exp \circ h(a)}{\int (\exp \circ h) \mathrm{d}\mu}, \quad a \in A.$$

Concentration inequality

Lemma (Massart, 2007)

Let $(T_i)_{i=1}^n$ *be a collection of real independant random variables. Assume there exist two positive constants* v *and* w *such that*

$$\sum_{i=1}^{n} \mathbb{E}T_i^2 \leq v,$$

and for any integer $k \geq 3$*,*

$$\sum_{i=1}^{n} \mathbb{E}[(T_i)_+^k] \leq \frac{k!}{2} v w^{k-2}.$$

Then, for any $\gamma \in (0, \frac{1}{w})$*,*

$$\mathbb{E}\left[\exp\left(\gamma \sum_{i=1}^{n} (T_i - \mathbb{E}T_i)\right)\right] \leq \exp\left(\frac{v\gamma^2}{2(1 - w\gamma)}\right).$$