Risk bounds for linear regression

Jean-Yves Audibert[1,2] & Olivier Catoni[3]

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

October 2009

Least squares regression

- Training data = n input-output pairs :

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

- A new input X comes
- General goal: predict the corresponding output Y
- Probabilistic assumption :

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

from some unknown distribution P

- Prediction function: $f : \mathcal{X} \to \mathbb{R}$
- Risk: $R(f) = \mathbb{E}\left[Y - f(X)\right]^2$

Linear least squares

- $\varphi_1, \dots, \varphi_d$ 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
- $\mathcal{F} = \left\{ f_\theta = \sum_{j=1}^d \theta_j \varphi_j ; \theta = (\theta_1, \dots, \theta_d) \in \Theta \right\}$
- Goal: predict as well as $f^* \in \operatorname{argmin}_{f \in \mathcal{F}} R(f)$ (which is possibly different from $f^{(\mathrm{reg})} : x \mapsto \mathbb{E}(Y|X = x)$)

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 $\hat{f}^{(\text{ols})} \in \text{argmin}_{f \in \mathcal{F}_{\text{lin}}} \frac{1}{n} \sum_{i=1}^{n} [Y_i - f(X_i)]^2$.
- if $\sup_{x \in \mathcal{X}} \mathbf{Var}\,(Y|X = x) = \sigma^2 < +\infty$ and $f^{(\text{reg})} = f^*_{\text{lin}}$, we have

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

- $\mathbb{E} R(\hat{f}^{(\text{ols})}) - R(f^*_{\text{lin}}) = \mathbb{E}\left[\hat{f}^{(\text{ols})}(X) - f^*_{\text{lin}}(X) \right]^2$.
- 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*

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

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

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

for some numerical constant κ.

Theorem (Birgé, Massart, 1998)

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

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

for some positive constants A and M. Let

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

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 $1 - \epsilon$:

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

where κ is a positive constant depending only on M.

Projection estimator

Theorem (Tsybakov, 2003)

Let $\phi_1, \ldots, \phi_d$ 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

- $R(\hat{f}^{(\mathrm{erm})}) - R(f^*) = O\big(\frac{d\log(2+n/d)+\log(\epsilon^{-1})}{n}\big)$ 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 ?

Ridge regression and empirical risk minimization

Theorem

Let $\lambda \geq 0$ *and* $\tilde{f} \in \arg\min_{f_\theta \in \mathcal{F}} \left\{ R(f_\theta) + \lambda \|\theta\|^2 \right\}$.
Assume $\mathbb{E}\left[\|\varphi(X)\|^4\right] < +\infty$ *and* $\sup_{x \in \mathcal{X}} \mathbb{E}\{[Y - \tilde{f}(X)]^2 | X = x\} \leq \sigma^2$.
Let $\nu_1, \ldots, \nu_d$ *be the eigenvalues of* Q*, and*

$$D = \sum_{i=1}^{d} \frac{\nu_i}{\nu_i + \lambda} \mathbf{1}_{\nu_i > 0} = \mathrm{Tr}\left[(Q + \lambda I)^{-1} Q\right] = \mathbb{E}\left\{\|(Q + \lambda I)^{-1/2} \varphi(X)\|^2\right\}.$$

For any $\epsilon > 0$*, there is* n_ϵ *s.t. for any* $n \geq n_\epsilon$*, with proba. at least* $1 - \epsilon$*,*

$$R(\hat{f}_\lambda^{(\text{ridge})}) + \lambda \|\hat{\theta}^{(\text{ridge})}\|^2 \leq \min_{f_\theta \in \mathcal{F}} \left\{ R(f_\theta) + \lambda \|\theta\|^2 \right\} + \sigma^2 \frac{30 D + 1000 \log(3\epsilon^{-1})}{n}.$$

A simple tight risk bound

Theorem

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

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

For an appropriate (randomized) estimator, for any $\epsilon > 0$*, with probability at least* $1 - \epsilon$*, we have*

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

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 $[0, +\infty)$.

Core of the PAC-Bayesian approach

- Let $\chi : \mathcal{F} \to \mathbb{R}$ be an empirical process (for instance: $\chi(f) = R(f) - r(f)$ with $r(f) = \frac{1}{n}\sum_{i=1}^{n}[Y_i - f(X_i)]^2$)

$$\mathbb{E}\exp\left(\sup_{\rho}\left\{\mathbb{E}_{\rho(df)}\chi(f) - K(\rho,\pi')\right\}\right) = \mathbb{E}_{\pi'(df)}\mathbb{E}\exp\left(\chi(f)\right).$$

- Different from the standard approach based on the analysis of $\sup_{f\in\mathcal{F}}\chi(f)$.
- 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
- The bound holds for any prior and posterior
  $\longrightarrow$ different from the usual Bayesian approach

Choice of the empirical process

- Consider $\check{r} : \mathcal{F} \to \mathbb{R}$ be an observable process such that for any $f \in \mathcal{F}$, we have

$$\mathbb{E} \exp\left(\chi(f)\right) \leq 1$$

for $\chi(f) = \lambda[R(f) - \check{r}(f)]$ and some $\lambda > 0$. For instance:

$$\check{r}(f) = -\frac{1}{\lambda} \sum_{i=1}^{n} \log\left(1 - \frac{\lambda}{n}[Y_i - f(X_i)]^2\right).$$

- for any $\epsilon > 0$, with probability at least $1 - \epsilon$, for any distribution ρ on $\mathcal{F}$, we have

$$\mathbb{E}_{\rho(df)} R(f) \leq \mathbb{E}_{\rho(df)} \check{r}(f) + \frac{K(\rho, \pi') + \log(\epsilon^{-1})}{\lambda}.$$

- $\pi'_{-\lambda\check{r}}$ minimizes the righthand-side

The resulting sophisticated PAC-Bayes algorithm

- π uniform distribution on $\mathcal{F}$ (with Θ bounded)
- $\lambda > 0$
- $W_i(f, f') = \frac{\lambda}{n}\Big\{\big[Y_i - f(X_i)\big]^2 - \big[Y_i - f'(X_i)\big]^2\Big\}$
- $\hat{\mathcal{E}}(f) = \log \mathbb{E}_{\pi(df')} \frac{1}{\prod_{i=1}^n [1 - W_i(f,f') + \frac{1}{2} W_i(f,f')^2]}$
- 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

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

PAC-Bayesian localization

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

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

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
- Solution (Catoni, 2003): apply basic bound to $\pi_{-\beta R}$, expand $K(\rho, \pi_{-\beta R})$:

$$K(\rho, \pi_{-\beta R}) = K(\rho, \pi) + \log \left(\int \pi(df) \exp[-\beta R(f)] \right) + \beta \int \rho(df) R(f),$$

and develop additional empirical bounds to control the non observable terms

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

For linear least squares

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

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

- Let $0<\lambda<(2\sigma+H)^{-2}$, $\eta=\lambda(2\sigma+H)^2$, and $\epsilon>0$
- Let $\mathcal{I}(\beta) = -\log \mathbb{E}_{\pi(df)} \exp\big\{-\beta[R(f)-R(f^*)]\big\}$
- For $0\le\gamma\le\lambda n(1-\eta)$, with proba. at least $1-\epsilon$,

$$[\lambda n(1-\eta)-\gamma][R(\hat{f})-R(f^*)] \le 2\mathcal{I}\big(\lambda n(1+\eta)\big)-2\mathcal{I}\big(\gamma\big)+2\log(2\epsilon^{-1}),$$

- Without vs with localization: $\gamma=0$ vs $\gamma=Cn$
  $\mathcal{I}\big(Cn\big)\approx d\log n$ vs $\mathcal{I}\big(\beta n\big)-\mathcal{I}\big(\alpha n\big)\approx d\log(\beta/\alpha)$.

Conclusion

- For any $\epsilon > 0$, with probability at least $1 - \epsilon$, for any distribution ρ on $\mathcal{F}$, we have

$$\mathbb{E}_{\rho(df)} R(f) \leq -\frac{1}{\lambda} \mathbb{E}_{\rho(df)} \sum_{i=1}^{n} \log\left(1 - \frac{\lambda}{n}[Y_i - f(X_i)]^2\right) + \frac{K(\rho, \pi') + \log(\epsilon^{-1})}{\lambda}.$$

- Main result: $\frac{d}{n}$ convergence rate in deviations under minimal moment assumption
- Key tools:
  localized PAC-Bayesian bounds + soft truncation