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

May 2009

- Linear aggregation: $\mathcal{F} = \mathcal{F}_{\text{lin}} = \text{span}\{\varphi_1, \dots, \varphi_d\}$ 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 $\sup_{x \in \mathcal{X}} \text{Var}\left(Y | X = x\right) = \sigma^2 < +\infty$ 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*

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

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

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

for some numerical constant κ*.*

Theorem (Catoni, 2004)

Let $\mathcal{F}' \subset \mathcal{F}_{\text{lin}}$ satisfying for some positive constants a, M, M':

- *there exists $f_0 \in \mathcal{F}'$ s.t. for any $x \in \mathcal{X}$,*

$$\mathbb{E}\Big\{\exp\Big[a\big|Y - f_0(X)\big|\Big] \,\Big|\, X = x\Big\} \leq M.$$

- *for any $f_1, f_2 \in \mathcal{F}'$, $\sup_{x\in\mathcal{X}} |f_1(x) - f_2(x)| \leq M'$.*

Let $Q = \mathbb{E}\big[\varphi(X)\varphi(X)^T\big]$ and $\hat{Q} = \big[\frac{1}{n}\sum_{i=1}^n \varphi(X_i)\varphi(X_i)^T\big]$. If $\det Q \neq 0$, then there exist positive constants C_1 and C_2 s.t. with probability at least $1-\epsilon$, as soon as

$$\left\{ f \in \mathcal{F}_{\text{lin}} : r(f) \leq r(\hat{f}^{(\text{ols})}) + C_1 \frac{d}{n} \right\} \subset \mathcal{F}',$$

we have

$$R(\hat{f}^{(\text{ols})}) - R(f^*_{\text{lin}}) \leq C_2 \frac{d + \log(\epsilon^{-1}) + \log(\frac{\det \hat{Q}}{\det Q})}{n}.$$

Theorem (Alquier, 2008)

Let $q_{\min}$ be the smallest eigenvalue of $Q = \mathbb{E}\left[\varphi(X)\varphi(X)^T\right]$.
Let f_0 and H such that $\|f^*_{\text{lin}} - f_0\|_\infty \leq H$.
Assume that there exists $C > 0$ such that $|Y| \leq C$.
Then for an appropriate randomized estimator requiring the knowledge of f_0, H and C, for any $\epsilon > 0$ with probability at least $1 - \epsilon$, we have

$$R(\hat{f}) - R(f^*_{\text{lin}}) \leq \kappa(H^2 + C^2)\frac{d\log(3q_{\min}^{-1}) + \log(\epsilon^{-1})}{n}.$$

Theorem (Bartlett, Bousquet, Mendelson, 2005)

Assume that for some positive constants H and C,

$$\sup_{\theta \in \Theta} \|\theta\| \leq 1,$$

$$\|\varphi(x)\| \leq H, \qquad \forall x \in \mathcal{X}$$

$$|Y| \leq C.$$

Let $\nu_1 \geq \cdots \geq \nu_d$ be the eigenvalues of $Q = \mathbb{E}\left[\varphi(X)\varphi(X)^T\right]$. With probability at least $1 - \epsilon$, we have

$$R(\hat{f}^{(\mathrm{erm})}) - R(f^*) \leq \kappa (H+C)^2 \frac{\min_{0 \leq h \leq d} \left(h + \sqrt{\frac{n}{(H+C)^2} \sum_{i>h} \nu_i}\right) + \log(\epsilon^{-1})}{n}$$

$$\leq \kappa (H+C)^2 \frac{d + \log(\epsilon^{-1})}{n},$$

where κ is a numerical constant.

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

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

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 $1 - \epsilon$:

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

where κ is a positive constant depending only on M.

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

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

Theorem (Caponnetto, De Vito, 2007)

$$\hat{f}_\lambda^{(\text{ridge})} \in \underset{\{f_\theta;\, \theta \in \mathbb{R}^d\}}{\textit{argmin}} \ \frac{1}{n}\sum_{i=1}^{n}[Y_i - f_\theta(X_i)]^2 + \lambda\|\theta\|^2.$$

Let $q_{\min}$ be the smallest eigenvalue of $Q = \mathbb{E}\left[\varphi(X)\varphi(X)^T\right]$.
Let $\mathcal{K} = \sup_{x\in\mathcal{X}} \sum_{j=1}^{d} \varphi_j(x)^2 = \|\|\varphi\|^2\|_\infty$.
Recall $f_{\text{lin}}^* = \sum_{j=1}^{d} \theta_j^* \varphi_j$. Let $0 < \epsilon < 1/2$ and $\mathcal{L}_\epsilon = \log^2(\epsilon^{-1})$.
Assume that for any $x \in \mathcal{X}$,

$$\mathbb{E}\left(e^{|Y - f_{\text{lin}}^*(X)|/A} | X = x\right) \le M.$$

For $\lambda = (\mathcal{K} d \mathcal{L}_\epsilon)/n$, if $\lambda \le q_{\min}$, with probability at least $1-\epsilon$:

$$R(\hat{f}_\lambda^{(\text{ridge})}) - R(f_{\text{lin}}^*) \le \kappa \mathcal{L}_\epsilon \frac{d}{n}\left(A^2 + \frac{\lambda}{q_{\min}}\mathcal{K}\mathcal{L}_\epsilon\|\theta^*\|^2\right)$$

for some positive constant κ depending only on M.

$$\hat{f}_\lambda^{(\text{lasso})} \in \underset{\{f_\theta;\, \theta \in \mathbb{R}^d\}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - f_\theta(X_i)\big)^2 + \lambda \|\theta\|_1.$$

- As the L^2 penalty, the L^1 penalty shrinks the coefficients.
- It allows to select relevant variables (i.e., find the j's such that $\theta_j^* \neq 0$).
- Assume that $f^{(\text{reg})}$ is a linear combination of only $d^* \ll d$ variables/functions φ_j's, then *under strong conditions on the eigenvalues of submatrices of Q*, the risk of the Lasso estimator for λ of order $\sqrt{(\log d)/n}$ is of order $(d^* \log d)/n$.
- From a model selection approach, the assumptions can be weakened.

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

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* $\mathbb{E}\left\{ \|\varphi(X)\|^2 \left[\tilde{f}(X) - Y\right]^2 \right\} < +\infty$.
Let $\nu_1, \ldots, \nu_d$ *be the eigenvalues of* Q*, and* $Q_\lambda = Q + \lambda I$*. Let*

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

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

$$R(\hat{f}_\lambda^{(\mathrm{ridge})}) + \lambda \|\hat{\theta}^{(\mathrm{ridge})}\|^2 \leq \min_{f_\theta \in \mathcal{F}} \left\{ R(f_\theta) + \lambda \|\theta\|^2 \right\}$$
$$+ \frac{30\, \mathbb{E}\left\{ \|Q_\lambda^{-1/2} \varphi(X)\|^2 \left[\tilde{f}(X) - Y\right]^2 \right\}}{\mathbb{E}\left\{ \|Q_\lambda^{-1/2} \varphi(X)\|^2 \right\}} \frac{D}{n}$$
$$+ 1000 \sup_{v \in \mathbb{R}^d} \frac{\mathbb{E}\left[\langle v, \varphi(X) \rangle^2 \left[\tilde{f}(X) - Y\right]^2 \right]}{\mathbb{E}(\langle v, \varphi(X) \rangle^2) + \lambda \|v\|^2} \frac{\log(3\epsilon^{-1})}{n}.$$

Corollary

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

$$R(\hat{f}_\lambda^{(\text{ridge})}) \leq R(f^*_{\text{lin}}) + \lambda \|\theta^*\|^2 + \operatorname{ess\,sup} \mathbb{E}\{[Y - \tilde{f}(X)]^2 | X\} \frac{30D + 1000 \log(3\epsilon^{-1})}{n}$$

$$D = \sum_{i=1}^{d} \frac{\nu_i}{\nu_i + \lambda} \mathbf{1}_{\nu_i > 0} = \operatorname{Tr}\left[(Q + \lambda I)^{-1} Q\right] = \text{ effective ridge dimension}$$

Theorem

Let $d' = \mathrm{rank}(Q)$*. Assume* $\mathbb{E}\{[Y - f^*(X)]^4\} < +\infty$ *and*

$$B = \sup_{f \in \mathrm{span}\{\varphi_1,\dots,\varphi_d\} - \{0\}} \|f\|_\infty^2 / \mathbb{E}[f(X)^2] < +\infty.$$

Consider the (unique) function $\hat{f}^{(\mathrm{erm})} : x \mapsto \langle \hat{\theta}^{(\mathrm{erm})}, \varphi(x) \rangle$ *on* $\mathcal{F}$ *for which* $\hat{\theta}^{(\mathrm{erm})} \in \mathrm{span}\{\varphi(X_1), \dots, \varphi(X_n)\}$*.*
For any values of ϵ *and* n *such that* $2/n \le \epsilon \le 1$ *and*

$$n > 1280 B^2 \left[3Bd' + \log(2\epsilon^{-1}) + \frac{16B^2 d'^2}{n} \right],$$

with probability at least $1 - \epsilon$*,*

$$R(\hat{f}^{(\mathrm{erm})}) - R(f^*) \le 1920\, B \sqrt{\mathbb{E}[Y - f^*(X)]^4} \left[\frac{3Bd' + \log(2\epsilon^{-1})}{n} + \left(\frac{4Bd'}{n} \right)^2 \right].$$

Let

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

$\hat{\mathcal{E}}(f) \approx \lambda \sum_{i=1}^n [Y_i - f(X_i)]^2 + \log \int \pi(df') \exp\Big\{-\lambda \sum_{i=1}^n \big[Y_i - f'(X_i)\big]^2\Big\}$,
We consider the “posterior” distribution $\hat{\pi}$ on the set $\mathcal{F}$ with density:

$$\frac{d\hat{\pi}}{d\pi}(f) = \frac{\exp[-\hat{\mathcal{E}}(f)]}{\int \exp[-\hat{\mathcal{E}}(f')]\pi(df')}.$$

$$\frac{d\hat{\pi}}{d\pi}(f) \approx \frac{\exp\{-\lambda \sum_{i=1}^n [Y_i - f(X_i)]^2\}}{\int \exp\{-\lambda \sum_{i=1}^n [Y_i - f'(X_i)]^2\}\pi(df')}.$$

Theorem

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

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

Let $\lambda = \frac{1}{3(2\sigma+H)^2}$ *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*

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