Dimension-free PAC-Bayesian bounds

Olivier Catoni
Olivier.Catoni@ensae.fr
http://ocatoni.perso.math.cnrs.fr/

CREST, CNRS — UMR 9194
Université Paris–Saclay

Nips Workshop 2017
(Almost) 50 shades of Bayesian Learning: PAC-Bayesian trends and insights

December 9, 2017

Joint work with
Ilaria Giulini

Laboratoire de Probabilités et Modèles Aléatoires
Université Paris Diderot
giulini@math.univ-paris-diderot.fr

Topics to be covered

Three related questions

1. Given $X_1, \ldots, X_n$, n independent copies of the r. v. $X \in \mathbb{R}^d$, estimate $\mathbb{E}(X)$?
2. Given $M_1, \ldots, M_n$, n independent copies of the random matrix $M \in \mathbb{R}^{p \times q}$, estimate $\mathbb{E}(M)$ in operator norm ?
3. Given $(X_1, Y_1), \ldots, (X_n, Y_n)$, n independent copies of the couple of random variables $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$, estimate $\arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}[(Y - \langle \theta, X \rangle)^2]$?

Dimension-free assumptions

In case

1. $\mathbb{E}(\|X\|^2) < \infty$,
2. $\mathbb{E}(\|M\|_{\mathrm{HS}}^2) < \infty$,
3. $\mathbb{E}(\|X\|^4) < \infty$ and $\mathbb{E}(\|X\|^2 Y^2) < \infty$.

Approach

Directional estimates

1. Estimate $\mathbb{E}(\langle \theta, X \rangle)$ for any $\theta \in \mathbb{S}_d = \{\theta \in \mathbb{R}^d \,:\, \|\theta\| = 1\}$.
2. Estimate $\mathbb{E}(\langle \xi, M\theta \rangle)$ for any $\theta \in \mathbb{S}_q$ and any $\xi \in \mathbb{S}_p$.
3.

$$\begin{aligned} \text{Put } R(\theta) &= \langle \theta, \mathbb{E}(XX^\top)\theta \rangle - 2\langle \theta, \mathbb{E}(YX) \rangle \\ &= \mathbb{E}[(Y - \langle \theta, X \rangle)^2] - \mathbb{E}(Y^2) \end{aligned}$$

and estimate the Gram matrix $\mathbb{E}(XX^\top)$ and the mean vector $\mathbb{E}(YX)$.

PAC-Bayesian bound

General purpose inequality

For any prior probability measure $\mu \in \mathcal{M}_+^1(\mathbb{R}^d)$, for any $\delta \in [0,1]$, with probability at least $1-\delta$, for any $\rho \in \mathcal{M}_+^1(\mathbb{R}^d)$,

$$\int \frac{1}{n}\sum_{i=1}^{n} f(\theta', X_i)\,\mathrm{d}\rho(\theta') \leq \int \log\Big\{\mathbb{E}\Big[\exp(f(\theta', X))\Big]\Big\}\,\mathrm{d}\rho(\theta') + \frac{\mathcal{K}(\rho,\mu) + \log(\delta^{-1})}{n},$$

where $$\mathcal{K}(\rho,\mu) = \begin{cases} \int \log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\mu}\right)\mathrm{d}\rho, & \rho \ll \mu, \\ +\infty, & \text{otherwise.} \end{cases}$$

PAC-Bayesian bound

Special choices : Gaussian posteriors for non-Gaussian data

- Of ρ and μ: $\rho_\theta = \mathcal{N}(\theta, \beta^{-1} I_d)$, $\theta \in \mathbb{R}^d$, and $\mu = \rho_0$. Remark that $\mathcal{K}(\rho, \mu) = \dfrac{\beta}{2} \|\theta\|^2$ (complexity measured by parameter norm, independently of its linear dimension d).
- Of f: most obvious choice is $f(\theta', X_i) = \lambda \langle \theta', X_i \rangle$, but leads to hypotheses on exponential moments $\mathbb{E}[\exp(\lambda \langle \theta', X \rangle)]$. Rather use an influence function $f(\theta', X_i) = \psi(\lambda \langle \theta', X_i \rangle)$, where
  - $-\log(1 - t + t^2/2) \leq \psi(t) \leq \log(1 + t + t^2/2)$,
  - $\psi(-t) = -\psi(t)$,
  - ψ is bounded,
  - $\displaystyle\int \psi(\lambda \langle \theta', X_i \rangle) \, \mathrm{d}\rho_\theta(\theta')$ can be computed.

An influence function that ticks all the boxes

Let us compute !

Choose $\psi(t) = \begin{cases} t - t^3/6, & -\sqrt{2} \leq t \leq \sqrt{2}, \\ 2\sqrt{2}/3, & t > \sqrt{2}, \\ -2\sqrt{2}/3, & t < -\sqrt{2}. \end{cases}$

Introduce $\varphi(m,\sigma) = \mathbb{E}[\psi(m + \sigma W)]$, where $W \sim \mathcal{N}(0,1)$ is a standard normal. It can be computed from the normal distribution function $F(a) = \mathbb{P}(W \leq a)$ as $\varphi(m,\sigma) = m(1 - \sigma^2/2) - m^3/6 + r(m,\sigma)$, where

$$
\begin{aligned}
r(m,\sigma) = {} & \frac{2\sqrt{2}}{3}\left[F\left(\frac{-\sqrt{2}+m}{\sigma}\right) - F\left(\frac{-\sqrt{2}-m}{\sigma}\right)\right] - (m - m^3/6)\left[F\left(\frac{-\sqrt{2}+m}{\sigma}\right) + F\left(\frac{-\sqrt{2}-m}{\sigma}\right)\right] \\
& +\sigma\,\frac{(1-m^2/2)}{\sqrt{2\pi}}\left[\exp\left(-\frac{1}{2}\left(\frac{\sqrt{2}+m}{\sigma}\right)^2\right) - \exp\left(-\frac{1}{2}\left(\frac{\sqrt{2}-m}{\sigma}\right)^2\right)\right] + \frac{m\sigma^2}{2}\left\{F\left(\frac{-\sqrt{2}-m}{\sigma}\right) + F\left(\frac{-\sqrt{2}+m}{\sigma}\right)\right. \\
& \left. + \frac{1}{\sqrt{2\pi}}\left[\frac{(\sqrt{2}+m)}{\sigma}\exp\left[-\frac{1}{2}\left(\frac{\sqrt{2}+m}{\sigma}\right)^2\right] + \frac{(\sqrt{2}-m)}{\sigma}\exp\left[-\frac{1}{2}\left(\frac{\sqrt{2}-m}{\sigma}\right)^2\right]\right]\right\} \\
& + \frac{\sigma^3}{6\sqrt{2\pi}}\left\{\left[\left(\frac{\sqrt{2}-m}{\sigma}\right)^2 + 2\right]\exp\left[-\frac{1}{2}\left(\frac{\sqrt{2}-m}{\sigma}\right)^2\right] - \left[\left(\frac{\sqrt{2}+m}{\sigma}\right)^2 + 2\right]\exp\left[-\frac{1}{2}\left(\frac{\sqrt{2}+m}{\sigma}\right)^2\right]\right\}.
\end{aligned}
$$

An influence function that ticks all the boxes

Computing Gaussian perturbations

$$\int \psi\big(\lambda\langle\theta', X_i\rangle\big)\,\mathrm{d}\rho_{\theta}(\theta') = \varphi\left(\lambda\langle\theta, X_i\rangle, \frac{\lambda\|X\|}{\sqrt{\beta}}\right).$$

Turning exponentials into polynomials

$$\begin{aligned}
&\int \log\Big\{\mathbb{E}\Big[\exp\Big(\psi\big(\lambda\langle\theta', X\rangle\big)\Big)\Big]\Big\}\,\mathrm{d}\rho_{\theta}(\theta') \\
&\qquad \leq \int \log\Big[1 + \lambda\mathbb{E}(\langle\theta', X\rangle + \frac{\lambda^2}{2}\mathbb{E}(\langle\theta', X\rangle^2)\Big]\,\mathrm{d}\rho_{\theta}(\theta') \\
&\qquad\qquad \leq \lambda\mathbb{E}(\langle\theta, X\rangle) + \frac{\lambda^2}{2}\mathbb{E}(\langle\theta, X\rangle^2) + \frac{\lambda^2\mathbb{E}(\|X\|^2)}{\beta}.
\end{aligned}$$

The job can be done !

At this stage, we see that

- We will be able to estimate $\langle \theta, \mathbb{E}(X) \rangle$.
- The bound will not involve the dimension d but only the two moments

$$\sup_{\theta \in \mathbb{S}_d} \mathbb{E}(\langle \theta, X \rangle^2) \leq \mathbb{E}(\|X\|^2).$$

Putting things together

PAC-Bayesian inequality

With probability at least $1-\delta$, for any $\theta \in \mathbb{S}_d$,

$$\mathcal{E}(\theta) = \frac{1}{n\lambda}\sum_{i=1}^{n} \varphi(\lambda\langle\theta, X_i\rangle, \lambda\|X_i\|/\sqrt{\beta})$$

$$\leq \langle\theta, \mathbb{E}(X)\rangle + \frac{\lambda}{2}\left[\mathbb{E}(\langle\theta, X\rangle^2) + \frac{\mathbb{E}(\|X\|^2)}{\beta}\right] + \frac{\beta + 2\log(\delta^{-1})}{2n\lambda}.$$

Assumptions and optimized choices

- Assume that $\mathbb{E}(\|X\|^2) \leq T < \infty$ and $\sup_{\theta\in\mathbb{S}_d} \mathbb{E}(\langle\theta, X\rangle^2) \leq v \leq T < \infty$, where v and T are known.
- Choose $\lambda = \sqrt{\frac{2\log(\delta^{-1})}{nv}}$ and $\beta = \sqrt{\frac{2T\log(\delta^{-1})}{v}}$.

Putting things together

Non asymptotic confidence region

With probability at least $1 - \delta$,

$$\sup_{\theta \in \mathbb{S}_d} \left| \mathcal{E}(\theta) - \langle \theta, \mathbb{E}(X) \rangle \right| \leq \sqrt{\frac{T}{n}} + \sqrt{\frac{2v \log(\delta^{-1})}{n}}.$$

For comparison, when X is a Gaussian vector,

$$\left\| \tfrac{1}{n} \textstyle\sum_{i=1}^n X_i - \mathbb{E}(X) \right\| \leq$$

$$\sqrt{\frac{\mathbb{E}\big(\| X - \mathbb{E}(X) \|^2 \big)}{n}} + \sqrt{\frac{2 \sup_{\theta \in \mathbb{S}_d} \mathbb{E}\big(\langle \theta, X - \mathbb{E}(X) \rangle^2 \big) \log(\delta^{-1})}{n}}$$

We have lost

- Centering in the definition of T and v,
- adaptivity in T and v,
- and the c. r. is no more a ball, but the cts are the same !

Putting things together

Estimator

On the event of probability at least $1 - \delta$ defined by the PAC-Bayesian inequality, we can find $\widehat{m} = \widehat{m}_{v,T}(X_1, \dots, X_n) \in \mathbb{R}^d$ such that

$$\sup_{\theta \in \mathbb{S}_d} \left| \mathcal{E}(\theta) - \langle \theta, \widehat{m} \rangle \right| \leq \sqrt{\frac{T}{n}} + \sqrt{\frac{2v \log(\delta^{-1})}{n}},$$

and therefore such that

$$\| \widehat{m} - \mathbb{E}(X) \| \leq 2 \left(\sqrt{\frac{T}{n}} + \sqrt{\frac{2v \log(\delta^{-1})}{n}} \right).$$

(The constant 2 can be lowered to $\sqrt{3}$ by setting $\widehat{m}$ to the middle of a diameter of the confidence region.)

Centering through sample splitting

Assumptions

Put $m = \mathbb{E}(X)$ and assume that for known b, v' and T',
$\|m\|^2 \leq b < \infty$, $\mathbb{E}(\|X - m\|^2) \leq T' < \infty$, and
$\sup_{\theta \in \mathbb{S}_d} \mathbb{E}(\langle \theta, X - m \rangle^2) \leq v' \leq T' < \infty$.

Sample splitting. Put

- $\widetilde{m}_1 = \widehat{m}_{v'+b,\, T'+b}(X_1, \dots, X_k)$
- and $\widetilde{m}_2 = \widehat{m}_{v'+A/k,\, T'+A/k}(X_{k+1} - \widetilde{m}_1, \dots, X_n - \widetilde{m}_1)$, where $A = 4\Big(\sqrt{T' + b} + \sqrt{2(v' + b)\log(\delta^{-1})}\Big)^2$.
- With probability at least $1 - 2\delta$, $\|\widetilde{m}_1 - m\|^2 \leq A/k$ and $\|\widetilde{m}_2 - m\| \leq 2\left(\sqrt{\frac{T'+A/k}{n-k}} + \sqrt{\frac{2(v'+A/k)\log(\delta^{-1})}{n-k}}\right)$

$$\underset{\substack{n\to\infty, k\to\infty \\ k/n\to 0}}{\sim} 2\left(\sqrt{\frac{T'}{n}} + \sqrt{\frac{2v'\log(\delta^{-1})}{n}}\right).$$

Mean matrix estimate

Estimator

- $(M_1, \dots, M_n)$ n independent copies of $M \in \mathbb{R}^{p\times q}$.
- Put $\nu_\xi = \mathcal{N}(\xi, \gamma^{-1} I_p)$, $\xi \in \mathbb{R}^p$ and $\rho = \mathcal{N}(\theta, \beta^{-1} I_q)$, $\theta \in \mathbb{R}^q$.
- Define $\mathcal{E}(\xi,\theta) = \dfrac{1}{\lambda n} \displaystyle\sum_{i=1}^n \psi\Big(\lambda \langle \xi', M_i \theta' \rangle\Big) \, \mathrm{d}\nu_\xi(\xi') \, \mathrm{d}\rho_\theta(\theta')$

Let us compute ! Consider $W_q \sim \mathcal{N}(0, I_q) \in \mathbb{R}^q$.

$$\begin{aligned}
\mathcal{E}(\xi,\theta) = \frac{1}{n} \sum_{i=1}^n & \langle \xi, M_i \theta \rangle - \frac{\lambda^2}{6} \langle \xi, M_i \theta \rangle^3 \\
& - \frac{\lambda^2}{2\beta} \langle \xi, M_i \theta \rangle \| M_i \theta \|^2 - \frac{\lambda^2}{2\gamma} \langle \xi, M_i \theta \rangle \| M_i^\top \xi \|^2 \\
& - \frac{\lambda^2}{2\beta\gamma} \langle \xi, M_i \theta \rangle \| M_i \|_{\mathrm{HS}}^2 - \frac{\lambda^2}{\beta\gamma} \langle \xi, M_i M_i^\top M_i \theta \rangle \\
+ \frac{1}{\lambda} \mathbb{E}\Big[& r\Big(\lambda \langle M_i^\top \xi, \theta + \gamma^{-1/2} W_q \rangle, \lambda \beta^{-1/2} \| M_i (\theta + \gamma^{-1/2} W_q) \| \Big) \Big].
\end{aligned}$$

PAC-Bayesian inequality

With probability at least $1 - \delta$, for any $(\xi, \theta) \in \mathbb{R}^p \times \mathbb{R}^q$,

$$
\begin{aligned}
\mathcal{E}(\xi,\theta) \le \lambda^{-1} \int \log\Big\{\mathbb{E}\Big[\exp\Big(\psi\big(\lambda\langle \xi', M\theta'\rangle\big)\Big)\Big]\Big\}\, \mathrm{d}\nu_\xi(\xi')\mathrm{d}\rho_\theta(\theta') \\
+ \frac{\mathcal{K}(\nu_\xi, \nu_0)}{n\lambda} + \frac{\mathcal{K}(\rho_\theta, \rho_0)}{n\lambda} + \frac{\log(\delta^{-1})}{n\lambda} \\
\le \mathbb{E}(\langle \xi, M\theta\rangle) + \frac{\lambda}{2}\Bigg[\mathbb{E}(\langle \xi, M\theta\rangle^2) + \frac{\mathbb{E}(\|M\theta\|^2)}{\beta} \\
+ \frac{\mathbb{E}(\|M^\top \xi\|^2)}{\gamma} + \frac{\mathbb{E}(\|M\|_{\mathrm{HS}}^2)}{\beta\gamma}\Bigg] + \frac{\beta + \gamma + 2\log(\delta^{-1})}{2n\lambda}.
\end{aligned}
$$

Confidence region

Assumptions: for known v, t, u, T

$$\mathbb{E}(\|M\|_{\mathrm{HS}}^2) \le T < \infty, \qquad \|\mathbb{E}(M^\top M)\|_\infty \le t \le T < \infty,$$
$$\|\mathbb{E}(MM^\top)\|_\infty \le u \le T < \infty, \qquad \sup_{\xi \in \mathbb{S}_p, \theta \in \mathbb{S}_q} \mathbb{E}(\langle \xi, M\theta \rangle^2) \le v < \infty.$$

Choices

$$\lambda = \sqrt{\frac{\beta + \gamma + 2\log(\delta^{-1})}{n(v + t/\beta + u/\gamma + T/(\beta\gamma))}}, \ \beta = \gamma = 2\max\left\{\frac{t+u}{v}, \sqrt{\frac{T}{v}}\right\}$$

Confidence region: with probability at least $1 - \delta$,

$$\sup_{\xi \in \mathbb{S}_p, \xi \in \mathbb{S}_q} \left| \mathcal{E}(\xi, \theta) - \langle \xi, \mathbb{E}(M)\theta \rangle \right|$$
$$\le \sqrt{\frac{2v}{n}\left(2\log(\delta^{-1}) + 4\max\left\{\frac{t+u}{v}, \sqrt{\frac{T}{v}}\right\}\right)}.$$

Mean matrix estimator

With probability at least $1-\delta$,

we can find $\widehat{m}$ within the confidence region and

$$\|\widehat{m}-\mathbb{E}(M)\|_\infty \leq 2\sqrt{\frac{2v}{n}\left(2\log(\delta^{-1})+4\max\left\{\frac{t+u}{v},\sqrt{\frac{T}{v}}\right\}\right)}.$$

Adaptive estimators

Question:

Is it possible to adapt to the values of the constants that were assumed to be known for the previous estimators, because they were used to set their parameters ?

Approach:

- Introduce the asymmetric influence function defined on the positive real line $\psi(t) = \begin{cases} t - t^2/2, & 0 \leq t \leq 1, \\ 1/2, & 1 \leq t, \end{cases}$
- and estimate separately positive and negative parts.

Lemma: For any $t \in \mathbb{R}_+$,

$$-\log(1 - t + t^2) \leq \psi(t) \leq \log(1 + t).$$

Estimating the mean of the positive part

Directional estimator

Consider a discrete set $\Lambda \in \mathbb{R}_+$ and a probability $\mu \in \mathcal{M}_+^1(\Lambda)$. Define

$$\mathcal{E}_+(\theta) = \sup_{\Lambda \in \Lambda} \frac{1}{n\lambda} \sum_{i=1}^{n} \int \psi\big(\lambda \langle \theta', X_i \rangle_+\big) \, \mathrm{d}\rho_{\theta}(\theta') - \frac{\beta + 2\log(\delta^{-1}\mu(\lambda)^{-1})}{2n\lambda}$$

PAC-Bayesian inequality. With probability at least $1 - 2\delta$,

$$\int \mathbb{E}(\langle \theta', X \rangle_+) \, \mathrm{d}\rho_{\theta}(\theta') - \inf_{\lambda \in \Lambda} \left\{ \lambda \int \mathbb{E}(\langle \theta', X \rangle_+^2) \, \mathrm{d}\rho_{\theta}(\theta') + \frac{\beta + 2\log(\delta^{-1}\mu(\lambda)^{-1})}{\lambda n} \right\} \leq \mathcal{E}_+(\theta) \leq \int \mathbb{E}(\langle \theta', X \rangle_+) \, \mathrm{d}\rho_{\theta}(\theta').$$

Putting the positive and the negative parts together

Confidence region

Define $\mathcal{E}(\theta) = \mathcal{E}_+(\theta) - \mathcal{E}_+(-\theta)$. With probability at least $1 - 2\delta$

$$\langle \theta, \mathbb{E}(X) \rangle - \mathcal{E}(\theta) \leq B_+(\theta) = \inf_{\lambda \in \Lambda} \left\{ \lambda \int \mathbb{E}\left(\langle \theta', X \rangle_+^2\right) \mathrm{d}\rho_\theta(\theta') + \frac{\beta + 2 \log(\delta^{-1} \mu(\lambda)^{-1})}{\lambda n} \right\}.$$

Estimator

Define $\widehat{m} \in \arg\min_{m \in \mathbb{R}^d} \sup_{\theta \in \mathbb{S}_d} \{ \langle \theta, m \rangle - \mathcal{E}(\theta) \}$. With probability at least $1 - 2\delta$,

$$\| \widehat{m} - \mathbb{E}(X) \| \leq \inf_{\lambda \in \Lambda} \left\{ 2\lambda \left(\sup_{\theta \in \mathbb{S}_d} \mathbb{E}\left(\langle \theta, X \rangle^2\right) + \frac{\mathbb{E}(\|X\|^2)}{\beta} \right) + \frac{2\beta + 4 \log(\delta^{-1} \mu(\lambda)^{-1})}{n\lambda} \right\}.$$

Let's compute !

Choices

- $\beta = 2\log(\delta^{-1})$,
- $\Lambda = \left\{ \lambda_k = \dfrac{\exp(k)}{\sigma\sqrt{n}}, k \in \mathbb{Z} \right\}$,
- $\mu(\lambda_k) = \frac{1}{2(|k|+1)(|k|+2)}$.

Result

With probability at least $1 - 2\delta$,
$\|\widehat{m} - \mathbb{E}(X)\| \leq 4C\sqrt{2(2v\log(\delta^{-1}) + T)/n}$, where
$v = \sup_{\theta \in \mathbb{S}_d} \mathbb{E}(\langle \theta, X \rangle^2)$, $T = \mathbb{E}(\|X\|^2)$, and

$$C = \cosh\left(\frac{1}{2}\right) + \frac{\exp(1/2)}{2\log(\delta^{-1})} \log\left[\frac{1}{\sqrt{2}} \left| \log\left(\frac{2v\log(\delta^{-1}) + T}{8\sigma^2 \log(\delta^{-1})^2} \right) \right| + \frac{5}{\sqrt{2}} \right] \underset{\text{typically}}{\leq} 2.$$

Adaptive mean matrix estimate

Executive summary

Using $\mathcal{E}_+(\xi,\theta) = \sup_{\lambda \in \Lambda} \biggl\{ \frac{1}{n\lambda} \sum_{i=1}^n \int \psi\bigl[\lambda \langle \xi', M_i \theta' \rangle_+\bigr] \, \mathrm{d}\nu_\xi(\xi') \, \mathrm{d}\rho_\theta(\theta') - \frac{\beta + \gamma + 2\log(\delta^{-1}\mu(\lambda)^{-1})}{2n\lambda} \biggr\}$, we get with probability at least $1 - 2\delta$ that

$$\bigl| \langle \xi, \mathbb{E}(M)\theta \rangle - \mathcal{E}(\xi,\theta) \bigr| \\ \leq C \sqrt{\frac{1+\chi}{n} \biggl(2\log(\delta^{-1}) v + \frac{t+u}{\chi} + \frac{T}{2\log(\delta^{-1})\chi^2} \biggr)},$$

where $v = \mathbb{E}(\langle \xi, M\theta \rangle^2)$, $t = \mathbb{E}(\|M\theta\|^2)$, $u = \mathbb{E}(\|M^\top \xi\|^2)$ and $T = \mathbb{E}(\|M\|_{\mathrm{HS}}^2)$.

Also $$\|\mathbb{E}(M) - \widehat{m}\| \leq 2 \sup_{\xi \in \mathbb{S}_q, \theta \in \mathbb{S}_p} \bigl(\cdots \bigr).$$

Adaptive Gram matrix estimate

Question

Given $X_1, \dots, X_n$, n independent copies of $X \in \mathbb{R}^d$, estimate $G = \mathbb{E}(XX^\top)$? $\rightarrow$ estimate $\mathbb{E}(\langle \theta, X\rangle^2) = \langle \theta, G\theta\rangle$, $\theta \in \mathbb{S}_d$.

PAC-Bayesian bound. With probability at least $1 - 2\delta$,

for any $\theta \in \mathbb{S}_d$,

$$\sup_{\lambda \in \mathbb{R}_+, \beta \in B} \mathbb{E}(\langle \theta, X\rangle^2) - \lambda \mathbb{E}(\langle \theta, X\rangle^4) - \frac{6}{\beta}\mathbb{E}(\|X\|^2\langle \theta, X\rangle^2)$$
$$- \frac{4\mathbb{E}(\|X\|^4)}{\lambda\beta^2} - \frac{\beta}{n} - \frac{2\log(\mu(\beta)^{-1}\delta^{-1})}{n\lambda} \leq \mathcal{E}(\theta)$$
$$\stackrel{\text{def}}{=} \sup_{\lambda \in \mathbb{R}_+, \beta \in B} \frac{1}{n\lambda}\sum_{i=1}^n \left[\int \psi(\langle \theta', X_i\rangle^2)\, \mathrm{d}\rho_{\sqrt{\lambda}\theta}(\theta') - \log\left(1 + \frac{\|X_i\|^2}{\beta}\right)\right]$$
$$- \frac{\beta}{2n} - \frac{\log(\mu(\beta)^{-1}\delta^{-1})}{n\lambda} - \frac{\mathbb{E}(\|X\|^4)}{\lambda\beta^2} \leq \mathbb{E}(\langle \theta, X\rangle^2).$$

$\rightarrow$ we still need a known bound for $\mathbb{E}(\|X\|^4)$, but not for $\mathbb{E}(\langle \theta, X\rangle^4)$.

Confidence region

Assumptions and choices

- $\mathbb{E}(\|X\|^4) \leq T < \infty$,
- $\beta \in \{\beta_k = \sqrt{10\,Tn\exp(-k)} \;:\; k \in \mathbb{N}\}$
- $\mu(\beta_k) = (k+1)^{-1}(k+2)^{-1}$.

With probability at least $1 - 2\delta$, for any $\theta \in \mathbb{S}_d$,

$$\mathcal{E}(\theta) \leq \langle \theta, G\theta \rangle \leq \mathcal{E}(\theta) + B(\theta), \text{ where}$$

$$B(\theta) = 2\sqrt{\frac{\mathbb{E}(\langle \theta, X\rangle^4)}{n}} \left\{ 3.3\left(\frac{T}{\mathbb{E}(\langle \theta, X\rangle^4)}\right)^{1/4} \right.$$

$$\left. + \sqrt{4\log\left(\frac{1}{2}\log\left(\frac{T}{\mathbb{E}(\langle \theta, X\rangle^4)}\right) + \frac{5}{2}\right) + 2\log(\delta^{-1})} \right\}.$$

Gram matrix estimate

Choice

$$\text{Let } \widehat{G} \in \arg\min \Big\{ \sup_{\theta \in \mathbb{S}_d} \langle \theta, M\theta \rangle - \mathcal{E}(\theta) \;:\; M \in \mathbb{R}^{d \times d},$$

$$M = M^\top, 0 \leq \inf_{\theta \in \mathbb{S}_d} \langle \theta, M\theta \rangle - \mathcal{E}(\theta) \Big\}.$$

Estimation error in operator norm

With probability at least $1 - 2\delta$,

$$\| G - \widehat{G} \|_\infty \leq \sup_{\theta \in \mathbb{S}_d} B(\theta).$$

Linear least squares ridge regression

Question

- Given $(X_1, Y_1), \ldots, (X_n, Y_n)$, n independent copies of $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$,
- and optionally a regularization parameter $\lambda \in \mathbb{R}_+$,
- estimate $\arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}[(\langle \theta, X \rangle - Y)^2] + \lambda \|\theta\|^2$.

Linear least squares ridge regression

Approach

- Put $R_\lambda(\theta) = \langle \theta, (G + \lambda I)\theta \rangle - 2\langle \theta, V \rangle$, where $G = \mathbb{E}(XX^\top)$ and $V = \mathbb{E}(YX)$.
- Remark that
  $\arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}[(\langle \theta, X \rangle - Y)^2] + \lambda \|\theta\|^2 = \arg\min_{\theta \in \mathbb{R}^d} R_\lambda(\theta).$
- Assuming that $\mathbb{E}(\|X\|^4) < \infty$ and $\mathbb{E}(Y^2 \|X\|^2) < \infty$,
- compute estimators $\widehat{G}$ and $\widehat{V}$ such that with probability at least $1 - \delta$,
  $$\|G - \widehat{G}\|_\infty \leq \epsilon = \mathcal{O}\left(\sqrt{\frac{\log(\delta^{-1})}{n}} \right)$$
  $$\text{and } \|V - \widehat{V}\| \leq \eta = \mathcal{O}\left(\sqrt{\frac{\log(\delta^{-1})}{n}} \right).$$
- Consider
  $$\widehat{R}_\lambda(\theta) = \langle \theta, (\widehat{G} + \lambda I)\theta \rangle - 2\langle \theta, \widehat{V} \rangle.$$

Regression on a compact parameter set

Slow rate

Let $\Theta \subset \mathbb{R}^d$ be a compact subset. Consider any $\widehat{\theta} \in \arg\min_{\theta \in \Theta} \widehat{R}_\lambda$. With probability at least $1 - \delta$,

$$R_\lambda(\widehat{\theta}) - \inf_{\Theta} R_\lambda \le 2\|\Theta\|(\epsilon\|\Theta\| + 2\eta) = \mathcal{O}\left(\sqrt{\frac{\log(\delta^{-1})}{n}} \right),$$

where $\|\Theta\| = \sup\left\{ \|\theta\| \,:\, \theta \in \Theta \right\}$.

Confidence region

Approach

Remark that with probability at least $1-\delta$, for any $\theta,\xi \in \mathbb{R}^d$,

$$R_\lambda(\xi) - R_\lambda(\theta) \leq \gamma(\theta,\xi) \overset{\text{def}}{=} \widehat{R}_\lambda(\xi) - \widehat{R}_\lambda(\theta) \\ + \epsilon\|\xi-\theta\|^2 + 2\|\xi-\theta\|(\epsilon\|\theta\|+\eta),$$

so that the subdifferential of γ defines the confidence region

$$0 \in \frac{\partial}{\partial\xi}_{|\xi=\theta_\lambda} \gamma(\theta_\lambda,\xi), \qquad \text{where } \theta_\lambda \in \arg\min_{\theta\in\mathbb{R}^d} R_\lambda(\theta).$$

Doing the computations

Consider $\widehat{\theta}_\lambda \in \arg\min_{\theta\in\mathbb{R}^d} \widehat{R}_\lambda$. With probability at least $1-\delta$,

$$\theta_\lambda \in \widehat{\Theta}_\lambda = \left\{\theta \in \mathbb{R}^d \,:\, \|(\widehat{G}+\lambda I)(\theta-\widehat{\theta}_\lambda)\| \leq \|\theta\|\epsilon + \eta\right\}.$$

Estimator, fast rate and slow rate

Choice of an estimator

$$\widetilde{\theta} \in \arg\min\Big\{\|\theta\| \ : \ \theta \in \widehat{\Theta}_\lambda\Big\}.$$

Fast rate. With probability at least $1-\delta$

$$\|(G+\lambda I)(\widetilde{\theta}-\theta_\lambda)\|^2 \leq 4(\epsilon\|\theta_\lambda\|+\eta)^2.$$

Slow rate. With probability at least $1-\delta$

$$R_\lambda(\widetilde{\theta}) - R_\lambda(\theta_\lambda) = \langle \widetilde{\theta}-\theta_\lambda, (G+\lambda I)(\widetilde{\theta}-\theta_\lambda)\rangle \leq \frac{4}{\sigma_d+\lambda}\Big(\epsilon\|\theta_\lambda\|+\eta\Big)^2,$$

where σ_d is the smallest eigenvalue of G. For small values of λ and σ_d, the following bound is meaningful.

$$R_0(\widetilde{\theta}_\lambda) - R_0(\theta_0) \leq (\|\theta_0\|+1/2)\big((2\epsilon+\eta)\|\theta_0\|+\eta\big).$$

Sparse recovery

Sparse submodels

- Let $\mathcal{L}$ be a family of linear subspaces of $\mathbb{R}^d$.
- Assume that $\|\theta_\lambda\| \le A < \infty$, where A is known.
- Consider the global confidence region

$$\widehat{\Theta}_\lambda = \left\{\theta \in \mathbb{R}^d \,:\, \|(\widehat{G} + \lambda I)(\theta - \widehat{\theta}_\lambda)\| \le \epsilon\|\theta\| + \eta, \|\theta\| \le A\right\}.$$

Model selector

- Put $\widehat{\mathcal{L}} = \left\{L \in \mathcal{L} \,:\, L \cap \widehat{\Theta}_\lambda \neq \varnothing\right\}$.
- Choose $\widehat{L} = \arg\max\left\{\widehat{\sigma}_L \,:\, L \in \widehat{\mathcal{L}}\right\}$, where $\widehat{\sigma}_L = \inf_{\theta \in \mathbb{S}_d \cap L}\|\widehat{G}\theta\|$.
- Define $\widetilde{\theta} \in \arg\min\left\{\|\theta\| \,:\, \theta \in \widehat{L} \cap \widehat{\Theta}_\lambda\right\}$.

Sparse recovery

Sparse convergence rate

- Assume that $\theta_\lambda \in L_*$.
- Define $\sigma_* = \inf\left\{\sigma_{L+\mathbb{R}\theta_\lambda} \;:\; L \in \mathcal{L}, \sigma_L \geq \sigma_{L_*} - 2\epsilon\right\}$, where $\sigma_L = \inf_{\theta \in \mathbb{S}_d \cap L} \|G\theta\|$.
- With probability at least $1 - \delta$,

$$(\sigma_* + \lambda)\|\widehat{\theta} - \theta_\lambda\| \leq \|(G + \lambda I)(\widehat{\theta} - \theta_\lambda)\| \leq 2(\epsilon A + \eta)$$

and

$$R_\lambda(\widehat{\theta}) - R_\lambda(\theta_\lambda) \leq \frac{4}{\lambda + \sigma_*}(\epsilon A + \eta)^2.$$

Nested models

In the case when $\mathcal{L} = \left\{L_1 \subset L_2 \subset \cdots \subset L_K\right\}$, we can take $\sigma_* = \sigma_{L_*}$.