New PAC-Bayesian bounds for k-means algorithms

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

CREST, CNRS — UMR 9194
Université Paris–Saclay

Three-day meeting of statisticians in Paris
I.H.P., July 18-20 2022

Joint work with Gautier Appert

The k-means criterion

Consider a random variable $X \in H$, where H is a separable Hilbert space and the quantization problem

$$\inf_{c \in H^k} \mathbb{P}_X \Big(\min_{j \in [\![1, k]\!]} \|X - c_j\|^2 \Big).$$

Given a sample $\overline{X} = (X_1, \dots, X_n)$ made of n independent copies of X, we want an estimator $\widehat{c}(\overline{X}) \in H^k$ such that

$$\mathbb{P}_X \Big(\min_{j \in [\![1, k]\!]} \|X - \widehat{c}_j\|^2 \Big).$$

is small. Since it is a r. v. we can bound either its mean or its deviations. The aim of this talk is to present a series of ideas that lead to new bounds and new estimators.

A min-linear criterion

The first thing we propose is to rewrite the criterion as a min-linear problem.

$$\begin{aligned}\min_{j\in[\![1,k]\!]} \|X - c_j\|^2 &= \min_{j\in[\![1,k]\!]} \|X\|^2 + \|c_j\|^2 - 2\langle X, c_j\rangle \\ &= \min_{j\in[\![1,k]\!]} \|W_1\|^2/4 + \langle \theta_j, W\rangle,\end{aligned}$$

where $\theta_j = \left(c_j, \gamma^{-1}\|c_j\|^2\right), W = (-2X, \gamma) \in\in H \times \mathbb{R}$ and $W_1 = -2X \in H$ is the first component of W.

A more general loss function

More generally we will consider a loss function

$$f(\theta, w) \in \mathbb{R}, \quad \theta \in H^m, w \in H,$$

such that

$$\left| f(\theta', w) - f(\theta, w) \right| \leq \left(a + b \|w\|^{\alpha_2} \right) \max_{j \in [\![1,k]\!]} \left| \sum_{\ell=1}^{m} A_{j,\ell} \langle \theta'_\ell - \theta_\ell, w \rangle \right|^{\alpha_1},$$

$$\alpha_1 \in]0,1], \alpha_2 \in \mathbb{R}_+, A \in \mathbb{R}^{k \times m}.$$

The minimization problem

$$\inf_{\theta \in \Theta} \mathbb{P}_W \left[f(\theta, W) \right]$$

covers in particular the case

$$\inf_{c \in H^k} \mathbb{P}_X \left(\mu(X) \min_{j \in [\![1,k]\!]} \|X - c_j\|^{2\alpha_1} \right), \quad \alpha_1 \in]0,1], \tag{1}$$

$$\text{where } \mu(X) \in [0, a' + b' \|X\|^{\alpha_2}], \quad a', b', \alpha_2 \in \mathbb{R}_+. \tag{2}$$

Structured k-means

We use k centers depending on T parameters.

$$\inf_{\xi \in \Xi} \mathbb{P}_X\Big(\mu(X) \min_{j \in [\![1,k]\!]} \Big\| X - \sum_{t=1}^{T} B_{j,t} \xi_t \Big\|^{2\alpha_1} \Big)$$

The (θ, W) parametrization is given by $m = 2T + T(T-1)/2$,
$W = \big(-2X, \gamma\big) \in H \times \mathbb{R}$, and for $1 \leq t \leq T$, $1 \leq s < T$,

$$\begin{aligned}
A_{j,t} &= B_{j,t}, & \theta_t &= (\xi_t, 0), \\
A_{j,T+t} &= B_{j,t}^2, & \theta_{T+j} &= \Big(0, \gamma^{-1} \|\xi_t\|^2\Big), \\
A_{j,2T+t(t-1)/2+s-1} &= 2B_{j,t}B_{j,s}, & \theta_{2T+t(t-1)/2+s-1} &= \Big(0, \gamma^{-1} \langle \xi_t, \xi_s \rangle\Big)
\end{aligned}$$

With this choice of coordinates

$$\Big\| X - \sum_{t=1}^{T} B_{j,t}\, \xi_t \Big\|^{2\alpha_1} = \Big| \|W_1\|^2/4 + \Big\langle \sum_{\ell=1}^{m} A_{j,\ell}\, \theta_\ell, W \Big\rangle \Big|^{\alpha_1}$$

Bound the excess risk

Consider a non random reference value of the parameter $\theta^\star$ and work on the excess risk

$$h_0(\theta, w) = f(\theta, w) - f(\theta^\star, w).$$

Consider a statistical sample $\overline{W} = (W_1, \ldots, W_n)$ made of n independent copies of $W \in H$. From a bound in expectation

$$\mathbb{P}_{\overline{W}}\Big[\sup_{\theta \in \Theta}\Big((\mathbb{P}_W h_0)(\theta) - B(\theta, \overline{W})\Big)\Big] \leq 0 \tag{3}$$

and the ϵ-minimizer

$$B\big(\widehat{\theta}, \overline{W}\big) \leq \inf_{\theta \in \Theta} B\big(\theta, \overline{W}\big) + \epsilon,$$

we get

$$\mathbb{P}_{\overline{W}}\Big[\mathbb{P}_W\big[f\big(\widehat{\theta}, W\big)\big]\Big] \leq \inf_{\theta^\star \in \Theta}\Big[\mathbb{P}_W\big[f(\theta^\star, W)\big] + \mathbb{P}_{\overline{W}}\big[B\big(\theta^\star, \overline{W}\big)\big]\Big] + \epsilon.$$

when the choice of $\theta^\star$ is optimal.

Deviation bounds

From a deviation bound

$$\mathbb{P}_{\overline{W}}\Big[\sup_{\theta\in\Theta}\Big((\mathbb{P}_W h_0)(\theta) - B(\theta, \overline{W}) - \log(\delta^{-1})\Big) \leq 0\Big] \geq 1 - \delta. \tag{4}$$

and the ϵ-minimizer

$$B(\widehat{\theta}, \overline{W}) \leq \inf_{\theta\in\Theta} B(\theta, \overline{W}) + \epsilon,$$

we get

$$\mathbb{P}_{\overline{W}}\Big[\mathbb{P}_W\big[f(\widehat{\theta}, W)\big] \leq \inf_{\theta^\star\in\Theta} \mathbb{P}_W\big[f(\theta^\star, W)\big] + B\big(\theta^\star, \overline{W}\big) + \epsilon\Big] \geq 1 - \delta.$$

when $\theta^\star$ is optimal.

Deduce everything from exponential moments

We will deduce both bounds in expectation and deviation bounds from bounds on exponential moments of the form

$$\mathbb{P}_{\overline{W}}\left\{\exp\left[\lambda\left(\sup_{\theta\in\Theta}(\mathbb{P}_W h_0)(\theta) - B(\theta,\overline{W})\right)\right]\right\} \leq 1, \tag{5}$$

where λ is a positive real exponent. This implies (3) and

$$\mathbb{P}_{\overline{W}}\left[\sup_{\theta\in\Theta}\left((\mathbb{P}_W h_0)(\theta) - B(\theta,\overline{W}) - \log(\delta^{-1})/\lambda\right) \leq 0\right] \geq 1-\delta. \tag{6}$$

Lemma on the expectation of the supremum of Gaussian random variables

Let α be some positive real exponent, and let ϵ_j, $1 \leq j \leq k$ be k centered Gaussian random variables with variances σ_j^2, $1 \leq j \leq k$. We do not assume independence nor the fact that the vector $(\epsilon_1, \ldots, \epsilon_k)$ has a joint Gaussian distribution. Let $m_j \in \mathbb{R}$, $1 \leq j \leq k$ be mean parameters. Assume that $k \geq \exp(\alpha - 1)$.

$$\mathbb{P}_{\epsilon_1,\ldots,\epsilon_k}\left(\max_{j\in[\![1,k]\!]}\left|m_j + \epsilon_j\right|^{\alpha}\right) \leq \left(\sqrt{2\log(2k)}\max_{j\in[\![1,k]\!]}\sigma_j + \max_{j\in[\![1,k]\!]}|m_j|\right)^{\alpha}.$$

Gaussian perturbations

- Assume w.l.o.g. that $H = \ell^2$.
- Let $\rho_{\theta' \mid \theta} = \bigotimes_{\ell=1}^{m} \left(\bigotimes_{i \in \mathbb{N}} \mathcal{N}(\theta_{\ell,i}, \sigma_\ell^2/\beta) \right) : (\mathbb{R}^{\mathbb{N}})^m \to \mathcal{M}_+^1\big((\mathbb{R}^{\mathbb{N}})^m\big)$.
- Let $\langle \theta, w \rangle = \begin{cases} \lim\limits_{s \to +\infty} \sum\limits_{i=0}^{s} \theta_i w_i, & \text{when } \overline{\lim\limits_{s \to +\infty}} \sum\limits_{i=0}^{s} \theta_i w_i = \underline{\lim\limits_{s \to +\infty}} \sum\limits_{i=0}^{s} \theta_i w_i \in \mathbb{R}, \\ 0, & \text{otherwise} \end{cases}$

  be a non bilinear but measurable extension of the scalar product from ℓ^2 to $\mathbb{R}^{\mathbb{N}}$.
- The linear operator ρ operates on suitably integrable functions of θ and w according to the rule

$$(\rho f)(\theta, w) = \rho_{\theta' \mid \theta}\big(f(\theta', w)\big).$$

PAC-Bayesian lemma

Consider the increasing function $g(t) = \frac{2}{t^2}\left[\exp(t) - 1 - t\right], \quad t \in \mathbb{R}$ defined by continuity at $t = 0$, where $g(0) = 1$. For any measurable bounded real valued function $h(w)$, $w \in H$, such that $\sup_{w \in H} |h(w)| \leq \eta$, for any positive exponent λ,

$$\mathbb{P}_{\overline{W}}\left\{\exp\left[n\lambda\left(\overline{\mathbb{P}}_W - \mathbb{P}_W\right)h - n\frac{\lambda^2}{2}g(2\eta\lambda)\mathbb{P}_W\left[\left(h - \mathbb{P}_W h\right)^2\right]\right]\right\} \leq 1.$$

If $h(\theta, w) \in \mathbb{R}$, $\theta \in H^m, w \in H$ depends also on θ and if $\sup_{\theta \in H^m, w \in H}\left|h(\theta, w)\right| \leq \eta$,

$$\mathbb{P}_{\overline{W}}\left\{\exp\left[\sup_{\rho \in \mathcal{M}_+^1(\Theta)} n\lambda\left(\overline{\mathbb{P}}_W - \mathbb{P}_W\right)\rho h \right.\right.$$
$$\left.\left. - n\frac{\lambda^2}{2}g(2\eta\lambda)\rho\mathbb{P}_W\left[\left(h - \mathbb{P}_W h\right)^2\right] - \mathcal{K}(\rho, \pi)\right]\right\} \leq 1.$$

Multi-scale decomposition of the excess risk

The decomposition

$$h_0 = \left(\mathrm{I} - \rho + \sum_{q=1}^{p} \left(\rho^{2^{q-1}} - \rho^{2^q}\right) + \rho^{2^p}\right) h_0, \tag{7}$$

can be written as

$$h_0 = h_{p+1} + \sum_{q=0}^{p} \left(h_q - h_{q+1}\right), \tag{8}$$

where

$$h_0(\theta, w) = f(\theta, w) - f(\theta^\star, w), \quad \theta \in H^m, w \in H,$$

$$\text{(with implicit dependence on } \theta^\star\text{)},$$

$$h_q = \rho^{2^{q-1}} h_0 = \rho\left(2^{-q+1}\beta\right) h_0, \quad 1 \le q \le p+1.$$

Thresholds

We will also need a sequence of truncation operators T_q using threshold levels $\eta_q > 0$. They are defined as

$$T_q(z) = \min\{\eta_q, z\}, \quad z \in \mathbb{R}_+,$$
$$T_q(w) = T_q(\|w\|)\frac{w}{\|w\|}, \quad w \in H,$$
$$(T_q f)(\theta, w) = f\big(\theta, T_q(w)\big), \quad \theta \in H^m, w \in H,$$

and the operator $T_q f$ acting on functions is linear (although the truncation acting on vectors or positive real numbers is not). Moreover, T_q commutes with ρ :

$$T_q \rho = \rho T_q.$$

We also have the composition rules

$$\mathbb{P}_W \rho = \rho \mathbb{P}_W, \quad \text{and} \quad \mathbb{P}_W T_q = \mathbb{P}_{T_q(W)}.$$

At stage $p+1$, we will need the non linear threshold operator T_{p+1} defined as

$$\big(T_{p+1} h\big)(\theta, w) = \min\Big\{\eta_{p+1}, \max\big\{-\eta_{p+1}, h(\theta, w)\big\}\Big\},$$

Bounding the expected excess risk

We will prove a bound that compares the expected excess risk $\mathbb{P}_W h_0$ with the possibly truncated empirical excess risk $\overline{\mathbb{P}}_W T_0 h_0$. We decompose the risk into

$$\mathbb{P}_W h_0 = \mathbb{P}_W (\mathrm{I} - T_0) h_0 + (\mathbb{P}_W - \overline{\mathbb{P}}_W) T_0 h_0 + \overline{\mathbb{P}}_W T_0 h_0,$$

leading to

$$\mathbb{P}_W h_0 = \mathbb{P}_W (\mathrm{I} - T_0) h_0 + (\mathbb{P}_W - \overline{\mathbb{P}}_W) T_0 \left(h_{p+1} + \sum_{q=0}^{p} (h_q - h_{q+1}) \right) + \overline{\mathbb{P}}_W T_0 h_0. \quad (9)$$

We will analyze each term of this decomposition separately.

Bounding $A_{q,0} = \left(\mathbb{P}_W - \overline{\mathbb{P}}_W\right) T_0 \left(h_q - h_{q+1}\right)$

Let us deal first with the case when $q \geq 1$. Decompose further $A_{q,0}$ into

$$A_{q,0} = A_{q,1} + A_{q,2},$$

where

$$\begin{aligned} A_{q,1} &= \left(\mathbb{P}_W - \overline{\mathbb{P}}_W\right) T_0 (\mathrm{I} - T_q) \left(h_q - h_{q+1}\right) \\ A_{q,2} &= \left(\mathbb{P}_W - \overline{\mathbb{P}}_W\right) T_q \left(h_q - h_{q+1}\right). \end{aligned}$$

As $h_q = \rho_q h_0$,

$$h_q - h_{q+1} = \rho_q (\mathrm{I} - \rho_q) h_0 = \rho_q (\mathrm{I} - \rho_q) f,$$

since $h_0(\theta, w) = f(\theta, w) - f(\theta^\star, w)$, where $f(\theta^\star, w)$ does not depend on θ. Thus

$$A_{q,2} = \left(\mathbb{P}_W - \overline{\mathbb{P}}_W\right) \rho_q (\mathrm{I} - \rho_q) T_q f.$$

Moreover

$$
\begin{aligned}
\left|(\mathrm{I}-\rho_q)T_q f\right| &= \left|\rho(\beta_q)_{\theta' \mid \theta}\left[f(\theta, T_q(w)) - f(\theta', T_q(w))\right]\right| \\
&\leq \rho(\beta_q)_{\theta' \mid \theta}\left[\left|f(\theta, T_q(w)) - f(\theta', T_q(w))\right|\right],
\end{aligned}
$$

so that

$$
\begin{aligned}
&\left|(\mathrm{I}-\rho_q)T_q f\right| \\
&\quad \leq \left(a + bT_q(\|w\|)^{\alpha_2}\right)\rho(\beta_q)_{\theta' \mid \theta}\left(\max_{j\in[\![1,k]\!]}\left|\left\langle T_q(w), \sum_{\ell=1}^{m} A_{j,\ell}\left(\theta_\ell - \theta'_\ell\right)\right\rangle\right|^{\alpha_1}\right) \\
&\quad \leq \left(a + bT_q(\|w\|)^{\alpha_2}\right)\left(\sqrt{2\log(2k)/\beta}\,\sigma_\star T_q(\|w\|)\right)^{\alpha_1},
\end{aligned}
$$

where

$$
\sigma_\star = \max_{j\in[\![1,k]\!]} \sqrt{\sum_{\ell=1}^{m} A_{j,\ell}^2 \sigma_\ell^2}.
$$

Introduce the bounds

$$B_{q,0}(w) = B_{q,1}(\|w\|)$$

where

$$B_{q,1}(t) = \bigl(a + bt^{\alpha_2}\bigr)\bigl(2\log(2k)/\beta_q\bigr)^{\alpha_1/2}\bigl(\sigma_\star t\bigr)^{\alpha_1}, \quad t \in \mathbb{R}_+.$$

We get

$$\Bigl|(\mathrm{I} - \rho_q)T_q f\Bigr| \leq T_q B_{q,0}.$$

Let us put

$$K_q(\theta) = \sum_{\ell=1}^m \frac{\beta_q}{2\sigma_\ell^2}\|\theta_\ell - \widetilde{\theta}_\ell\|^2 = \mathcal{K}\bigl(\rho(\beta_q)_{\theta' \mid \theta}, \rho(\beta_q)_{\theta' \mid \theta = \widetilde{\theta}}\bigr), \quad \theta \in H^m,$$

where $\widetilde{\theta} \in H^m$ is a non random reference that may or may not be equal to $\theta^\star$. For any $\lambda_q > 0$,

$$\mathbb{P}_{\overline{W}}\Bigl\{\exp\Bigl[\sup_{\theta \in \Theta} n\lambda A_{q,2} - n\frac{\lambda^2}{2}g\bigl(2\lambda B_{q,1}(\eta_q)\bigr)\mathbb{P}_W\bigl(B_{q,0}(W)^2\bigr) - K_q(\theta)\Bigr]\Bigr\} \leq 1.$$

Let us now bound

$$A_{q,1} = \left(\mathbb{P}_W - \overline{\mathbb{P}}_W\right) T_0 (\mathrm{I} - T_q) \rho_q (\mathrm{I} - \rho_q) f.$$

We can write

$$\begin{aligned}
|A_{q,1}| \leq \left(\mathbb{P}_W + \overline{\mathbb{P}}_W\right) T_0 (\mathrm{I} + T_q) \left[\mathbb{1}\left(\|W\| \geq \eta_q\right) \rho_q \left|(\mathrm{I} - \rho_q) f\right|\right] \\
\leq 2\left(\mathbb{P}_W + \overline{\mathbb{P}}_W\right) T_0 \left[\mathbb{1}\left(\|W\| \geq \eta_q\right) B_{q,0}\right] \\
\leq 2\left(\mathbb{P}_W + \overline{\mathbb{P}}_W\right) \left[T_0 B_{q,0}^2 / B_{q,1}(\eta_q)\right].
\end{aligned}$$

Bounding $A_{0,0}$

In the case when $q = 0$,

$$A_{0,0} = \left(\mathbb{P}_W - \overline{\mathbb{P}}_W\right) T_0 (\mathrm{I} - \rho) f,$$

so that

$$\left|A_{0,0}\right| \leq \left(\mathbb{P}_W + \overline{\mathbb{P}}_W\right) T_0 \left|(\mathrm{I} - \rho) f\right| \leq \left(\mathbb{P}_W + \overline{\mathbb{P}}_W\right) T_0 B_{1,0}.$$

Let us now come to

$$A_{-1,0} = \mathbb{P}_W(\mathrm{I} - T_0)h_0.$$

Remark that

$$\left|h_0(\theta, w)\right| \leq B_{-1,0}(\theta, w),$$

where

$$B_{-1,0}(\theta, w) = B_{-1,1}(\theta, \|w\|),$$

where

$$B_{-1,1}(\theta, t) = \left(a + bt^{\alpha_2}\right)\left(t \max_{j\in[\![1,k]\!]} \left\|\sum_{\ell=1}^{m} A_{j,\ell}\left(\theta_\ell - \theta_\ell^\star\right)\right\|\right)^{\alpha_1}$$

is not decreasing in $t \in \mathbb{R}_+$. Therefore,

$$\begin{aligned}
\left|A_{-1,0}\right| = \left|\mathbb{P}_W(\mathrm{I} - T_0)h_0\right| &\leq \mathbb{P}_W\left(\left|(\mathrm{I} - T_0)h_0\right|\right) \\
&= \mathbb{P}_W\left(\left|(\mathrm{I} - T_0)\left[\mathbb{1}\left(\|w\| \geq \eta_0\right)h_0\right]\right|\right) \\
&\leq \mathbb{P}_W\left((\mathrm{I} + T_0)\left[\mathbb{1}\left(\|w\| \geq \eta_0\right)|h_0|\right]\right) \\
&\leq \mathbb{P}_W\left((\mathrm{I} + T_0)\left[\mathbb{1}\left(\|w\| \geq \eta_0\right)B_{-1,0}\right]\right) \\
&\leq 2\mathbb{P}_W\left(\left[\mathbb{1}\left(\|w\| \geq \eta_0\right)B_{-1,0}\right]\right) \\
&\leq 2\mathbb{P}_W\left(B_{-1,0}(\theta, W)^2 / B_{-1,1}(\theta, \eta_0)\right).
\end{aligned}$$

Finally, let us bound

$$A_{p+1,3} = \left(\mathbb{P}_W - \overline{\mathbb{P}}_W\right) T_0 h_{p+1} = \left(\mathbb{P}_W - \overline{\mathbb{P}}_W\right) \rho_{p+1} T_0 h_0.$$

Introduce the non linear threshold operator T_{p+1} defined by the equation

$$\left(T_{p+1} h\right)(\theta, w) = \min\Big\{\eta_{p+1}, \max\big\{-\eta_{p+1}, h(\theta, w)\big\}\Big\},$$

where $\eta_{p+1} > 0$ is some threshold level. We can decompose $A_{p+1,3}$ into

$$A_{p+1,3} = A_{p+1,4} + A_{p+1,5},$$

where

$$A_{p+1,4} = \left(\mathbb{P}_W - \overline{\mathbb{P}}_W\right) \rho_{p+1} T_{p+1} T_0 h_0$$

and

$$A_{p+1,5} = \left(\mathbb{P}_W - \overline{\mathbb{P}}_W\right) \rho_{p+1} (\mathrm{I} - T_{p+1}) T_0 h_0.$$

Remark first that

$$\begin{aligned}\left|A_{p+1,5}\right| &\le \left(\mathbb{P}_W + \overline{\mathbb{P}}_W\right)\rho_{p+1}\left|(\mathrm{I} - T_{p+1})T_0 h_0\right| \\ &= \left(\mathbb{P}_W + \overline{\mathbb{P}}_W\right)\rho_{p+1}\left[\left(|T_0 h_0| - \eta_{p+1}\right)_+\right] \\ &\le \left(\mathbb{P}_W + \overline{\mathbb{P}}_W\right)\rho_{p+1}\left[\mathbb{1}\left(|T_0 h_0| \ge \eta_{p+1}\right)|T_0 h_0|\right] \\ &\le \frac{1}{\eta_{p+1}}\left(\mathbb{P}_W + \overline{\mathbb{P}}_W\right)\rho_{p+1}\left(|T_0 h_0|^2\right).\end{aligned}$$

Define $f^\star(\theta, w) = f(\theta^\star, w)$ depending on w only. Remark that

$$\begin{aligned}\rho_{p+1}\left[|T_0 h_0|^2\right] &= T_0\rho_{p+1}\left[(f - f^\star)^2\right] \\ &\le T_0\rho_{p+1}\left[\left(a + b\|w\|^{\alpha_2}\right)^2\Big(\max_{j\in[\![1,k]\!]}\Big|\Big\langle w, \sum_{\ell=1}^{m} A_{j,\ell}\left(\theta_\ell - \theta_\ell^\star\right)\Big\rangle\Big|\Big)^{2\alpha_1}\right] \\ &\le T_0\left(a+b\|w\|^{\alpha_2}\right)^2\left(\left(2\log(2k)/\beta_{p+1}\right)^{1/2}\sigma_\star\|w\| + \max_{j\in[\![1,k]\!]}\Big|\Big\langle w, \sum_{\ell=1}^{m} A_{j,\ell}\left(\theta_\ell - \theta_\ell^\star\right)\Big\rangle\Big|\right)^{2\alpha_1} \\ &\le T_0 B_{p+1,0},\end{aligned}$$

where

$$
\begin{aligned}
B_{p+1,0}(\theta, w) = \big(a + b\|w\|^{\alpha_2}\big)^2 \|w\|^{2\alpha_1} \\
\times \left(\big(2\log(2k)/\beta_{p+1}\big)^{1/2} \sigma_\star + \max_{j \in [\![1,k]\!]} \Big\| \sum_{\ell=1}^{m} A_{j,\ell} \big(\theta_\ell - \theta_\ell^\star\big) \Big\| \right)^{2\alpha_1}.
\end{aligned}
$$

Therefore

$$
\big|A_{p+1,5}\big| \leq \big(\mathbb{P}_W + \overline{\mathbb{P}}_W\big)\big(T_0 B_{p+1,0}/\eta_{p+1}\big).
$$

Low frequency component

The last term to bound is $A_{p+1,4}$. We get for any $\lambda_{p+1} > 0$

$$\mathbb{P}_{\overline{W}}\left\{\exp\left[\sup_{\theta\in\Theta} n\lambda_{p+1}A_{p+1,4} - n\frac{\lambda_{p+1}^2}{2}g\left(2\lambda_{p+1}\eta_{p+1}\right)\mathbb{P}_W\rho_{p+1}\left((T_0h_0)^2\right) - K_{p+1}(\theta)\right]\right\} \leq 1,$$

so that

$$\mathbb{P}_{\overline{W}}\left\{\exp\left[\sup_{\theta\in\Theta}\left(n\lambda_{p+1}A_{p+1,4} - n\frac{\lambda_{p+1}^2}{2}g\left(2\lambda_{p+1}\eta_{p+1}\right)\mathbb{P}_W\left(T_0B_{p+1,0}\right) - K_{p+1}(\theta)\right)\right]\right\} \leq 1.$$

Summary

We have written

$$\mathbb{P}_W h_0 = \overline{\mathbb{P}}_W T_0 h_0 + A_{-1,0} + A_{0,0} + A_{p+1,4} + A_{p+1,5} + \sum_{q=1}^{p} \left(A_{q,1} + A_{q,2} \right)$$

and provided bounds for each $A_{q,\ell}$, either almost sure bounds or exponential moment bounds.

Based on the bounds

$$
\begin{aligned}
B_{q,0}(w) &= B_{q,1}(\|w\|) \\
\text{where } B_{q,1}(t) &= \left(a + bt^{\alpha_2}\right)\left(2\log(2k)/\beta_q\right)^{\alpha_1/2}\left(\sigma_\star t\right)^{\alpha_1}, \quad t \in \mathbb{R}_+ \\
B_{q,0}(w) &= \xi(\|w\|)\mathcal{S}_3^{\alpha_1}\beta_q^{-\alpha_1/2}, \\
B_{-1,0}(\theta, w) &= \left(a + b\|w\|^{\alpha_2}\right)\left(\|w\| \max_{j\in[\![1,k]\!]}\left\|\sum_{\ell=1}^{m} A_{j,\ell}\left(\theta_\ell - \theta_\ell^\star\right)\right\|\right)^{\alpha_1} \\
&= \xi(\|w\|)\mathcal{S}_2^{\alpha_1}, \\
B_{p+1,0}(\theta, w) &= \left(a + b\|w\|^{\alpha_2}\right)^2\|w\|^{2\alpha_1} \\
&\qquad \times \left(\left(2\log(2k)/\beta_{p+1}\right)^{1/2}\sigma_\star + \max_{j\in[\![1,k]\!]}\left\|\sum_{\ell=1}^{m} A_{j,\ell}\left(\theta_\ell - \theta_\ell^\star\right)\right\|\right)^{2\alpha_1}, \\
&= \xi(\|w\|)^2\left(\mathcal{S}_3\beta_{p+1}^{-1/2} + \mathcal{S}_2\right)^{2\alpha_1}
\end{aligned}
$$

$$
\text{and } K_q(\theta) = \sum_{\ell=1}^{m} \frac{\beta_q}{2\sigma_\ell^2}\|\theta_\ell - \widetilde{\theta}_\ell\|^2,
$$

What we proved

We proved that

$$\mathbb{P}_{\overline{W}}\Big\{\exp\Big[\sup_{\theta\in\Theta} n\lambda A_{q,2} - n\frac{\lambda^2}{2} g\big(2\lambda B_{q,1}(\eta_q)\big)\mathbb{P}_W\big(B_{q,0}(W)^2\big) - K_q(\theta)\Big]\Big\} \le 1,$$

$$\mathbb{P}_{\overline{W}}\Big\{\exp\Big[\sup_{\theta\in\Theta}\Big(n\lambda A_{p+1,4} - n\frac{\lambda^2}{2} g\big(2\lambda \eta_{p+1}\big)\mathbb{P}_W\big(T_0 B_{p+1,0}\big) - K_{p+1}(\theta)\Big)\Big]\Big\} \le 1.$$

We also bounded the remaining terms by

$$\begin{aligned}
|A_{q,1}| &\le 2\big(\mathbb{P}_W + \overline{\mathbb{P}}_W\big)T_0\big[\mathbb{1}(\|W\| \ge \eta_q)B_{q,0}\big] \\
&\le 2\big(\mathbb{P}_W + \overline{\mathbb{P}}_W\big)\big[T_0 B_{q,0}(W)^2/B_{q,1}(\eta_q)\big] \\
\big|A_{0,0}\big| &\le \big(\mathbb{P}_W + \overline{\mathbb{P}}_W\big)T_0 B_{1,0}. \\
\big|A_{-1,0}\big| &\le 2\mathbb{P}_W\Big(B_{-1,0}(\theta, W)^2/B_{-1,1}(\theta,\eta_0)\Big), \\
\big|A_{p+1,5}\big| &\le \big(\mathbb{P}_W + \overline{\mathbb{P}}_W\big)\big(T_0 B_{p+1,0}/\eta_{p+1}\big)
\end{aligned}$$

Bound in expectation

Assume that $n \geq 2\mathcal{S}_1$ and set $\eta_0 = +\infty$. We get
$\mathbb{P}_{\overline{W}}\left\{\sup_{\theta\in\Theta} \mathbb{P}_W h_0 - \overline{\mathbb{P}}_W h_0 - \gamma\right\} \leq 0$, where

$$\gamma = 2\left[\left(\frac{g(1)}{2} + 8\right)^{1/2} \mathbb{P}_W\left(\xi(\|W\|)^2\right)^{1/2} + \mathbb{P}_W\left(\xi(\|W\|)\right)\right] \times \mathcal{S}_3^{\alpha_1}\left(\frac{\log(2n/\mathcal{S}_1)}{\log(2)}\right)^{\alpha_1}\left(\frac{\mathcal{S}_1}{n}\right)^{\alpha_1/2} + 2\left[\left(\frac{g(1)}{2} + 4\right)\mathbb{P}_W\left(\xi(\|W\|)^2\right)\left(\mathcal{S}_3 + \mathcal{S}_2\right)^{2\alpha_1}\frac{\mathcal{S}_1}{n}\right]^{1/2}$$

$$\text{and } \xi(t) = \left(a + bt^{\alpha_2}\right)t^{\alpha_1}, \quad \mathcal{S}_1 = \sup_{\theta\in\Theta}\sum_{\ell=1}^{m}\frac{\|\theta_\ell - \widetilde{\theta}_\ell\|^2}{2\sigma_\ell^2}$$

$$\mathcal{S}_2 = \sup_{\theta\in\Theta}\max_{j\in[\![1,k]\!]}\left\|\sum_{\ell=1}^{m} A_{j,\ell}(\theta_\ell - \theta_\ell^\star)\right\|$$

$$\text{and } \mathcal{S}_3 = \left(2\log(2k)\right)^{1/2}\max_{j\in[\![1,k]\!]}\sqrt{\sum_{\ell=1}^{m} A_{j,\ell}^2\sigma_\ell^2}.$$

Proposition

With the above definitions, consider an ϵ-minimizer of the empirical risk $\widehat{\theta}(\overline{W}) \in \Theta$, that is an estimator satisfying $\mathbb{P}_{\overline{W}}$ almost surely

$$\overline{\mathbb{P}}\big(f(\widehat{\theta}, W)\big) \leq \inf_{\theta \in \Theta} \overline{\mathbb{P}}\big(f(\theta, W)\big) + \epsilon.$$

Its mean excess risk satisfies

$$\mathbb{P}_{\overline{W}}\Big[\mathbb{P}\big(f(\widehat{\theta}, W)\big)\Big] \leq \inf_{\theta \in \Theta} \mathbb{P}\big(f(\theta, W)\big) + \gamma + \epsilon.$$

Deviation bounds

Introduce the increasing function

$$\widetilde{g}(t) = \frac{1}{t}\bigl[\exp(t) - 1\bigr].$$

Remark that with probability at least $1 - \delta$

$$\overline{\mathbb{P}}_W\bigl(\xi(\|T_0(W)\|)^2\bigr) \leq \widetilde{g}\bigl(\lambda_0 \xi(\eta_0)^2\bigr)\mathbb{P}\bigl(\xi(\|W\|)^2\bigr) + \frac{\log(\delta^{-1})}{n\lambda_0}$$

Take

$$\lambda_0 = \xi(\eta_0)^{-2},$$

and choose η_0 such that

$$\frac{\log(2/\delta)}{n}\xi(\eta_0)^2 = \widetilde{g}(1)\mathbb{P}\bigl(\xi(W)^2\bigr),$$

so that with probability at least $1 - \delta/2$

$$\overline{\mathbb{P}}_W\bigl(\xi(\|T_0(W)\|)^2\bigr) \leq 2\widetilde{g}(1)\mathbb{P}\bigl(\xi(\|W\|)^2\bigr).$$

Choosing thresholds

Choose as previously η_q and η_{p+1} such that

$$B_{q,1}(\eta_q) = \frac{1}{2\lambda_q} \text{ and } \eta_{p+1} = \frac{1}{2\lambda_{p+1}}.$$

We get with probability at least $1 - \delta/2$

$$\begin{aligned}
|A_{q,1}| &\leq 4(1 + 2\widetilde{g}(1))\lambda_q \mathbb{P}_W\left[T_0 B_{q,0}(W)^2\right], \\
|A_{0,0}| &\leq \left(1 + \sqrt{2\widetilde{g}(1)}\right)\mathbb{P}_W\left(B_{1,0}^2\right)^{1/2}, \\
\left|A_{-1,0}\right| &\leq 2\mathbb{P}_W\left(B_{-1,0}^2 / B_{-1,1}(\eta_0)\right), \\
|A_{p+1,5}| &\leq 2\left(1 + 2\widetilde{g}(1)\right)\lambda_{p+1}\mathbb{P}_W\left(T_0 B_{p+1,0}\right).
\end{aligned}$$

PAC-Bayesian inequality

We also have

$$\mathbb{P}_{\overline{W}}\Big\{\exp\Big[n\lambda \sup_{\theta\in\Theta} \sum_{q=1}^{p} A_{q,2} - \frac{\lambda_q}{2} g(1)\mathbb{P}_W\big(B_{q,0}^2\big) - \frac{\beta_q \mathcal{S}_1}{n\lambda_q} \\ + A_{p+1,4} - \frac{\lambda_{p+1}}{2} g(1)\mathbb{P}_W\big(B_{p+1,0}\big) - \frac{\beta_{p+1}\mathcal{S}_1}{n\lambda_{p+1}}\Big]\Big\} \leq 1,$$

where

$$\frac{1}{\lambda} = \frac{1}{\lambda_{p+1}} + \sum_{q=1}^{p} \frac{1}{\lambda_q}.$$

Proposition

Assume that $\mathbb{P}\left[\xi\left(\|W\|\right)^2\right]^{1/2} \leq B$, where B is known and that $n \geq 2\mathcal{S}_1$. Consider the threshold η_0 such that

$$\frac{\log(2/\delta)}{n}\xi(\eta_0)^2 = (e-1)B.$$

Let $\widehat{\theta} \in \Theta$, be such that

$$\overline{\mathbb{P}}_W\big(f(\widehat{\theta}, T_0 W)\big) \leq \inf_{\theta\in\Theta} \overline{\mathbb{P}}\big(f(\theta, T_0 W)\big) + \epsilon.$$

With probability at least $1-\delta$, its excess risk is such that

$$\mathbb{P}_W\big(f(\widehat{\theta}, W)\big) \leq \inf_{\theta\in\Theta} \mathbb{P}_W\big(f(\theta, W)\big) + B\gamma + \epsilon$$

$$\begin{aligned}
\text{where } \gamma &= 12\, p^{\alpha_1}\mathcal{S}_3^{\alpha_1}\left(\mathcal{S}_1/n\right)^{\alpha_1/2} + 7\left(\mathcal{S}_3 + \mathcal{S}_2/p\right)^{\alpha_1} p^{-(1-\alpha_1)}\left(\mathcal{S}_1/n\right)^{1/2} \\
&\quad + \left[(9(2^{\alpha_1/2}-1)^{-1}+7)\mathcal{S}_3^{\alpha_1}p^{\alpha_1} + 12\,\mathcal{S}_2^{\alpha_1}\right]\left(\log(2/\delta)/n\right)^{1/2} \\
&= \mathcal{O}\Big(\log\left(n/\mathcal{S}_1\right)^{\alpha_1}\left(\mathcal{S}_1/n\right)^{\alpha_1/2}\Big), \text{ with } p = \left\lceil \log\left(n/\mathcal{S}_1\right)/\log(2)\right\rceil.
\end{aligned}$$