Means and k-means : dimension free PAC-Bayesian bounds for robust estimators

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

CREST, CNRS — UMR 9194
Université Paris–Saclay

Conference on robustness and privacy 2021
March 22, 2021

Joint work with

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

and Gautier Appert
SAMM, Université Paris 1 Panthéon-Sorbonne
`gautier.appert.chess@gmail.com`

General purpose

Illustrate the use of PAC-Bayesian inequalities to derive dimension free concentration and complexity bounds.

A general formulation of PAC-Bayesian inequalities

Consider $h : \mathcal{T} \times \mathcal{W} \to \mathbb{R}$ measurable, $\pi \in \mathcal{M}_+^1(\mathcal{T})$ a prior on the parameter, $W \in \mathcal{W}$ a random variable and $(W_1, \dots, W_n)$, a sample made of n independent copies of it. Let $\overline{\mathbb{P}}_W = \frac{1}{n} \sum_{i=1}^n \delta_{W_i}$. For any $\lambda > 0$

$$\mathbb{P}_{W_1,\dots,W_n} \Biggl\{ \exp \Biggl[\sup_{\rho \in \mathcal{M}_+^1(\mathcal{T})} \sup_{\eta \in \mathbb{N}} \biggl\{ \int \min \Bigl\{ \eta, -n\lambda \overline{\mathbb{P}}_W \bigl[h(\theta', W) \bigr] \\ - n \log \Bigl[\mathbb{P}_W \Bigl(\exp \bigl[-\lambda h(\theta', W) \bigr] \Bigr) \Bigr] \Bigr\} \, \mathrm{d}\rho(\theta') - \mathcal{K}(\rho, \pi) \biggr\} \Biggr] \Biggr\} \leq 1.$$

Usage

$$\mathbb{P}\bigl[\exp(X)\bigr] \leq 1 \Rightarrow \begin{cases} \mathbb{P}(X) \leq 0 & \text{complexity bound} \\ \text{and } \mathbb{P}\bigl[X \leq \log(\delta^{-1})\bigr] \geq 1 - \delta & \text{concentration bound.} \end{cases}$$

General purpose

Illustrate the use of PAC-Bayesian inequalities to derive dimension free concentration and complexity bounds.

A general formulation of PAC-Bayesian inequalities

Consider $h : \mathcal{T} \times \mathcal{W} \to \mathbb{R}$ measurable, $\pi \in \mathcal{M}_+^1(\mathcal{T})$ a prior on the parameter, $W \in \mathcal{W}$ a random variable and $(W_1, \dots, W_n)$, a sample made of n independent copies of it. Let $\overline{\mathbb{P}}_W = \frac{1}{n}\sum_{i=1}^n \delta_{W_i}$. For any $\lambda > 0$

$$\mathbb{P}_{W_1,\dots,W_n}\Bigg\{\exp\Bigg[\sup_{\rho \in \mathcal{M}_+^1(\mathcal{T})}\Bigg\{\int -n\lambda\overline{\mathbb{P}}_W\big[h(\theta', W)\big] - n\log\Big[\mathbb{P}_W\Big(\exp\big[-\lambda h(\theta', W)\big]\Big)\Big]\,\mathrm{d}\rho(\theta') - \mathcal{K}(\rho, \pi)\Bigg\}\Bigg]\Bigg\} \leq 1.$$

Usage

$$\mathbb{P}\big[\exp(X)\big] \leq 1 \Rightarrow \begin{cases} \mathbb{P}(X) \leq 0 & \text{complexity bound} \\ \text{and } \mathbb{P}\big[X \leq \log(\delta^{-1})\big] \geq 1 - \delta & \text{concentration bound.} \end{cases}$$

General purpose

Illustrate the use of PAC-Bayesian inequalities to derive dimension free concentration and complexity bounds.

A general formulation of PAC-Bayesian inequalities

Consider $h : \mathcal{T} \times \mathcal{W} \to \mathbb{R}$ measurable, $\pi \in \mathcal{M}_+^1(\mathcal{T})$ a prior on the parameter, $W \in \mathcal{W}$ a random variable and $(W_1, \dots, W_n)$, a sample made of n independent copies of it. Let $\overline{\mathbb{P}}_W = \frac{1}{n} \sum_{i=1}^n \delta_{W_i}$. For any $\lambda > 0$

$$\mathbb{P}_{W_1,\dots,W_n} \Biggl\{ \exp \Biggl[\sup_{\rho \in \mathcal{M}_+^1(\mathcal{T})} \biggl\{ \int -n\lambda \overline{\mathbb{P}}_W \bigl[h(\theta', W) \bigr] - n \log \Bigl[\mathbb{P}_W \Bigl(\exp \bigl[-\lambda h(\theta', W) \bigr] \Bigr) \Bigr] \, \mathrm{d}\rho(\theta') - \mathcal{K}(\rho, \pi) \biggr\} \Biggr] \Biggr\} \leq 1.$$

Usage

$$\mathbb{P}\bigl[\exp(X)\bigr] \leq 1 \Rightarrow \begin{cases} \mathbb{P}(X) \leq 0 & \text{complexity bound} \\ \text{and } \mathbb{P}\bigl[X \leq \log(\delta^{-1})\bigr] \geq 1 - \delta & \text{concentration bound.} \end{cases}$$

Mean estimation and Gaussian concentration

Aim

Given n independent copies $(X_1, \dots, X_n) \in \mathbb{R}^{d \times n}$ of $X \in \mathbb{R}^d$, estimate $\mathbb{P}(X)$.

Gaussian concentration

Assume that $\mathbb{P}_X = \mathcal{N}(m, \Sigma)$. With probability at least $1 - \delta$

$$\left\| \overline{\mathbb{P}}(X) - \mathbb{P}(X) \right\| \leq \sqrt{\frac{\mathbf{Tr}(\Sigma)}{n}} + \sqrt{\frac{2\|\Sigma\|_{\mathrm{op}} \log(\delta^{-1})}{n}}.$$

PAC-Bayesian proof

Write $\|\overline{\mathbb{P}}(X)\| = \sup_{\theta, \|\theta\|=1} \int \overline{\mathbb{P}}\big(\langle \theta', X \rangle\big) \, \mathrm{d}\rho_\theta(\theta')$, where $\rho_\theta = \mathcal{N}(\theta, \sigma^2 I_d)$.
Assume w.l.o.g. that $\mathbb{P}(X) = 0$. Take $\pi = \rho_0$, to get w.p. at least $1 - \delta$

$$\left\| \overline{\mathbb{P}}(X) \right\| \leq \frac{\lambda}{2} \sup_{\theta, \|\theta\|=1} \mathbb{P}_X\big(\langle \theta, X \rangle^2\big) + \frac{\lambda \sigma^2 \mathbb{P}(\|X\|^2)}{2} + \frac{1}{2n\lambda\sigma^2} + \frac{\log(\delta^{-1})}{n\lambda}$$

and optimize the choice of λ and σ^2.

Robust confidence region

To get a sub-Gaussian estimate of $\langle \theta, \mathbb{P}(X) \rangle$, from a PAC-B. inequality, assuming that X is already nearly centered, we need

$$\log\Big[\mathbb{P}_X\Big(\exp\big[-\lambda h(\theta', X)\big]\Big)\Big] \leq -\lambda \mathbb{P}_X\big(\langle \theta', X\rangle\big) + \frac{\lambda^2}{2}\mathbb{P}_X\Big(\langle \theta', X\rangle^2\Big).$$

This is the case when

$$-\lambda h(\theta', X) \leq \log\bigg(1 - \lambda\langle \theta', X\rangle + \frac{\lambda^2}{2}\langle \theta', X\rangle^2\bigg)$$

and we can take $h(\theta', X) = \lambda^{-1}\psi\big(\lambda\langle \theta', X\rangle\big)$, where $\psi(t) = T_{\sqrt{2}}(t) - \frac{1}{6}T_{\sqrt{2}}(t)^3$, with $T_s(t) = \min\big\{s, \max\{-s, t\}\big\}$.

Robust confidence region

Assume that $\mathbb{P}(\|X\|^2) < \infty$.

W. p. at least $1 - \delta$ for any θ, $\|\theta\| = 1$,

$$\langle \theta, \mathbb{P}(X) \rangle - \underbrace{\lambda^{-1} \int \overline{\mathbb{P}}_X \big(\psi \big(\lambda \langle \theta', X \rangle \big) \, \mathrm{d}\rho_\theta(\theta') }_{= \mathcal{E}_{\lambda,\sigma}(\theta)} \leq \frac{\lambda}{2} \mathbb{P}_X \big(\langle \theta, X \rangle^2 \big) + \frac{\lambda \sigma^2 \mathbb{P}\big(\|X\|^2 \big)}{2} + \frac{1}{2n\lambda\sigma^2} + \frac{\log(\delta^{-1})}{n\lambda}.$$

Let $G = \mathbb{P}_X\big(XX^\top\big)$. For optimal values of λ and σ, w. p. at least $1 - \delta$,

$$\sup_{\theta, \|\theta\|=1} \Big| \langle \theta, \mathbb{P}(X) \rangle - \mathcal{E}_{\lambda,\sigma}(\theta) \Big| \leq \sqrt{\frac{\mathbf{Tr}(G)}{n}} + \sqrt{\frac{2\|G\|_{\mathrm{op}} \log(\delta^{-1})}{n}}$$

Robust estimator

Mean estimator

If $\widehat{m} \in \arg\min_m \left[\sup_{\theta, \|\theta\|=1} \Big(\langle \theta, m \rangle - \mathcal{E}_{\lambda,\sigma}(\theta) \Big) \right]$, w. p. at least $1-\delta$,

$$\|\widehat{\mu} - \mathbb{P}(X)\| \leq 2 \left(\sqrt{\frac{\mathbf{Tr}(G)}{n}} + \sqrt{\frac{2\|G\|_{\mathrm{op}} \log(\delta^{-1})}{n}} \right).$$

Note that the lack of centering, resulting in the presence of the Gram matrix G in place of the covariance matrix Σ is not crucial and can be fixed using a split sample scheme.

A simpler estimator

Threshold the norm

Consider $\psi(t) = \min\{1, t\}, t \in \mathbb{R}_+$ and put $Y_i = \frac{\psi\big(\lambda\|X_i\|\big)}{\lambda\|X_i\|} X_i$. Define the estimator $\widehat{m} = \overline{\mathbb{P}}_Y(Y)$.

As $0 \leq 1 - \dfrac{\psi(t)}{t} \leq \inf_{p \geq 1} \dfrac{t^p}{p+1}\left(\dfrac{p}{p+1}\right)^p$,

$$\|\mathbb{P}(Y) - \mathbb{P}(X)\| \leq \inf_{p \geq 1} \frac{\lambda^p}{p+1}\left(\frac{p}{p+1}\right)^p \sup_{\theta, \|\theta\|=1} \mathbb{P}_X\big(\|X\|^p \langle \theta, X - m\rangle_-\big)$$

$$+ \inf_{p \geq 2} \frac{\lambda^p}{p+1}\left(\frac{p}{p+1}\right)^p \mathbb{P}\big(\|X\|^p\big)\|m\|.$$

A simpler estimator

Control the exponential moment

$$\begin{aligned}
&\int \log\Big(\mathbb{P}_Y\big[\exp\big(\mu\lambda\langle\theta', Y - \mathbb{P}(Y)\rangle\big)\big]\Big)\,\mathrm{d}\rho_\theta(\theta') \\
&\qquad \leq \log\bigg[\int \mathbb{P}_Y\big[\exp\big(\mu\lambda\langle\theta', Y - \mathbb{P}(Y)\rangle\big)\big]\,\mathrm{d}\rho_\theta(\theta')\bigg] \\
&\qquad = \log\bigg[\mathbb{P}_Y\bigg(\exp\bigg[\underbrace{\mu\lambda\langle\theta, Y - \mathbb{P}(Y)\rangle}_{\leq 2\mu} + \underbrace{\frac{\mu^2\lambda^2\sigma^2}{2}\|Y - \mathbb{P}(Y)\|^2}_{\leq 2\mu^2\sigma^2}\bigg]\bigg)\bigg] \\
&\qquad \leq g_2(2\mu)\frac{\mu^2\lambda^2}{2}\mathbb{P}_Y\big(\langle\theta, Y - \mathbb{P}(Y)\rangle^2\big) \\
&\qquad\qquad + \exp(2\mu)g_1\big(2\mu^2\sigma^2\big)\frac{\mu^2\lambda^2\sigma^2}{2}\mathbb{P}_Y\big(\|Y - \mathbb{P}(Y)\|^2\big),
\end{aligned}$$

where $g_1(t) = t^{-1}\big[\exp(t) - 1\big]$ and $g_2(t) = 2t^{-2}\big[\exp(t) - 1 - t\big]$ are increasing from $g_1(0) = g_2(0) = 1$.

Generalization bound

Assume at least that $\mathbb{P}_X\big(\|X\|^2\big) < \infty$.

Consider $\Sigma = \mathbb{P}_X\Big[\big(X - \mathbb{P}(X)\big)\big(X - \mathbb{P}(X)\big)^\top\Big]$ and put $\lambda = 4\sqrt{\dfrac{2\log(\delta^{-1})}{1.2\,\|\Sigma\|_{\mathrm{op}}\, n}}$.

With probability at least $1 - \delta$,

$$\|\widehat{m} - \mathbb{P}(X)\| \leq \sqrt{\frac{4\,\mathbf{Tr}(\Sigma)}{n}} + \sqrt{\frac{2.4\,\|\Sigma\|_{\mathrm{op}} \log(\delta^{-1})}{n}} + \inf_{p\geq 1} \frac{C_p}{n^{p/2}} + \inf_{p \geq 2} \frac{C'_p}{n^{p/2}},$$

where

$$C_p = \frac{1}{p+1}\left(\frac{4p}{p+1}\right)^p \left(\frac{2\log(\delta^{-1})}{1.2\,\|\Sigma\|_{\mathrm{op}}}\right)^{p/2} \sup_{\theta \in \mathbb{S}_d} \mathbb{P}_X\Big(\|X\|^p \langle \theta, X - \mathbb{P}(X)\rangle_-\Big),$$

$$C'_p = \frac{1}{p+1}\left(\frac{4p}{p+1}\right)^p \left(\frac{2\log(\delta^{-1})}{1.2\,\|\Sigma\|_{\mathrm{op}}}\right)^{p/2} \mathbb{P}_X\Big(\|X\|^p\Big)\big\|\mathbb{P}(X)\big\|$$

$$\times \left(1 + \big\|\mathbb{P}(X)\big\|\sqrt{\frac{0.6\log(\delta^{-1})}{\|\Sigma\|_{\mathrm{op}}\, n}}\right).$$

The quadratic k-means criterion

Aim

Given a r.v. $X \in H$ a separable Hilbert space, minimize

$$\mathcal{R}(c_1, \dots, c_k) = \mathbb{P}_X\Big(\min_{j \in [\![1,k]\!]} \|X - c_j\|^2 \Big), \quad (c_1, \dots, c_k) \in H^k,$$

in view of $(X_1, \dots, X_n)$, made of n independent copies of X.

Interpretation in terms of probabilities

Consider $(X, Y) \in H \times \mathbb{R}^{\mathbb{N}}$, where $\mathbb{P}_{Y \mid X} = \bigotimes_{i \in \mathbb{N}} \mathcal{N}\big(\langle X, e_i \rangle, \sigma^2\big)$. Define $Q^{(c)} \in \mathcal{M}_+^1\big(H \times \mathbb{R}^{\mathbb{N}}\big)$ by $Q_X^{(c)} = \mathbb{P}_X$ and $Q_{Y \mid X}^{(c)} = \bigotimes_{i \in \mathbb{N}} \mathcal{N}\big(\langle c_{\ell(X)}, e_i \rangle, \sigma^2\big)$, where $\ell(X) = \min\Big\{ \arg \min_{j \in [\![1,k]\!]} \|X - c_j\| \Big\}$.

$$\mathcal{R}(c) = 2\sigma^2 \mathcal{K}\big(Q_{X,Y}^{(c)}, \mathbb{P}_{X,Y}\big).$$

A robust criterion

Obtained by optimizing Q_X

$$\begin{aligned}
\mathcal{R}(c) &\overset{\text{def}}{=} \mathbb{P}_X\Big(\min_{j\in[\![1,k]\!]} \|X - c_j\|^2\Big) \\
&\geq 2\sigma^2 \inf_{Q\in\mathcal{M}_+^1(H\times\mathbb{R}^{\mathbb{N}})\,:\,Q_{Y\,|\,X}=Q^{(c)}_{Y\,|\,X}} \mathcal{K}\big(Q_{X,\,Y}, \mathbb{P}_{X,\,Y}\big) \\
&= -2\sigma^2 \log \mathbb{P}_X\bigg[\exp\Big(-\frac{1}{2\sigma^2}\min_{j\in[\![1,k]\!]} \|X - c_j\|^2\Big)\bigg] \\
&\geq 2\sigma^2 \mathbb{P}_X\bigg[1 - \exp\Big(-\frac{1}{\sigma^2}\min_{j\in[\![1,k]\!]} \|X - c_j\|^2\Big)\bigg] \overset{\text{def}}{=} \mathcal{C}(c).
\end{aligned}$$

Robust Lloyd's algorithm

An exponential weights update scheme

For any $c \in H^k$, define updated centers $c' \in H^k$ as

$$c'_j = \frac{\mathbb{P}_{X \mid \ell_c(X)=j}\left[X \exp\left(-\frac{1}{2\sigma^2}\|X - c_j\|^2\right)\right]}{\mathbb{P}_{X \mid \ell_c(X)=j}\left[\exp\left(-\frac{1}{2\sigma^2}\|X - c_j\|^2\right)\right]},$$

where $\ell_c(x) = \min\left\{\arg\min_{j \in [\![1,k]\!]} \|x - c_j\|\right\}$. It is such that $\mathcal{C}(c') \leq \mathcal{C}(c)$. More precisely

$$-\log\left(1 - \frac{1}{2\sigma^2}\mathcal{C}(c')\right) \leq -\log\left(1 - \frac{1}{2\sigma^2}\mathcal{C}(c)\right) - \frac{1}{2\sigma^2} Q_X^*\left(\|c'_{\ell_c(X)} - c_{\ell_c(X)}\|^2\right),$$

where $\dfrac{\mathrm{d}Q_X^*}{\mathrm{d}\mathbb{P}_X} = Z^{-1} \exp\left(-\dfrac{1}{2\sigma^2}\|X - c_{\ell_c(X)}\|^2\right)$.

A linear interpretation of the robust criterion

Using the kernel trick

There is a mapping $\Psi : H \longrightarrow \mathcal{H}$ another separable Hilbert space, such that

$$\exp\left(-\frac{1}{2\sigma^2}\|x-y\|^2\right) = \langle \Psi(x), \Psi(y) \rangle_{\mathcal{H}}, \qquad x, y \in H.$$

Putting $\theta_j = -\Psi(c_j)$ and $W = \Psi(X)$, we obtain that

$$\mathcal{C}(c) = 2\sigma^2 \left[1 + \mathbb{P}_W \left(\min_{j \in [\![1,k]\!]} \langle \theta_j, W \rangle_{\mathcal{H}} \right) \right],$$

where θ_j and W belong to the unit sphere of $\mathcal{H}$.

PAC-Bayesian bound for some linear k-means criterion

Observable upper bound

Let $W \in H$ be a random vector in a separable Hilbert space and let $(W_1, \dots, W_n)$ be n independent copies of W. Let $\Theta \in H^k$ be a bounded measurable set of parameters. Let $\|\Theta\| = \sup\left\{ \left(\sum_{j=1}^{k} \|\theta_j\|^2 \right)^{1/2} \;:\; \theta \in \Theta \right\} < \infty$.

Assume that $\mathbb{P}_W\left(\min_{j \in [\![1,k]\!]} \langle \theta_j, W \rangle \in [a, b] \text{ for all } \theta \in \Theta \right) = 1$ and that $\|W\|_\infty \overset{\text{def}}{=} \operatorname{ess\,sup}_{\mathbb{P}_W} \|W\| < \infty$. For any $k \geq 2$, any $n \geq 2k$ and any $\delta \in]0, 1[$, w. p. at least $1 - \delta$, for any $\theta \in \Theta$,

$$(\mathbb{P}_W - \overline{\mathbb{P}}_W)\left(\min_{j \in [\![1,k]\!]} \langle \theta_j, W \rangle \right) \leq \left(\frac{\log(n/k)}{\log(2)} \sqrt{\frac{8 \log(k)}{n}} + 2\sqrt{\frac{\log(k)}{n}} \right) \|\Theta\| \, \|W\|_\infty$$

$$+ \sqrt{\frac{(\sqrt{2} + 1)\left(k(b-a)^2 + 2\log(ek)\|W\|_\infty^2 \|\Theta\|^2\right)}{n}} + \sqrt{\frac{\log(\delta^{-1})}{2n}}(b - a).$$

PAC-Bayesian bound for some linear k-means criterion

Excess risk upper bound

For any $\theta^* \in \Theta$, w. p. at least $1-\delta$, for any $\theta \in \Theta$,

$$
\begin{aligned}
&\big(\mathbb{P}_W - \overline{\mathbb{P}}_W\big)\Big(\min_{j\in[\![1,k]\!]} \langle \theta_j, W\rangle - \min_{j\in[\![1,k]\!]} \langle \theta_j^*, W\rangle\Big) \\
&\qquad \leq B_n \stackrel{\text{def}}{=} \left(\frac{\log(n/k)}{\log(2)}\sqrt{\frac{8\log(k)}{n}} + 2\sqrt{\frac{\log(k)}{n}}\right)\|\Theta\|\,\|W\|_\infty \\
&\quad + \sqrt{\frac{(\sqrt{2}+1)\big(k(b-a)^2 + 2\log(ek)\|W\|_\infty^2\|\Theta\|^2\big)}{n}} + \sqrt{\frac{2\log(\delta^{-1})}{n}}(b-a).
\end{aligned}
$$

PAC-Bayesian bound for some linear k-means criterion

Consequences for ϵ minimizers

Assume that the estimator $\widehat{\theta}(W_1, \dots, W_n) \in \Theta$ is such that

$$\overline{\mathbb{P}}_W\Big(\min_{j \in [\![1,k]\!]} \langle \widehat{\theta}_j, W \rangle\Big) \leq \inf_{\theta \in \Theta} \overline{\mathbb{P}}_W\Big(\min_{j \in [\![1,k]\!]} \langle \theta_j, W \rangle\Big) + \epsilon, \quad \mathbb{P}_{W_1,\dots,W_n} \text{ a. s.}$$

W. p. at least $1 - \delta$

$$\mathbb{P}_W\Big(\min_{j \in [\![1,k]\!]} \langle \widehat{\theta}_j, W \rangle - \inf_{\theta \in \Theta} \mathbb{P}_W\Big(\min_{j \in [\![1,k]\!]} \langle \theta_j, W \rangle\Big) \leq B_n + \epsilon.$$

PAC-Bayesian bound for some linear k-means criterion

Expected risk of ϵ minimizers

$$
\begin{aligned}
\mathbb{P}_{W_1,\dots,W_n}&\left[\mathbb{P}_W\left(\min_{j\in[\![1,k]\!]}\langle\widehat{\theta}_j, W\rangle\right) - \inf_{\theta\in\Theta}\mathbb{P}_W\left(\min_{j\in[\![1,k]\!]}\langle\theta_j, W\rangle\right)\right] \\
&\leq \left(\frac{\log(n/k)}{\log(2)}\sqrt{\frac{8\log(k)}{n}} + 2\sqrt{\frac{\log(k)}{n}}\right)\|\Theta\|\,\|W\|_\infty \\
&\qquad + \sqrt{\frac{(\sqrt{2}+1)\big(k(b-a)^2 + 2\log(ek)\|W\|_\infty^2\|\Theta\|^2\big)}{n}} + \epsilon.
\end{aligned}
$$

Consequences for the robust k-means criterion

Uniform deviations of the empirical robust criterion

Let $\overline{\mathcal{C}}(c) = 2\sigma^2 \overline{\mathbb{P}}_X \left[1 - \exp\left(-\frac{1}{2\sigma^2} \min_{j \in [\![1,k]\!]} \|X - c_j\|^2 \right) \right], \quad c \in H^k.$

For any $k \geq 2$, any $n \geq 2k$, any $\delta \in]0,1[$, w. p. at least $1-\delta$, for any $c \in H^k$,

$$\begin{aligned}
\mathcal{C}(c) - \overline{\mathcal{C}}(c) \leq 2\sigma^2 \Bigg(& \frac{\log(n/k)}{\log(2)} \sqrt{\frac{8k\log(k)}{n}} + 2\sqrt{\frac{k\log(k)}{n}} \\
& + \sqrt{\frac{(\sqrt{2}+1)k\big(3 + 2\log(k)\big)}{n}} + \sqrt{\frac{\log(\delta^{-1})}{2n}} \Bigg) \\
& = \sigma^2 \mathcal{O}\left(\log\left(\frac{n}{k}\right) \sqrt{\frac{k\log(k)}{n}} + \sqrt{\frac{\log(\delta^{-1})}{n}} \right).
\end{aligned}$$

Consequences for the robust k-means criterion

Excess risk bound

For any $c^* \in H^k$, w. p. at least $1 - \delta$, for any $c \in H^k$,

$$\mathcal{C}(c) - \mathcal{C}(c^*) - \overline{\mathcal{C}}(c) + \overline{\mathcal{C}}(c^*) \leq B_n \overset{\text{def}}{=} 2\sigma^2 \left(\frac{\log(n/k)}{\log(2)} \sqrt{\frac{8k\log(k)}{n}} + 2\sqrt{\frac{k\log(k)}{n}} + \sqrt{\frac{(\sqrt{2}+1)k\big(3 + 2\log(k)\big)}{n}} + \sqrt{\frac{2\log(\delta^{-1})}{n}} \right).$$

Consequences for the robust k-means criterion

Consequences for ϵ-minimizers

If $\widehat{c}(X_1, \ldots, X_n)$ is such that $\overline{\mathcal{C}}(\widehat{c}) \leq \inf_{c \in H^k} \overline{\mathcal{C}}(c) + \epsilon$, $\mathbb{P}_{X_1,\ldots,X_n}$ a.s., w. p. at least $1 - \delta$,

$$\mathcal{C}(\widehat{c}) - \inf_{c \in H^k} \mathcal{C}(c) \leq B_n + \epsilon.$$

In expectation,

$$\mathbb{P}_{X_1,\ldots,X_n}\big(\mathcal{C}(\widehat{c})\big) \leq \inf_{c \in H^k} \mathcal{C}(c) + 2\sigma^2 \left(\frac{\log(n/k)}{\log(2)} \sqrt{\frac{8k \log(k)}{n}} + 2\sqrt{\frac{k \log(k)}{n}} + \sqrt{\frac{(\sqrt{2}+1)k\big(3 + 2\log(k)\big)}{n}} \right) + \epsilon.$$

Consequences for the usual k-means criterion

In expectation

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

Assume that $\mathbb{P}\big(\|X\| \leq B\big) = 1.$

For any $k \geq 2$, any $n \geq 2k$, any estimator $\widehat{c} \in \arg\min_{c \in H^k} \overline{\mathcal{R}}(c)$,

$$\mathbb{P}_{X_1,\dots,X_n}\Big(\mathcal{R}(\,\widehat{c}\,) \Big) \leq \inf_{c \in H^k} \mathcal{R}(c) + 16\, B^2 \log\left(\frac{n}{k} \right) \sqrt{\frac{k \log(k)}{n}}.$$

Proof of the linear k-means bounds

Gaussian perturbations

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

  be a non bilinear but measurable extension of the scalar product from ℓ^2 to $\mathbb{R}^{\mathbb{N}}$.
- Introduce $f(\theta, w) = \min_{j \in [\![1, k]\!]} \langle \theta_j, w \rangle, \quad \theta \in (\mathbb{R}^{\mathbb{N}})^k, w \in \mathbb{R}^{\mathbb{N}}$
- and the centered loss function $\overline{f}(\theta, w) = f(\theta, w) - \mathbb{P}_W\big(f(\theta, W)\big)$.

Proof of the linear k-means bounds

PAC-Bayesian chaining

- Write

$$(\mathbb{P}_W - \overline{\mathbb{P}}_W) f(\theta, W) = \bigl(\mathbb{P}_W - \overline{\mathbb{P}}_W\bigr)\bigl(\delta_{\theta' \mid \theta} - \underbrace{\rho_{\theta' \mid \theta}}_{\text{small perturbation}}\bigr) f(\theta', W)$$
$$+ \sum_{q=1}^{p} \bigl(\mathbb{P}_W - \overline{\mathbb{P}}_W\bigr)\bigl(\rho^{2^{q-1}}_{\theta' \mid \theta} - \underbrace{\rho^{2^q}_{\theta' \mid \theta}}_{\text{chain of intermediate scales}}\bigr) f(\theta', W)$$
$$+ \bigl(\mathbb{P}_W - \overline{\mathbb{P}}_W\bigr) \underbrace{\rho^{2^p}_{\theta' \mid \theta}}_{\text{big perturbation}} f(\theta', W).$$

- Remark that

$$\bigl(\delta_{\theta' \mid \theta} - \rho_{\theta' \mid \theta}\bigr) f(\theta', W) = \rho_{\theta' \mid \theta}\Bigl(\min_{j \in [\![1,k]\!]} \langle \theta_j, W\rangle - \min_{j \in [\![1,k]\!]} \langle \theta'_j, W\rangle\Bigr)$$
$$\leq \rho_{\theta' \mid \theta}\Bigl(\max_{j \in [\![1,k]\!]} \underbrace{\langle \theta_j - \theta'_j, W\rangle}_{\text{Gaussian}/\rho}\Bigr) \leq \sqrt{2\log(k)}\beta \, \|W\|_\infty.$$

Proof of the linear k-means bounds

Chaining inequalities

- From the PAC-Bayesian inequality applied to $h(\theta, w) = \bigl(\delta_{\theta' \mid \theta} - \rho_{\theta' \mid \theta}\bigr)\overline{f}(\theta', w)$,

$$\mathbb{P}_{W_1,\dots,W_n}\Biggl\{\exp \sup_{\theta \in (\ell^2)^k} \Biggl[n\lambda\bigl(\mathbb{P}_W - \overline{\mathbb{P}}_W\bigr)\bigl(\rho_{\theta' \mid \theta} - \rho^2_{\theta' \mid \theta}\bigr) f(\theta', W) - n\rho_{\theta' \mid \theta}\biggl[\log\biggl(\mathbb{P}_W\Bigl[\exp\Bigl(-\lambda\bigl(\delta_{\theta'' \mid \theta'} - \rho_{\theta'' \mid \theta'}\bigr)\overline{f}(\theta'', W)\Bigr)\Bigr]\biggr)\biggr] - \frac{\|\theta\|^2}{2\beta^2}\Biggr]\Biggr\} \leq 1.$$

- This gives

$$\mathbb{P}_{W_1,\dots,W_n}\Biggl[\sup_{\theta \in \Theta} \bigl(\mathbb{P}_W - \overline{\mathbb{P}}_W\bigr)\bigl(\rho_{\theta' \mid \theta} - \rho^2_{\theta' \mid \theta}\bigr) f(\theta', W)\Biggr] \leq 4\lambda\beta^2 \log(k) \|W\|_\infty^2 + \frac{\|\Theta\|^2}{2n\lambda\beta^2} \underset{\lambda_{\text{opt}}}{=} \|W\|_\infty \|\Theta\| \sqrt{\frac{8\log(k)}{n}}.$$

Proof of the linear k-means bounds

Bounding the biggest perturbation

- Consider $\psi(x) = \begin{cases} \log\left(1 + x + x^2/2\right), & x \geq 0, \\ -\log(1 - x + x^2/2), & x \leq 0, \end{cases}$
- and $\widetilde{f}(\theta, W) = f(\theta, W) - \dfrac{a+b}{2}$.
- Remark that

$$\left(\mathbb{P}_W - \overline{\mathbb{P}}_W\right)\rho_{\theta' \mid \theta} f(\theta', W) = \\ \rho_{\theta' \mid \theta}\left[\mathbb{P}_W \widetilde{f}(\theta', W) - \overline{\mathbb{P}}_W\left(\lambda^{-1}\psi\left[\lambda \widetilde{f}(\theta', W)\right]\right)\right] \\ + \underbrace{\rho_{\theta' \mid \theta}\overline{\mathbb{P}}_W\left[\lambda^{-1}\psi\left[\lambda \widetilde{f}(\theta', W)\right] - \widetilde{f}(\theta', W)\right]}_{\substack{\leq \frac{\lambda}{2(1+\sqrt{2})}\left[(b-a)^2/4 + 2\log(ek)\|W\|_\infty^2 \beta^2\right] \\ \text{since } |x - \psi(x)| \leq \frac{x^2}{4(1+\sqrt{2})}, \quad x \in \mathbb{R}.}}.$$

Proof of the linear k-means bounds

PAC-Bayesian inequality with an influence function

- Take $h(\theta, w) = \lambda^{-1}\psi\left[\lambda\widetilde{f}(\theta, w)\right]$ to obtain

$$\mathbb{P}_{W_1,\dots,W_n}\Bigg\{\sup_{\theta\in\Theta}\exp\Bigg[-n\lambda\rho_{\theta' \mid \theta}\overline{\mathbb{P}}_W\Big(\lambda^{-1}\psi\left[\lambda\widetilde{f}(\theta', W)\right]\Big) - n\rho_{\theta' \mid \theta}\Bigg[\log\Big(\mathbb{P}_W\Big[\exp\Big(-\psi\left[\lambda\widetilde{f}(\theta', W)\right]\Big)\Big]\Big)\Bigg] - \frac{\|\theta\|^2}{2\beta^2}\Bigg]\Bigg\} \leq 1.$$

- Use $\psi(x) \leq \log\big(1 + x + x^2/2\big), \quad x \in \mathbb{R}$ to deduce

$$\mathbb{P}_{W_1,\dots,W_n}\Bigg\{\sup_{\theta\in\Theta}\rho_{\theta' \mid \theta}\Big[\mathbb{P}_W\Big(\widetilde{f}(\theta', W)\Big) - \overline{\mathbb{P}}_W\Big(\lambda^{-1}\psi\left[\lambda\widetilde{f}(\theta', W)\right]\Big)\Big]\Bigg\} \leq \lambda\Big[(b-a)^2/4 + 2\log(ek)\|W\|_\infty^2\beta^2\Big] + \frac{\|\Theta\|^2}{2n\lambda\beta^2}.$$

Proof of the linear k-means bounds

Putting all together

- For the biggest perturbation we get

$$\mathbb{P}_{W_1,\dots,W_n}\left\{\sup_{\theta\in\Theta}\left(\mathbb{P}_W-\overline{\mathbb{P}}_W\right)\rho^{2p}_{\theta'\mid\theta}f(\theta',W)\right\} \underset{\lambda_{\mathrm{opt}}}{\leq} \sqrt{\frac{(\sqrt{2}+1)\left(2^{-p}\beta^{-2}(b-a)^2+8\log(ek)\|W\|_\infty^2\right)\|\Theta\|^2}{4n}}.$$

- Putting all together

$$\mathbb{P}_{W_1,\dots,W_n}\left\{\sup_{\theta\in\Theta}(\mathbb{P}_W-\overline{\mathbb{P}}_W)f(\theta,W)\right\}\leq 2\sqrt{2\log(k)}\beta\|W\|_\infty + \sqrt{\frac{(\sqrt{2}+1)\left(2^{-p}\beta^{-2}(b-a)^2+8\log(ek)\|W\|_\infty^2\right)\|\Theta\|^2}{4n}} + p\|W\|_\infty\|\Theta\|\sqrt{\frac{8\log(k)}{n}}.$$

- Choose $\beta=\|\Theta\|/\sqrt{2n}$ and $p=\lfloor\log(n/k)/\log(2)\rfloor$.

Proof of the linear k-means bounds

Deviations

According to the bounded difference inequality, with probability at least $1 - \delta$

$$\sup_{\theta \in \Theta} \big(\mathbb{P}_W - \overline{\mathbb{P}}_W\big) f(\theta, W)$$

$$\leq \mathbb{P}_{W_1, \dots, W_n} \left\{ \sup_{\theta \in \Theta} \big(\mathbb{P}_W - \overline{\mathbb{P}}_W\big) f(\theta, W) \right\} + \sqrt{\frac{2 \log(\delta^{-1})}{n}} (b - a).$$

References

[AC21] Gautier Appert and Olivier Catoni. "New bounds for k-means and information k-means." In: *arXiv preprint arXiv:2101.05728* (2021). URL: https://arxiv.org/abs/2101.05728.

[Cat12] Olivier Catoni. "Challenging the empirical mean and empirical variance: A deviation study." In: *Ann. Inst. H. Poincaré Probab. Statist.* 48.4 (Nov. 2012), pp. 1148–1185. DOI: 10.1214/11-AIHP454. URL: https://doi.org/10.1214/11-AIHP454.

[CG17a] Olivier Catoni and Ilaria Giulini. "Dimension free PAC-Bayesian bounds for the estimation of the mean of a random vector." In: *the Nips 2017 Workshop : (Almost) 50 shades of Bayesian learning : PAC-Bayesian trends and insights*. 2017, pp. 1–4. eprint: https://bguedj.github.io/nips2017/pdf/PAC-Bayes_2017_paper_1.pdf. URL: https://bguedj.github.io/nips2017/50shadesbayesian.html.

[CG17b] Olivier Catoni and Ilaria Giulini. "Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression." In: *arXiv preprint arXiv:1712.02747* (2017). URL: https://arxiv.org/abs/1712.02747.