PAC-Bayes bounds

Pierre Alquier

ESSEC BUSINESS SCHOOL

Machine Learning Summer School
Okinawa Institute of Science and Technology, March 2024

OIST

Cergy (near Paris)

- Kamelia DAUDEL
- Olga KLOPP
- Marie KRATZ
- Guillaume CHEVILLON
- Roberto RENO
- Jeroen ROMBOUTS
- Vincenzo ESPOSITO VINZI
- Mohamed NDAOUD
- Guillaume LECUE
- Pierre JACOB

Singapore

- Jeremy HENG
- Pierre ALQUIER

Information :

- contact : alquier@essec.edu
- webpage : https ://pierrealquier.github.io/

This lecture will be based on :

Alquier, P. (2024). User-friendly Introduction to PAC-Bayes bounds. *Foundations and Trends© in Machine Learning*.

(link to preliminary arXiv version + slides on my webpage).

Many thanks to Richard Cariño III who helped with the drawings !

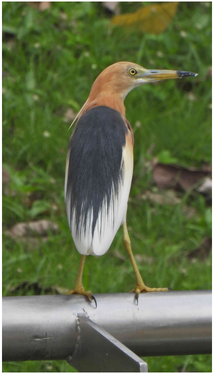
1. PAC-Bayes bounds : introduction
   - Generalization bounds and PAC-Bayes
   - Minimization of the PAC-Bayes bound
   - A zoo of PAC-Bayes bounds

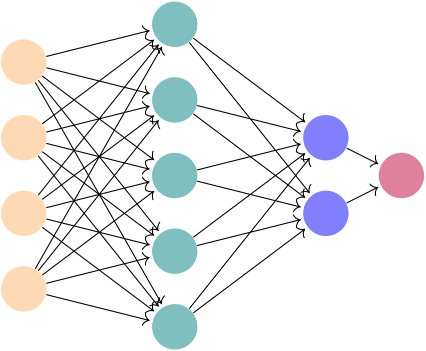
2. PAC-Bayes and Mutual Information bounds
   - Excess risk bounds
   - Fast rates
   - Mutual information bounds

- Objects $x \in \mathcal{X}$, labels $y \in \mathcal{Y}$.

- Objects $x \in \mathcal{X}$, labels $y \in \mathcal{Y}$.
- Predictor : function $f_\theta : \mathcal{X} \to \mathcal{Y}$ indexed by $\theta \in \Theta$.

- Objects $x \in \mathcal{X}$, labels $y \in \mathcal{Y}$.
- Predictor : function $f_\theta : \mathcal{X} \to \mathcal{Y}$ indexed by $\theta \in \Theta$.





- dog
- shark
- heron
- cat
- ...

- Prediction error measured through loss function ℓ :

$$\ell\Big(y, f_\theta(x)\Big).$$

- Prediction error measured through loss function ℓ :

$$\ell\Big(y, f_\theta(x)\Big).$$

- Risk :

$$R(\theta) := \mathbb{E}_{(X,Y)\sim P}\left[\ell\Big(Y, f_\theta(X)\Big)\right].$$

where P is the probability distribution of pairs object-label we want to learn to classify.

- Prediction error measured through loss function ℓ :

$$\ell\Big(y, f_\theta(x)\Big).$$

- Risk :

$$R(\theta) := \mathbb{E}_{(X,Y)\sim P}\Big[\ell\Big(Y, f_\theta(X)\Big)\Big].$$

where P is the probability distribution of pairs object-label we want to learn to classify.

- Objective :

$$R^* = \inf_{\theta\in\Theta} R(\theta).$$

- Prediction error measured through loss function ℓ :

$$\ell\Big(y, f_\theta(x)\Big).$$

- Risk :

$$R(\theta) := \mathbb{E}_{(X,Y)\sim P}\left[\ell\Big(Y, f_\theta(X)\Big)\right].$$

  where P is the probability distribution of pairs object-label we want to learn to classify.

- Objective :

$$R^* = \inf_{\theta\in\Theta} R(\theta).$$

- Data $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. from P. Empirical risk :

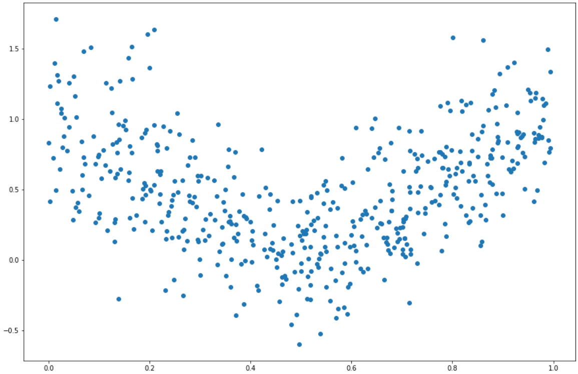
$$R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell\Big(Y_i, f_\theta(X_i)\Big).$$

Toy example :

- X uniform on $[0, 1]$,
- $Y = |2X - 1| + \epsilon$.

Toy example :

- X uniform on $[0, 1]$,
- $Y = |2X - 1| + \epsilon$.



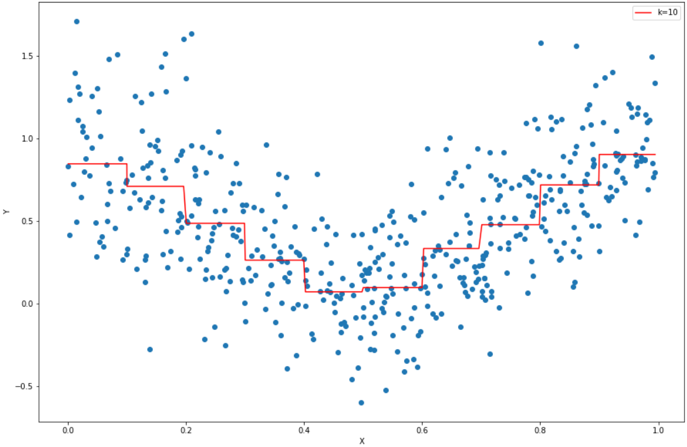
- Prediction by regular histogram with k-bins.
- $\ell(y, f_\theta(x)) = (y - f_\theta(x))^2$.

- Prediction by regular histogram with k-bins.
- $\ell(y, f_\theta(x)) = (y - f_\theta(x))^2$.

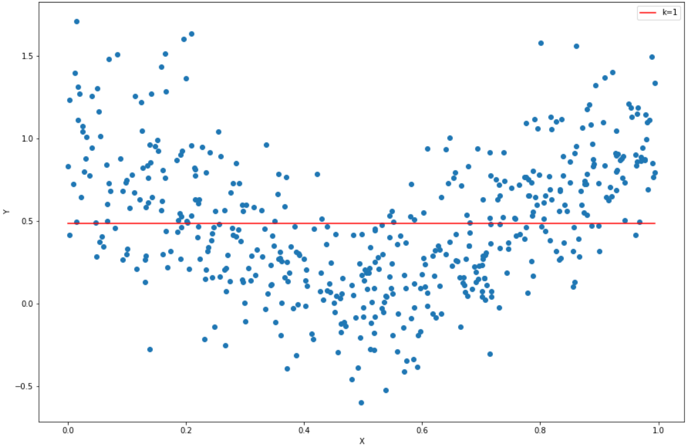
That is,

$$f_\theta(x) = \begin{cases} \theta_1 \text{ if } x \in [0, 1/k), \\ \theta_2 \text{ if } x \in [1/k, 2/k), \\ \vdots \\ \theta_k \text{ if } x \in [(k-1)/k, 1]. \end{cases}$$

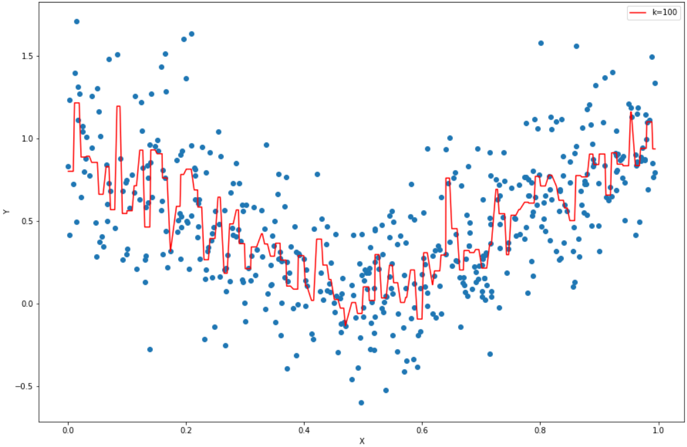
- Prediction by regular histogram with k-bins.
- $\ell(y, f_\theta(x)) = (y - f_\theta(x))^2$.

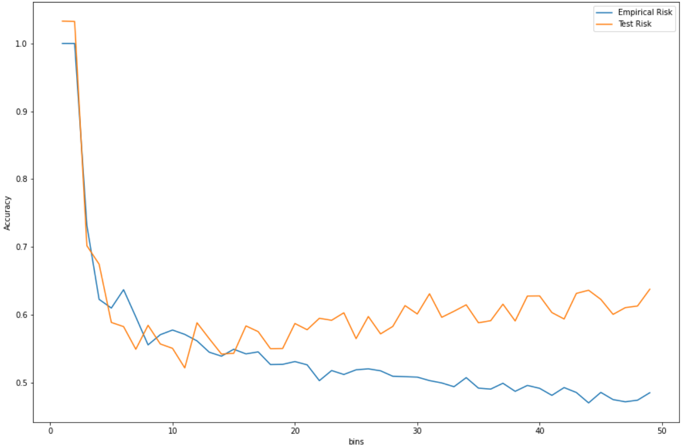


- Prediction by regular histogram with k-bins.
- $\ell(y, f_\theta(x)) = (y - f_\theta(x))^2$.



- Prediction by regular histogram with k-bins.
- $\ell(y, f_\theta(x)) = (y - f_\theta(x))^2$.




Empirical Risk
Test Risk
1.0
0.9
0.8
0.7
0.6
0.5
Accuracy
0
10
20
30
40
50
bins

Law of large numbers : for a fixed θ,

$$R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell\Big(Y_i, f_\theta(X_i)\Big) \xrightarrow[n\to\infty]{} R(\theta).$$

Law of large numbers : for a fixed θ,

$$R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell\Big(Y_i, f_\theta(X_i) \Big) \xrightarrow[n\to\infty]{} R(\theta).$$

But $\hat{\theta} = \hat{\theta}((X_1, Y_1), \dots, (X_n, Y_n)) = \hat{\theta}(\mathcal{S})$ learnt from data.

Law of large numbers : for a fixed θ,

$$R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell\Big(Y_i, f_\theta(X_i)\Big) \xrightarrow[n\to\infty]{} R(\theta).$$

But $\hat{\theta} = \hat{\theta}((X_1, Y_1), \dots, (X_n, Y_n)) = \hat{\theta}(\mathcal{S})$ learnt from data.

Can we quantify $R(\hat{\theta}) - R_n(\hat{\theta})$ when $\hat{\theta}$ is learnt ?

Law of large numbers : for a fixed θ,

$$R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell\Big(Y_i, f_\theta(X_i) \Big) \xrightarrow[n\to\infty]{} R(\theta).$$

But $\hat{\theta} = \hat{\theta}((X_1, Y_1), \dots, (X_n, Y_n)) = \hat{\theta}(\mathcal{S})$ learnt from data.

Can we quantify $R(\hat{\theta}) - R_n(\hat{\theta})$ when $\hat{\theta}$ is learnt ?

Various approaches :

Law of large numbers : for a fixed θ,

$$R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell\Big(Y_i, f_\theta(X_i)\Big) \xrightarrow[n\to\infty]{} R(\theta).$$

But $\hat{\theta} = \hat{\theta}((X_1, Y_1), \dots, (X_n, Y_n)) = \hat{\theta}(\mathcal{S})$ learnt from data.

Can we quantify $R(\hat{\theta}) - R_n(\hat{\theta})$ when $\hat{\theta}$ is learnt ?

Various approaches :

- Vapnik-Chervonenkis theory,

Law of large numbers : for a fixed θ,

$$R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell\Big(Y_i, f_\theta(X_i)\Big) \xrightarrow[n\to\infty]{} R(\theta).$$

But $\hat{\theta} = \hat{\theta}((X_1, Y_1), \dots, (X_n, Y_n)) = \hat{\theta}(\mathcal{S})$ learnt from data.

Can we quantify $R(\hat{\theta}) - R_n(\hat{\theta})$ when $\hat{\theta}$ is learnt ?

Various approaches :

- Vapnik-Chervonenkis theory,
- algorithmic stability,

Law of large numbers : for a fixed θ,

$$R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell\Big(Y_i, f_\theta(X_i)\Big) \xrightarrow[n\to\infty]{} R(\theta).$$

But $\hat{\theta} = \hat{\theta}((X_1, Y_1), \dots, (X_n, Y_n)) = \hat{\theta}(\mathcal{S})$ learnt from data.

Can we quantify $R(\hat{\theta}) - R_n(\hat{\theta})$ when $\hat{\theta}$ is learnt ?

Various approaches :

- Vapnik-Chervonenkis theory,
- algorithmic stability,
- information bounds : MDL, PAC-Bayes, etc.

Information we can get from these theories :

Information we can get from these theories :

1. "generalization bound"

$$R(\hat{\theta}) \leq R_n(\hat{\theta}) + \text{ data-dependent terms.}$$

Information we can get from these theories :

1. "generalization bound"

$$R(\hat{\theta}) \leq R_n(\hat{\theta}) + \text{ data-dependent terms.}$$

2. "excess risk bound"

$$R(\hat{\theta}) \leq R^* + \text{ rate of convergence.}$$

Assumption for whole lecture

Unless specified otherwise, $0 \leq \ell \leq 1$ and data is i.i.d. from P.

Assumption for whole lecture

Unless specified otherwise, $0 \leq \ell \leq 1$ and data is i.i.d. from P.

Vapnik-Chervonenkis – classification ($\mathcal{Y} = \{0, 1\}$)

With probability at least $1 - \delta$ on the data, for any $\hat{\theta}$ learnt from the data,

$$R(\hat{\theta}) \leq R_n(\hat{\theta}) + \sqrt{\frac{8d \log\left(\frac{2en}{d}\right) + 8 \log\left(\frac{4}{\delta}\right)}{n}}$$

where d : the VC-dimension of the set of classifiers $(f_\theta, \theta \in \Theta)$.

Statistical estimation / ERM etc.

Statistical estimation / ERM etc.

$$\text{data} \longrightarrow \text{estimator}$$

$$(\mathcal{X} \times \mathcal{Y})^n \longrightarrow \Theta$$

$$\mathcal{S} \longmapsto \hat{\theta} = \hat{\theta}(\mathcal{S})$$

Statistical estimation / ERM etc.

$$\text{data} \longrightarrow \text{estimator}$$

$$(\mathcal{X} \times \mathcal{Y})^n \longrightarrow \Theta$$

$$\mathcal{S} \longmapsto \hat{\theta} = \hat{\theta}(\mathcal{S})$$

Randomized estimators :

Statistical estimation / ERM etc.

$$\text{data} \longrightarrow \text{estimator}$$

$$(\mathcal{X} \times \mathcal{Y})^n \longrightarrow \Theta$$

$$\mathcal{S} \longmapsto \hat{\theta} = \hat{\theta}(\mathcal{S})$$

Randomized estimators :

$$(\mathcal{X} \times \mathcal{Y})^n \longrightarrow \mathcal{M}(\Theta) \dashrightarrow \Theta$$

$$\mathcal{S} \longmapsto \hat{\rho} = \hat{\rho}(\mathcal{S}) \overset{\theta \sim \hat{\rho}}{\dashrightarrow} \theta$$

- Randomized estimator inspired by Bayesian statistics, but it is a more general notion.

- Randomized estimator inspired by Bayesian statistics, but it is a more general notion.
- For each new pair object-label $(x, y) \sim P$, we can draw a predictor $\theta \sim \hat{\rho}$. We incur a loss $\ell(y, f_\theta(x))$.

- Randomized estimator inspired by Bayesian statistics, but it is a more general notion.
- For each new pair object-label $(x, y) \sim P$, we can draw a predictor $\theta \sim \hat{\rho}$. We incur a loss $\ell(y, f_\theta(x))$.
- If we repeat this for each new object to classify, our average loss will converge to

$$\mathbb{E}_{\theta\sim\hat{\rho}}\mathbb{E}_{(x,y)\sim P}\ell(y, f_\theta(x)) = \mathbb{E}_{\theta\sim\hat{\rho}}[R(\theta)].$$

McAllester's PAC-Bayes bound

Fix a prior distribution $\pi \in \mathcal{M}(\Theta)$. With probability at least $1-\delta$ on the data $\mathcal{S}$, for any probability distribution ρ learnt on the data,

$$\mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \mathbb{E}_{\theta\sim\rho}[R_n(\theta)] + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{2n}}.$$

McAllester's PAC-Bayes bound

Fix a prior distribution $\pi \in \mathcal{M}(\Theta)$. With probability at least $1-\delta$ on the data $\mathcal{S}$, for any probability distribution ρ learnt on the data,

$$\mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \mathbb{E}_{\theta\sim\rho}[R_n(\theta)] + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{2n}}.$$

$\mathrm{KL}(\rho\|\pi) =$ Küllback-Leibler divergence between ρ and π

McAllester's PAC-Bayes bound

Fix a prior distribution $\pi \in \mathcal{M}(\Theta)$. With probability at least $1-\delta$ on the data $\mathcal{S}$, for any probability distribution ρ learnt on the data,

$$\mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \mathbb{E}_{\theta\sim\rho}[R_n(\theta)] + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{2n}}.$$

$\mathrm{KL}(\rho\|\pi)$ = Küllback-Leibler divergence between ρ and π

- ρ can be learnt on the data, so if we have a randomized estimator $\hat{\rho}$ in mind, we can apply the bound to $\rho = \hat{\rho}$.

McAllester's PAC-Bayes bound

Fix a prior distribution $\pi \in \mathcal{M}(\Theta)$. With probability at least $1-\delta$ on the data $\mathcal{S}$, for any probability distribution ρ learnt on the data,

$$\mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \mathbb{E}_{\theta\sim\rho}[R_n(\theta)] + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{2n}}.$$

$\mathrm{KL}(\rho\|\pi) =$ Küllback-Leibler divergence between ρ and π

- ρ can be learnt on the data, so if we have a randomized estimator $\hat{\rho}$ in mind, we can apply the bound to $\rho = \hat{\rho}$.
- we will see later that the bound is helpful to define good randomized estimators $\hat{\rho}$.

$\mathrm{KL}(\rho\|\pi)$ = Küllback-Leibler divergence between ρ and π

- discrete case :

$$\mathrm{KL}(\rho\|\pi) = \sum_{\theta\in\Theta} \rho(\theta) \log \frac{\rho(\theta)}{\pi(\theta)}$$

and $\mathrm{KL}(\rho\|\pi) = \infty$ if for some θ, $\pi(\theta) = 0$ and $\rho(\theta) > 0$.

$\mathrm{KL}(\rho\|\pi) =$ Küllback-Leibler divergence between ρ and π

- discrete case :

$$\mathrm{KL}(\rho\|\pi) = \sum_{\theta\in\Theta} \rho(\theta) \log \frac{\rho(\theta)}{\pi(\theta)}$$

  and $\mathrm{KL}(\rho\|\pi) = \infty$ if for some θ, $\pi(\theta) = 0$ and $\rho(\theta) > 0$.

- general case

$$\mathrm{KL}(\rho\|\pi) = \mathbb{E}_{\theta\sim\rho}\left[\log \frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta)\right]$$

  and $\mathrm{KL}(\rho\|\pi) = \infty$ is ρ has no density $\frac{\mathrm{d}\rho}{\mathrm{d}\pi}$ w.r.t. π...

$\mathrm{KL}(\rho\|\pi)$ = Küllback-Leibler divergence between ρ and π

- discrete case :

$$\mathrm{KL}(\rho\|\pi) = \sum_{\theta\in\Theta} \rho(\theta) \log \frac{\rho(\theta)}{\pi(\theta)}$$

  and $\mathrm{KL}(\rho\|\pi) = \infty$ if for some θ, $\pi(\theta) = 0$ and $\rho(\theta) > 0$.

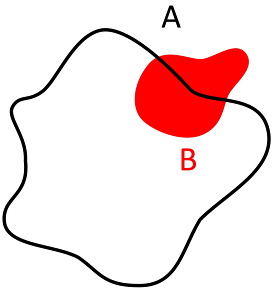
- general case

$$\mathrm{KL}(\rho\|\pi) = \mathbb{E}_{\theta\sim\rho}\left[\log \frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta)\right]$$

  and $\mathrm{KL}(\rho\|\pi) = \infty$ is ρ has no density $\frac{\mathrm{d}\rho}{\mathrm{d}\pi}$ w.r.t. π...

$\mathrm{KL}(\rho\|\pi) \geq 0$ and $\mathrm{KL}(\rho\|\pi) = 0 \Leftrightarrow \rho = \pi$.

Intuition on KL :



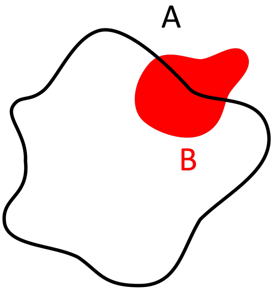
- π uniform on A

$$\pi(\theta) = \frac{1_A(\theta)}{\mathcal{V}(A)}$$

- ρ uniform on B

$$\rho(\theta) = \frac{1_B(\theta)}{\mathcal{V}(B)}$$

Intuition on KL :



- π uniform on A

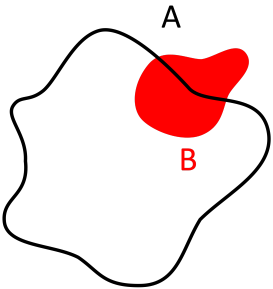
$$\pi(\theta) = \frac{1_A(\theta)}{\mathcal{V}(A)}$$

- ρ uniform on B

$$\rho(\theta) = \frac{1_B(\theta)}{\mathcal{V}(B)}$$

$\frac{\mathrm{d}\rho}{\mathrm{d}\pi}$ not defined here.

Intuition on KL :



- π uniform on A

$$\pi(\theta) = \frac{1_A(\theta)}{\mathcal{V}(A)}$$

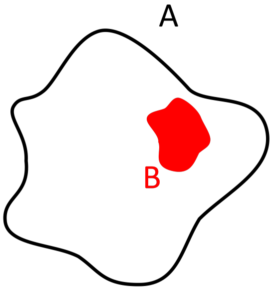
- ρ uniform on B

$$\rho(\theta) = \frac{1_B(\theta)}{\mathcal{V}(B)}$$

$\frac{\mathrm{d}\rho}{\mathrm{d}\pi}$ not defined here.

$$B \not\subseteq A \Rightarrow \mathrm{KL}(\rho\|\pi) = +\infty.$$

Intuition on KL :



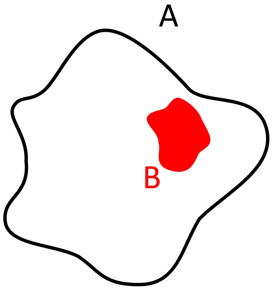
- π uniform on A

$$\pi(\theta) = \frac{1_A(\theta)}{\mathcal{V}(A)}$$

- ρ uniform on B

$$\rho(\theta) = \frac{1_B(\theta)}{\mathcal{V}(B)}$$

Intuition on KL :



- π uniform on A

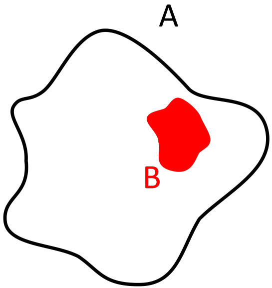
$$\pi(\theta) = \frac{1_A(\theta)}{\mathcal{V}(A)}$$

- ρ uniform on B

$$\rho(\theta) = \frac{1_B(\theta)}{\mathcal{V}(B)}$$

$$B \subseteq A \Rightarrow \frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta) = \frac{\mathcal{V}(A) 1_B(\theta)}{\mathcal{V}(B)}$$

Intuition on KL :



- π uniform on A

$$\pi(\theta) = \frac{1_A(\theta)}{\mathcal{V}(A)}$$

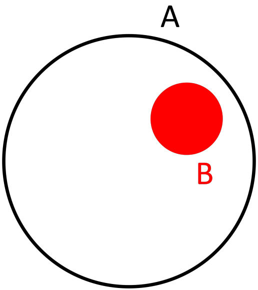
- ρ uniform on B

$$\rho(\theta) = \frac{1_B(\theta)}{\mathcal{V}(B)}$$

$$B \subseteq A \Rightarrow \frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta) = \frac{\mathcal{V}(A) 1_B(\theta)}{\mathcal{V}(B)}$$

$$\mathrm{KL}(\rho\|\pi) = \mathbb{E}_{\theta\sim\rho}\left[\log \frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta)\right] = \log \frac{\mathcal{V}(A)}{\mathcal{V}(B)}.$$

Intuition on KL :

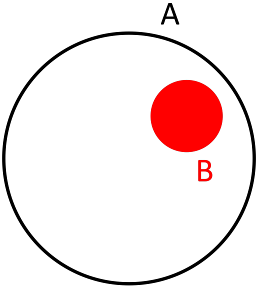


$B_d(x,r)$ ball centered on x, with radius r in $\mathbb{R}^d$

$$\mathcal{V}(B_d(x,r)) = \frac{r^d \pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)}$$

- π uniform on $A = B_d(0,C)$
- ρ uniform on $B = B_d(\theta_0,\epsilon)$

Intuition on KL :



$B_d(x,r)$ ball centered on x, with radius r in $\mathbb{R}^d$

$$\mathcal{V}(B_d(x,r)) = \frac{r^d \pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)}$$

- π uniform on $A = B_d(0, C)$
- ρ uniform on $B = B_d(\theta_0, \epsilon)$

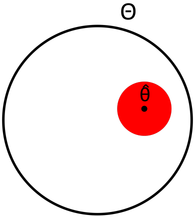
$$\mathrm{KL}(\rho \| \pi) = \log \frac{\mathcal{V}(A)}{\mathcal{V}(B)} = d \log \frac{C}{\epsilon}.$$

McAllester's PAC-Bayes bound

$$\mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \mathbb{E}_{\theta\sim\rho}[R_n(\theta)] + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{2n}}.$$

McAllester's PAC-Bayes bound

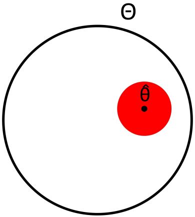
$$\mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \mathbb{E}_{\theta\sim\rho}[R_n(\theta)] + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{2n}}.$$



- π uniform on $\Theta = B_d(0, C)$
- $\rho = \hat{\rho}$ uniform on $B_d(\hat{\theta}, \epsilon)$

McAllester's PAC-Bayes bound

$$\mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \mathbb{E}_{\theta\sim\rho}[R_n(\theta)] + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{2n}}.$$



- π uniform on $\Theta = B_d(0, C)$
- $\rho = \hat{\rho}$ uniform on $B_d(\hat{\theta}, \epsilon)$

$$\mathbb{E}_{\theta\sim\hat{\rho}}[R(\theta)] \leq \mathbb{E}_{\theta\sim\hat{\rho}}[R_n(\theta)] + \sqrt{\frac{d\log\frac{C}{\epsilon} + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{2n}}.$$

Toy classification example :

- $X_i \in [-1, 1]$,

Toy classification example :

- $X_i \in [-1, 1]$,
- classifiers $(f_\theta)_{\theta \in [-1,1]}$ given by

$$f_\theta(x) = \begin{cases} 0 \text{ if } x \leq \theta \\ 1 \text{ if } x > \theta. \end{cases}$$

Toy classification example :

- $X_i \in [-1, 1]$,
- classifiers $(f_\theta)_{\theta \in [-1,1]}$ given by

$$f_\theta(x) = \begin{cases} 0 \text{ if } x \leq \theta \\ 1 \text{ if } x > \theta. \end{cases}$$

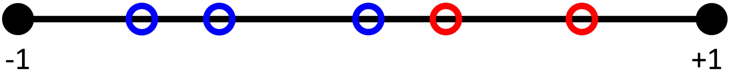
- $Y_i = f_{\theta^*}(X_i)$.

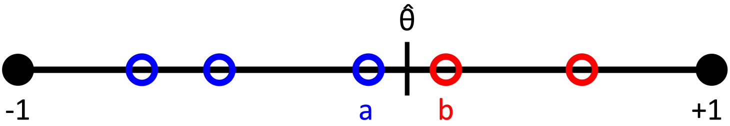
Toy classification example :

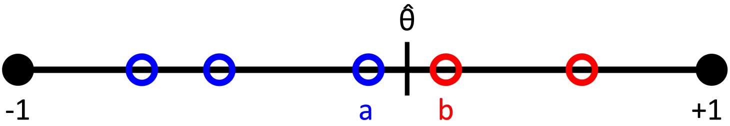
- $X_i \in [-1, 1]$,
- classifiers $(f_\theta)_{\theta \in [-1,1]}$ given by

$$f_\theta(x) = \begin{cases} 0 \text{ if } x \leq \theta \\ 1 \text{ if } x > \theta. \end{cases}$$

- $Y_i = f_{\theta^*}(X_i)$.

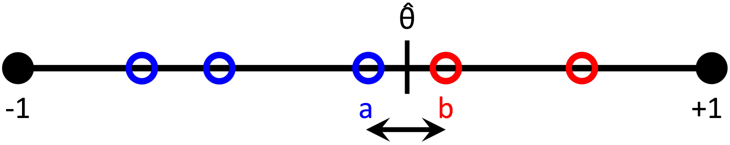


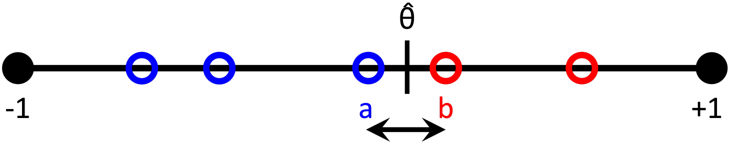

θ̂
-1
a
b
+1



Vapnik-type bound :

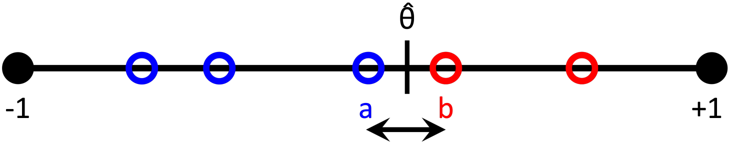
$$R(\hat{\theta}) \leq \sqrt{\frac{8 \log(2en) + 8 \log\left(\frac{4}{\delta}\right)}{n}}$$


θ̂
-1
a
b
+1



PAC-Bayes :

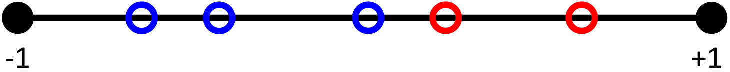
- π uniform on $[-1,1]$,
- $\hat{\rho}$ uniform on $[a,b]$.

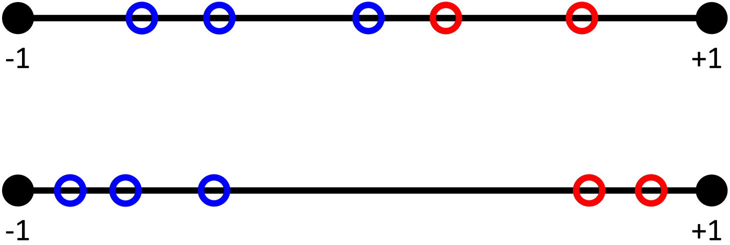


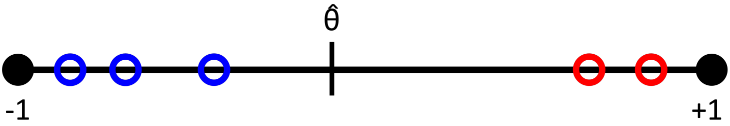
PAC-Bayes :

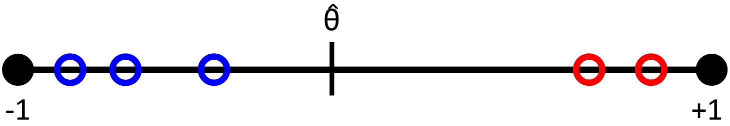
- π uniform on $[-1, 1]$,
- $\hat{\rho}$ uniform on $[a, b]$.

$$\mathbb{E}_{\theta\sim\hat{\rho}}[R(\theta)] \leq \sqrt{\frac{\log\frac{2}{b-a} + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{2n}}.$$


-1
+1

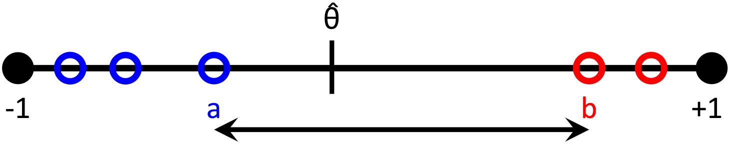

-1
+1
-1
+1

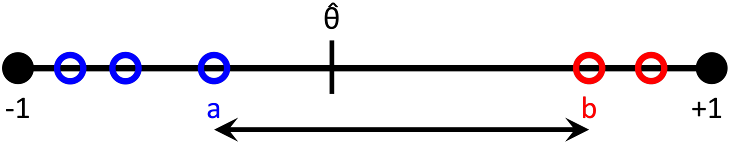

$\hat{\theta}$
-1
+1



Vapnik-type bound :

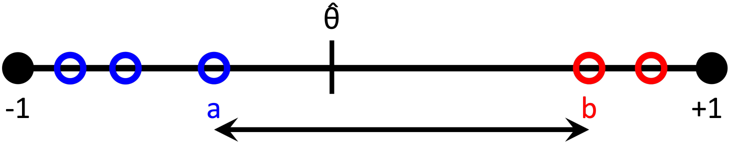
$$R(\hat{\theta}) \leq \sqrt{\frac{8\log\left(2\mathrm{e}n\right) + 8\log\left(\frac{4}{\delta}\right)}{n}}$$


θ̂
-1
a
b
+1



PAC-Bayes :

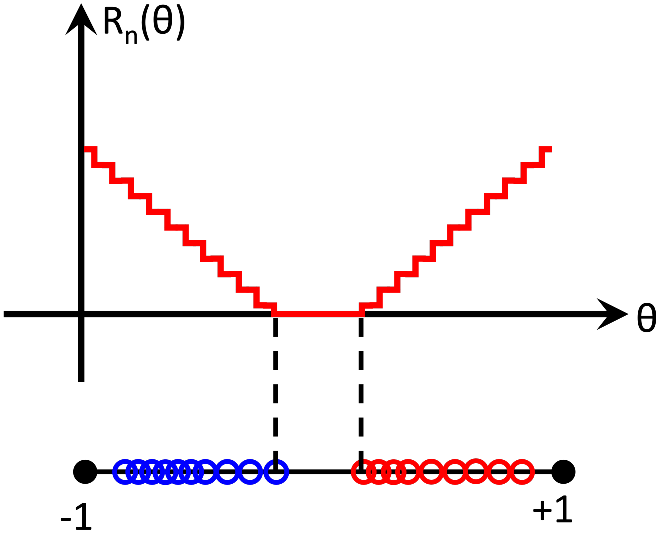
- π uniform on $[-1, 1]$,
- $\hat{\rho}$ uniform on $[a, b]$.

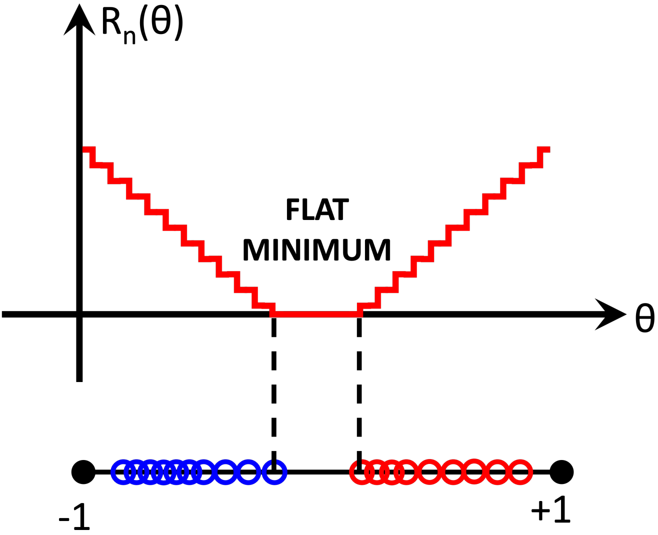


PAC-Bayes :

- π uniform on $[-1, 1]$,
- $\hat{\rho}$ uniform on $[a, b]$.

$$\mathbb{E}_{\theta\sim\hat{\rho}}[R(\theta)] \leq \sqrt{\frac{\log\frac{2}{b-a} + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{2n}}.$$


Rn(θ)
θ
-1
+1


$R_n(\theta)$
FLAT
MINIMUM
θ
-1
+1

1. PAC-Bayes bounds : introduction
   - Generalization bounds and PAC-Bayes
   - Minimization of the PAC-Bayes bound
   - A zoo of PAC-Bayes bounds

2. PAC-Bayes and Mutual Information bounds
   - Excess risk bounds
   - Fast rates
   - Mutual information bounds

McAllester's PAC-Bayes bound

Fix prior $\pi \in \mathcal{M}(\Theta)$. With proba. at least $1-\delta$, $\forall \rho \in \mathcal{M}(\Theta)$,

$$\mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \mathbb{E}_{\theta\sim\rho}[R_n(\theta)] + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{2n}}.$$

McAllester's PAC-Bayes bound

Fix prior $\pi \in \mathcal{M}(\Theta)$. With proba. at least $1-\delta$, $\forall \rho \in \mathcal{M}(\Theta)$,

$$\mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \mathbb{E}_{\theta\sim\rho}[R_n(\theta)] + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{2n}}.$$

$$\sqrt{\frac{a}{b}} = \inf_{\lambda>0}\left\{\frac{a}{\lambda} + \frac{\lambda}{4b}\right\}.$$

McAllester's PAC-Bayes bound

Fix prior $\pi \in \mathcal{M}(\Theta)$. With proba. at least $1-\delta$, $\forall \rho \in \mathcal{M}(\Theta)$,

$$\mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \mathbb{E}_{\theta\sim\rho}[R_n(\theta)] + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{2n}}.$$

$$\sqrt{\frac{a}{b}} = \inf_{\lambda>0}\left\{\frac{a}{\lambda} + \frac{\lambda}{4b}\right\}.$$

$$\mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \mathbb{E}_{\theta\sim\rho}[R_n(\theta)] + \inf_{\lambda>0}\left\{\frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{\lambda} + \frac{\lambda}{8n}\right\}.$$

$$\forall \lambda > 0,\ \mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \mathbb{E}_{\theta\sim\rho}[R_n(\theta)] + \frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{\lambda} + \frac{\lambda}{8n}.$$

$$\forall \lambda > 0, \mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \mathbb{E}_{\theta\sim\rho}[R_n(\theta)] + \frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{\lambda} + \frac{\lambda}{8n}.$$

Definition - Gibbs posterior

$$\hat{\pi}_\lambda(\mathrm{d}\theta) = \frac{\exp(-\lambda R_n(\theta))}{\mathbb{E}_{\vartheta\sim\pi}[\exp(-\lambda R_n(\vartheta))]}\pi(\mathrm{d}\theta).$$

$$\forall \lambda > 0,\ \mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \mathbb{E}_{\theta\sim\rho}[R_n(\theta)] + \frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{\lambda} + \frac{\lambda}{8n}.$$

Definition - Gibbs posterior

$$\hat{\pi}_\lambda(\mathrm{d}\theta) = \frac{\exp(-\lambda R_n(\theta))}{\mathbb{E}_{\vartheta\sim\pi}[\exp(-\lambda R_n(\vartheta))]}\pi(\mathrm{d}\theta).$$

Theorem

$$\hat{\pi}_\lambda = \operatorname*{arg\,min}_{\rho\in\mathcal{M}(\Theta)} \left\{ \mathbb{E}_{\theta\sim\rho}[R_n(\theta)] + \frac{\mathrm{KL}(\rho\|\pi)}{\lambda} \right\}.$$

Proof :

$$
\begin{aligned}
0 &\leq \mathrm{KL}(\rho \| \hat{\pi}_\lambda) \\
&= \mathbb{E}_{\theta\sim\rho}\left[\log \frac{\mathrm{d}\rho}{\mathrm{d}\hat{\pi}_\lambda}(\theta)\right] \\
&= \mathbb{E}_{\theta\sim\rho}\left[\log \frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta) - \log \frac{\mathrm{d}\hat{\pi}_\lambda}{\mathrm{d}\pi}(\theta)\right] \\
&= \mathbb{E}_{\theta\sim\rho}\left[\log \frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta) + \lambda R_n(\theta) + \log \mathbb{E}_{\vartheta\sim\pi}[\exp(-\lambda R_n(\vartheta)]\right] \\
&= \mathrm{KL}(\rho\|\pi) + \lambda \mathbb{E}_{\theta\sim\rho}[R_n(\theta)] + \log \mathbb{E}_{\vartheta\sim\pi}[\exp(-\lambda R_n(\vartheta)].
\end{aligned}
$$

Consequence of the PAC-Bayes bound

Fix prior $\pi \in \mathcal{M}(\Theta)$. With proba. at least $1-\delta$, $\forall \lambda > 0$,

$$\mathbb{E}_{\theta \sim \hat{\pi}_\lambda}[R(\theta)] \leq \frac{-\log \mathbb{E}_{\vartheta \sim \pi}[\exp(-\lambda R_n(\vartheta))] + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{\lambda} + \frac{\lambda}{8n}.$$

Consequence of the PAC-Bayes bound

Fix prior $\pi \in \mathcal{M}(\Theta)$. With proba. at least $1-\delta$, $\forall \lambda > 0$,

$$\mathbb{E}_{\theta \sim \hat{\pi}_\lambda}[R(\theta)] \leq \frac{-\log \mathbb{E}_{\vartheta \sim \pi}[\exp(-\lambda R_n(\vartheta))] + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{\lambda} + \frac{\lambda}{8n}.$$

- in simple cases, we can sample from $\hat{\pi}_\lambda$ by standard Monte Carlo / MCMC techniques,

Consequence of the PAC-Bayes bound

Fix prior $\pi \in \mathcal{M}(\Theta)$. With proba. at least $1-\delta$, $\forall \lambda > 0$,

$$\mathbb{E}_{\theta\sim\hat{\pi}_\lambda}[R(\theta)] \leq \frac{-\log \mathbb{E}_{\vartheta\sim\pi}[\exp(-\lambda R_n(\vartheta))] + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{\lambda} + \frac{\lambda}{8n}.$$

- in simple cases, we can sample from $\hat{\pi}_\lambda$ by standard Monte Carlo / MCMC techniques,
- the minimization in λ can be tricky.

Approximate minimization of the PAC-Bayes bound.

$$\forall \lambda > 0,\ \mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \mathbb{E}_{\theta\sim\rho}[R_n(\theta)] + \frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{\lambda} + \frac{\lambda}{8n}.$$

$$\forall \lambda > 0, \ \mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \mathbb{E}_{\theta\sim\rho}[R_n(\theta)] + \frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{\lambda} + \frac{\lambda}{8n}.$$

Alternative approach : optimize ρ in a smaller set $\mathcal{F} \subsetneq \mathcal{M}(\Theta)$.

$$\forall \lambda > 0,\ \mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \mathbb{E}_{\theta\sim\rho}[R_n(\theta)] + \frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{\lambda} + \frac{\lambda}{8n}.$$

Alternative approach : optimize ρ in a smaller set $\mathcal{F} \subsetneq \mathcal{M}(\Theta)$.

Definition - variational approximation of Gibbs posterior

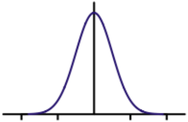
$$\tilde{\rho}_\lambda = \arg\min_{\rho\in\mathcal{F}} \left\{ \mathbb{E}_{\theta\sim\rho}[R_n(\theta)] + \frac{\mathrm{KL}(\rho\|\pi)}{\lambda} \right\}.$$

$$\forall \lambda > 0,\ \mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \mathbb{E}_{\theta\sim\rho}[R_n(\theta)] + \frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{\lambda} + \frac{\lambda}{8n}.$$

Alternative approach : optimize ρ in a smaller set $\mathcal{F} \subsetneq \mathcal{M}(\Theta)$.

Definition - variational approximation of Gibbs posterior

$$\tilde{\rho}_\lambda = \arg\min_{\rho\in\mathcal{F}} \left\{ \mathbb{E}_{\theta\sim\rho}[R_n(\theta)] + \frac{\mathrm{KL}(\rho\|\pi)}{\lambda} \right\}.$$



Example : $\rho = \mathcal{N}(\mu,\Sigma)$, optimize (μ,Σ).

Example : Gaussian prior π, and we optimize a Gaussian posterior ρ :

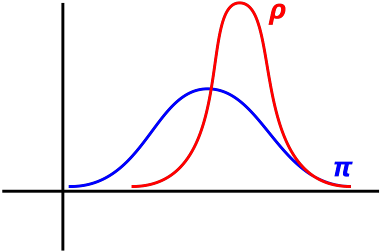
$$\pi = \mathcal{N}(\mu_0, \Sigma_0) \text{ and } \rho = \mathcal{N}(\mu_1, \Sigma_1) \text{ in } \mathbb{R}^d.$$

Example : Gaussian prior π, and we optimize a Gaussian posterior ρ :

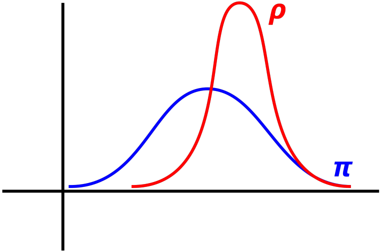
$$\pi = \mathcal{N}(\mu_0, \Sigma_0) \text{ and } \rho = \mathcal{N}(\mu_1, \Sigma_1) \text{ in } \mathbb{R}^d.$$

$$\begin{aligned} \mathrm{KL}(\rho \| \pi) = \frac{1}{2} \Bigg[& \mathrm{tr}(\Sigma_1 {\Sigma_0}^{-1}) - d \\ & + (\mu_1 - \mu_0)^T {\Sigma_0}^{-1} (\mu_1 - \mu_0) + \log \frac{\det \Sigma_0}{\det \Sigma_1} \Bigg]. \end{aligned}$$

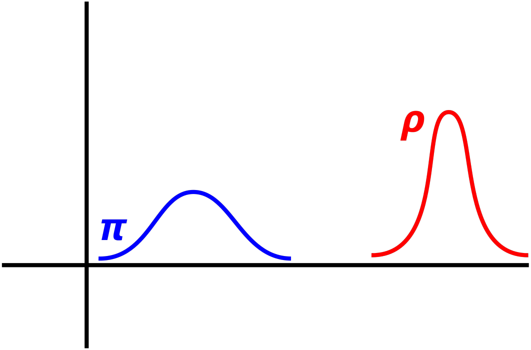
$\pi = \mathcal{N}(\mu_0, \Sigma_0)$ and $\rho = \mathcal{N}(\mu_1, \Sigma_1)$ in $\mathbb{R}$.

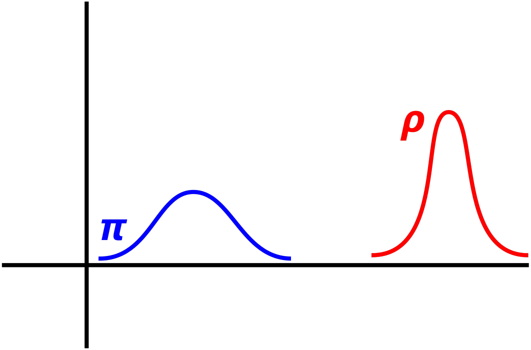


$\pi = \mathcal{N}(\mu_0, \Sigma_0)$ and $\rho = \mathcal{N}(\mu_1, \Sigma_1)$ in $\mathbb{R}$.



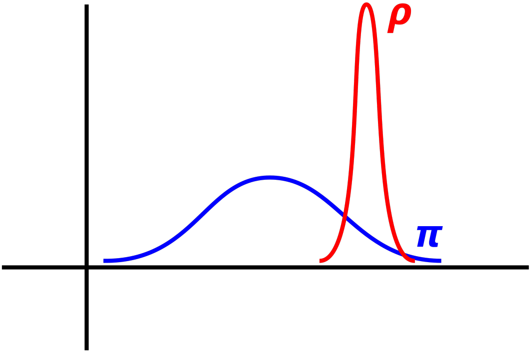
$$\mathrm{KL}(\rho\|\pi) = \frac{1}{2}\left[\frac{\Sigma_1}{\Sigma_0} - 1 + \frac{(\mu_0 - \mu_1)^2}{\Sigma_0} + \log\frac{\Sigma_0}{\Sigma_1}\right].$$

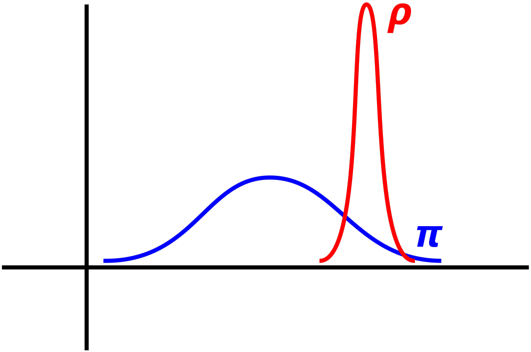

π
ρ



If μ_1 goes far away from μ_0 to ∞,

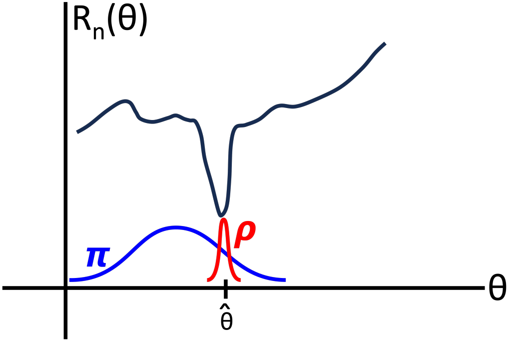
$$\mathrm{KL}(\rho \| \pi) \sim \frac{(\mu_0 - \mu_1)^2}{2\Sigma_0} \to \infty.$$


ρ
π



If $\Sigma_1 \to 0$,

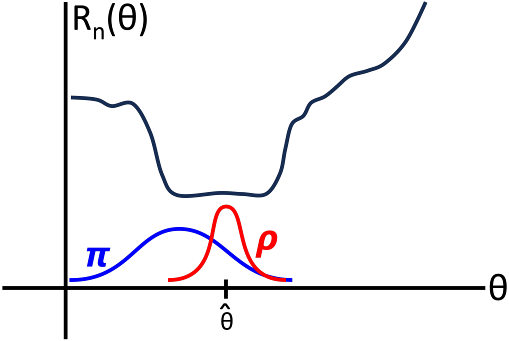
$$\mathrm{KL}(\rho \| \pi) \sim \frac{1}{2} \log \frac{\Sigma_0}{\Sigma_1} \to \infty.$$



With a sharp minimum, to keep

$$\mathbb{E}_{\theta\sim\mathcal{N}(\hat{\theta},\Sigma_1)}[R_n(\theta)] \sim R_n(\hat{\theta}),$$

Σ_1 should be small, and thus $\mathrm{KL}(\rho\|\pi)$ will be large.



With a flat minimum,

$$\mathbb{E}_{\theta\sim\mathcal{N}(\hat{\theta},\Sigma_1)}[R_n(\theta)] \sim R_n(\hat{\theta})$$

for Σ_1 "not so small", thus $\mathrm{KL}(\rho\|\pi)$ does not have to be large.

$$\rho = \rho_{\mu_1,\Sigma_1} = \mathcal{N}(\mu_1, \Sigma_1) = \mathcal{N}(\mu_1, UU^T).$$

$$\min_{\mu_1, U} \left\{ \mathbb{E}_{\theta\sim\mathcal{N}(\mu_1, UU^T)}[R_n(\theta)] + \frac{\mathrm{KL}(\rho\|\pi)}{\lambda} \right\}.$$

$$\rho = \rho_{\mu_1,\Sigma_1} = \mathcal{N}(\mu_1, \Sigma_1) = \mathcal{N}(\mu_1, UU^T).$$

$$\min_{\mu_1, U} \left\{ \mathbb{E}_{\theta\sim\mathcal{N}(\mu_1, UU^T)}[R_n(\theta)] + \frac{\mathrm{KL}(\rho\|\pi)}{\lambda} \right\}.$$

$$\mathbb{E}_{\theta\sim\mathcal{N}(\mu_1, UU^T)}[R_n(\theta)] = \mathbb{E}_{\xi\sim\mathcal{N}(0,I)}[R_n(\mu_1 + U\xi)].$$

$$\rho = \rho_{\mu_1,\Sigma_1} = \mathcal{N}(\mu_1, \Sigma_1) = \mathcal{N}(\mu_1, UU^T).$$

$$\min_{\mu_1, U} \left\{ \mathbb{E}_{\theta\sim\mathcal{N}(\mu_1, UU^T)}[R_n(\theta)] + \frac{\mathrm{KL}(\rho\|\pi)}{\lambda} \right\}.$$

$$\mathbb{E}_{\theta\sim\mathcal{N}(\mu_1, UU^T)}[R_n(\theta)] = \mathbb{E}_{\xi\sim\mathcal{N}(0,I)}[R_n(\mu_1 + U\xi)].$$

Stochastic Gradient Algorithm

Random initialization of μ_1 and U, then iterate :

- sample $\xi \sim \mathcal{N}(0, I)$,
- update

$$\begin{cases} \mu_1 \leftarrow \mu_1 - \eta \frac{\partial}{\partial \mu_1}[R_n(\mu_1 + U\xi) + \mathrm{KL}(\rho_{\mu_1,\Sigma_1}\|\pi)] \\ U \leftarrow U - \eta \frac{\partial}{\partial U}[R_n(\mu_1 + U\xi) + \mathrm{KL}(\rho_{\mu_1,\Sigma_1}\|\pi)] \end{cases}$$

Application : generalization bounds for deep learning.

Train a neural network for classification (0-1 loss).

Train a neural network for classification (0-1 loss).

Vapnik-type bound usually lead to something larger than 1, for example :

$$R(\hat{\theta}) \leq 35.4$$

Train a neural network for classification (0-1 loss).

Vapnik-type bound usually lead to something larger than 1, for example :

$$R(\hat{\theta}) \leq 35.4$$

As $R(\hat{\theta}) = \mathbb{P}(Y \neq f_{\hat{\theta}}(X))$, the bound brings no information (vacuous).

Train a neural network for classification (0-1 loss).

Vapnik-type bound usually lead to something larger than 1, for example :

$$R(\hat{\theta}) \leq 35.4$$

As $R(\hat{\theta}) = \mathbb{P}(Y \neq f_{\hat{\theta}}(X))$, the bound brings no information (vacuous).

Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data

Gintare Karolina Dziugaite
Department of Engineering
University of Cambridge

Daniel M. Roy
Department of Statistical Sciences
University of Toronto

Abstract

One of the defining properties of deep learning is that models are chosen to have many more parameters than available training data. In light of this capacity for overfitting, it is remarkable that simple algorithms like SGD reliably return solutions with low test error. One roadblock to explaining these phenomena in terms of implicit regularization, structural properties of the solution, and/or easiness of the data is that many learning bounds are quantitatively vacuous when applied to networks learned by SGD in this "deep learning" regime. Logically, in order to explain generalization, we need nonvacuous bounds. We return to an idea by Langford and Caruana (2001), who used PAC-Bayes bounds to compute nonvacuous numerical bounds on generalization error for *stochastic* two-layer two-hidden-unit neural networks via a sensitivity analysis. By optimizing the PAC-Bayes bound directly, we are able to extend their approach and obtain nonvacuous generalization bounds for deep stochastic neural network classifiers with millions of parameters trained on only tens of thousands of examples. We connect our findings to recent and old work on flat minima and MDL-based explanations of generalization.

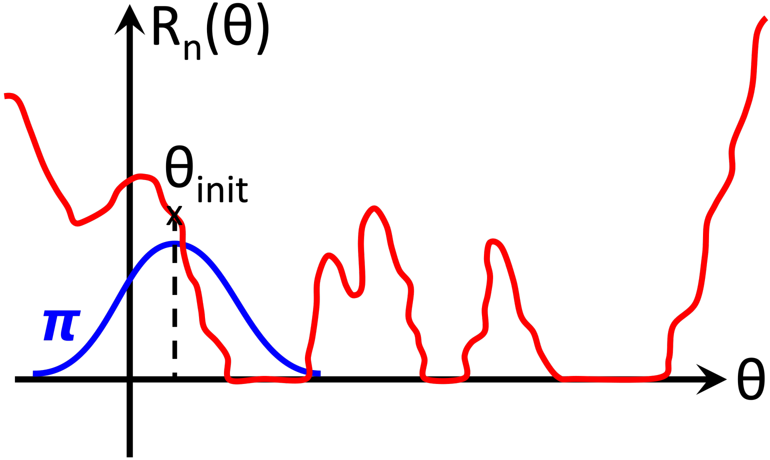
1 INTRODUCTION

By optimizing a PAC-Bayes bound, we show that it is possible to compute nonvacuous numerical bounds on the generalization error of deep *stochastic* neural networks with millions of parameters, despite the training data sets being one or more orders of magnitude smaller than the number of parameters. To our knowledge, these are the first explicit and nonvacuous numerical bounds computed for trained neural networks in the modern deep learning regime where the number of network parameters eclipses the number of training examples.

The bounds we compute are data dependent, incorporating millions of components optimized numerically to identify a large region in weight space with low average empirical error around the solution obtained by stochastic gradient descent (SGD). The data dependence is essential: indeed, the VC dimension of neural networks is typically bounded below by the number of parameters, and so one needs as many training data as parameters before (uniform) PAC bounds are nonvacuous, i.e., before the generalization error falls below 1. To put this in concrete terms, on MNIST, having even 72 hidden units in a fully connected first layer yields vacuous PAC bounds.

Evidently, we are operating far from the worst case: observed generalization cannot be explained in terms the regularizing effect of the size of the neural network alone. This is an old observation, and one that attracted considerable theoretical attention two decades ago: Bartlett [Bar97; Bar98] showed that, in large (sigmoidal) neural networks, when the learned weights are small in magnitude, the fat-shattering dimension is more important than the VC dimension for characterizing generalization. In particular, Bartlett established classification error bounds in terms of the empirical margin and the fat-shattering dimension, and then gave fat-shattering bounds for neural networks in terms of the *magnitudes* of the weights and the depth of the network alone. Improved norm-based bounds were obtained using Rademacher and Gaussian complexity by Bartlett and Mendelson [BM02] and Koltchinskii and Panchenko [KP02].

These norm-based bounds are the foundation of our current understanding of neural network generalization. It is widely accepted that these bounds explain observed generalization, at least "qualitatively" and/or when the weights are explicitly regularized. Indeed, recent work by Neyshabur, Tomioka, and Srebro [NTS14] puts forth


$R_n(\theta)$
θ_{init}
x
π
θ

Combine many ideas to get tighter bounds :

- prior centered at the (random) initialization of SGD, θ_{init}.

Combine many ideas to get tighter bounds :

- prior centered at the (random) initialization of SGD, $\theta_{\rm init}$.
- "multi-scale" prior :

$$\pi = \sum_{\sigma \in S} p(\sigma) \mathcal{N}(\theta_{\rm init}, \sigma^2 I).$$

Combine many ideas to get tighter bounds :

- prior centered at the (random) initialization of SGD, $\theta_{\rm init}$.
- "multi-scale" prior :

$$\pi = \sum_{\sigma \in S} p(\sigma) \mathcal{N}(\theta_{\rm init}, \sigma^2 I).$$

- replace $R_n(\theta)$ by convex surrogate.
- ...

Experiment	T-600	T-1200	T-300^2	T-600^2	T-1200^2	T-600^3	R-600
Train error	0.001	0.002	0.000	0.000	0.000	0.000	0.007
Test error	0.018	0.018	0.015	0.016	0.015	0.013	0.508
SNN train error	0.028	0.027	0.027	0.028	0.029	0.027	0.112
SNN test error	0.034	0.035	0.034	0.033	0.035	0.032	0.503
PAC-Bayes bound	0.161	0.179	0.170	0.186	0.223	0.201	1.352
KL divergence	5144	5977	5791	6534	8558	7861	201131
# parameters	471k	943k	326k	832k	2384k	1193k	472k
VC dimension	26m	56m	26m	66m	187m	121m	26m

Table 1: Results for experiments on binary class variant of MNIST. SGD is either trained on (T) true labels or (R) random labels. The network architecture is expressed as N^L, indicating L hidden layers with N nodes each. Errors are classification error. The reported VC dimension is the best known upper bound (in millions) for ReLU networks. The SNN error rates are tight upper bounds (see text for details). The PAC-Bayes bounds upper bound the test error with probability 0.965.

Results taken from :

Dzuigaite, G. K. and Roy, D. M. (2017). Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data. *UAI*.

More recent results (among others !) :

Pérez-Ortiz, M., Rivasplata, O., Shawe-Taylor, J. and Szepesvári, C. (2021). Tighter risk certificates for neural networks. *Journal of Machine Learning Research*.

Clerico, E., Farghly, T., Deligiannidis, G., Guedj, B. and Doucet, A. (2022). *Generalisation under gradient descent via deterministic PAC-Bayes*. ArXiv preprint arXiv :2209.02525.

1. PAC-Bayes bounds : introduction
   - Generalization bounds and PAC-Bayes
   - Minimization of the PAC-Bayes bound
   - A zoo of PAC-Bayes bounds

2. PAC-Bayes and Mutual Information bounds
   - Excess risk bounds
   - Fast rates
   - Mutual information bounds

PAC-Bayesian Model Averaging

David A. McAllester
AT&T Shannon Labs
180 Park Avenue
Florham Park, NJ 07932-0971
dmac@research.att.com

Abstract

PAC-Bayesian learning methods combine the informative priors of Bayesian methods with distribution-free PAC guarantees. Building on earlier methods for PAC-Bayesian model selection, this paper presents a method for PAC-Bayesian model averaging. The method constructs an optimized weighted mixture of concepts analogous to a Bayesian posterior distribution. Although the main result is stated for bounded loss, a preliminary analysis for unbounded loss is also given.

1 INTRODUCTION

A PAC-Bayesian approach to machine learning attempts to combine the advantages of both PAC and Bayesian approaches [12, 8]. The Bayesian approach has the advantage of using arbitrary domain knowledge in the form of a Bayesian prior. The PAC approach has the advantage that one can prove guarantees for generalization error without assuming the truth of the prior. A PAC-Bayesian approach combines the features of the PAC and Bayesian approaches — it bases the bias of the learning algorithm on an arbitrary prior distribution, thus allowing the incorporation of domain knowledge, and yet provides a guarantee on generalization error that is independent of any truth of the prior.

PAC-Bayesian approaches are related to structural risk minimization (SRM) [6]. Here we interpret this broadly as describing any learning algorithm optimizing a tradeoff between the "complexity", "structure", or "prior probability" of the concept or model and the "goodness of fit", "description length", or "likelihood" of the training data. Under this interpretation of SRM, Bayesian algorithms which select a concept of maximum posterior probability (MAP algorithms) are viewed as a kind of SRM algorithm. Various approaches to SRM are compared both theoretically and experimentally by Kearns et al. in [6]. They give experimental evidence that Bayesian and MDL algorithms tend to over-fit in experimental settings where the Bayesian assumptions fail. A PAC-Bayesian approach uses a prior distribution analogous to that used in MAP or MDL but provides a theoretical guarantee against over-fitting independent of the truth of the prior.

Earlier work on PAC-Bayesian algorithms has focused on model selection — selecting either a single concept or a uniformly weighted set of concepts. Here we consider nonuniform model averaging, i.e., selecting a weighted mixture of the concepts.

Model averaging is empirically important in certain applications. For example, in statistical language modeling for speech recognition one "smooths" a trigram model with a bigram model and smooths the bigram model with a unigram model. This smoothing is essential for minimizing the cross entropy between, say, the model and a test corpus of newspaper sentences. It turns out that smoothing in statistical language modeling is more naturally formulated as model averaging than as model selection. A smoothed language model is very large — it contains a full trigram model, a full bigram model and a full unigram model as parts. If one uses MDL to select the structure of a language model, selecting model parameters with maximum likelihood, the resulting structure is much smaller than that of a smoothed trigram model. Furthermore, the MDL model performs quite badly. However, a smoothed trigram model can be theoretically derived as a compact representation of a Bayesian mixture of an exponential number of (smaller) suffix tree models [10].

Model averaging can also be applied to decision trees. A common method of constructing decision trees is to first build an overly large tree which over-fits the training data and then prune the tree in some way so as to get a smaller tree that does not over-fit the data [11, 5]. An alternative to pruning is to construct a weighted mixture of the subtrees of the original over-fit tree. It is possible to construct a concise representation of a weighting over exponentially many different subtrees [3, 9, 4].

This paper proves a new PAC-Bayesian theorem giving a bound on the generalization error of weighted mixtures. A weighted mixture which gives too much weight to models with low prior probability will over-fit the

COLT '99 7/99 Santa Cruz, CA, USA

Seminal paper, that contains the bound stated earlier today.

PAC-Bayesian Model Averaging

David A. McAllester
AT&T Shannon Labs
180 Park Avenue
Florham Park, NJ 07932-0971
dmac@research.att.com

Abstract

PAC-Bayesian learning methods combine the informative priors of Bayesian methods with distribution-free PAC guarantees. Building on earlier methods for PAC-Bayesian model selection, this paper presents a method for PAC-Bayesian model averaging. The method constructs an optimized weighted mixture of concepts analogous to a Bayesian posterior distribution. Although the main result is stated for bounded loss, a preliminary analysis for unbounded loss is also given.

1 INTRODUCTION

A PAC-Bayesian approach to machine learning attempts to combine the advantages of both PAC and Bayesian approaches [12, 8]. The Bayesian approach has the advantage of using arbitrary domain knowledge in the form of a Bayesian prior. The PAC approach has the advantage that one can prove guarantees for generalization error without assuming the truth of the prior. A PAC-Bayesian approach combines the features of the PAC and Bayesian approaches — it bases the bias of the learning algorithm on an arbitrary prior distribution, thus allowing the incorporation of domain knowledge, and yet provides a guarantee on generalization error that is independent of any truth of the prior.

PAC-Bayesian approaches are related to structural risk minimization (SRM) [6]. Here we interpret this broadly as describing any learning algorithm optimizing a tradeoff between the "complexity", "structure", or "prior probability" of the concept or model and the "goodness of fit", "description length", or "likelihood" of the training data. Under this interpretation of SRM, Bayesian algorithms which select a concept of maximum posterior probability (MAP algorithms) are viewed as a kind of SRM algorithm. Various approaches to SRM are compared both theoretically and experimentally by Kearns et al. in [6]. They give experimental evidence that Bayesian and MDL algorithms tend to over-fit in experimental settings where the Bayesian assumptions fail. A PAC-Bayesian approach uses a prior distribution analogous to that used in MAP or MDL but provides a theoretical guarantee against over-fitting independent of the truth of the prior.

Earlier work on PAC-Bayesian algorithms has focused on model selection — selecting either a single concept or a uniformly weighted set of concepts. Here we consider nonuniform model averaging, i.e., selecting a weighted mixture of the concepts.

Model averaging is empirically important in certain applications. For example, in statistical language modeling for speech recognition one "smooths" a trigram model with a bigram model and smooths the bigram model with a unigram model. This smoothing is essential for minimizing the cross entropy between, say, the model and a test corpus of newspaper sentences. It turns out that smoothing in statistical language modeling is more naturally formulated as model averaging than as model selection. A smoothed language model is very large — it contains a full trigram model, a full bigram model and a full unigram model as parts. If one uses MDL to select the structure of a language model, selecting model parameters with maximum likelihood, the resulting structure is much smaller than that of a smoothed trigram model. Furthermore, the MDL model performs quite badly. However, a smoothed trigram model can be theoretically derived as a compact representation of a Bayesian mixture of an exponential number of (smaller) suffix tree models [10].

Model averaging can also be applied to decision trees. A common method of constructing decision trees is to first build an overly large tree which over-fits the training data and then prune the tree in some way so as to get a smaller tree that does not over-fit the data [11, 5]. An alternative to pruning is to construct a weighted mixture of the subtrees of the original over-fit tree. It is possible to construct a concise representation of a weighting over exponentially many different subtrees [3, 9, 4].

This paper proves a new PAC-Bayesian theorem giving a bound on the generalization error of weighted mixtures. A weighted mixture which gives too much weight to models with low prior probability will over-fit the

COLT '99 7/99 Santa Cruz, CA, USA

Seminal paper, that contains the bound stated earlier today.

Since then, various bounds published :

- tighter,
- with less assumptions (i.i.d, bounded loss),
- easier to optimize,
- ...

Catoni's PAC-Bayes bound, 2003

Fix $\lambda > 0$ and π. With proba. at least $1 - \delta$ on $\mathcal{S}$, for any $\hat{\rho}$,

$$\mathbb{E}_{\theta \sim \hat{\rho}}[R(\theta)] \leq \mathbb{E}_{\theta \sim \hat{\rho}}[R_n(\theta)] + \frac{\mathrm{KL}(\hat{\rho} \| \pi) + \log \frac{1}{\delta}}{\lambda} + \frac{\lambda}{8n}.$$

Catoni's PAC-Bayes bound, 2003

Fix $\lambda > 0$ and π. With proba. at least $1-\delta$ on $\mathcal{S}$, for any $\hat{\rho}$,

$$\mathbb{E}_{\theta\sim\hat{\rho}}[R(\theta)] \leq \mathbb{E}_{\theta\sim\hat{\rho}}[R_n(\theta)] + \frac{\mathrm{KL}(\hat{\rho}\|\pi) + \log\frac{1}{\delta}}{\lambda} + \frac{\lambda}{8n}.$$

"De-randomized" PAC-Bayes bound, 2003

- Fix $\lambda > 0$, π and a randomized estimator $\hat{\rho}$.
- Sample $\hat{\theta} \sim \hat{\rho}(\mathcal{S})$.

With probability at least $1-\delta$ on $(\mathcal{S}, \hat{\theta})$,

$$R(\hat{\theta}) \leq R_n(\hat{\theta}) + \frac{\log\frac{\mathrm{d}\hat{\rho}}{d\pi}(\hat{\theta}) + \log\frac{1}{\delta}}{\lambda} + \frac{\lambda}{8n}.$$

Catoni's PAC-Bayes bound, 2003

Fix $\lambda > 0$ and π. With proba. at least $1 - \delta$ on $\mathcal{S}$, for any $\hat{\rho}$,

$$\mathbb{E}_{\theta\sim\hat{\rho}}[R(\theta)] \leq \mathbb{E}_{\theta\sim\hat{\rho}}[R_n(\theta)] + \frac{\mathrm{KL}(\hat{\rho}\|\pi) + \log\frac{1}{\delta}}{\lambda} + \frac{\lambda}{8n}.$$

"De-randomized" PAC-Bayes bound, 2003

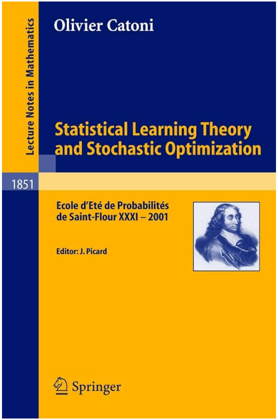
- Fix $\lambda > 0$, π and a randomized estimator $\hat{\rho}$.
- Sample $\hat{\theta} \sim \hat{\rho}(\mathcal{S})$.

With probability at least $1 - \delta$ on $(\mathcal{S}, \hat{\theta})$,

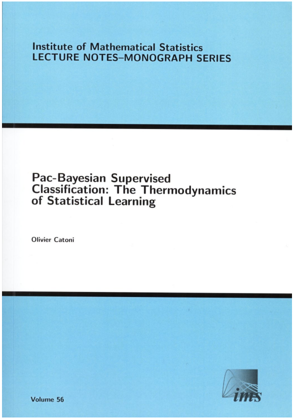
$$R(\hat{\theta}) \leq R_n(\hat{\theta}) + \frac{\log\frac{\mathrm{d}\hat{\rho}}{d\pi}(\hat{\theta}) + \log\frac{1}{\delta}}{\lambda} + \frac{\lambda}{8n}.$$

Catoni, O. (2003). *A PAC-Bayesian approach to adaptive classification*. Preprint LPMA 840.

Classical references :



Connections with information theory and MDL.



Very tight bounds, applications to Support Vector Machines.

Seeger, M. (2002). PAC-Bayesian generalisation error bounds for Gaussian process classification. *Journal of Machine Learning Research*.

Maurer, A. (2004). *A note on the PAC-Bayesian theorem*. Arxiv preprint arXiv :cs/0411099.

Tolstikhin, I. and Seldin, Y. (2013). PAC-Bayes-empirical-Bernstein inequality. *NeurIPS*.

Seeger, M. (2002). PAC-Bayesian generalisation error bounds for Gaussian process classification. *Journal of Machine Learning Research*.

Maurer, A. (2004). *A note on the PAC-Bayesian theorem*. Arxiv preprint arXiv :cs/0411099.

Tolstikhin, I. and Seldin, Y. (2013). PAC-Bayes-empirical-Bernstein inequality. *NeurIPS*.

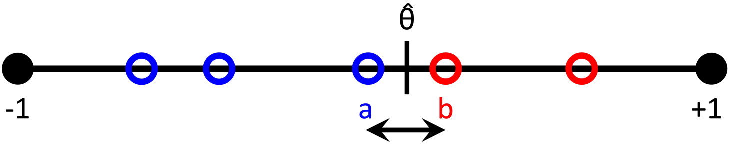
Tolstikhin and Seldin's PAC-Bayes bound, 2013

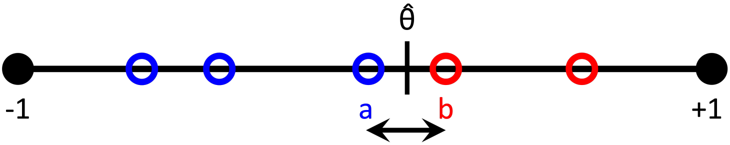
With proba. at least $1-\delta$, for any ρ,

$$\mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \mathbb{E}_{\theta\sim\rho}[R_n(\theta)] + \sqrt{2\mathbb{E}_{\theta\sim\rho}[R_n(\theta)]\frac{\mathrm{KL}(\rho\|\pi)+\log\frac{2\sqrt{n}}{\delta}}{n}} + 2\frac{\mathrm{KL}(\rho\|\pi)+\log\frac{2\sqrt{n}}{\delta}}{n}.$$

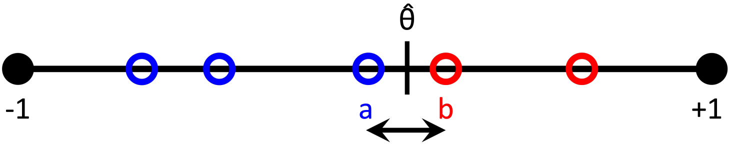
Consequence : if $\mathbb{E}_{\theta\sim\rho}[R_n(\theta)] = 0$,

$$\mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \cancel{\mathbb{E}_{\theta\sim\rho}[R_n(\theta)]} \cancel{+ \sqrt{2\mathbb{E}_{\theta\sim\rho}[R_n(\theta)]\frac{\mathrm{KL}(\rho\|\pi) + \log\frac{2\sqrt{n}}{\delta}}{n}}} + 2\frac{\mathrm{KL}(\rho\|\pi) + \log\frac{2\sqrt{n}}{\delta}}{n}.$$


θ̂
-1
a
b
+1



$$\mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \cancel{\sqrt{\frac{\log \frac{2}{b-a} + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{2n}}}.$$



$$\mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \cancel{\sqrt{\frac{\log \frac{2}{b-a} + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{2n}}}.$$

$$\mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq 2\frac{\log \frac{2}{b-a} + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{n}.$$

Bound in expectation → bound on the weighted majority vote :

Germain, P., Lacasse, A., Laviolette, F., Marchand, M. and Roy, J.-F. (2015). Risk bounds for the majority vote : from a PAC-Bayesian analysis to a learning algorithm *Journal of Machine Learning Research*.

Bound in expectation $\rightarrow$ bound on the weighted majority vote :

Germain, P., Lacasse, A., Laviolette, F., Marchand, M. and Roy, J.-F. (2015). Risk bounds for the majority vote : from a PAC-Bayesian analysis to a learning algorithm *Journal of Machine Learning Research*.

Tight bound that allows to recover all the above, and more :

Germain, P., Lacasse, A., Laviolette, F. and Marchand, M. (2009). PAC-Bayesian learning of linear classifiers. *ICML*.

Bound in expectation $\rightarrow$ bound on the weighted majority vote :

Germain, P., Lacasse, A., Laviolette, F., Marchand, M. and Roy, J.-F. (2015). Risk bounds for the majority vote : from a PAC-Bayesian analysis to a learning algorithm *Journal of Machine Learning Research*.

Tight bound that allows to recover all the above, and more :

Germain, P., Lacasse, A., Laviolette, F. and Marchand, M. (2009). PAC-Bayesian learning of linear classifiers. *ICML*.

François Laviolette (1962-2021).

Tutorials :

McAllester, D. (2013). *A PAC-Bayesian tutorial with a dropout bound*. ArXiv preprint arXiv :1307.2118.

Van Erven, T. (2014). *PAC-Bayes mini-tutorial : a continuous union bound*. ArXiv preprint arXiv :1405.1580.

Guedj, B. (2019). *A primer on PAC-Bayesian learning*. ArXiv preprint arXiv :1901.05353.

1. PAC-Bayes bounds : introduction
   - Generalization bounds and PAC-Bayes
   - Minimization of the PAC-Bayes bound
   - A zoo of PAC-Bayes bounds

2. PAC-Bayes and Mutual Information bounds
   - Excess risk bounds
   - Fast rates
   - Mutual information bounds

Recap :

- Data : $\mathcal{S} = ((X_1, Y_1), \dots, (X_n, Y_n))$.

Recap :

- Data : $\mathcal{S} = ((X_1, Y_1), \dots, (X_n, Y_n))$.
- Risk : $R(\theta) := \mathbb{E}_{(X,Y)\sim P}\left[\ell\Big(Y, f_\theta(X)\Big)\right]$.

Recap :

- Data : $\mathcal{S} = ((X_1, Y_1), \dots, (X_n, Y_n))$.
- Risk : $R(\theta) := \mathbb{E}_{(X,Y)\sim P}\left[\ell\left(Y, f_\theta(X)\right)\right]$.
- Oracle risk : $R^* = \inf_{\theta\in\Theta} R(\theta)$.

Recap :

- Data : $\mathcal{S} = ((X_1, Y_1), \dots, (X_n, Y_n))$.
- Risk : $R(\theta) := \mathbb{E}_{(X,Y)\sim P}\left[\ell\left(Y, f_\theta(X)\right)\right]$.
- Oracle risk : $R^* = \inf_{\theta\in\Theta} R(\theta)$.
- Empirical risk : $R_n(\theta) = \frac{1}{n}\sum_{i=1}^n \ell\left(Y_i, f_\theta(X_i)\right)$.

Recap :

- Data : $\mathcal{S} = ((X_1, Y_1), \dots, (X_n, Y_n))$.
- Risk : $R(\theta) := \mathbb{E}_{(X,Y)\sim P}\left[\ell\left(Y, f_\theta(X)\right)\right]$.
- Oracle risk : $R^* = \inf_{\theta\in\Theta} R(\theta)$.
- Empirical risk : $R_n(\theta) = \frac{1}{n}\sum_{i=1}^n \ell\left(Y_i, f_\theta(X_i)\right)$.

"Randomized estimators"

data	proba. distribution	parameter

$$\mathcal{S} \longmapsto \hat{\rho} = \hat{\rho}(\mathcal{S}) \overset{\theta\sim\hat{\rho}}{\dashrightarrow} \theta$$

Recap :

- Data : $\mathcal{S} = ((X_1, Y_1), \dots, (X_n, Y_n))$.
- Risk : $R(\theta) := \mathbb{E}_{(X,Y)\sim P}\left[\ell\Big(Y, f_\theta(X)\Big)\right]$.
- Oracle risk : $R^* = \inf_{\theta\in\Theta} R(\theta)$.
- Empirical risk : $R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell\Big(Y_i, f_\theta(X_i)\Big)$.

"Randomized estimators"

data	proba. distribution	parameter

$$\mathcal{S} \longmapsto \hat{\rho} = \hat{\rho}(\mathcal{S}) \xrightarrow{\theta\sim\hat{\rho}} \theta$$

Today, we study "excess-risk bounds", that is :

$$\mathbb{E}_{\theta\sim\hat{\rho}}[R(\theta)] \leq R^* + \dots \text{ (rate of convergence).}$$

Catoni's PAC-Bayes bound, 2003

Fix $\lambda > 0$ and π. With proba. at least $1 - \delta$ on $\mathcal{S}$, for any $\hat{\rho}$,

$$\mathbb{E}_{\theta \sim \hat{\rho}}[R(\theta)] \leq \mathbb{E}_{\theta \sim \hat{\rho}}[R_n(\theta)] + \frac{\mathrm{KL}(\hat{\rho} \| \pi) + \log \frac{1}{\delta}}{\lambda} + \frac{\lambda}{8n}.$$

Catoni's PAC-Bayes bound, 2003

Fix $\lambda > 0$ and π. With proba. at least $1-\delta$ on $\mathcal{S}$, for any $\hat{\rho}$,

$$\mathbb{E}_{\theta\sim\hat{\rho}}[R(\theta)] \leq \mathbb{E}_{\theta\sim\hat{\rho}}[R_n(\theta)] + \frac{\mathrm{KL}(\hat{\rho}\|\pi) + \log\frac{1}{\delta}}{\lambda} + \frac{\lambda}{8n}.$$

"De-randomized" PAC-Bayes bound, 2003

- Fix $\lambda > 0$, π and a randomized estimator $\hat{\rho}$.
- Sample $\hat{\theta} \sim \hat{\rho}(\mathcal{S})$.

With probability at least $1-\delta$ on $(\mathcal{S},\hat{\theta})$,

$$R(\hat{\theta}) \leq R_n(\hat{\theta}) + \frac{\log\frac{\mathrm{d}\hat{\rho}}{d\pi}(\hat{\theta}) + \log\frac{1}{\delta}}{\lambda} + \frac{\lambda}{8n}.$$

Catoni's PAC-Bayes bound in expectation, 2003

- Fix $\lambda > 0$, π and a randomized estimator $\hat{\rho}$.

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\rho}}[R(\theta)]\right] \leq \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\rho}}[R_n(\theta)] + \frac{\mathrm{KL}(\hat{\rho}\|\pi)}{\lambda} + \frac{\lambda}{8n}\right].$$

Catoni, O. (2003). *A PAC-Bayesian approach to adaptive classification*. Preprint LPMA 840.

Zhang, T. (2006). Information-theoretic upper and lower bounds for statistical estimation *IEEE Transactions on Information Theory*.

Catoni's PAC-Bayes bound in expectation, 2003

- Fix $\lambda > 0$, π and a randomized estimator $\hat{\rho}$.

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\rho}}[R(\theta)]\right] \leq \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\rho}}[R_n(\theta)] + \frac{\mathrm{KL}(\hat{\rho}\|\pi)}{\lambda} + \frac{\lambda}{8n}\right].$$

Catoni, O. (2003). *A PAC-Bayesian approach to adaptive classification*. Preprint LPMA 840.

Zhang, T. (2006). Information-theoretic upper and lower bounds for statistical estimation *IEEE Transactions on Information Theory*.

Note : sometimes refered to as "MAC-Bayes" for "Mean Approximately Correct"...

Reminder – Gibbs posterior

$$\hat{\pi}_\lambda = \operatorname*{arg\,min}_{\rho \in \mathcal{M}(\Theta)} \left\{ \mathbb{E}_{\theta\sim\rho}[R_n(\theta)] + \frac{\mathrm{KL}(\rho\|\pi)}{\lambda} \right\}$$

$$\hat{\pi}_\lambda(\mathrm{d}\theta) = \frac{\exp(-\lambda R_n(\theta))}{\mathbb{E}_{\vartheta\sim\pi}[\exp(-\lambda R_n(\vartheta)]} \pi(\mathrm{d}\theta).$$

Reminder – Gibbs posterior

$$\hat{\pi}_\lambda = \arg\min_{\rho \in \mathcal{M}(\Theta)} \left\{ \mathbb{E}_{\theta\sim\rho}[R_n(\theta)] + \frac{\mathrm{KL}(\rho\|\pi)}{\lambda} \right\}$$

$$\hat{\pi}_\lambda(\mathrm{d}\theta) = \frac{\exp(-\lambda R_n(\theta))}{\mathbb{E}_{\vartheta\sim\pi}[\exp(-\lambda R_n(\vartheta))]}\pi(\mathrm{d}\theta).$$

Consequence of PAC-Bayes bound in expectation :

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\pi}_\lambda}[R(\theta)]\right] &\leq \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\pi}_\lambda}[R_n(\theta)] + \frac{\mathrm{KL}(\hat{\pi}_\lambda\|\pi)}{\lambda} + \frac{\lambda}{8n}\right] \\
&= \mathbb{E}_{\mathcal{S}} \inf_\rho \left[\mathbb{E}_\rho[R_n(\theta)] + \frac{\mathrm{KL}(\rho\|\pi)}{\lambda} + \frac{\lambda}{8n}\right] \\
&\leq \inf_\rho \mathbb{E}_{\mathcal{S}} \left[\mathbb{E}_\rho[R_n(\theta)] + \frac{\mathrm{KL}(\rho\|\pi)}{\lambda} + \frac{\lambda}{8n}\right].
\end{aligned}$$

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\pi}_\lambda}[R(\theta)]\right] \leq \inf_{\rho} \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\rho}[R_n(\theta)] + \frac{\mathrm{KL}(\rho\|\pi)}{\lambda} + \frac{\lambda}{8n}\right].$$

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\pi}_\lambda}[R(\theta)]\right] \leq \inf_{\rho} \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\rho}[R_n(\theta)] + \frac{\mathrm{KL}(\rho\|\pi)}{\lambda} + \frac{\lambda}{8n}\right].$$

$$\begin{aligned}\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\rho}[R_n(\theta)] &= \mathbb{E}_{\rho}[\mathbb{E}_{\mathcal{S}} R_n(\theta)] \text{ (Fubini-Tonelli)} \\ &= \mathbb{E}_{\rho}[R(\theta)].\end{aligned}$$

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\pi}_\lambda}[R(\theta)]\right] \leq \inf_{\rho} \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\rho}[R_n(\theta)] + \frac{\mathrm{KL}(\rho\|\pi)}{\lambda} + \frac{\lambda}{8n}\right].$$

$$\begin{aligned}\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\rho}[R_n(\theta)] &= \mathbb{E}_{\rho}[\mathbb{E}_{\mathcal{S}} R_n(\theta)] \text{ (Fubini-Tonelli)} \\ &= \mathbb{E}_{\rho}[R(\theta)].\end{aligned}$$

Catoni's PAC-Bayes oracle bound, 2003

- Fix $\lambda > 0$, π, and let $\hat{\pi}_\lambda$ be the Gibbs posterior.

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\pi}_\lambda}[R(\theta)]\right] \leq \inf_{\rho\in\mathcal{M}(\Theta)} \left[\mathbb{E}_{\theta\sim\rho}[R(\theta)] + \frac{\mathrm{KL}(\rho\|\pi)}{\lambda} + \frac{\lambda}{8n}\right].$$

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\pi}_\lambda}[R(\theta)]\right] \leq \inf_{\rho} \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\rho}[R_n(\theta)] + \frac{\mathrm{KL}(\rho\|\pi)}{\lambda} + \frac{\lambda}{8n}\right].$$

$$\begin{aligned}\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\rho}[R_n(\theta)] &= \mathbb{E}_{\rho}[\mathbb{E}_{\mathcal{S}} R_n(\theta)] \text{ (Fubini-Tonelli)} \\ &= \mathbb{E}_{\rho}[R(\theta)].\end{aligned}$$

Catoni's PAC-Bayes oracle bound, 2003

- Fix $\lambda > 0$, π, and let $\hat{\pi}_\lambda$ be the Gibbs posterior.

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\pi}_\lambda}[R(\theta)]\right] \leq \inf_{\rho\in\mathcal{M}(\Theta)} \left[\mathbb{E}_{\theta\sim\rho}[R(\theta)] + \frac{\mathrm{KL}(\rho\|\pi)}{\lambda} + \frac{\lambda}{8n}\right].$$

- In this result, we use $\hat{\pi}_\lambda$ as our randomized estimator.

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\pi}_\lambda}[R(\theta)]\right] \leq \inf_{\rho} \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\rho}[R_n(\theta)] + \frac{\mathrm{KL}(\rho\|\pi)}{\lambda} + \frac{\lambda}{8n}\right].$$

$$\begin{aligned}\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\rho}[R_n(\theta)] &= \mathbb{E}_{\rho}[\mathbb{E}_{\mathcal{S}} R_n(\theta)] \text{ (Fubini-Tonelli)} \\ &= \mathbb{E}_{\rho}[R(\theta)].\end{aligned}$$

Catoni's PAC-Bayes oracle bound, 2003

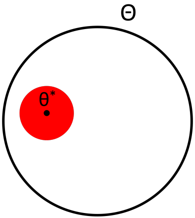
- Fix $\lambda > 0$, π, and let $\hat{\pi}_\lambda$ be the Gibbs posterior.

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\pi}_\lambda}[R(\theta)]\right] \leq \inf_{\rho\in\mathcal{M}(\Theta)} \left[\mathbb{E}_{\theta\sim\rho}[R(\theta)] + \frac{\mathrm{KL}(\rho\|\pi)}{\lambda} + \frac{\lambda}{8n}\right].$$

- In this result, we use $\hat{\pi}_\lambda$ as our randomized estimator.
- But to explicit the right-hand side, we can substitute anything to ρ in the infimum...

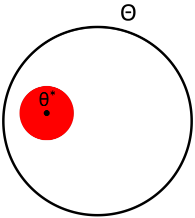
Illustration :

Illustration :



- π uniform on $\Theta = B_d(0, C)$
- ρ uniform on $B_d(\theta^*, \epsilon)$ where $R(\theta^*) = R^*$.

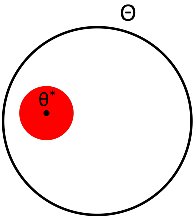
Illustration :



- π uniform on $\Theta = B_d(0, C)$
- ρ uniform on $B_d(\theta^*, \epsilon)$ where $R(\theta^*) = R^*$.

$$\mathrm{KL}(\rho\|\pi) = d \log \frac{C}{\epsilon}.$$

Illustration :



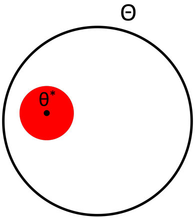
- π uniform on $\Theta = B_d(0, C)$
- ρ uniform on $B_d(\theta^*, \epsilon)$ where $R(\theta^*) = R^*$.

$$\mathrm{KL}(\rho \| \pi) = d \log \frac{C}{\epsilon}.$$

Assume $\theta \mapsto \ell(y, f_\theta(x))$ is L-Lipschitz around θ^*, that is :

$$|\ell(y, f_\theta(x)) - \ell(y, f_{\theta^*}(x))| \leq L \|\theta - \theta^*\|.$$

Illustration :



- π uniform on $\Theta = B_d(0, C)$
- ρ uniform on $B_d(\theta^*, \epsilon)$ where $R(\theta^*) = R^*$.

$$\mathrm{KL}(\rho\|\pi) = d \log \frac{C}{\epsilon}.$$

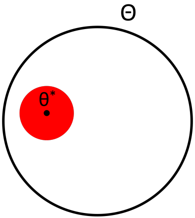
Assume $\theta \mapsto \ell(y, f_\theta(x))$ is L-Lipschitz around θ^*, that is :

$$|\ell(y, f_\theta(x)) - \ell(y, f_{\theta^*}(x))| \leq L\|\theta - \theta^*\|.$$

By taking expectations :

$$R(\theta) - R^* = R(\theta) - R(\theta^*) \leq L\|\theta - \theta^*\|.$$

Illustration :



- π uniform on $\Theta = B_d(0, C)$
- ρ uniform on $B_d(\theta^*, \epsilon)$ where $R(\theta^*) = R^*$.

$$\mathrm{KL}(\rho \| \pi) = d \log \frac{C}{\epsilon}.$$

Assume $\theta \mapsto \ell(y, f_\theta(x))$ is L-Lipschitz around θ^*, that is :

$$|\ell(y, f_\theta(x)) - \ell(y, f_{\theta^*}(x))| \leq L\|\theta - \theta^*\|.$$

By taking expectations :

$$R(\theta) - R^* = R(\theta) - R(\theta^*) \leq L\|\theta - \theta^*\|.$$

Thus $\mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq R^* + L\epsilon.$

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\pi}_\lambda}[R(\theta)]\right] \leq \inf_{\epsilon>0}\left[R^* + \epsilon + \frac{d\log\frac{C}{\epsilon}}{\lambda} + \frac{\lambda}{8n}\right].$$

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\pi}_\lambda}[R(\theta)]\right] \leq \inf_{\epsilon>0}\left[R^* + \epsilon + \frac{d\log\frac{C}{\epsilon}}{\lambda} + \frac{\lambda}{8n}\right].$$

The bound is exactly minimized for $\epsilon = d/\lambda$:

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\pi}_\lambda}[R(\theta)]\right] \leq R^* + \frac{d}{\lambda}\left(1 + \log\frac{C\lambda}{d}\right) + \frac{\lambda}{8n}.$$

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\pi}_\lambda}[R(\theta)]\right] \leq \inf_{\epsilon>0}\left[R^* + \epsilon + \frac{d\log\frac{C}{\epsilon}}{\lambda} + \frac{\lambda}{8n}\right].$$

The bound is exactly minimized for $\epsilon = d/\lambda$:

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\pi}_\lambda}[R(\theta)]\right] \leq R^* + \frac{d}{\lambda}\left(1 + \log\frac{C\lambda}{d}\right) + \frac{\lambda}{8n}.$$

In this case, we can calibrate the Gibbs posterior with $\lambda = \sqrt{n/d}$ which leads to

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\pi}_\lambda}[R(\theta)]\right] \leq R^* + \mathcal{O}\left(\sqrt{\frac{d}{n}}\log\frac{n}{d}\right).$$

Reminder : Catoni's PAC-Bayes oracle bound

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\pi}_\lambda}[R(\theta)]\right] \leq \inf_{\rho\in\mathcal{M}(\Theta)}\left[\mathbb{E}_{\theta\sim\rho}[R(\theta)] + \frac{\mathrm{KL}(\rho\|\pi)}{\lambda} + \frac{\lambda}{8n}\right].$$

Recap on the previous example :

- π uniform on $B_d(0, C)$,
- ρ uniform on $B_d(\theta^*, \epsilon)$,
- L-Lipschitz loss function.

$$\Rightarrow \mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq R^* + L\epsilon.$$

Reminder : Catoni's PAC-Bayes oracle bound

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\pi}_\lambda}[R(\theta)]\right] \leq \inf_{\rho\in\mathcal{M}(\Theta)}\left[\mathbb{E}_{\theta\sim\rho}[R(\theta)] + \frac{\mathrm{KL}(\rho\|\pi)}{\lambda} + \frac{\lambda}{8n}\right].$$

Recap on the previous example :

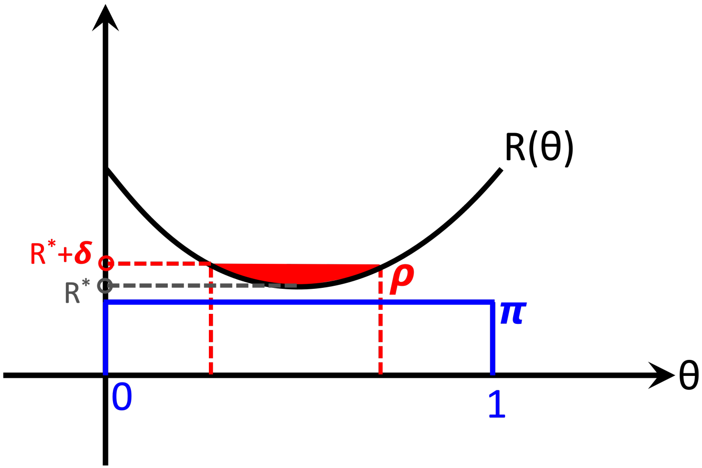
- π uniform on $B_d(0, C)$,
- ρ uniform on $B_d(\theta^*, \epsilon)$,
- L-Lipschitz loss function.

$$\Rightarrow \mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq R^* + L\epsilon.$$

More generally, we can consider in the PAC-Bayes oracle bound :

$$\rho = \pi_\delta := \text{ restriction of } \pi \text{ to } \{\theta : R(\theta) \leq R^* + \delta\}.$$

$$\rho = \pi_\delta := \text{ restriction of } \pi \text{ to } \{\theta : R(\theta) \leq R^* + \delta\}.$$



$$
\begin{aligned}
\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\pi}_\lambda}[R(\theta)]\right] & \leq \inf_{\delta>0}\left[\mathbb{E}_{\theta\sim\pi_\delta}[R(\theta)] + \frac{\mathrm{KL}(\pi_\delta\|\pi)}{\lambda} + \frac{\lambda}{8n}\right] \\
& \leq \inf_{\delta>0}\left[R^* + \delta + \frac{\log\frac{1}{\pi\{\theta: R(\theta)\leq R^*+\delta\}}}{\lambda} + \frac{\lambda}{8n}\right].
\end{aligned}
$$

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\pi}_\lambda}[R(\theta)]\right] \leq \inf_{\delta>0}\left[\mathbb{E}_{\theta\sim\pi_\delta}[R(\theta)] + \frac{\mathrm{KL}(\pi_\delta\|\pi)}{\lambda} + \frac{\lambda}{8n}\right]$$
$$\leq \inf_{\delta>0}\left[R^* + \delta + \frac{\log\frac{1}{\pi\{\theta:R(\theta)\leq R^*+\delta\}}}{\lambda} + \frac{\lambda}{8n}\right].$$

In the previous example,

$$\log\frac{1}{\pi\{\theta : R(\theta) \leq R^* + \delta\}} \leq d \log\frac{C}{\delta}$$

and we obtained a bound in $\sqrt{d/n}\log(n/d)$.

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\pi}_\lambda}[R(\theta)]\right] \leq \inf_{\delta>0}\left[\mathbb{E}_{\theta\sim\pi_\delta}[R(\theta)] + \frac{\mathrm{KL}(\pi_\delta\|\pi)}{\lambda} + \frac{\lambda}{8n}\right]$$
$$\leq \inf_{\delta>0}\left[R^* + \delta + \frac{\log\frac{1}{\pi\{\theta:R(\theta)\leq R^*+\delta\}}}{\lambda} + \frac{\lambda}{8n}\right].$$

In the previous example,

$$\log\frac{1}{\pi\{\theta : R(\theta)\leq R^*+\delta\}} \leq d\log\frac{C}{\delta}$$

and we obtained a bound in $\sqrt{d/n}\log(n/d)$.

Definition : the prior mass condition is satisfied if there are $C, D > 0$ such that, for any $\delta > 0$ small enough,

$$\log\frac{1}{\pi\{\theta : R(\theta)\leq R^*+\delta\}} \leq D\log\frac{C}{\delta}.$$

Theorem - excess risk bound

- Assume the prior mass condition with $C, D > 0$.
- Fix $\lambda = \sqrt{n/D}\log(D/n)$, and let $\hat{\pi}_\lambda$ be the Gibbs posterior.

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\pi}_\lambda}[R(\theta)]\right] \leq R^* + \mathcal{O}\left(\sqrt{\frac{D}{n}}\log\frac{n}{D}\right).$$

Reminder – Tolstikhin and Seldin's PAC-Bayes bound, 2013

With proba. at least $1-\delta$, for any ρ,

$$\mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \mathbb{E}_{\theta\sim\rho}[R_n(\theta)] + \sqrt{2\mathbb{E}_{\theta\sim\rho}[R_n(\theta)]\frac{\mathrm{KL}(\rho\|\pi)+\log\frac{2\sqrt{n}}{\delta}}{n}} + 2\frac{\mathrm{KL}(\rho\|\pi)+\log\frac{2\sqrt{n}}{\delta}}{n}.$$

Reminder – Tolstikhin and Seldin's PAC-Bayes bound, 2013

With proba. at least $1-\delta$, for any ρ,

$$\mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \mathbb{E}_{\theta\sim\rho}[R_n(\theta)] + \sqrt{2\mathbb{E}_{\theta\sim\rho}[R_n(\theta)] \frac{\mathrm{KL}(\rho\|\pi) + \log\frac{2\sqrt{n}}{\delta}}{n}} + 2\frac{\mathrm{KL}(\rho\|\pi) + \log\frac{2\sqrt{n}}{\delta}}{n}.$$

$\rightarrow$ if there is a perfect predictor θ^*, $R_n(\theta^*) = R(\theta^*) = 0$, then using the previous approach (prior mass condition) :

$$\mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \underbrace{R^*}_{=0} + \mathcal{O}\left(\frac{D}{n}\log\frac{n}{D}\right).$$

Reminder – Tolstikhin and Seldin's PAC-Bayes bound, 2013

With proba. at least $1-\delta$, for any ρ,

$$\mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \mathbb{E}_{\theta\sim\rho}[R_n(\theta)] + \sqrt{2\mathbb{E}_{\theta\sim\rho}[R_n(\theta)]\frac{\mathrm{KL}(\rho\|\pi)+\log\frac{2\sqrt{n}}{\delta}}{n}} + 2\frac{\mathrm{KL}(\rho\|\pi)+\log\frac{2\sqrt{n}}{\delta}}{n}.$$

$\rightarrow$ if there is a perfect predictor θ^*, $R_n(\theta^*) = R(\theta^*) = 0$, then using the previous approach (prior mass condition) :

$$\mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \underbrace{R^*}_{=0} + \mathcal{O}\left(\frac{D}{n}\log\frac{n}{D}\right).$$

This can happen beyond the case $R_n(\theta^*) = R(\theta^*) = 0$!

Example 1 : classification, $\ell(y, f_\theta(x)) = 1_{y \neq f_\theta(x)}$.

Example 1 : classification, $\ell(y, f_\theta(x)) = 1_{y \neq f_\theta(x)}$.

Recall :

- $\eta(x) = \mathbb{P}(Y = 1 | X = x)$,
- the "Bayes classifier" $f^*(x) = 1_{\eta(x) \geq 1/2}$.

Example 1 : classification, $\ell(y, f_\theta(x)) = 1_{y \neq f_\theta(x)}$.

Recall :

- $\eta(x) = \mathbb{P}(Y = 1 | X = x)$,
- the "Bayes classifier" $f^*(x) = 1_{\eta(x) \geq 1/2}$.

Mammen and Tsybakov margin assumption :

- $\mathbb{P}(|\eta(X) - 1/2| < \tau) = 0$ for some small enough $\tau > 0$.
- there is θ^* such that $f^* = f_{\theta^*}$.

Mammen, E. and Tsybakov, A. B. (1999). Smooth discrimination analysis. *The Annals of Statistics*.

Tsybakov, A. B. (2003). Optimal rates of aggregation. *COLT*.

Example 1 : classification, $\ell(y, f_\theta(x)) = 1_{y \neq f_\theta(x)}$.

Recall :

- $\eta(x) = \mathbb{P}(Y = 1 | X = x)$,
- the "Bayes classifier" $f^*(x) = 1_{\eta(x) \geq 1/2}$.

Mammen and Tsybakov margin assumption :

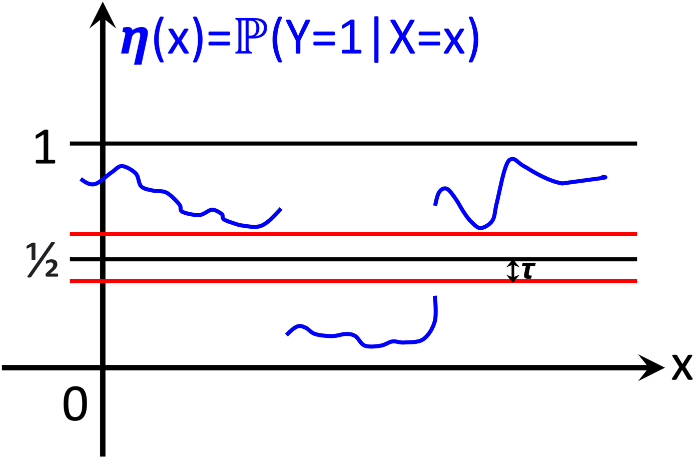
- $\mathbb{P}(|\eta(X) - 1/2| < \tau) = 0$ for some small enough $\tau > 0$.
- there is θ^* such that $f^* = f_{\theta^*}$.

Mammen, E. and Tsybakov, A. B. (1999). Smooth discrimination analysis. *The Annals of Statistics*.

Tsybakov, A. B. (2003). Optimal rates of aggregation. *COLT*.

Under the margin assumption, they prove fast rates in $\frac{1}{n}$ for various predictors.

$\mathbb{P}(|\eta(X) - 1/2| < \tau) = 0$ for some small enough $\tau > 0$.



Example 2 : “strongly convex, Lipschitz loss”.

Example 2 : “strongly convex, Lipschitz loss”.

For short, put $g(\theta) = g_{x,y}(\theta) = \ell(y, f_\theta(x))$.

Example 2 : "strongly convex, Lipschitz loss".

For short, put $g(\theta) = g_{x,y}(\theta) = \ell(y, f_\theta(x))$.

Assume there are $L, \alpha > 0$ such that, for any x, y, there is a $\delta(\cdot,\cdot) = \delta_{x,y}(\cdot,\cdot) \geq 0$ with

$$\forall \theta,\ \frac{g(\theta) + g(\theta^*)}{2} - g\left(\frac{\theta + \theta^*}{2}\right) \geq \frac{1}{2\alpha}\delta^2(\theta, \theta^*),$$

$$\text{and } g(\theta) - g(\theta^*) \leq L\delta(\theta, \theta^*).$$

Example 2 : "strongly convex, Lipschitz loss".

For short, put $g(\theta) = g_{x,y}(\theta) = \ell(y, f_\theta(x))$.

Assume there are $L, \alpha > 0$ such that, for any x, y, there is a $\delta(\cdot,\cdot) = \delta_{x,y}(\cdot,\cdot) \geq 0$ with

$$\forall \theta, \ \frac{g(\theta) + g(\theta^*)}{2} - g\left(\frac{\theta+\theta^*}{2}\right) \geq \frac{1}{2\alpha}\delta^2(\theta,\theta^*),$$

$$\text{and } g(\theta) - g(\theta^*) \leq L\delta(\theta,\theta^*).$$

Bartlett, P. L., Jordan, M. I. and McAuliffe, J. D. (2003). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*.

$\rightarrow$ fast rates also in this case.

Definition – Bernstein condition

Bernstein condition is satisfied with constant K if

$$\mathbb{E}\left[\left(\ell(Y,f_\theta(X))-\ell(Y,f_{\theta^*}(X))\right)^2\right] \leq K\,\underbrace{\mathbb{E}\left[\ell(Y,f_\theta(X))-\ell(Y,f_{\theta^*}(X))\right]}_{=R(\theta)-R^*}.$$

Mammen & Tsybakov margin assumption $\Longrightarrow$ Bernstein condition $\Longrightarrow$ fast rates

Bartlett et al. convexity condition $\Longrightarrow$ Bernstein condition

Intuition : for a fixed $\theta \in \Theta$ we have

$$\begin{aligned}
\mathbb{E}\left[\left(R_n(\theta) - R(\theta)\right)^2\right] &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f_\theta(X_i))\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \underbrace{\mathrm{Var}\Big(\ell(Y_i, f_\theta(X_i))\Big)}_{=:v(\theta)}.
\end{aligned}$$

Intuition : for a fixed $\theta \in \Theta$ we have

$$\mathbb{E}\left[\left(R_n(\theta) - R(\theta)\right)^2\right] = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f_\theta(X_i))\right)$$
$$= \frac{1}{n^2}\sum_{i=1}^{n} \underbrace{\mathrm{Var}\Big(\ell(Y_i, f_\theta(X_i))\Big)}_{=:v(\theta)}.$$

By Jensen,

$$\mathbb{E}\left[\left|R_n(\theta) - R(\theta)\right|\right] \leq \sqrt{\mathbb{E}\left[\left(R_n(\theta) - R(\theta)\right)^2\right]} = \sqrt{\frac{v(\theta)}{n}}.$$

Intuition : for a fixed $\theta \in \Theta$ we have

$$\mathbb{E}\left[\left(R_n(\theta) - R(\theta)\right)^2\right] = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f_\theta(X_i))\right)$$
$$= \frac{1}{n^2}\sum_{i=1}^{n} \underbrace{\mathrm{Var}\Big(\ell(Y_i, f_\theta(X_i))\Big)}_{=:v(\theta)}.$$

By Jensen,

$$\mathbb{E}\left[\left|R_n(\theta) - R(\theta)\right|\right] \leq \sqrt{\mathbb{E}\left[\left(R_n(\theta) - R(\theta)\right)^2\right]} = \sqrt{\frac{v(\theta)}{n}}.$$

If θ and θ' have the same empirical risk $R_n(\theta) = R_n(\theta')$, their risks might differ by $1/\sqrt{n}$!

$$
\begin{aligned}
&\mathbb{E}\left[\left(R_n(\theta) - R_n(\theta^*) - (R(\theta) - R^*)\right)^2\right] \\
&= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\left[\ell(Y_i, f_\theta(X_i)) - \ell(Y_i, f_{\theta^*}(X_i))\right]\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\Big(\ell(Y_i, f_\theta(X_i)) - \ell(Y_i, f_{\theta^*}(X_i))\Big) \\
&\leq \frac{1}{n}\mathbb{E}\left[\left(\ell(Y, f_\theta(X)) - \ell(Y, f_{\theta^*}(X))\right)^2\right] \\
&\leq \frac{K}{n}[R(\theta) - R^*] \quad \text{(using Bernstein condition).}
\end{aligned}
$$

$$\begin{aligned}
&\mathbb{E}\left[\Big(R_n(\theta) - R_n(\theta^*) - (R(\theta) - R^*)\Big)^2\right] \\
&= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\Big[\ell(Y_i, f_\theta(X_i)) - \ell(Y_i, f_{\theta^*}(X_i))\Big]\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\Big(\ell(Y_i, f_\theta(X_i)) - \ell(Y_i, f_{\theta^*}(X_i))\Big) \\
&\leq \frac{1}{n}\mathbb{E}\Big[\big(\ell(Y, f_\theta(X)) - \ell(Y, f_{\theta^*}(X))\big)^2\Big] \\
&\leq \frac{K}{n}[R(\theta) - R^*] \quad \text{(using Bernstein condition).}
\end{aligned}$$

If θ and θ^* have the same empirical risk $R_n(\theta) = R_n(\theta^*)$,

$$\Big(R(\theta) - R^*\Big)^2 \leq \frac{K}{n}[R(\theta) - R^*] \Rightarrow R(\theta) - R^* \leq \frac{K}{n}.$$

PAC-Bayes oracle inequality under Bernstein condition

- Assume Bernstein condition is satisfied with constant K.

Put $\lambda = n/\max(2K,1)$,

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\pi}_\lambda}[R(\theta)] - R^*\right] \leq 2 \inf_{\rho\in\mathcal{M}(\Theta)} \left[\mathbb{E}_{\theta\sim\rho}[R(\theta)] - R^* + \frac{\max(2K,1)\mathrm{KL}(\rho\|\pi)}{n}\right].$$

PAC-Bayes oracle inequality under Bernstein condition

- Assume Bernstein condition is satisfied with constant K.

Put $\lambda = n/\max(2K,1)$,

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\pi}_\lambda}[R(\theta)] - R^*\right] \leq 2 \inf_{\rho\in\mathcal{M}(\Theta)} \left[\mathbb{E}_{\theta\sim\rho}[R(\theta)] - R^* + \frac{\max(2K,1)\mathrm{KL}(\rho\|\pi)}{n}\right].$$

- Assume moreover the prior mass condition with $C, D > 0$.

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\pi}_\lambda}[R(\theta)] - R^*\right] \leq \frac{2\max(2K,1)D}{n}\log\left(\frac{\mathrm{e}Cn}{D}\right).$$

Reminder – Bernstein condition

$$\mathbb{E}\left[\left(\ell(Y, f_\theta(X)) - \ell(Y, f_{\theta^*}(X))\right)^2\right] \leq K \underbrace{\mathbb{E}\left[\ell(Y, f_\theta(X)) - \ell(Y, f_{\theta^*}(X))\right]}_{=R(\theta)-R^*}.$$

Mammen & Tsybakov margin assumption $\Longrightarrow$ Bernstein condition

Bartlett et al. convexity condition $\Longrightarrow$ Bernstein condition

Bernstein condition $\Longrightarrow$ fast rates

$$
\begin{aligned}
&\mathbb{E}\left[\left(\ell(Y, f_\theta(X)) - \ell(Y, f_{\theta^*}(X))\right)^2\right] \\
&= \mathbb{E}\left[\underbrace{\left|\ell(Y, f_\theta(X)) - \ell(Y, f_{\theta^*}(X))\right|}_{\leq 1}\left|\ell(Y, f_\theta(X)) - \ell(Y, f_{\theta^*}(X))\right|\right] \\
&\leq \mathbb{E}\left[\left|\ell(Y, f_\theta(X)) - \ell(Y, f_{\theta^*}(X))\right|\right] \not\leq \mathbb{E}\left[\ell(Y, f_\theta(X)) - \ell(Y, f_{\theta^*}(X))\right].
\end{aligned}
$$

$$\mathbb{E}\left[\left(\ell(Y, f_\theta(X)) - \ell(Y, f_{\theta^*}(X))\right)^2\right]$$
$$= \mathbb{E}\left[\underbrace{\left|\ell(Y, f_\theta(X)) - \ell(Y, f_{\theta^*}(X))\right|}_{\leq 1}\left|\ell(Y, f_\theta(X)) - \ell(Y, f_{\theta^*}(X))\right|\right]$$
$$\leq \mathbb{E}\left[\left|\ell(Y, f_\theta(X)) - \ell(Y, f_{\theta^*}(X))\right|\right] \not< \mathbb{E}\left[\ell(Y, f_\theta(X)) - \ell(Y, f_{\theta^*}(X))\right].$$

If we assume that there is a "uniformly best" θ^*, that is, with probability 1 on (X, Y),

$$\ell(Y, f_{\theta^*}(X)) \leq \ell(Y, f_\theta(X))$$

then we obtain

$$\mathbb{E}\left[\left(\ell(Y, f_\theta(X)) - \ell(Y, f_{\theta^*}(X))\right)^2\right]$$
$$\leq \mathbb{E}\left[\left(\ell(Y, f_\theta(X)) - \ell(Y, f_{\theta^*}(X))\right)\right] = R(\theta) - R^*.$$

$$
\begin{aligned}
&\mathbb{E}\left[\left(\ell(Y, f_\theta(X)) - \ell(Y, f_{\theta^*}(X))\right)^2\right] \\
&= \mathbb{E}\left[\underbrace{\left|\ell(Y, f_\theta(X)) - \ell(Y, f_{\theta^*}(X))\right|}_{\leq 1}\left|\ell(Y, f_\theta(X)) - \ell(Y, f_{\theta^*}(X))\right|\right] \\
&\leq \mathbb{E}\left[\left|\ell(Y, f_\theta(X)) - \ell(Y, f_{\theta^*}(X))\right|\right] \nless \mathbb{E}\left[\ell(Y, f_\theta(X)) - \ell(Y, f_{\theta^*}(X))\right].
\end{aligned}
$$

If we assume that there is a "uniformly best" θ^*, that is, with probability 1 on (X, Y),

$$
\ell(Y, f_{\theta^*}(X)) \leq \ell(Y, f_\theta(X))
$$

then we obtain

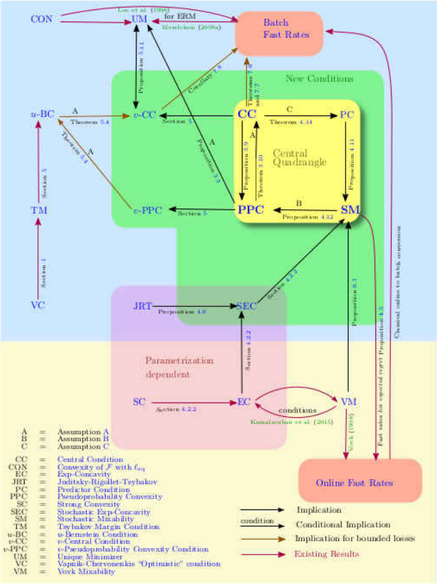
$$
\begin{aligned}
&\mathbb{E}\left[\left(\ell(Y, f_\theta(X)) - \ell(Y, f_{\theta^*}(X))\right)^2\right] \\
&\qquad\qquad \leq \mathbb{E}\left[\left(\ell(Y, f_\theta(X)) - \ell(Y, f_{\theta^*}(X))\right)\right] = R(\theta) - R^*.
\end{aligned}
$$

This is the case if $R^* = 0 \Rightarrow \ell(Y, f_{\theta^*}(X)) = 0$ with proba. 1.

Reminder – Bernstein condition

$$\mathbb{E}\left[\left(\ell(Y, f_\theta(X)) - \ell(Y, f_{\theta^*}(X))\right)^2\right] \leq K[R(\theta) - R^*].$$

$$\begin{tikzcd}
 & & \text{Mammen \& Tsybakov margin assumption} \arrow[d, Rightarrow] & \\
R^* = 0 \arrow[r, Rightarrow] \arrow[urr, Rightarrow, bend left] & \text{uniformly best } \theta^* \arrow[r, Rightarrow] & \text{Bernstein condition} \arrow[r, Rightarrow] & \text{fast rates} \\
 & & \text{Bartlett et al. convexity condition} \arrow[u, Rightarrow] &
\end{tikzcd}$$



Van Erven, T., Grünwald, P., Mehta, N., Reid, M. and Williamson, R. (2015). Fast Rates in Statistical and Online Learning. *JMLR*.

Example : linear regression with quadratic loss,

$$f_\theta(x) = \langle \theta, x \rangle$$

$$\ell(y, f_\theta(x)) = (y - \langle \theta, x \rangle)^2.$$

Example : linear regression with quadratic loss,

$$f_\theta(x) = \langle \theta, x \rangle$$

$$\ell(y, f_\theta(x)) = (y - \langle \theta, x \rangle)^2.$$

- We have to impose boundedness conditions on $\mathcal{Y} \subset \mathbb{R}$ and $\mathcal{X}, \Theta \subset \mathbb{R}^d$ to get $0 \leq \ell \leq 1$.
- We can check Bartlett *et al* condition with $\delta(\theta, \theta^*) = | \langle x, \theta - \theta^* \rangle |$.

Example : linear regression with quadratic loss,

$$f_\theta(x) = \langle \theta, x \rangle$$

$$\ell(y, f_\theta(x)) = (y - \langle \theta, x \rangle)^2.$$

- We have to impose boundedness conditions on $\mathcal{Y} \subset \mathbb{R}$ and $\mathcal{X}, \Theta \subset \mathbb{R}^d$ to get $0 \leq \ell \leq 1$.
- We can check Bartlett *et al* condition with $\delta(\theta, \theta^*) = |\langle x, \theta - \theta^* \rangle|$.

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\pi}_\lambda}[R(\theta)]\right] \leq R^* + \mathcal{O}\left(\frac{d}{n}\log\frac{n}{d}\right).$$

Example : linear regression with quadratic loss,

$$f_\theta(x) = \langle \theta, x \rangle$$

$$\ell(y, f_\theta(x)) = (y - \langle \theta, x \rangle)^2.$$

- We have to impose boundedness conditions on $\mathcal{Y} \subset \mathbb{R}$ and $\mathcal{X}, \Theta \subset \mathbb{R}^d$ to get $0 \leq \ell \leq 1$.
- We can check Bartlett *et al* condition with $\delta(\theta, \theta^*) = | \langle x, \theta - \theta^* \rangle |$.

$$\mathbb{E}_{\mathcal{S}} \left[\mathbb{E}_{\theta \sim \hat{\pi}_\lambda} [R(\theta)] \right] \leq R^* + \mathcal{O}\left(\frac{d}{n} \log \frac{n}{d} \right).$$

- Note that it is actually possible to get rid of some boundedness conditions, as well as to get rid of the log terms.

Catoni, O. (2004). *Statistical learning theory and Stochastic optimization*. Saint-Flour summer school on Probability Theory, Springer Lecture Notes in Mathematics.

Example : high-dimensional sparse linear regression with quadratic loss. That is, $d > n$ but θ^* has $d_0 \ll d$ non-zero components.

Example : high-dimensional sparse linear regression with quadratic loss. That is, $d > n$ but θ^* has $d_0 \ll d$ non-zero components.

- Sparsity inducing prior : under π,

$$\theta_j \begin{cases} = 0 \text{ with probability } p, \\ \sim \mathcal{U}[a, b] \text{ with probability } 1 - p. \end{cases}$$

Example : high-dimensional sparse linear regression with quadratic loss. That is, $d > n$ but θ^* has $d_0 \ll d$ non-zero components.

- Sparsity inducing prior : under π,

$$\theta_j \begin{cases} = 0 \text{ with probability } p, \\ \sim \mathcal{U}[a,b] \text{ with probability } 1-p. \end{cases}$$

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\pi}_\lambda}[R(\theta)]\right] \leq R^* + \mathcal{O}\left(\frac{d_0}{n}\log(d)\right).$$

Dalalyan, A. and Tsybakov, A. B. (2008). Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning*.

Alquier, P. and Lounici, K. (2011). PAC-Bayesian bounds for sparse regression estimation with exponential weights. *Electronic Journal of Statistics*.

More examples :

More examples :

- low-rank tensor estimation :

Suzuki, T. (2015). Convergence rate of Bayesian tensor estimator and its minimax optimality. *ICML*.

More examples :

- low-rank tensor estimation :

Suzuki, T. (2015). Convergence rate of Bayesian tensor estimator and its minimax optimality. *ICML*.

- deep learning :

Chérief-Abdellatif, B.-E. (2020). Convergence Rates of Variational Inference in Sparse Deep Learning. *ICML*.

Steffen, M. F. and Trabs, M. (2022). *PAC-Bayes training for neural networks : sparsity and uncertainty quantification*. ArXiv preprint arXiv :2204.12392.

More examples :

- low-rank tensor estimation :

Suzuki, T. (2015). Convergence rate of Bayesian tensor estimator and its minimax optimality. *ICML*.

- deep learning :

Chérief-Abdellatif, B.-E. (2020). Convergence Rates of Variational Inference in Sparse Deep Learning. *ICML*.

Steffen, M. F. and Trabs, M. (2022). *PAC-Bayes training for neural networks : sparsity and uncertainty quantification*. ArXiv preprint arXiv :2204.12392.

- quantum tomography (reconstructing the quantum state of a system from measurements) :

Mai, T. T. and Alquier, P. (2017). Pseudo-Bayesian quantum tomography with rank-adaptation. *Journal of Statistical Planning and Inference*.

- ...

(General) Bernstein condition

For $K > 0$ and $\gamma \in [0, 1]$,

$$\mathbb{E}\left[\left(\ell(Y, f_\theta(X)) - \ell(Y, f_{\theta^*}(X))\right)^2\right] \leq K[R(\theta) - R^*]^\gamma.$$

- So far, we studied $\gamma = 1$.

(General) Bernstein condition

For $K > 0$ and $\gamma \in [0, 1]$,

$$\mathbb{E}\left[\left(\ell(Y, f_{\theta}(X)) - \ell(Y, f_{\theta^*}(X))\right)^2\right] \leq K[R(\theta) - R^*]^{\gamma}.$$

- So far, we studied $\gamma = 1$.
- Always satisfied for $\gamma = 0$ (bounded loss).

(General) Bernstein condition

For $K > 0$ and $\gamma \in [0, 1]$,

$$\mathbb{E}\left[\left(\ell(Y, f_\theta(X)) - \ell(Y, f_{\theta^*}(X))\right)^2\right] \leq K[R(\theta) - R^*]^\gamma.$$

- So far, we studied $\gamma = 1$.
- Always satisfied for $\gamma = 0$ (bounded loss).
- In the general case,

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\pi}_\lambda}[R(\theta)] - R^*\right] \leq \mathcal{O}\left(\left(\frac{D}{N}\right)^{\frac{1}{2-\gamma}} \log\frac{N}{D}\right).$$

(General) Bernstein condition

For $K > 0$ and $\gamma \in [0, 1]$,

$$\mathbb{E}\left[\left(\ell(Y, f_\theta(X)) - \ell(Y, f_{\theta^*}(X))\right)^2\right] \leq K[R(\theta) - R^*]^\gamma.$$

- So far, we studied $\gamma = 1$.
- Always satisfied for $\gamma = 0$ (bounded loss).
- In the general case,

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta \sim \hat{\pi}_\lambda}[R(\theta)] - R^*\right] \leq \mathcal{O}\left(\left(\frac{D}{N}\right)^{\frac{1}{2-\gamma}} \log \frac{N}{D}\right).$$

- Mammen and Tsybakov proved a sufficient margin condition for $0 < \gamma < 1$:

$$\mathbb{P}(|\eta(X) - 1/2| < \tau) = \mathcal{O}\left(\tau^{\frac{1}{1-\gamma}}\right) \text{ for } \tau \to 0.$$

1. PAC-Bayes bounds : introduction
   - Generalization bounds and PAC-Bayes
   - Minimization of the PAC-Bayes bound
   - A zoo of PAC-Bayes bounds

2. PAC-Bayes and Mutual Information bounds
   - Excess risk bounds
   - Fast rates
   - Mutual information bounds

Reminder – Catoni's PAC-Bayes bound, 2003

Fix $\lambda > 0$ and π. With proba. at least $1 - \delta$ on $\mathcal{S}$, for any randomized estimator $\hat{\rho}$,

$$\mathbb{E}_{\theta\sim\hat{\rho}}[R(\theta)] \leq \mathbb{E}_{\theta\sim\hat{\rho}}[R_n(\theta)] + \frac{\mathrm{KL}(\hat{\rho}\|\pi) + \log\frac{1}{\delta}}{\lambda} + \frac{\lambda}{8n}.$$

Reminder – Catoni's PAC-Bayes bound, 2003

Fix $\lambda > 0$ and π. With proba. at least $1-\delta$ on $\mathcal{S}$, for any randomized estimator $\hat{\rho}$,

$$\mathbb{E}_{\theta\sim\hat{\rho}}[R(\theta)] \leq \mathbb{E}_{\theta\sim\hat{\rho}}[R_n(\theta)] + \frac{\mathrm{KL}(\hat{\rho}\|\pi) + \log\frac{1}{\delta}}{\lambda} + \frac{\lambda}{8n}.$$

For λ and π are fixed, this motivated the introduction of the Gibbs posterior $\hat{\rho} = \hat{\pi}_\lambda$, that minimizes the r.h.s.

Reminder – Catoni's PAC-Bayes bound, 2003

Fix $\lambda > 0$ and π. With proba. at least $1 - \delta$ on $\mathcal{S}$, for any randomized estimator $\hat{\rho}$,

$$\mathbb{E}_{\theta\sim\hat{\rho}}[R(\theta)] \leq \mathbb{E}_{\theta\sim\hat{\rho}}[R_n(\theta)] + \frac{\mathrm{KL}(\hat{\rho}\|\pi) + \log\frac{1}{\delta}}{\lambda} + \frac{\lambda}{8n}.$$

For λ and π are fixed, this motivated the introduction of the Gibbs posterior $\hat{\rho} = \hat{\pi}_\lambda$, that minimizes the r.h.s. Then, we applied the bound in expectation to derive rates of convergence :

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\pi}_\lambda}[R(\theta)]\right] \leq \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\pi}_\lambda}[R_n(\theta)] + \frac{\mathrm{KL}(\hat{\pi}_\lambda\|\pi)}{\lambda} + \frac{\lambda}{8n}\right].$$

Reminder – Catoni's PAC-Bayes bound, 2003

Fix $\lambda > 0$ and π. With proba. at least $1-\delta$ on $\mathcal{S}$, for any randomized estimator $\hat{\rho}$,

$$\mathbb{E}_{\theta\sim\hat{\rho}}[R(\theta)] \leq \mathbb{E}_{\theta\sim\hat{\rho}}[R_n(\theta)] + \frac{\mathrm{KL}(\hat{\rho}\|\pi) + \log\frac{1}{\delta}}{\lambda} + \frac{\lambda}{8n}.$$

For λ and π are fixed, this motivated the introduction of the Gibbs posterior $\hat{\rho} = \hat{\pi}_\lambda$, that minimizes the r.h.s. Then, we applied the bound in expectation to derive rates of convergence :

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\pi}_\lambda}[R(\theta)]\right] \leq \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\pi}_\lambda}[R_n(\theta)] + \frac{\mathrm{KL}(\hat{\pi}_\lambda\|\pi)}{\lambda} + \frac{\lambda}{8n}\right].$$

But... why did we keep the same λ and π ?

PAC-Bayes bound in expectation – v2.0

- Fix $\Lambda > 0$, Π and the randomized estimator $\hat{\rho}$ (for example $\hat{\rho} = \hat{\pi}_\lambda$).

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\rho}}[R(\theta)]\right] \leq \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\rho}}[R_n(\theta)] + \frac{\mathrm{KL}(\hat{\rho}\|\Pi)}{\Lambda} + \frac{\Lambda}{8n}\right].$$

PAC-Bayes bound in expectation – v2.0

- Fix $\Lambda > 0$, Π and the randomized estimator $\hat{\rho}$ (for example $\hat{\rho} = \hat{\pi}_\lambda$).

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\rho}}[R(\theta)]\right] \leq \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\rho}}[R_n(\theta)] + \frac{\mathrm{KL}(\hat{\rho}\|\Pi)}{\Lambda} + \frac{\Lambda}{8n}\right].$$

Thus,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\rho}}[R(\theta)]\right] \\
&\leq \inf_{\Lambda>0}\inf_{\Pi\in\mathcal{M}(\Theta)} \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\rho}}[R_n(\theta)] + \frac{\mathrm{KL}(\hat{\rho}\|\Pi)}{\Lambda} + \frac{\Lambda}{8n}\right] \\
&= \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\rho}}[R_n(\theta)]\right] + \inf_{\Lambda>0}\inf_{\Pi\in\mathcal{M}(\Theta)} \mathbb{E}_{\mathcal{S}}\left[\frac{\mathrm{KL}(\hat{\rho}\|\Pi)}{\Lambda} + \frac{\Lambda}{8n}\right].
\end{aligned}$$

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\rho}}[R_n(\theta)]\right] + \inf_{\Lambda>0} \inf_{\Pi\in\mathcal{M}(\Theta)} \mathbb{E}_{\mathcal{S}}\left[\frac{\mathrm{KL}(\hat{\rho}\|\Pi)}{\Lambda} + \frac{\Lambda}{8n}\right],$$

the infimum is reached, as shown by :

Catoni, O. (2007). *PAC-Bayesian supervised learning : the thermodynamics of statistical learning.* IMS lecture notes – monograph series.

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\rho}}[R_n(\theta)]\right] + \inf_{\Lambda>0} \inf_{\Pi\in\mathcal{M}(\Theta)} \mathbb{E}_{\mathcal{S}}\left[\frac{\mathrm{KL}(\hat{\rho}\|\Pi)}{\Lambda} + \frac{\Lambda}{8n}\right],$$

the infimum is reached, as shown by :

Catoni, O. (2007). *PAC-Bayesian supervised learning : the thermodynamics of statistical learning.* IMS lecture notes – monograph series.

$$\mathbb{E}_{\mathcal{S}}\mathrm{KL}(\hat{\rho}\|\Pi) = \mathbb{E}_{\mathcal{S}}\mathrm{KL}(\hat{\rho}\|\mathbb{E}_{\mathcal{S}}\hat{\rho}) + \underbrace{\mathrm{KL}(\mathbb{E}_{\mathcal{S}}\hat{\rho}\|\Pi)}_{=0 \text{ if } \Pi=\mathbb{E}_{\mathcal{S}}\hat{\rho}}.$$

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\rho}}[R_n(\theta)]\right] + \inf_{\Lambda>0} \inf_{\Pi\in\mathcal{M}(\Theta)} \mathbb{E}_{\mathcal{S}}\left[\frac{\mathrm{KL}(\hat{\rho}\|\Pi)}{\Lambda} + \frac{\Lambda}{8n}\right],$$

the infimum is reached, as shown by :

Catoni, O. (2007). *PAC-Bayesian supervised learning : the thermodynamics of statistical learning.* IMS lecture notes – monograph series.

$$\mathbb{E}_{\mathcal{S}}\mathrm{KL}(\hat{\rho}\|\Pi) = \mathbb{E}_{\mathcal{S}}\mathrm{KL}(\hat{\rho}\|\mathbb{E}_{\mathcal{S}}\hat{\rho}) + \underbrace{\mathrm{KL}(\mathbb{E}_{\mathcal{S}}\hat{\rho}\|\Pi)}_{=0 \text{ if } \Pi=\mathbb{E}_{\mathcal{S}}\hat{\rho}}.$$

- $\mathbb{E}_{\mathcal{S}}\hat{\rho} \in \mathcal{M}(\Theta)$ defined by $[\mathbb{E}_{\mathcal{S}}\hat{\rho}](E) = \mathbb{E}_{\mathcal{S}}[\hat{\rho}(E)]$.

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\rho}}[R_n(\theta)]\right] + \inf_{\Lambda>0}\inf_{\Pi\in\mathcal{M}(\Theta)}\mathbb{E}_{\mathcal{S}}\left[\frac{\mathrm{KL}(\hat{\rho}\|\Pi)}{\Lambda}+\frac{\Lambda}{8n}\right],$$

the infimum is reached, as shown by :

Catoni, O. (2007). *PAC-Bayesian supervised learning : the thermodynamics of statistical learning.* IMS lecture notes – monograph series.

$$\mathbb{E}_{\mathcal{S}}\mathrm{KL}(\hat{\rho}\|\Pi) = \mathbb{E}_{\mathcal{S}}\mathrm{KL}(\hat{\rho}\|\mathbb{E}_{\mathcal{S}}\hat{\rho}) + \underbrace{\mathrm{KL}(\mathbb{E}_{\mathcal{S}}\hat{\rho}\|\Pi)}_{=0 \text{ if } \Pi=\mathbb{E}_{\mathcal{S}}\hat{\rho}}.$$

- $\mathbb{E}_{\mathcal{S}}\hat{\rho} \in \mathcal{M}(\Theta)$ defined by $[\mathbb{E}_{\mathcal{S}}\hat{\rho}](E) = \mathbb{E}_{\mathcal{S}}[\hat{\rho}(E)]$.
- the first term in the r.h.s. has a nice interpretation...

Let $(U, V) \sim P$. Let P_U and P_V denote their marginals.

Let $(U, V) \sim P$. Let P_U and P_V denote their marginals. If U and V were independent, $P = P_U \otimes P_V$.

Let $(U, V) \sim P$. Let P_U and P_V denote their marginals. If U and V were independent, $P = P_U \otimes P_V$.

Mutual information between two random variables

$$\mathcal{I}(U, V) := \mathrm{KL}(P \| P_U \otimes P_V).$$

Let $(U, V) \sim P$. Let P_U and P_V denote their marginals. If U and V were independent, $P = P_U \otimes P_V$.

Mutual information between two random variables

$$\mathcal{I}(U, V) := \mathrm{KL}(P \| P_U \otimes P_V).$$

Note : $\mathcal{I}(U, V)$ depends on the distribution P of (U, V), not on (U, V). This is confusing... remember that $\mathbb{E}(U)$ is not a function of U !

Let $(U, V) \sim P$. Let P_U and P_V denote their marginals. If U and V were independent, $P = P_U \otimes P_V$.

Mutual information between two random variables

$$\mathcal{I}(U, V) := \mathrm{KL}(P \| P_U \otimes P_V).$$

Note : $\mathcal{I}(U, V)$ depends on the distribution P of (U, V), not on (U, V). This is confusing... remember that $\mathbb{E}(U)$ is not a function of U !

Proposition

$$\mathcal{I}(U, V) = \mathbb{E}_U\left[\mathrm{KL}(P_{V|U} \| P_V)\right].$$

Let $(U, V) \sim P$. Let P_U and P_V denote their marginals. If U and V were independent, $P = P_U \otimes P_V$.

Mutual information between two random variables

$$\mathcal{I}(U, V) := \mathrm{KL}(P \| P_U \otimes P_V).$$

Note : $\mathcal{I}(U, V)$ depends on the distribution P of (U, V), not on (U, V). This is confusing... remember that $\mathbb{E}(U)$ is not a function of U !

Proposition

$$\mathcal{I}(U, V) = \mathbb{E}_U \Big[\mathrm{KL}(P_{V|U} \| P_V) \Big].$$

Thus,

$$\mathbb{E}_{\mathcal{S}} \mathrm{KL}(\hat{\rho} \| \Pi) = \underbrace{\mathbb{E}_{\mathcal{S}} \mathrm{KL}(\hat{\rho} \| \mathbb{E}_{\mathcal{S}} \hat{\rho})}_{=:\mathcal{I}(\theta, \mathcal{S})} + \underbrace{\mathrm{KL}(\mathbb{E}_{\mathcal{S}} \hat{\rho} \| \Pi)}_{=0 \text{ if } \Pi = \mathbb{E}_{\mathcal{S}} \hat{\rho}}.$$

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\rho}}[R(\theta)]\right] \\
&\leq \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\rho}}[R_n(\theta)]\right] + \inf_{\Lambda>0} \inf_{\Pi\in\mathcal{M}(\Theta)} \mathbb{E}_{\mathcal{S}}\left[\frac{\mathrm{KL}(\hat{\rho}\|\Pi)}{\Lambda} + \frac{\Lambda}{8n}\right] \\
&= \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\rho}}[R_n(\theta)]\right] + \inf_{\Lambda>0} \left[\frac{\mathcal{I}(\theta,\mathcal{S})}{\Lambda} + \frac{\Lambda}{8n}\right].
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\rho}}[R(\theta)]\right] \\
&\leq \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\rho}}[R_n(\theta)]\right] + \inf_{\Lambda>0}\inf_{\Pi\in\mathcal{M}(\Theta)}\mathbb{E}_{\mathcal{S}}\left[\frac{\mathrm{KL}(\hat{\rho}\|\Pi)}{\Lambda}+\frac{\Lambda}{8n}\right] \\
&= \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\rho}}[R_n(\theta)]\right] + \inf_{\Lambda>0}\left[\frac{\mathcal{I}(\theta,\mathcal{S})}{\Lambda}+\frac{\Lambda}{8n}\right].
\end{aligned}
$$

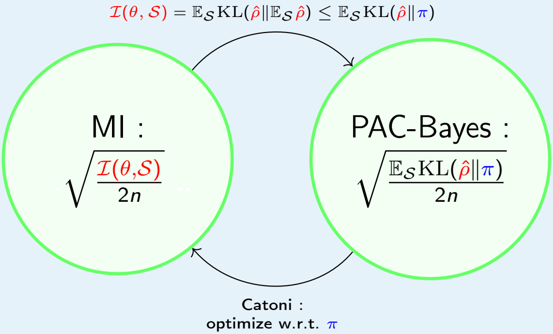
Mutual information bound

$$
\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\rho}}[R(\theta)]\right] \leq \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\rho}}[R_n(\theta)]\right] + \sqrt{\frac{\mathcal{I}(\theta,\mathcal{S})}{2n}}.
$$

Russo, D. and Zou, J. (2019). How much does your data exploration overfit ? controlling bias via information usage. *IEEE Transactions on Information Theory*.

Mutual information bound

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\rho}}[R(\theta)]\right] \leq \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\hat{\rho}}[R_n(\theta)]\right] + \sqrt{\frac{\mathcal{I}(\theta,\mathcal{S})}{2n}}.$$



Let us illustrate the improvements of MI over PAC-Bayes on a simple example :

- finite parameter set $\Theta = \{\theta_1, \dots, \theta_M\}$.
- $\hat{\rho} = \delta_{\hat{\theta}}$ the point mass on the ERM $\hat{\theta}$.

Let us illustrate the improvements of MI over PAC-Bayes on a simple example :

- finite parameter set $\Theta = \{\theta_1, \dots, \theta_M\}$.
- $\hat{\rho} = \delta_{\hat{\theta}}$ the point mass on the ERM $\hat{\theta}$.

$$\mathbb{E}_{\mathcal{S}}[R(\hat{\theta})] \leq \mathbb{E}_{\mathcal{S}}[R_n(\hat{\theta})] + \sqrt{\frac{\mathcal{I}(\hat{\theta}, \mathcal{S})}{2n}}.$$

Let us illustrate the improvements of MI over PAC-Bayes on a simple example :

- finite parameter set $\Theta = \{\theta_1, \dots, \theta_M\}$.
- $\hat{\rho} = \delta_{\hat{\theta}}$ the point mass on the ERM $\hat{\theta}$.

$$\mathbb{E}_{\mathcal{S}}[R(\hat{\theta})] \leq \mathbb{E}_{\mathcal{S}}[R_n(\hat{\theta})] + \sqrt{\frac{\mathcal{I}(\hat{\theta}, \mathcal{S})}{2n}}.$$

$$\mathbb{E}_{\mathcal{S}}[R_n(\hat{\theta})] = \mathbb{E}_{\mathcal{S}}[\inf_{\theta \in \Theta} R_n(\theta)] \leq \inf_{\theta \in \Theta} \mathbb{E}_{\mathcal{S}}[R_n(\theta)] = \inf_{\theta \in \Theta} R(\theta) = R^*.$$

Let us illustrate the improvements of MI over PAC-Bayes on a simple example :

- finite parameter set $\Theta = \{\theta_1, \dots, \theta_M\}$.
- $\hat{\rho} = \delta_{\hat{\theta}}$ the point mass on the ERM $\hat{\theta}$.

$$\mathbb{E}_{\mathcal{S}}[R(\hat{\theta})] \leq \mathbb{E}_{\mathcal{S}}[R_n(\hat{\theta})] + \sqrt{\frac{\mathcal{I}(\hat{\theta}, \mathcal{S})}{2n}}.$$

$$\mathbb{E}_{\mathcal{S}}[R_n(\hat{\theta})] = \mathbb{E}_{\mathcal{S}}[\inf_{\theta \in \Theta} R_n(\theta)] \leq \inf_{\theta \in \Theta} \mathbb{E}_{\mathcal{S}}[R_n(\theta)] = \inf_{\theta \in \Theta} R(\theta) = R^*.$$

MI bound for the ERM

$$\mathbb{E}_{\mathcal{S}}[R(\hat{\theta})] \leq R^* + \sqrt{\frac{\mathcal{I}(\hat{\theta}, \mathcal{S})}{2n}}.$$

Reminder – MI bound for the ERM

$$\mathbb{E}_{\mathcal{S}}[R(\hat{\theta})] \leq R^* + \sqrt{\frac{\mathcal{I}(\hat{\theta}, \mathcal{S})}{2n}}.$$

Reminder – MI bound for the ERM

$$\mathbb{E}_{\mathcal{S}}[R(\hat{\theta})] \leq R^* + \sqrt{\frac{\mathcal{I}(\hat{\theta}, \mathcal{S})}{2n}}.$$

PAC-Bayes : let π be uniform on Θ,

$$\begin{aligned} \mathcal{I}(\hat{\theta}, \mathcal{S}) &\leq \mathbb{E}_{\mathcal{S}} \mathrm{KL}(\delta_{\hat{\theta}} \| \pi) \\ &= \log(M). \end{aligned}$$

Reminder – MI bound for the ERM

$$\mathbb{E}_{\mathcal{S}}[R(\hat{\theta})] \leq R^* + \sqrt{\frac{\mathcal{I}(\hat{\theta}, \mathcal{S})}{2n}}.$$

PAC-Bayes : let π be uniform on Θ,

$$\begin{aligned}\mathcal{I}(\hat{\theta}, \mathcal{S}) &\leq \mathbb{E}_{\mathcal{S}} \mathrm{KL}(\delta_{\hat{\theta}} \| \pi) \\ &= \log(M).\end{aligned}$$

$$\mathbb{E}_{\mathcal{S}}[R(\hat{\theta})] \leq R^* + \sqrt{\frac{\log(M)}{2n}}.$$

Reminder – MI bound for the ERM

$$\mathbb{E}_{\mathcal{S}}[R(\hat{\theta})] \leq R^* + \sqrt{\frac{\mathcal{I}(\hat{\theta}, \mathcal{S})}{2n}}.$$

Reminder – MI bound for the ERM

$$\mathbb{E}_{\mathcal{S}}[R(\hat{\theta})] \leq R^* + \sqrt{\frac{\mathcal{I}(\hat{\theta},\mathcal{S})}{2n}}.$$

Catoni : $\pi(\theta) = \frac{\exp(-\alpha\Delta(\theta))}{\sum_{\theta\in\Theta}\exp(-\alpha\Delta(\theta))}$ where $\Delta(\theta) = R(\theta) - R^*$,

$$\mathcal{I}(\hat{\theta},\mathcal{S}) \leq \mathbb{E}_{\mathcal{S}}\mathrm{KL}(\delta_{\hat{\theta}}\|\pi) = \mathbb{E}_{\mathcal{S}}\left[\alpha\Delta(\hat{\theta}) + \log\sum_{\theta\in\Theta}\exp(-\alpha\Delta(\theta))\right]$$

Reminder – MI bound for the ERM

$$\mathbb{E}_{\mathcal{S}}[R(\hat{\theta})] \leq R^* + \sqrt{\frac{\mathcal{I}(\hat{\theta},\mathcal{S})}{2n}}.$$

Catoni : $\pi(\theta) = \frac{\exp(-\alpha\Delta(\theta))}{\sum_{\theta\in\Theta}\exp(-\alpha\Delta(\theta))}$ where $\Delta(\theta) = R(\theta) - R^*$,

$$\mathcal{I}(\hat{\theta},\mathcal{S}) \leq \mathbb{E}_{\mathcal{S}}\mathrm{KL}(\delta_{\hat{\theta}}\|\pi) = \mathbb{E}_{\mathcal{S}}\left[\alpha\Delta(\hat{\theta}) + \log\sum_{\theta\in\Theta}\exp(-\alpha\Delta(\theta))\right]$$

Put $\zeta = \log\sum_{\theta\in\Theta}\exp(-\alpha\Delta(\theta))$, we obtain the inequation :

$$\mathbb{E}_{\mathcal{S}}[\Delta(\hat{\theta})] \leq \sqrt{\frac{\alpha\mathbb{E}_{\mathcal{S}}[\Delta(\hat{\theta})] + \zeta}{2n}}.$$

$$\mathbb{E}_{\mathcal{S}}[\Delta(\hat{\theta})] \leq \sqrt{\frac{\alpha \mathbb{E}_{\mathcal{S}}[\Delta(\hat{\theta})] + \zeta}{2n}}.$$

$$\mathbb{E}_{\mathcal{S}}[\Delta(\hat{\theta})] \leq \sqrt{\frac{\alpha \mathbb{E}_{\mathcal{S}}[\Delta(\hat{\theta})] + \zeta}{2n}}.$$

$$\mathbb{E}_{\mathcal{S}}[\Delta(\hat{\theta})] \leq \frac{\alpha}{4n} + \frac{1}{2}\sqrt{\frac{1}{n}\left[\frac{\alpha^2}{4n} + 2\zeta\right]}.$$

$$\mathbb{E}_{\mathcal{S}}[\Delta(\hat{\theta})] \leq \sqrt{\frac{\alpha \mathbb{E}_{\mathcal{S}}[\Delta(\hat{\theta})] + \zeta}{2n}}.$$

$$\mathbb{E}_{\mathcal{S}}[\Delta(\hat{\theta})] \leq \frac{\alpha}{4n} + \frac{1}{2}\sqrt{\frac{1}{n}\left[\frac{\alpha^2}{4n} + 2\zeta\right]}.$$

$$\zeta = \log \sum_{\theta \in \Theta} \exp(-\alpha \Delta(\theta)) \leq \log(M)$$

$$\zeta = \log \left[1 + \sum_{\theta \neq \theta^*} \exp(-\alpha \Delta(\theta))\right] \leq M \exp\left(-\alpha \min_{\theta \neq \theta^*} \Delta(\theta)\right).$$

$$\mathbb{E}_{\mathcal{S}}[\Delta(\hat{\theta})] \leq \sqrt{\frac{\alpha \mathbb{E}_{\mathcal{S}}[\Delta(\hat{\theta})] + \zeta}{2n}}.$$

$$\mathbb{E}_{\mathcal{S}}[\Delta(\hat{\theta})] \leq \frac{\alpha}{4n} + \frac{1}{2}\sqrt{\frac{1}{n}\left[\frac{\alpha^2}{4n} + 2\zeta\right]}.$$

$$\zeta = \log \sum_{\theta \in \Theta} \exp(-\alpha \Delta(\theta)) \leq \log(M)$$

$$\zeta = \log \left[1 + \sum_{\theta \neq \theta^*} \exp(-\alpha \Delta(\theta))\right] \leq M \exp\left(-\alpha \min_{\theta \neq \theta^*} \Delta(\theta)\right).$$

Take $\alpha = 2\sqrt{n}$.

Recap : MI bound for the ERM on a finite Θ

Assume $\Theta = \{\theta_1, \dots, \theta_M\}$ and put $\Delta = \min_{\theta \neq \theta^*}[R(\theta) - R^*]$. Then

$$\mathbb{E}_{\mathcal{S}}[R(\hat{\theta})] \leq R^* + \sqrt{\frac{\frac{1}{2} + \min\left[M\exp(-\Delta\sqrt{2n}), \log(M)\right]}{2n}} + \frac{1}{2n}.$$

Starting from the PAC-Bayes bound in expectation, we can combine the improvements due to Bernstein assumption to the optimization with respect to the prior.

Starting from the PAC-Bayes bound in expectation, we can combine the improvements due to Bernstein assumption to the optimization with respect to the prior.

MI bound with Bernstein condition

- Assume Bernstein condition is satisfied with constant K.

Fix $\lambda = n/\max(2K,1)$, and $\hat{\rho}$, then

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\hat{\rho}}[R(\theta) - R^*] \leq 2\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\hat{\rho}}[R_n(\theta) - R^*] + \frac{\max(2K,1)\mathcal{I}(\theta,\mathcal{S})}{n}.$$

A recent survey/tutorials that covers MI bounds in depth, and their relation to PAC-Bayes bounds :

Hellström, F., Durisi, G., Guedj, B. and Raginsky, M. (2023). *Generalization bounds : Perspectives from information theory and PAC-Bayes*. Arxiv preprint arXiv :2309.04381.

Review of topics not covered in these slides.

Review of topics not covered in these slides.

Unbounded losses :

Haddouche, M. and Guedj, B. (2023). *PAC-Bayes Generalisation Bounds for Heavy-Tailed Losses through Supermartingales*. Transactions on Machine Learning Research.

Rodríguez-Gálvez, B., Thobaben, R. and Skoglund, M. (2023). *More PAC-Bayes bounds : From bounded losses, to losses with general tail behaviors, to anytime-validity*. ArXiv preprint arXiv :2306.12214

Review of topics not covered in these slides.

Unbounded losses :

Haddouche, M. and Guedj, B. (2023). *PAC-Bayes Generalisation Bounds for Heavy-Tailed Losses through Supermartingales*. Transactions on Machine Learning Research.

Rodríguez-Gálvez, B., Thobaben, R. and Skoglund, M. (2023). *More PAC-Bayes bounds : From bounded losses, to losses with general tail behaviors, to anytime-validity*. ArXiv preprint arXiv :2306.12214

Non i.i.d., time series...

Alquier, P. and Wintenberger, O. (2012). Model selection for weakly dependent time series forecasting. *Bernoulli.*

Alquier, P., Li. X. and Wintenberger, O. (2013). Prediction of time series by statistical learning : general losses and fast rates. *Dependence Modeling*.

Banerjee, I., Rao, V. A. and Honnappa, H. (2021). PAC-Bayes bounds on variational tempered posteriors for Markov models. *Entropy*.

Robust estimator (not Bayesian) studied by adding a random perturbation, and then using PAC-Bayes bounds.

Catoni, O. (2012). Challenging the empirical mean and empirical variance : a deviation study. *Annales de l'IHP*.

Catoni, O. and Giulini, I. (2017). Dimension free PAC-Bayesian bounds for the estimation of the mean of a random vector. *NeurIPS 2017 Workshop : (Almost) 50 Shades of Bayesian Learning : PAC-Bayesian trends and insights*.

Zhivotovskiy, N. (2024). Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle. *Electronic Journal of Probability*.

Robust estimator (not Bayesian) studied by adding a random perturbation, and then using PAC-Bayes bounds.

Catoni, O. (2012). Challenging the empirical mean and empirical variance : a deviation study. *Annales de l'IHP*.

Catoni, O. and Giulini, I. (2017). Dimension free PAC-Bayesian bounds for the estimation of the mean of a random vector. *NeurIPS 2017 Workshop : (Almost) 50 Shades of Bayesian Learning : PAC-Bayesian trends and insights*.

Zhivotovskiy, N. (2024). Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle. *Electronic Journal of Probability*.

Meta-learning.

Rothfuss, J., Fortuin, V., Josifoski, M. and Krause, A. (2021). PACOH : Bayes-optimal meta-learning with PAC-guarantees. *ICML*.

Riou, C., Alquier, P. and Chérief-Abdellatif, B.-E. (2023). *Bayes meets Bernstein at the Meta Level : an Analysis of Fast Rates in Meta-Learning with PAC-Bayes*. Arxiv preprint arXiv :2302.11709.

PAC-Bayes or MI bounds where $\mathrm{KL}(\rho\|\pi)$ is replaced by another $D(\rho, \pi)$.

Alquier, P. and Guedj, B. (2018). Simpler PAC-Bayesian bounds for hostile data. *Machine Learning*.

Neu, G. and Lugosi, G. (2022). Generalization Bounds via Convex Analysis. *ICML*.

PAC-Bayes or MI bounds where $\mathrm{KL}(\rho\|\pi)$ is replaced by another $D(\rho, \pi)$.

Alquier, P. and Guedj, B. (2018). Simpler PAC-Bayesian bounds for hostile data. *Machine Learning*.

Neu, G. and Lugosi, G. (2022). Generalization Bounds via Convex Analysis. *ICML*.

In particular, Wasserstein distance studied in :

Rodríguez-Gálvez, B., Bassi, G., Thobaben, R. and Skoglund, M. (2021). Tighter expected generalization error bounds via Wasserstein distance *NeurIPS*.

Clerico, E., Shidani, A., Deligiannidis, G. and Doucet, A. (2022). Chained Generalisation Bounds. *COLT*.

Viallard, P., Haddouche, M., Simsekli, U. and Guedj, B. (2023). Learning via Wasserstein-based high probability generalisation bounds. *NeurIPS*.

Neu, G. and Lugosi, G. (2023). *Online-to-PAC Conversions : Generalization Bounds via Regret Analysis*. Arxiv preprint arXiv :2305.19674.

終わり

C'est la fin.

The end.

$t = +\infty$.

終わり

C'est la fin.

The end.

$t = +\infty.$

Thank you !

ありがとう ございした。