# PAC-Bayes and contraction of the posterior

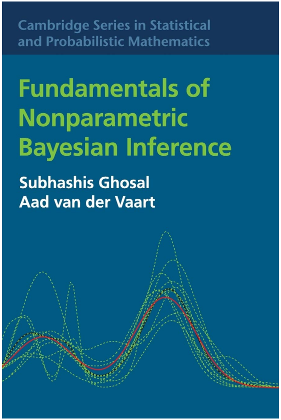
Pierre Alquier

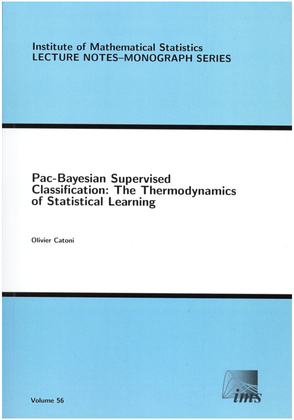
RIKEN

AIP
Center for
Advanced Intelligence Project

The (International) Bayes Club – Apr. 21, 2022

# Two worlds


Cambridge Series in Statistical and Probabilistic Mathematics
Fundamentals of Nonparametric Bayesian Inference
Subhashis Ghosal
Aad van der Vaart


Institute of Mathematical Statistics
LECTURE NOTES–MONOGRAPH SERIES
Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning
Olivier Catoni
Volume 56
ims

# Contents

1. Introduction
   - Contraction of the posterior
   - PAC-Bayes bounds
2. PAC-Bayes point of view on contraction (and vice-versa)
   - An empirical point of view on the prior mass condition
   - Variational approximations

# Contents

1 Introduction
- Contraction of the posterior
- PAC-Bayes bounds

2 PAC-Bayes point of view on contraction (and vice-versa)
- An empirical point of view on the prior mass condition
- Variational approximations

# Notations and setting

- $X_1, \dots, X_n$ i.i.d. from $P_0$,
- $(P_\theta, \theta \in \Theta)$ model, densities $p_\theta(x)$,
- prior $\pi$ on $\Theta$,
- posterior

$$\pi(\mathrm{d}\theta|\mathcal{S}) \propto \left(\prod_{i=1}^{n} p_\theta(X_i)\right) \pi(\mathrm{d}\theta).$$

Question : if $P_0 = P_{\theta_0}$, do we have

$$\mathbb{E}_{\mathcal{S}} \mathbb{P}_{\theta\sim\pi(\cdot|\mathcal{S})}[d(\theta,\theta_0) \leq r_n] \xrightarrow[n\to\infty]{} 1$$

for some

$$r_n \xrightarrow[n\to\infty]{} 0\,?$$

# Conditions for contraction

This can be proven under the following 2 assumptions :

$$\mathcal{B}(r) = \left\{ \theta \in \Theta : KL(P_{\theta_0}, P_\theta) \leq r \text{ and } \mathrm{Var}\left[ \log \frac{p_\theta(X_i)}{p_{\theta_0}(X_i)} \right] \leq r \right\}.$$

## Prior mass condition

The sequence $(r_n)$ satisfies

$$\pi[B(r_n)] \geq \mathrm{e}^{-dnr_n} \text{ that is } \log \pi[B(r_n)] \geq -dnr_n.$$

## Test condition

There is a sequence of tests $\phi_n = \phi_n(\mathcal{S}) \in [0,1]$ such that

$$\mathbb{E}_{\mathcal{S}} \phi_n \xrightarrow[n\to\infty]{} 0, \text{ and } \sup_{d(\theta,\theta_0)>r_n} \mathbb{E}_{\mathcal{S}\sim P_\theta^n}[1-\phi_n] = o\left(\mathrm{e}^{-(d+2)nr_n}\right).$$

# Tempered posteriors - $0 < \alpha < 1$

$$\hat{\pi}_{\alpha}(\mathrm{d}\theta) \propto \left( \prod_{i=1}^{n} p_{\theta}(X_i) \right)^{\alpha} \pi(\mathrm{d}\theta).$$

$\hat{\pi}_{\alpha}$ is more robust than $\pi(\cdot|\mathcal{S})$ to misspecifitation.

Grünwald, P. & Van Ommen, T. (2017). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian analysis*.

## The $\alpha$-Rényi divergence

$$D_{\alpha}(P, R) = \frac{1}{\alpha - 1} \log \int (\mathrm{d}P)^{\alpha} (\mathrm{d}R)^{1-\alpha}.$$

Among others, for $1/2 \leq \alpha$, link with Hellinger and Kullback :

$$\mathcal{H}^2(P, R) \leq D_{\alpha}(P, R) \xrightarrow[\alpha \nearrow 1]{} KL(P, R).$$

# Contraction of tempered posteriors

## Theorem

For any $r_n$ with $nr_n \to \infty$ satisfying the prior mass condition **only**, there is a known $C(d)$ such that

$$\mathbb{E}_{\mathcal{S}} \mathbb{P}_{\theta \sim \hat{\pi}_\alpha} \left( D_\alpha(P_\theta, P_{\theta_0}) \leq \frac{C(d) r_n}{1-\alpha} \right) \xrightarrow[n\to\infty]{} 1.$$

Bhattacharya, A., Pati, D. & Yang, Y. (2019). Bayesian fractional posteriors. *Annals of Statistics*.

# Contents

1 Introduction
- Contraction of the posterior
- PAC-Bayes bounds

2 PAC-Bayes point of view on contraction (and vice-versa)
- An empirical point of view on the prior mass condition
- Variational approximations

# Objective of PAC-Bayes bounds

- empirical risk

$$r(f) = \frac{1}{n} \sum_{i=1}^{n} \ell\Big( f(X_i), Y_i \Big)$$

- generalization risk

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

- randomized prediction / ensemble / ... : $f \sim \rho$,

compare $\mathbb{E}_{f\sim\rho}[R(f)]$ and $\mathbb{E}_{f\sim\rho}[r(f)]$.

In a first time, we only consider bounded losses $\ell(u, v) \in [0, 1]$.

# A generic PAC-Bayes bound

Let $\mathcal{S}$ denote the sample $\mathcal{S} = [(X_i, Y_i)]_{i=1}^n$.

**Theorem**

For any $\varepsilon > 0$, for any $\lambda > 0$,

$$\mathbb{P}_{\mathcal{S}}\left[\forall \rho,\ \mathbb{E}_{f\sim\rho}[R(f)] \leq \mathbb{E}_{f\sim\rho}[r(f)] + \frac{\lambda}{2n} + \frac{KL(\rho\|\pi) + \log\frac{1}{\varepsilon}}{\lambda}\right] \geq 1 - \varepsilon.$$

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

# Minimization of PAC-Bayes bounds

$$\mathbb{E}_{f\sim\rho}[R(f)] \leq \mathbb{E}_{f\sim\rho}[r(f)] + \frac{\lambda}{2n} + \frac{KL(\rho\|\pi) + \log\frac{1}{\varepsilon}}{\lambda}$$

This motivates the introduction of

$$\hat{\rho}_\lambda = \arg\min_\rho \left\{ \mathbb{E}_{f\sim\rho}[r(f)] + \frac{KL(\rho\|\pi)}{\lambda} \right\}.$$

$$\Rightarrow \hat{\rho}(\mathrm{d}f) \propto \exp(-\lambda r(f))\pi(\mathrm{d}f).$$

Question : how small can the bound be ?

# A bound in expectation

**Theorem**

For any (data-dependent) $\rho$ and for any $\lambda$,

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{f\sim\rho}[R(f)] \leq \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{f\sim\rho}[r(f)] + \frac{\lambda}{2n} + \frac{KL(\rho\|\pi)}{\lambda}\right]$$

and $\lambda = \mathbb{E}_{\mathcal{S}} KL(\rho\|\pi)/2n$ leads to

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{f\sim\rho}[R(f)] \leq \mathbb{E}_{\mathcal{S}}\mathbb{E}_{f\sim\rho}[r(f)] + \sqrt{\frac{2\mathbb{E}_{\mathcal{S}} KL(\rho\|\pi)}{n}}$$

**Important !** This it does not give a generalization certificate. But necessary to study the statistical properties of $\hat{\rho}_\lambda$.

# Generalization under $\hat{\rho}_\lambda$

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{f\sim\hat{\rho}_\lambda}[R(f)] \leq \mathbb{E}_{\mathcal{S}} \min_{\rho}\left[\mathbb{E}_{f\sim\rho}[r(f)] + \frac{\lambda}{2n} + \frac{KL(\rho\|\pi)}{\lambda}\right]$$
$$\leq \min_{\rho}\left[\mathbb{E}_{f\sim\rho}[R(f)] + \frac{\lambda}{2n} + \frac{KL(\rho\|\pi)}{\lambda}\right]$$

and take $\rho$ as $\pi$ restricted to $\{f : R(f) - \inf R \leq s\}$ :

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{f\sim\hat{\rho}_\lambda}[R(f)] \leq \inf R + s + \frac{\lambda}{2n} + \frac{\log \frac{1}{\pi\{f:R(f)-\inf R\leq s\}}}{\lambda}.$$

Prior mass condition : $\log \pi\{f : R(f) - \inf R \leq s\} \geq d \log(s)$,

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{f\sim\hat{\rho}_\lambda}[R(f)] \leq \inf R + s + \frac{\lambda}{2n} + \frac{d\log\frac{1}{s}}{\lambda}.$$

Optimize in $s$ and $\lambda$ :

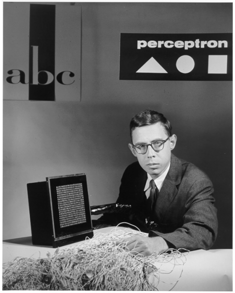
$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{f\sim\hat{\rho}_\lambda}[R(f)] \leq \inf R + \sqrt{\frac{2d}{n}\log\frac{n}{d}}.$$

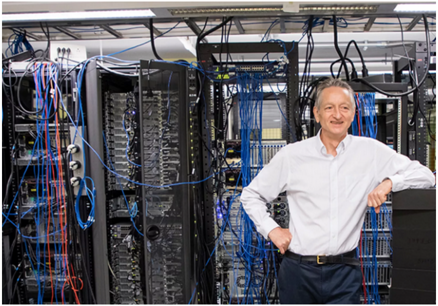
# PAC-Bayes : other topics

- unbounded losses : many important works, see references at the end.
- let's discuss briefly the optimality of the rates.

# An easy problem : find the best neural network

You have one data set $\mathcal{S}$ that you will use as a test set, and two classifiers.



$r(f_1) = 0.15$
$\color{red}{R(f_1)} = ?$

$r(f_2) = 0.01$
$\color{red}{R(f_2)} = ?$

# PAC-Bayes bound for classifier selection

More generally, $M$ classifiers $f_1, \dots, f_M$ :

- uniform prior : $\pi = \frac{1}{M}\sum_{i=1}^{M} \delta_{f_i}$
- $\hat{f} = \arg\min_f r(f)$ and $\rho = \delta_{\hat{f}}$

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{f\sim\rho}[R(f)] \leq \mathbb{E}_{\mathcal{S}}\mathbb{E}_{f\sim\rho}[r(f)] + \sqrt{\frac{2\mathbb{E}_{\mathcal{S}} KL(\rho\|\pi)}{n}}$$

$$\mathbb{E}_{\mathcal{S}} R(\hat{f}) \leq \mathbb{E}_{\mathcal{S}}[\min_f r(f)] + \sqrt{\frac{2\log(M)}{n}}$$

$$\mathbb{E}_{\mathcal{S}} R(\hat{f}) \leq \min_f R(f) + \sqrt{\frac{2\log(M)}{n}}$$

# Ask an undergrad student in statistics

Say $R(f_1) < R(f_2)$,

$$\mathbb{E}_{\mathcal{S}} R(\hat{f}) = \mathbb{E}_{\mathcal{S}} \left[ R(f_1) 1_{\hat{f}=f_1} + R(f_2) 1_{\hat{f}=f_2} \right]$$
$$\leq \mathbb{E}_{\mathcal{S}} \left[ R(f_1) + 1_{\hat{f}=f_2} \right]$$
$$= \min_f R(f) + \mathbb{P}_{\mathcal{S}}[r(f_2) - r(f_1) < 0]$$

and $r(f_2) - r(f_1) \rightsquigarrow \mathcal{N}\left(\Delta R, \frac{v}{n}\right)$ so

$$\mathbb{P}_{\mathcal{S}}[r(f_2) - r(f_1) < 0] \sim \Phi\left(\Delta R \sqrt{\frac{n}{v}}\right) \sim \frac{\exp\left(-\frac{n[\Delta R]^2}{v}\right)}{\Delta R \sqrt{2\pi \frac{n}{v}}},$$

$\Delta R = R(f_2) - R(f_1)$ and $v = R(f_2)[1 - R(f_2)] + R(f_1)[1 - R(f_1)] - 2\mathbb{P}(f_1(X) = f_2(X) \neq Y)$.

# Which is the largest ?

# Optimizing with respect to the prior

In practice, popular choices :

- $\rho = \delta_{\hat{\theta}}$,
- $\rho(f) \propto \exp(-\lambda r(f)) p(f)$
- $\dots$

Once $\rho$ is fixed, why not optimize with respect to $\pi$ ?

$$\mathbb{E}_{\mathcal{S}} \mathbb{E}_{f\sim\rho}[R(f)] \leq \mathbb{E}_{\mathcal{S}} \mathbb{E}_{f\sim\rho}[r(f)] + \sqrt{\frac{2\mathbb{E}_{\mathcal{S}} KL(\rho\|\pi)}{n}}$$

$$\mathbb{E}_{\mathcal{S}} KL(\rho\|\pi) = \underbrace{\mathbb{E}_{\mathcal{S}} KL(\rho\|\mathbb{E}_{\mathcal{S}}\rho)}_{=:\mathcal{I}(\rho,\mathcal{S})} + \underbrace{KL(\mathbb{E}_{\mathcal{S}}\rho\|\pi)}_{=0 \text{ if } \pi=\mathbb{E}_{\mathcal{S}}\rho}$$

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

# Mutual information bound

The corresponding bound was re-discovered (independently).

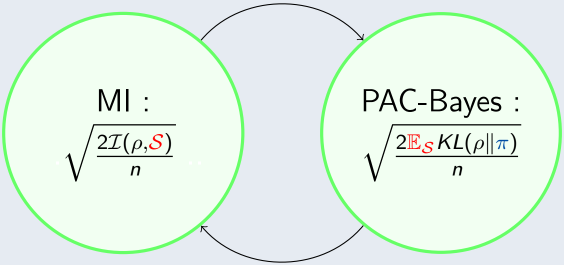
## Mutual information bound

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{f\sim\rho}[R(f)] \leq \mathbb{E}_{\mathcal{S}}\mathbb{E}_{f\sim\rho}[r(f)] + \sqrt{\frac{2\mathcal{I}(\rho,\mathcal{S})}{n}}$$

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

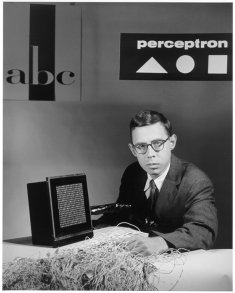
# PAC-Bayes and MI bounds

$$\mathcal{I}(\rho, \mathcal{S}) = \mathbb{E}_{\mathcal{S}} KL(\rho \| \mathbb{E}_{\mathcal{S}} \rho) \leq \mathbb{E}_{\mathcal{S}} KL(\rho \| \pi)$$

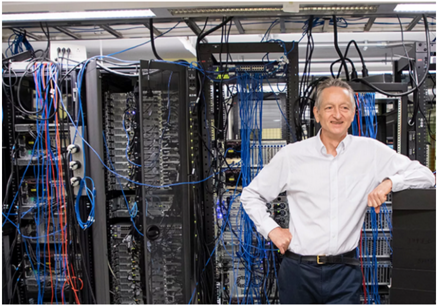



Catoni :
optimize w.r.t. $\pi$

# Classifier selection




$r(f_1) = 0.15$



$r(f_2) = 0.01$

# Application in the selection problem

Prior $\pi_{\alpha}(f) = \alpha \delta_{f_1} + (1-\alpha)\delta_{f_2}$.
Say $R(f_1) < R(f_2)$. For any $\alpha$,

$$\mathbb{E}_{\mathcal{S}} R(\hat{f}) \leq \min_f R(f) + \sqrt{\frac{2\mathbb{E}_{\mathcal{S}} KL(\rho \| \pi_{\alpha})}{n}}$$

$$= \min_f R(f) + \sqrt{\frac{2\mathbb{E}_{\mathcal{S}}\left[1_{\hat{f}=f_1} \log \frac{1}{\alpha} + 1_{\hat{f}=f_2} \log \frac{1}{1-\alpha}\right]}{n}}$$

$$\leq \min_f R(f) + \sqrt{\frac{2\left[\log \frac{1}{\alpha} + \Phi\left(\frac{n\Delta R}{2v}\right) \log \frac{1}{1-\alpha}\right]}{n}}$$

Take $\alpha = \exp\left[-\Phi\left(\frac{n\Delta R}{2v}\right)\right]$...

# Application in the selection problem

## Theorem

In the case of $M$ functions $f_1, \dots, f_M$, put

$$\Delta = \min_{i:R(f_i)\neq \min_f R(f)} R(f_i) - \min R(f).$$

Then

$$\mathbb{E}_{\mathcal{S}} R(\hat{f}) \leq \min_f R(f) + \frac{16}{n\Delta} \log\left(1 + M\mathrm{e}^{-\frac{n\Delta^2}{32}}\right)$$

For $\Delta \simeq 1\sqrt{n}$ we recover the $\sqrt{\log(M)/n}$ rate...

# Optimization of the prior : more cases

When $\rho(f) \propto \exp(-\lambda r(f))p(f)$, Catoni suggests to use the (almost optimal) "localized prior"

$$\pi_{-\beta R}(f) \propto \exp(-\beta R(f))p(f).$$

| situation | uniform prior | localized prior |
|---|---|---|
| $\dim(\Theta) = d$ | $\sqrt{\frac{d}{n} \log \frac{n}{d}}$ | $\sqrt{\frac{d}{n}}$ |
| (MA) $+ \dim(\Theta) = d$ | $\frac{d}{n} \log \frac{n}{d}$ | $\frac{d}{n}$ |

(MA) = margin assumption, includes noiseless classification

# Additional references

Arguments for generalized posteriors :

Bissiri, P. G., Holmes, C. C. & Walker, S. G. (2016). A general framework for updating belief distributions. *JRSS-B*.

Knoblauch, J., Jewson, J. & Damoulas, T. (2022). An Optimization-centric View on Bayes' Rule : Reviewing and Generalizing Variational Inference. *JMLR* (to appear).

Also note that the connection between contraction and PAC-Bayes was already used by many authors to study generalized posteriors :

Grünwald, P. D. & Mehta, N. A. (2020). Fast Rates for General Unbounded Loss Functions : From ERM to Generalized Bayes. *JMLR*.

Syring, N. & Martin, R. (2020). *Gibbs posterior concentration rates under sub-exponential type losses*. Preprint arXiv :2012.04505.

# Contents

1 Introduction
- Contraction of the posterior
- PAC-Bayes bounds

2 PAC-Bayes point of view on contraction (and vice-versa)
- An empirical point of view on the prior mass condition
- Variational approximations

# Bayesian deep learning

Deep neural networks :

- amazing practical performances,
- theory not yet complete.

Contraction of the posterior for Bayesian deep networks :

Polson, N. G. & Ročková, V. (2018). Posterior concentration for sparse deep learning. *NeurIPS*.

Chérief-Abdellatif, B. E. (2020). Convergence rates of variational inference in sparse deep learning. *ICML*.

beautiful results but do not really match the algorithm used in practice...

# Empirical prior mass

Among practitioners, consensus : "flat minima" lead to good generalization in deep learning. Tentative interpretation :

$$r(f^*) \text{ ``flat''} \leftrightarrow \{f : r(f) - r(f^*) \leq s\} \text{ is large}$$
$$\leftrightarrow \pi(\{f : r(f) - r(f^*) \leq s\}) \text{ is not too small.}$$

$$\mathbb{E}_{f\sim\hat{\rho}}[R(f)] \leq \min_{\rho} \left[ \mathbb{E}_{f\sim\rho}[r(f)] + \frac{\lambda}{2n} + \frac{KL(\rho\|\pi) + \log\frac{1}{\varepsilon}}{\lambda} \right]$$

take $\rho$ as $\pi$ restricted to $\{f : r(f) - r(f^*) \leq s\}$,

$$\mathbb{E}_{f\sim\hat{\rho}}[R(f)] \leq \min_{s} \left[ \underbrace{r(f^*)}_{=0} + s + \frac{\lambda}{2n} + \frac{\log\frac{1}{\pi(\{f:r(f)-r(f^*)\leq s\})} + \log\frac{1}{\varepsilon}}{\lambda} \right].$$

# PAC-Bayes and deep learning

In recent papers : minimization of PAC-Bayes bounds to train a neural network, leading to tight generalization certificates.

Dziugaite, G. K. and Roy, D. M. (2017). Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *UAI*.

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

| | Training method | Stch. Pred. 01 Err | Risk cert. $\ell^{01}$ | Bound used |
|---|---|---|---|---|
| D&R 2018 | SGLD ($\tau = 3e+3$) | 0.1200 | 0.2100<br>0.2600 | D&R18 Thm. 4.2<br>Lever et al. 2013 |
| D&R 2018 | SGLD ($\tau = 1e+5$) | 0.0600 | 0.6500<br>1.0000 | D&R18 Thm. 4.2<br>Lever et al. 2013 |
| This work | SGD + $f_{\text{quad}}$ | *0.0202* | **0.0279** | PAC-Bayes-kl |
| | SGD + $f_{\text{lambda}}$ | **0.0196** | 0.0354 | PAC-Bayes-kl |
| | SGD + $f_{\text{classic}}$ | 0.0230 | *0.0284* | PAC-Bayes-kl |

# Contents

1 Introduction
- Contraction of the posterior
- PAC-Bayes bounds

2 PAC-Bayes point of view on contraction (and vice-versa)
- An empirical point of view on the prior mass condition
- Variational approximations

# Variational approximations

Reminder :

- $X_1, \ldots, X_n$ i.i.d. from $P_0$,
- $(P_\theta, \theta \in \Theta)$ model, densities $p_\theta(x)$,
- prior $\pi$ on $\Theta$,
- tempered posterior $\hat{\pi}_\alpha(\mathrm{d}\theta) \propto \left(\prod_{i=1}^n p_\theta(X_i)\right)^\alpha \pi(\mathrm{d}\theta)$.

## Variational approximations

Let $\mathcal{F}$ be a set of (tractable) distributions,

$$\tilde{\pi}_\alpha = \arg\min_{\rho \in \mathcal{F}} \mathcal{K}(\rho, \pi_\alpha)$$
$$= \arg\min_{\rho \in \mathcal{F}} \left\{ -\alpha \int \frac{1}{n} \sum_{i=1}^n \log p_\theta(X_i) \rho(\mathrm{d}\theta) + \mathcal{K}(\rho, \pi) \right\}.$$

# PAC-Bayes bound for tempered posteriors

## Theorem

Alquier, P. & Ridgway, J. (2020). Concentration of tempered posteriors and of their variational approximations. *The Annals of Statistics*.

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\tilde{\pi}_\alpha} D_\alpha(P_\theta, P_{\theta_0}) \leq \inf_{\rho\in\mathcal{F}} \left[\frac{\alpha}{1-\alpha}\mathbb{E}_{\theta\sim\rho} KL(P_\theta, P_{\theta_0}) + \frac{KL(\rho,\pi)}{n(1-\alpha)}\right].$$

Assume that for any $n$, there is a $\rho_n \in \mathcal{F}$ such that

- $\mathbb{E}_{\theta\sim\rho_n} KL(P_\theta, P_{\theta_0}) \leq r_n$
- $KL(\rho_n,\pi) \leq n r_n$,

then

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\tilde{\pi}_\alpha} D_\alpha(P_\theta, P_{\theta_0}) \leq \frac{2\alpha r_n}{1-\alpha}.$$

# Further references :

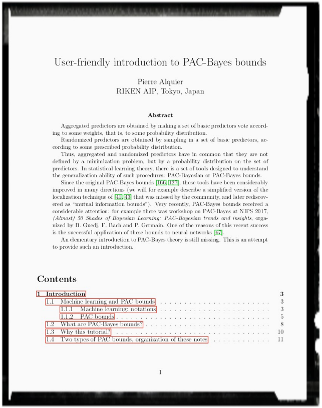
Application to mixture models, application to Markov chains :


- Chérief-Abdellatif, B.-E. & Alquier, P. (2018). Consistency of variational Bayes inference for estimation and model selection in mixtures. *Electronic Journal of Statistics*.
- Banerjee, I., Rao, V. A. & Honnappa, H. (2021). PAC-Bayes Bounds on Variational Tempered Posteriors for Markov Models. *Entropy*.


Allowing $\alpha = 1$ :


- Y. Yang, D. Pati & A. Bhattacharya (2020). $\alpha$-Variational Inference with Statistical Guarantees. *The Annals of Statistics*.
- F. Zhang & C. Gao (2020). Convergence Rates of Variational Posterior Distributions. *The Annals of Statistics*.
- Ohn, I. & Lin, L. (2021). *Adaptive variational Bayes : Optimality, computation and applications*. Preprint arXiv :2109.03204.

# Advertisement


User-friendly introduction to PAC-Bayes bounds

Pierre Alquier
RIKEN AIP, Tokyo, Japan

**Abstract**

Aggregated predictors are obtained by making a set of basic predictors vote according to some weights, that is, to some probability distribution.

Randomized predictors are obtained by sampling in a set of basic predictors, according to some prescribed probability distribution.

Thus, aggregated and randomized predictors have in common that they are not defined by a minimization problem, but by a probability distribution on the set of predictors. In statistical learning theory, there is a set of tools designed to understand the generalization ability of such procedures: PAC-Bayesian or PAC-Bayes bounds.

Since the original PAC-Bayes bounds [166, 127], these tools have been considerably improved in many directions (we will for example describe a simplified version of the localization technique of [41, 43] that was missed by the community, and later rediscovered as "mutual information bounds"). Very recently, PAC-Bayes bounds received a considerable attention: for example there was workshop on PAC-Bayes at NIPS 2017, *(Almost) 50 Shades of Bayesian Learning: PAC-Bayesian trends and insights*, organized by B. Guedj, F. Bach and P. Germain. One of the reasons of this recent success is the successful application of these bounds to neural networks [67].

An elementary introduction to PAC-Bayes theory is still missing. This is an attempt to provide such an introduction.

**Contents**

1 Introduction 3
1.1 Machine learning and PAC bounds . . . 3
1.1.1 Machine learning: notations . . . 3
1.1.2 PAC bounds . . . 5
1.2 What are PAC-Bayes bounds? . . . 8
1.3 Why this tutorial? . . . 10
1.4 Two types of PAC bounds, organization of these notes . . . 11

1

Alquier, P. (2021). *User-friendly introduction to PAC-Bayes bounds*. Preprint arXiv.

Discusses the topics above and

- unbounded losses,
- non i.i.d. observations,
- ...

and provides references.

# La fin

終わり

ありがとう ございます。