Empirical PAC-Bayes bounds for Markov chains

Pierre Alquier

ESSEC
BUSINESS SCHOOL

MCQMC
The University of Edinburgh, June 2026

Cergy (near Paris)

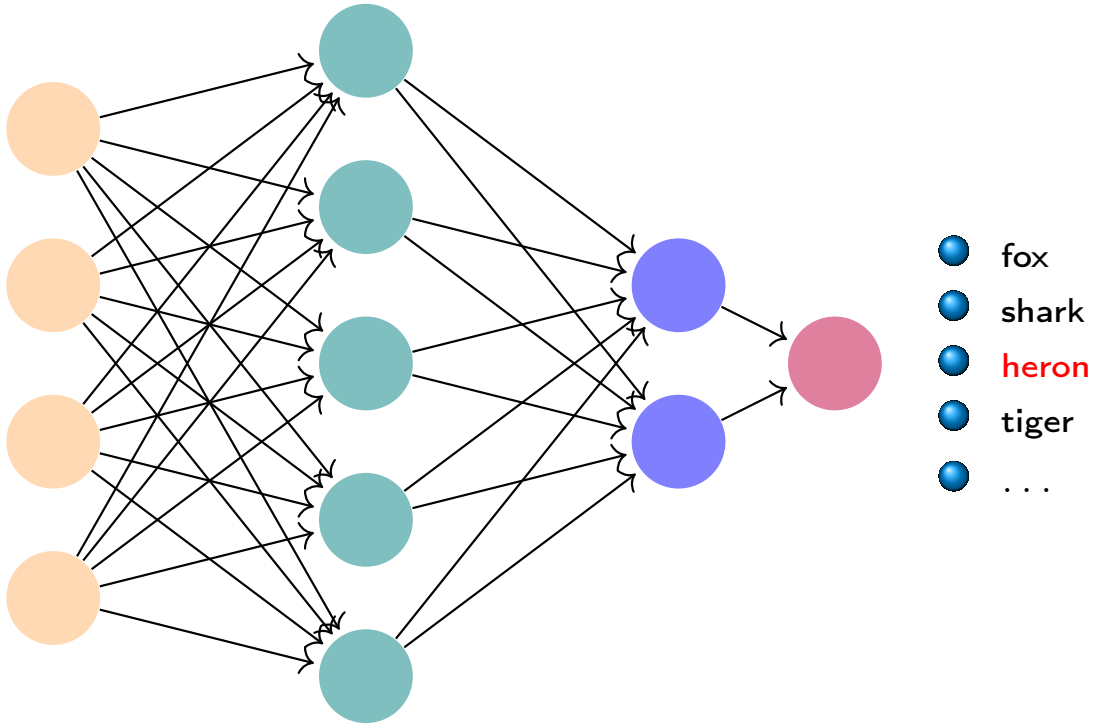
- Kamelia DAUDEL
- Olga KLOPP
- Marie KRATZ
- Roberto RENO
- Pierre JACOB
- Guillaume CHEVILLON
- Jeroen ROMBOUTS
- Mohamed NDAOUD
- Guillaume LECUE
- Mikołaj KASPRZAK
- Maria ALLAYIOTI
- Vincenzo ESPOSITO VINZI

Singapore

- Jeremy HENG
- Pierre ALQUIER

Rabat

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



- 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 :

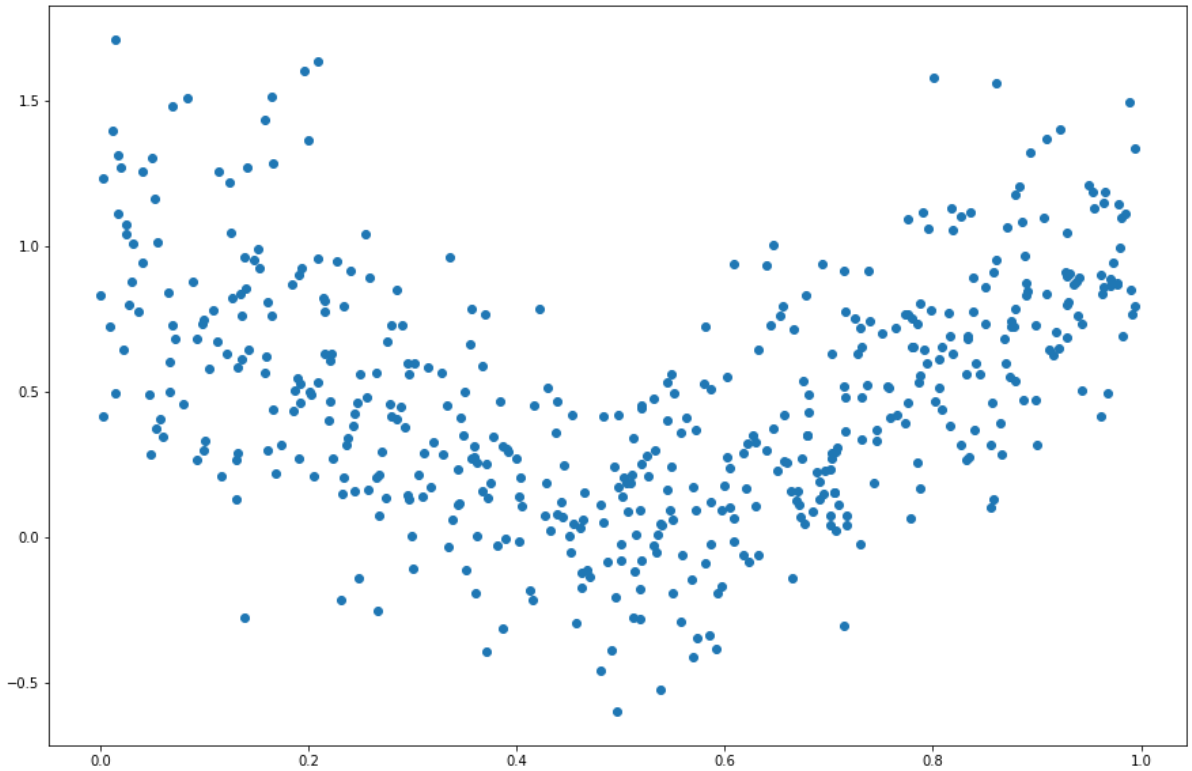
$$R^* = \inf_{\theta\in\Theta} R(\theta).$$

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

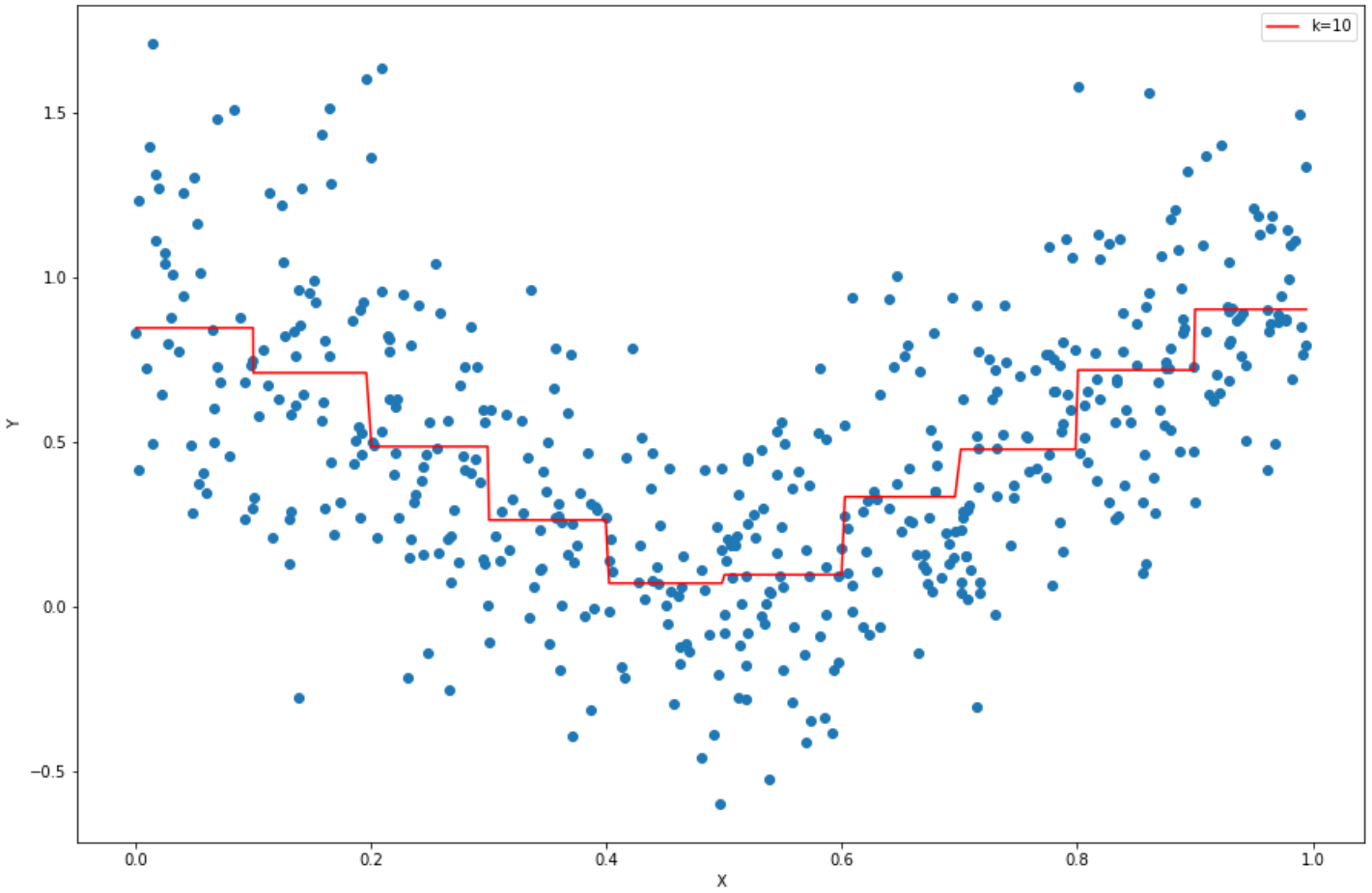
$$R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell\Big(Y_i, f_\theta(X_i)\Big).$$

Toy example :

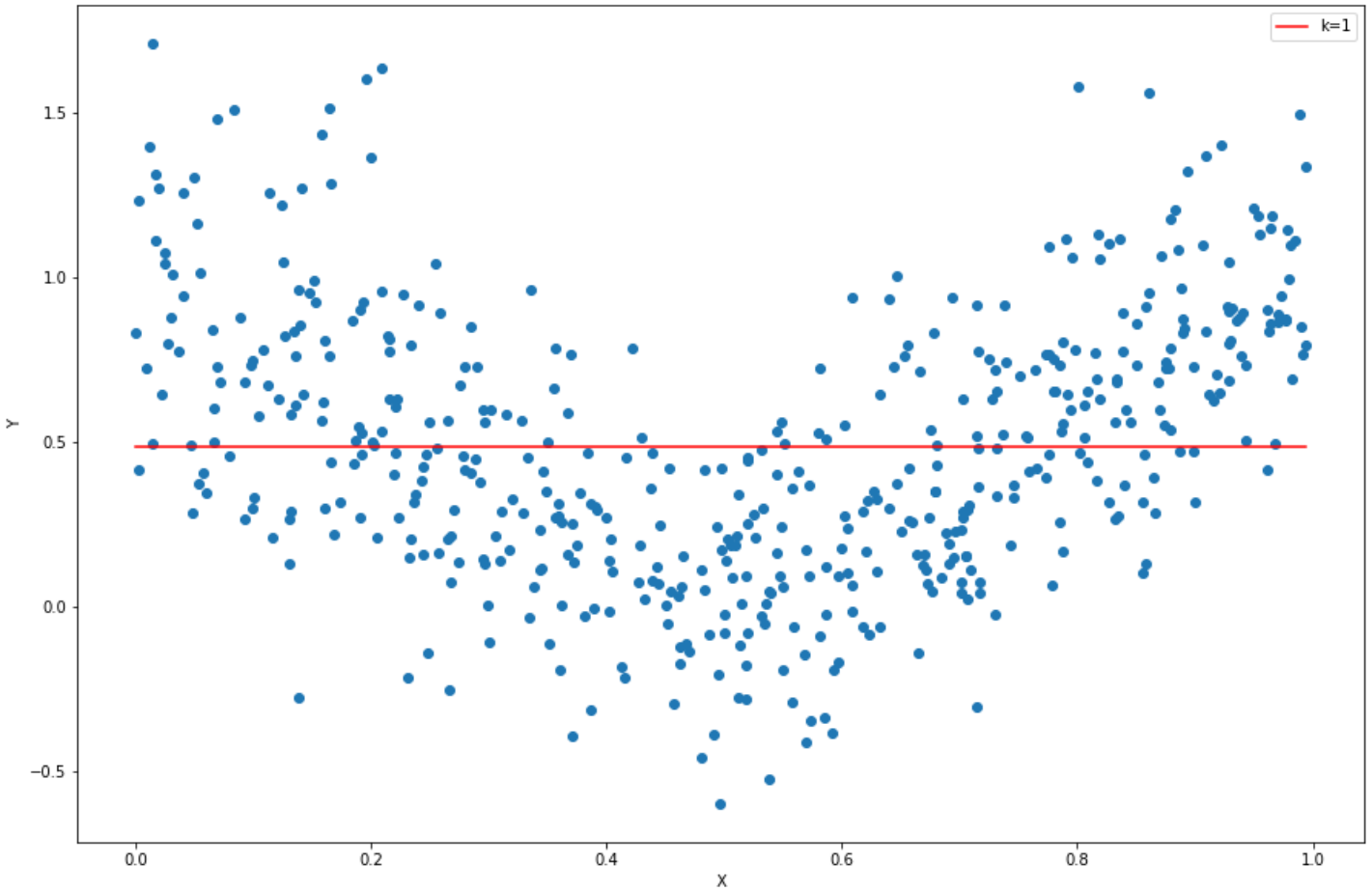
- X uniform on $[0, 1]$,
- $Y = |2X - 1| + \epsilon$.



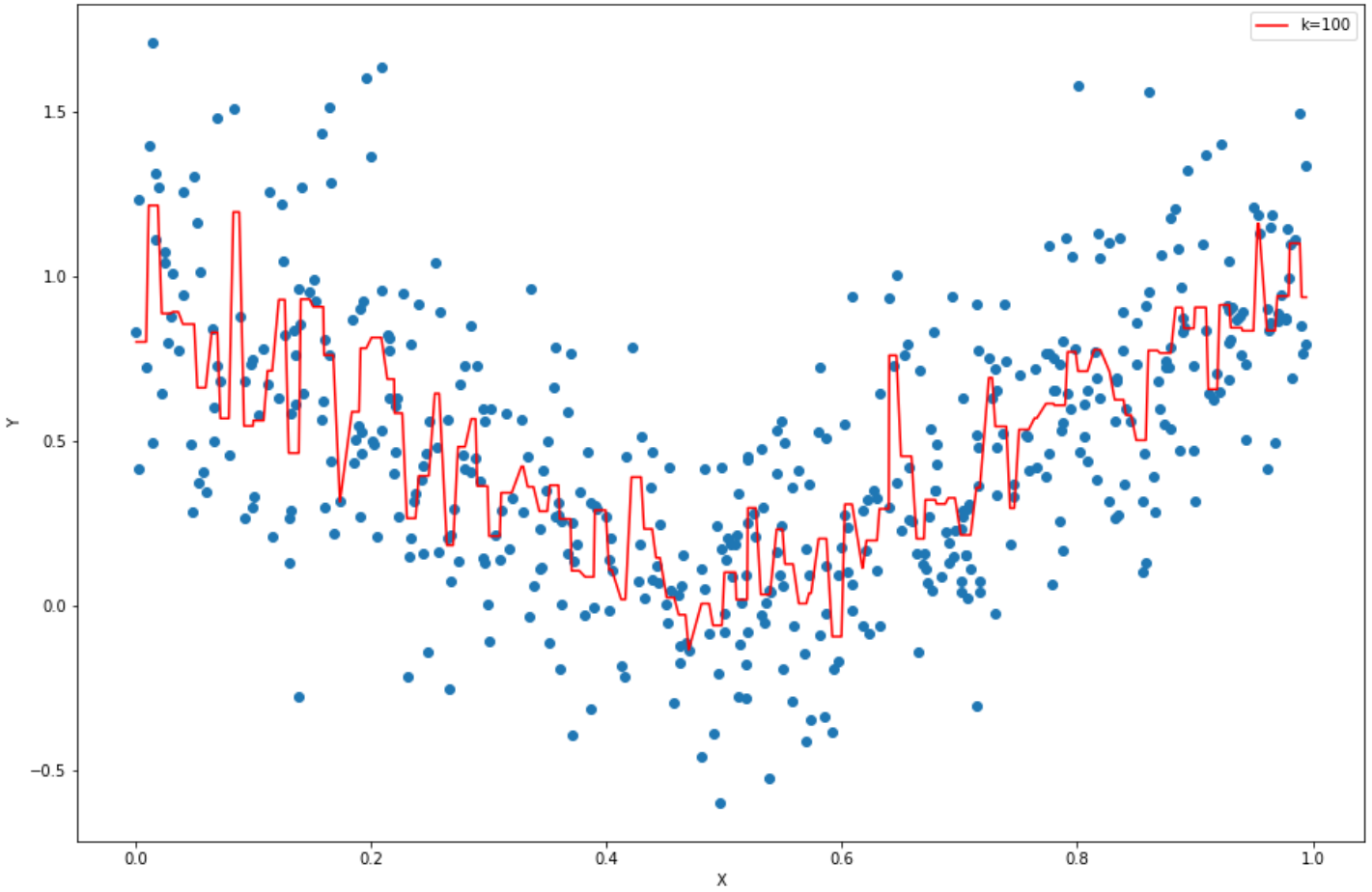
- Prediction by regular histogram with k-bins.
- $\ell(y, f_\theta(x)) = (y - f_\theta(x))^2$.

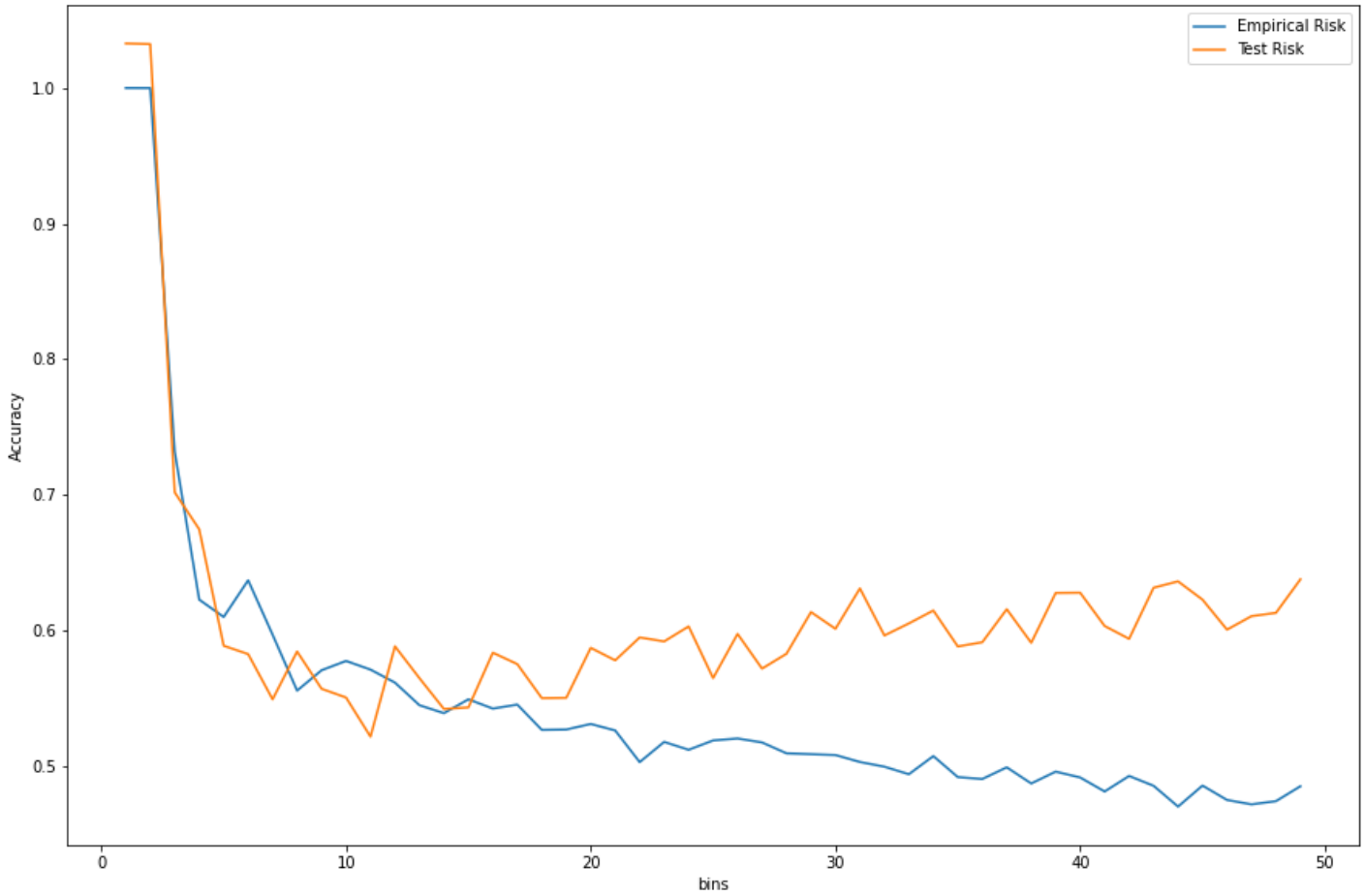


- Prediction by regular histogram with k-bins.
- $\ell(y, f_\theta(x)) = (y - f_\theta(x))^2$.



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




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

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 $\hat{\theta} = \hat{\theta}((X_1, Y_1), \ldots, (X_n, Y_n)) = \hat{\theta}(\mathcal{S})$ 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.

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.

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

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

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

Randomized estimators :

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

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

McAllester's PAC-Bayes bound

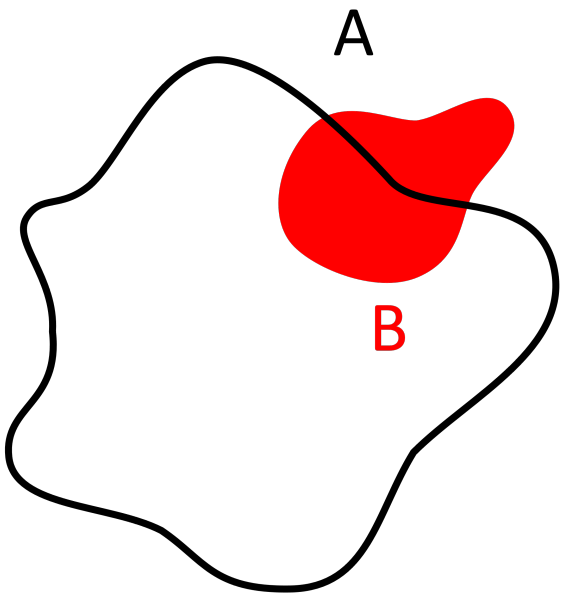
Fix a prior distribution $\pi \in \mathcal{M}(\Theta)$. With probability at least $1-\delta$ on the data $\mathcal{S}$, for any probability distribution ρ learnt on the data,

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

$\mathrm{KL}(\rho\|\pi)$ = 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 $\hat{\rho}$.

Intuition on KL :



- π uniform on A

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

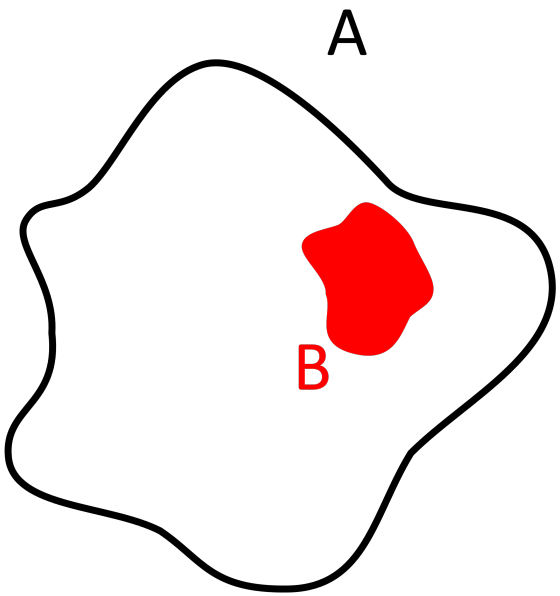
- ρ uniform on B

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

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

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

Intuition on KL :



- π uniform on A

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

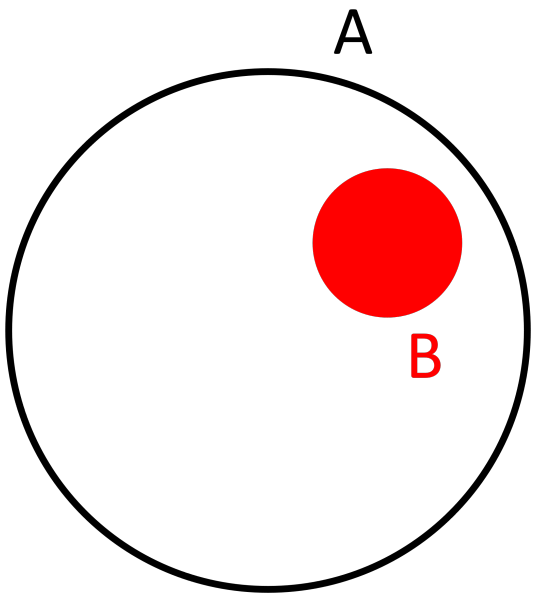
- ρ uniform on B

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

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

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

Intuition on KL :



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

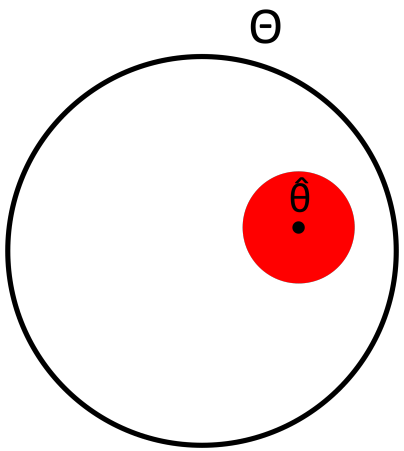
$$\mathcal{V}(B_d(x,r)) = \frac{r^d \pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)}$$

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

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

McAllester's PAC-Bayes bound

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



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

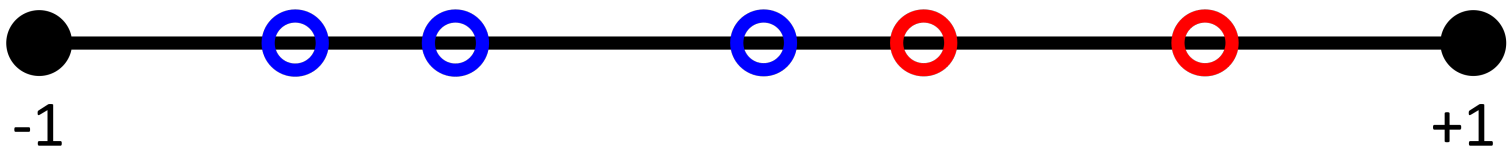
$$\mathbb{E}_{\theta\sim\hat{\rho}}[R(\theta)] \leq \mathbb{E}_{\theta\sim\hat{\rho}}[R_n(\theta)] + \sqrt{\frac{d\log\frac{C}{\epsilon} + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{2n}}.$$

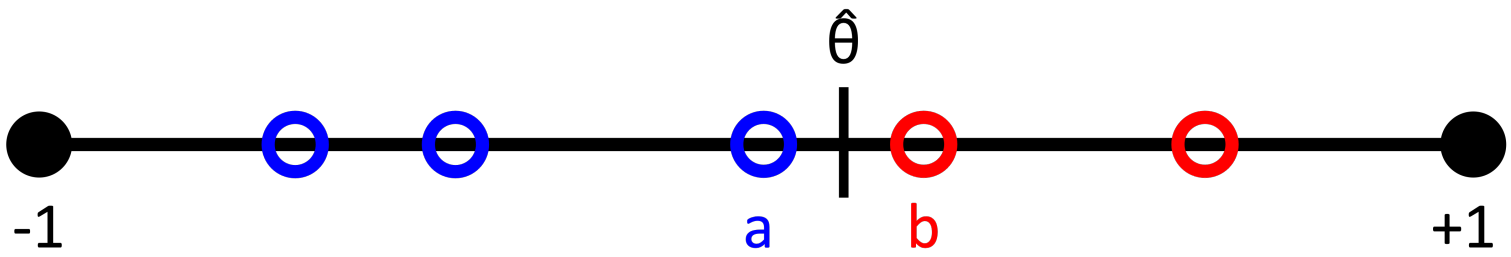
Toy classification example :

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

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

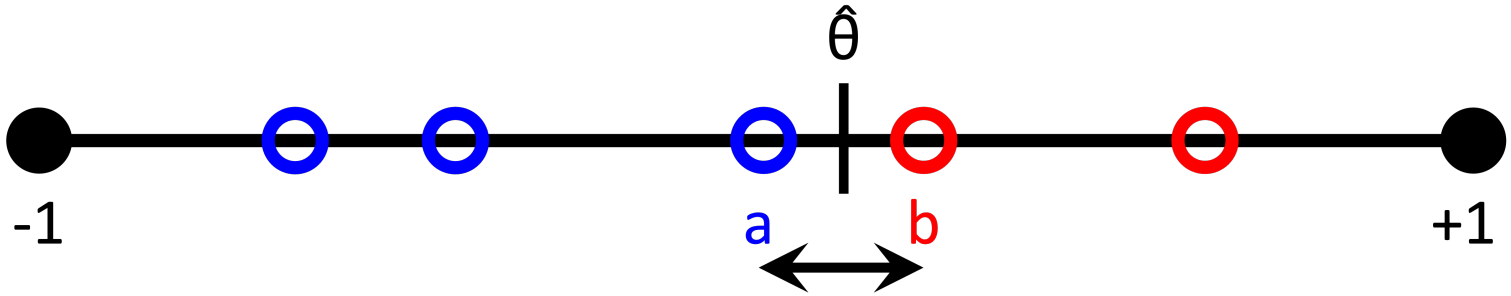
- $Y_i = f_{\theta^*}(X_i)$.





Vapnik-type bound :

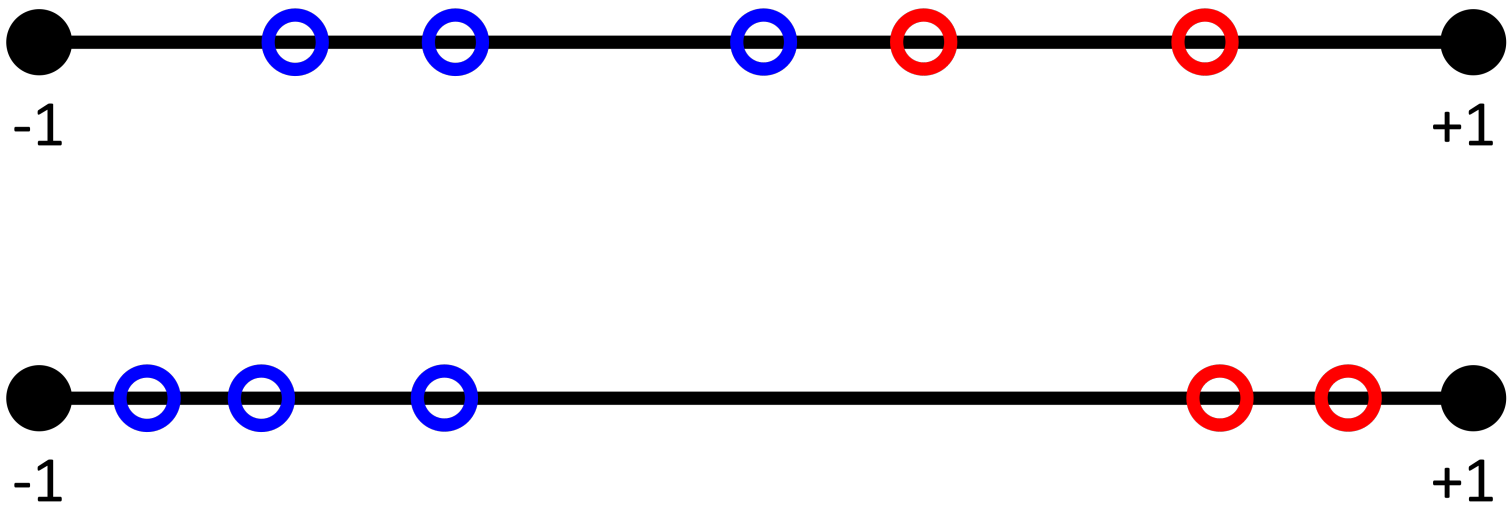
$$R(\hat{\theta}) \leq \sqrt{\frac{8\log(2en) + 8\log\left(\frac{4}{\delta}\right)}{n}}$$

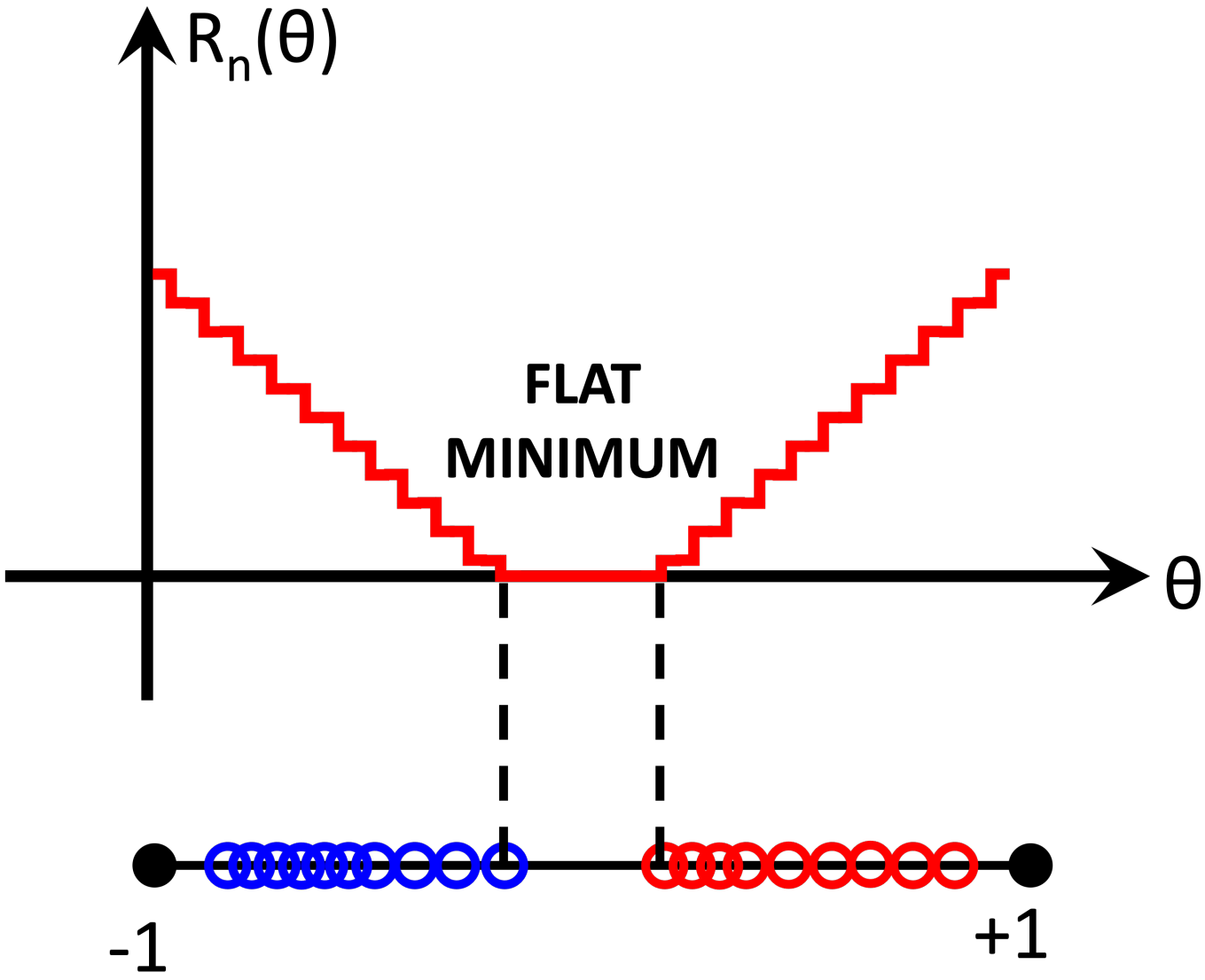


PAC-Bayes :

- 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


$R_n(\theta)$
FLAT
MINIMUM
θ
-1
+1

Minimization of the PAC-Bayes bound.

McAllester's PAC-Bayes bound

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

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

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

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

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

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

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

Sampling from $\hat{\pi}_\lambda$ by Monte-Carlo techniques...

Example : quantile regression.

$$\ell_\tau(y,\hat{y}) = (y-\hat{y})(\tau - 1_{\{y\leq\hat{y}\}}).$$

$$\hat{\pi}_\lambda(\mathrm{d}\theta) = \frac{\exp(-\frac{\lambda}{n}\sum_{i=1}^n \ell_\tau(Y_i,\theta^T X_i))}{Z_\lambda}\pi(\mathrm{d}\theta).$$

Algorithms :

- importance sampling

Alquier, P. and Li, X. (2012). Prediction of quantiles by statistical learning and application to GDP forecasting. *15th International Conference on Discovery Science*.

- Lanvegin Monte-Carlo

$$\begin{aligned}\theta_{s+1} &= \theta_s - \eta\nabla\log\hat{\pi}_\lambda(\theta_s) + \sqrt{2\eta}W_s \\ &= \theta_s + \eta\lambda\nabla R_n(\theta_s) - \eta\nabla\log\pi(\theta) + \sqrt{2\eta}W_s\end{aligned}$$

Mai, T. T. (2025). A sparse PAC-Bayesian approach for high-dimensional quantile prediction. *Statistics and Computing*.

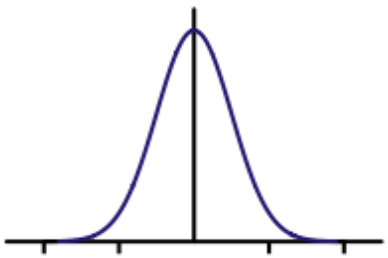
Approximate minimization of the PAC-Bayes bound.

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

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

Definition - variational approximation of Gibbs posterior

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



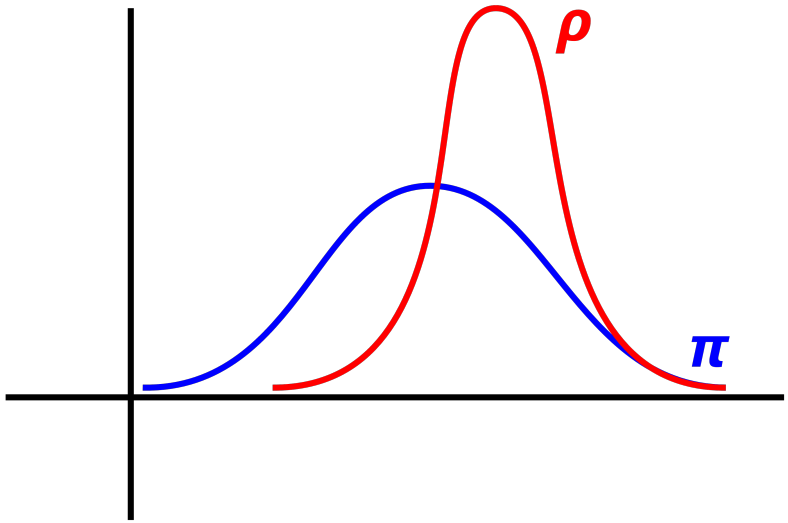
Example : $\rho = \mathcal{N}(\mu, \Sigma)$, optimize (μ, Σ).

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

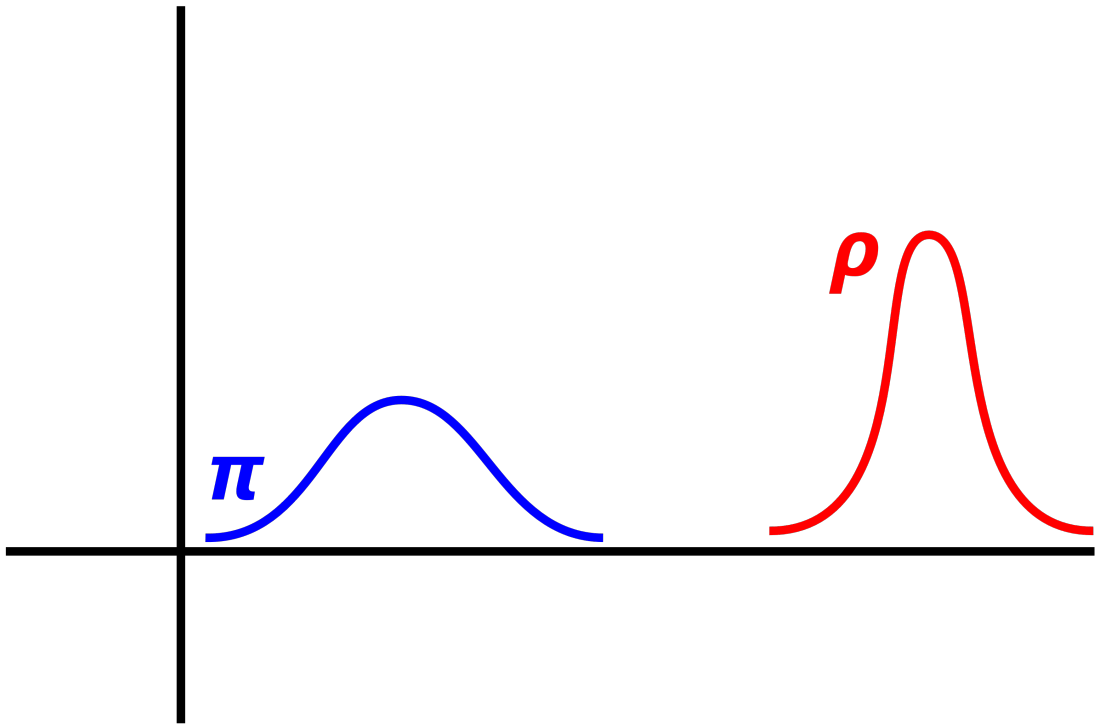
$$\pi = \mathcal{N}(\mu_0, \Sigma_0) \text{ and } \rho = \mathcal{N}(\mu_1, \Sigma_1) \text{ in } \mathbb{R}^d.$$

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

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

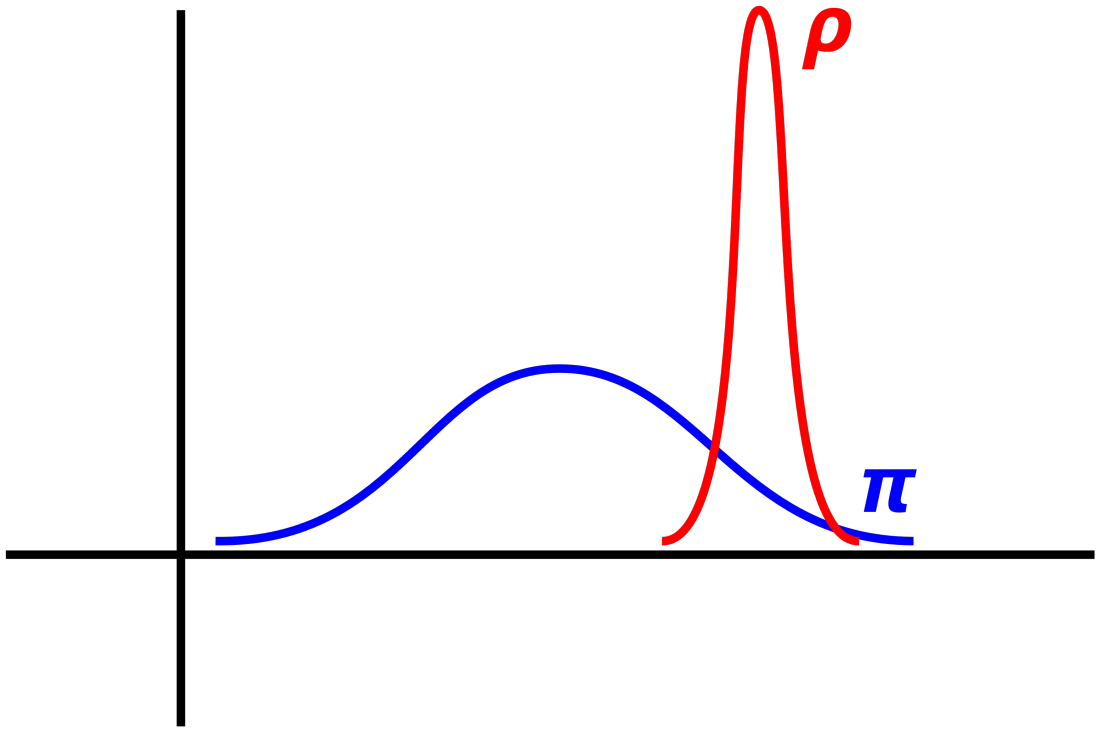


$$\mathrm{KL}(\rho\|\pi) = \frac{1}{2}\left[\frac{\Sigma_1}{\Sigma_0} - 1 + \frac{(\mu_0 - \mu_1)^2}{\Sigma_0} + \log\frac{\Sigma_0}{\Sigma_1}\right].$$



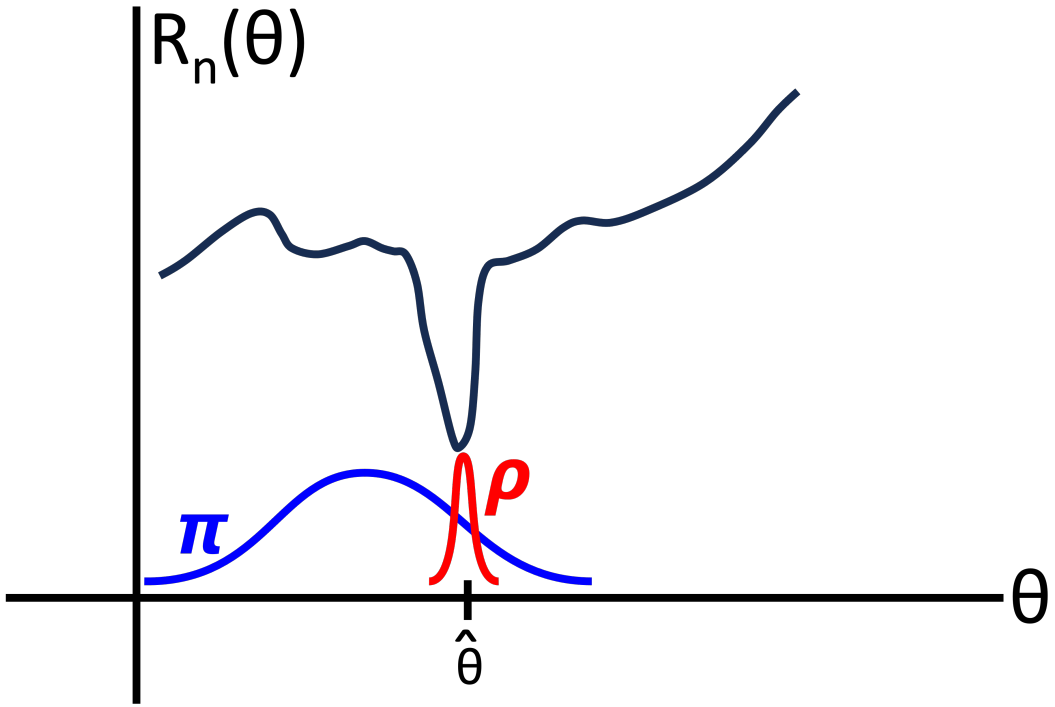
If μ_1 goes far away from μ_0 to ∞,

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



If $\Sigma_1 \to 0$,

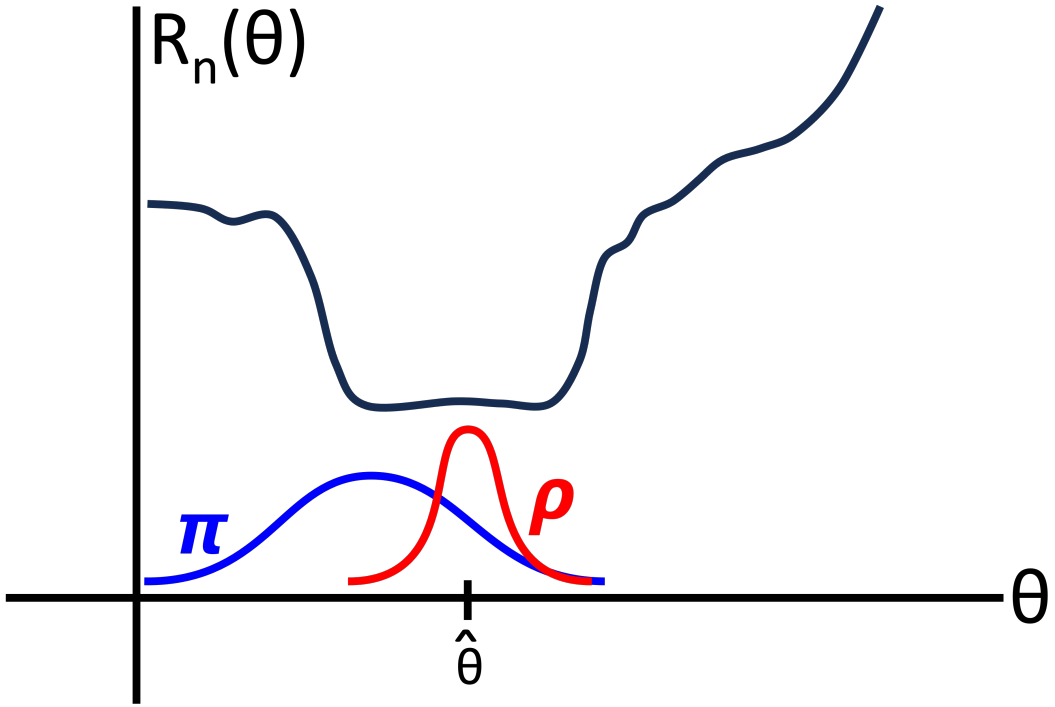
$$\mathrm{KL}(\rho\|\pi) \sim \frac{1}{2}\log\frac{\Sigma_0}{\Sigma_1} \to \infty.$$



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.

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

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

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

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).

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

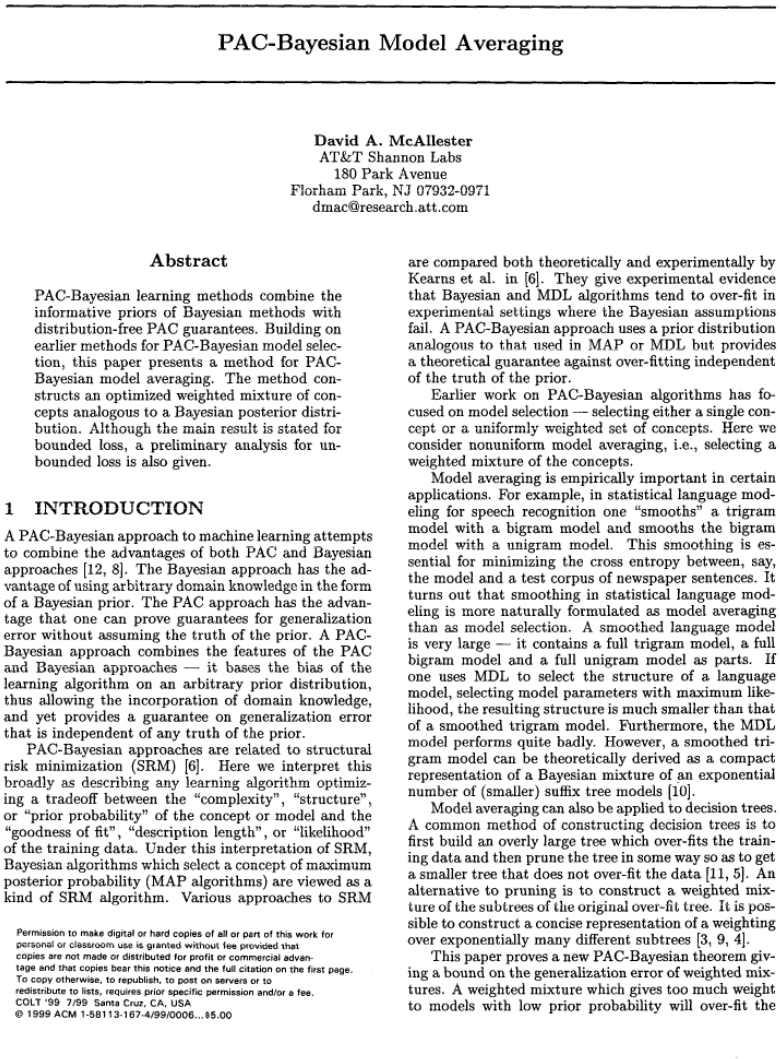
Experiment	T-600	T-1200	T-300^2	T-600^2	T-1200^2	T-600^3	R-600
Train error	0.001	0.002	0.000	0.000	0.000	0.000	0.007
Test error	0.018	0.018	0.015	0.016	0.015	0.013	0.508
SNN train error	0.028	0.027	0.027	0.028	0.029	0.027	0.112
SNN test error	0.034	0.035	0.034	0.033	0.035	0.032	0.503
PAC-Bayes bound	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*.

Extensions.



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

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, 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*.


Fard, M. M., Pineau, J. and Szepesvári, C. (2012). PAC-Bayesian policy evaluation for reinforcement learning. *UAI*.


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


Alquier, P. and Li, X. (2012). Prediction of quantiles by statistical learning and application to GDP forecasting. *15th International Conference on Discovery Science*.


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


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


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


Haussmann, M., Gerwinn, S., Look, A., Rakitsch, B. and Kandemir, M. (2021). Learning partially known stochastic dynamics with empirical PAC Bayes. *AISTATS*.


Eringis, D., Leth, J.-J., Tan, Z.-H., Wisniewski, R., Esfahani, A. F. and Petreczky, M. (2021). PAC-Bayesian theory for stochastic LTI systems. *60th IEEE Conference on Decision and Control*.


Eringis, D., Leth, J.-J., Tan, Z.-H., Wisniewski, R. and Petreczky, M. (2024). PAC-Bayesian error bound, via Rényi divergence, for a class of linear time-invariant state-space models. *ICML*.


Boroujeni, M.G., Galimberti, C.L., Krause, A. and Ferrari-Trecate, G. (2025). *PAC-Bayesian Optimal Control with Stability and Generalization Guarantees.* ArXiv preprint arXiv :2512.02858.

Definition – φ-mixing coefficient

$$\varphi(\mathcal{A},\mathcal{B}) = \sup_{A\in\mathcal{A},B\in\mathcal{B}} \Big|\mathbb{P}(B|A) - \mathbb{P}(B)\Big|.$$

$$\varphi_\ell := \sup_{t\in\mathbb{N}} \varphi\Big(\sigma(\dots,X_{t-1},X_t),\sigma(X_{t+\ell},X_{t+\ell+1},\dots)\Big).$$

Theorem – essentially from Alquier and Li (2012)

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

$$\mathbb{E}_{\theta\sim\hat{\rho}}[R(\theta)]$$
$$\leq \mathbb{E}_{\theta\sim\hat{\rho}}[R_n(\theta)] + \frac{\mathrm{KL}(\hat{\rho}\|\pi) + \log\frac{1}{\delta}}{\lambda} + \frac{\lambda\big(1+2\sum_{\ell=1}^{\infty}\varphi_\ell\big)^2}{8n}.$$

Karagulyan, V. and Alquier, P. (2026). Empirical PAC-Bayes bounds for Markov chains. *AISTATS*.

- Loss $\ell\Big(y, f_\theta(x)\Big)$.
- Objects $X_1, \ldots, X_n$ is a Markov chain with transition kernel P :

$$\mathbb{P}(X_{i+1} \in A | X_i = x) = P(x, A) = \int_A P(x, \mathrm{d}y).$$

- Labels $Y_i | X_i = x \sim Q(x, \cdot)$.
- Empirical risk : $R_n(\theta) = \frac{1}{n} \sum_{i=1}^n \ell\Big(Y_i, f_\theta(X_i)\Big)$.
- Stationary distribution π : $\pi P = \pi$.
- Risk : $R(\theta) := \mathbb{E}_{X \sim \pi} \mathbb{E}_{Y \sim Q(X, \cdot)} \Big[\ell\Big(Y, f_\theta(X)\Big)\Big]$.

Definition : reversal kernel

We put

$$P^*(u, \mathrm{d}v) = \frac{P(v, \mathrm{d}u)\pi(\mathrm{d}v)}{\pi(\mathrm{d}u)}.$$

- interpretation : in the stationary regime, the distribution of (X_t, X_{t+1}) is

$$\pi(x_t)P(x_t, \mathrm{d}x_{t+1}) = P^*(x_{t+1}, \mathrm{d}x_t)\pi(x_{t+1}).$$

- note that P^*P is also a Markov kernel.

Definition

For a Markov kernel M, 1 is an eigenvalue of M. Let m_1 denote its multiplicity. Define :

$$\gamma(M) := \begin{cases} 0 \text{ if } m_1 > 1, \\ \\ 1 - \sup\{\lambda \in \mathrm{sp}(M),\ \lambda < 1\} \text{ otherwise.} \end{cases}$$

Definition – pseudo spectral-gap

$$\gamma_{\mathrm{ps}} := \max_{k \geq 1} \frac{\gamma\left((P^*)^k P^k\right)}{k}.$$

Paulin, D. (2015). Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability*.

Theorem

Fix $0 < \lambda < \frac{n}{20}$ and a prior μ. With probability at least $1 - \delta$ on $\mathcal{S}$, for any $\hat{\rho}$,

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

Corollary

Assume we have an estimator $\hat{\gamma}_{\mathrm{ps}}$ of γ_{ps} such that

$$\mathbb{P}\left(\left|\frac{\hat{\gamma}_{\mathrm{ps}}}{\gamma_{\mathrm{ps}}} - 1\right| > \epsilon\right) \leq \alpha(n, \epsilon, \gamma_{\mathrm{ps}}) =: \alpha.$$

With probability at least $1 - \delta - \alpha$ on $\mathcal{S}$, for any $\hat{\rho}$,

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

Example : finite state space $\{1, \dots, d\}$


Wolfer, G. and Kontorovich, A. (2019). Estimating the mixing time of ergodic Markov chains. *COLT*.


Wolfer, G. and Kontorovich, A. (2024). Improved estimation of relaxation time in nonreversible Markov chains. *The Annals of Applied Probability*.

$$N^k(i,j) = \sum_{\ell=1}^{\lfloor \frac{n-1}{k} \rfloor} 1\left[X_{1+(\ell-1)k} = i, X_{1+\ell k} = j\right],$$

$$N^k(i) = \sum_{\ell=1}^{\lfloor \frac{n-1}{k} \rfloor} 1\left[X_{1+(\ell-1)k} = i\right], \text{ and } \hat{L}^k(i,j) = \frac{N^k(i,j)}{\sqrt{N^k(i)N^k(j)}}.$$

Note : $\hat{L}^k$ estimates $L^k = \mathrm{diag}(\pi)^{\frac{1}{2}} P^k \mathrm{diag}(\pi)^{-\frac{1}{2}}$.

Fix $K > 0$, put $\hat{\gamma}_{\mathrm{ps},K} = \max\left\{\gamma((\hat{L}^k)^T \hat{L}^k)/k, 1 \le k \le K\right\}$.

Proposition (Wolfer and Kontorovich)

For $K = \lceil \frac{2}{\epsilon} \rceil$, we have

$$\mathbb{P}\left(\left| \frac{\hat{\gamma}_{\mathrm{ps},K}}{\gamma_{\mathrm{ps}}} - 1 \right| > \epsilon \right) \leq \alpha(n, \epsilon, \gamma_{\mathrm{ps}}) = \alpha$$

where

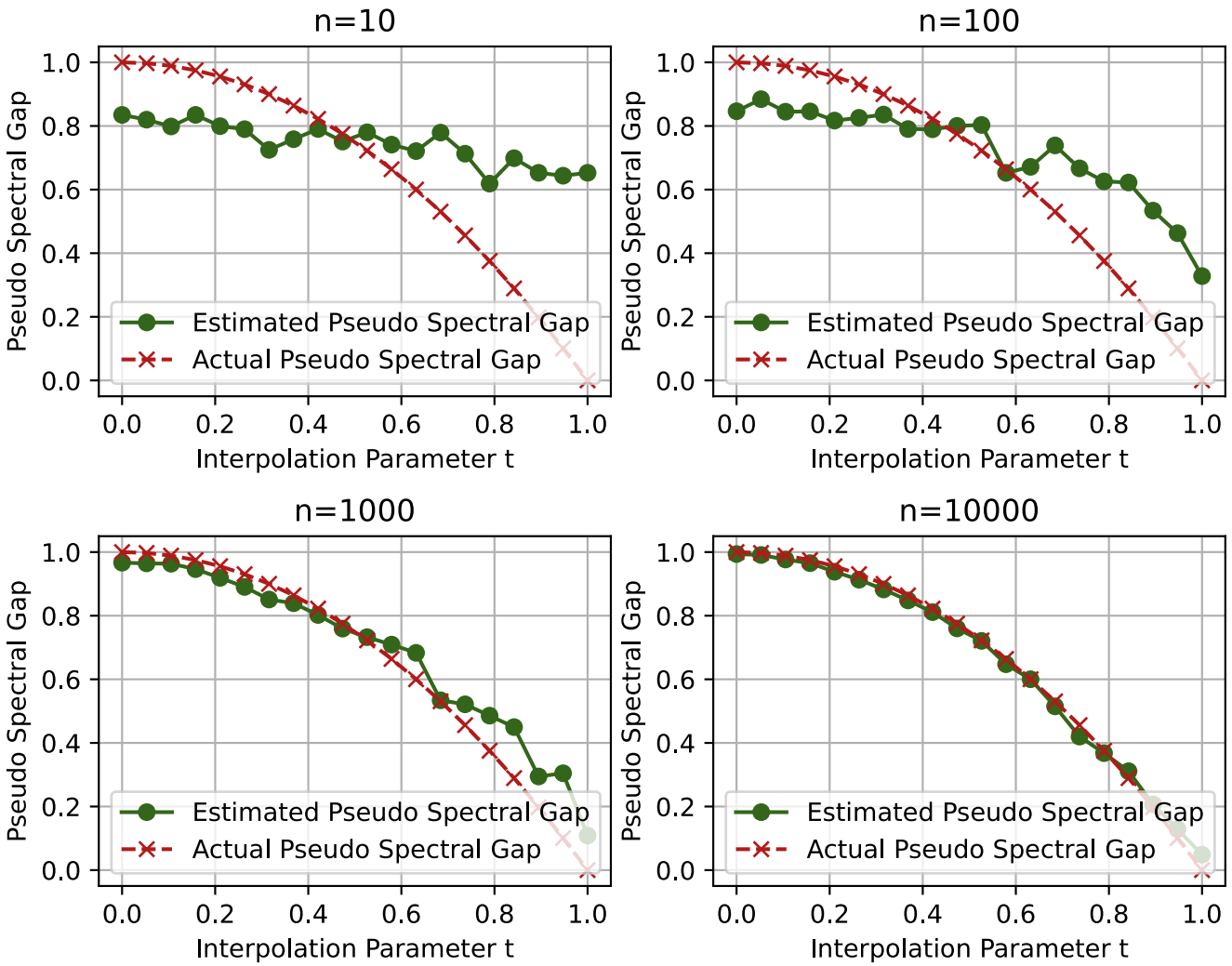
$$\alpha = \frac{d\|P\|_\pi \min(d, \|P\|_\pi)}{\epsilon \gamma_{\mathrm{ps}} \min_{1 \leq i \leq d} \pi_i} \exp\left(-n\epsilon^2 \gamma_{\mathrm{ps}}^2 \min_{1 \leq i \leq d} \pi_i \min\left(\gamma_{\mathrm{ps}}, \frac{1}{d \min(d, \|P\|_\pi)} \right) \right).$$

Corollary

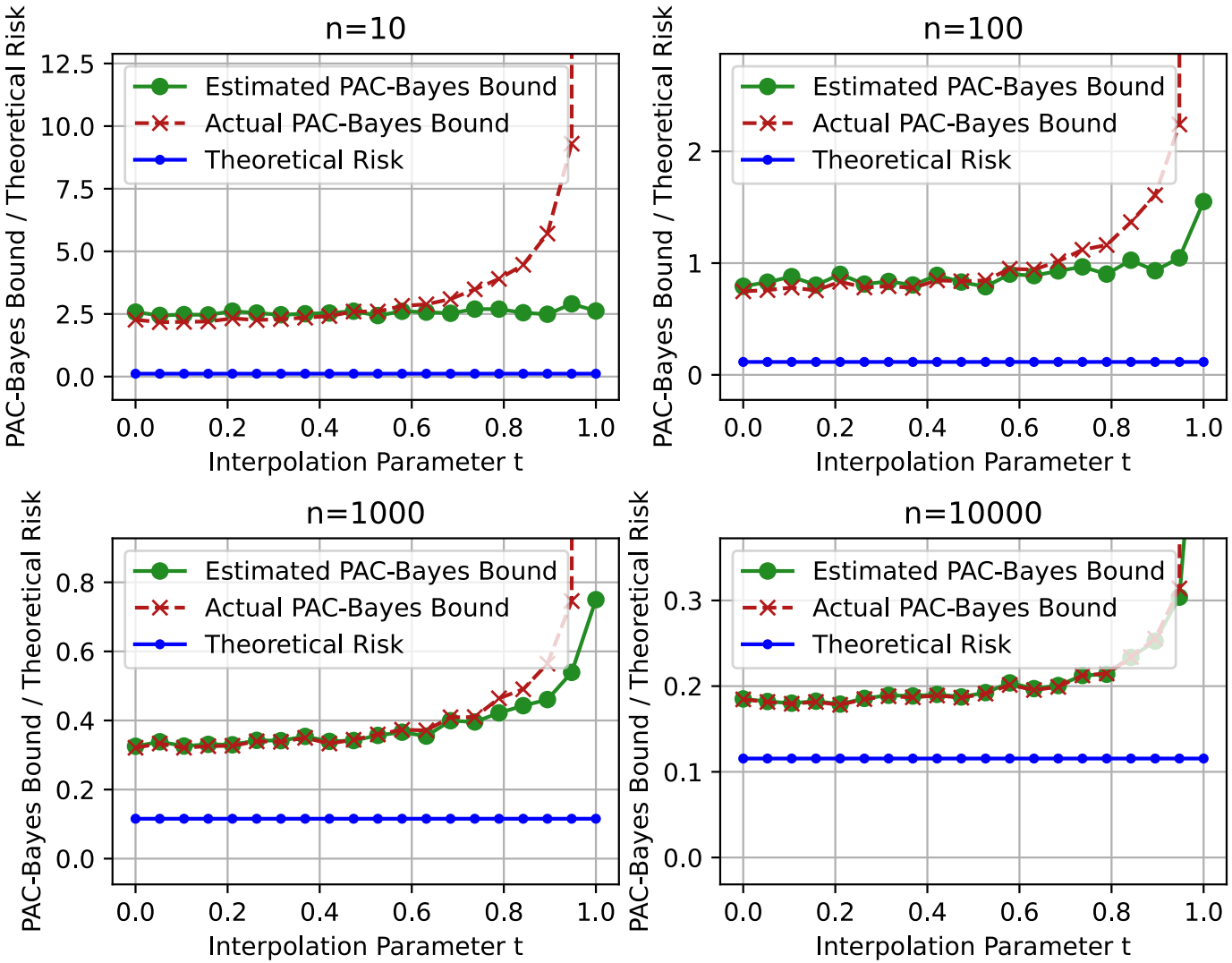
With probability at least $1 - \delta - \alpha$ on $\mathcal{S}$, for any $\hat{\rho}$,

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

dim=20


n=10
n=100
n=1000
n=10000
Pseudo Spectral Gap
Interpolation Parameter t
Estimated Pseudo Spectral Gap
Actual Pseudo Spectral Gap
0.0
0.2
0.4
0.6
0.8
1.0

dim=20


n=10
n=100
n=1000
n=10000
Estimated PAC-Bayes Bound
Actual PAC-Bayes Bound
Theoretical Risk
PAC-Bayes Bound / Theoretical Risk
Interpolation Parameter t

- Infinite case : not possible to estimate γ_{ps} *in general*.
- However, under additional assumptions, this can be feasible. For example :

Proposition

Assume there is $a \in [-1, 1]$ such that

$$X_{t+1} = aX_t + \varepsilon_t, \text{ with } \varepsilon_t \sim \mathcal{N}(0, 1).$$

Then

$$\gamma_{\mathrm{ps}} = 1 - a^2.$$

- Going beyond the Markov case ?

Theorem – essentially from Alquier and Li (2012)

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

$$\mathbb{E}_{\theta\sim\hat{\rho}}[R(\theta)] \leq \mathbb{E}_{\theta\sim\hat{\rho}}[R_n(\theta)] + \frac{\mathrm{KL}(\hat{\rho}\|\pi) + \log\frac{1}{\delta}}{\lambda} + \frac{\lambda\left(1+2\sum_{\ell=1}^{\infty}\varphi_\ell\right)^2}{8n}.$$

Can we estimate the whole sequence $(\varphi_\ell)_{\ell\geq 0}$ or even simply the sum $\Phi = \sum_{\ell=1}^{\infty}\varphi_\ell$?
Research in progress by G. Wolfer, based on :

Khaleghi, A. and Lugosi, G. (2023). Inferring the mixing properties of a stationary ergodic process from a single sample-path. *IEEE Transactions on Information Theory*.

This seems to be a difficult problem !

Thank you !

Our project "SafeTime : Safe Decisions via Adaptive Learning of Time Series" will be funded from July 1st by

Initiative of excellence

possible post-doc position in 2027, reach me if you are interested !