How to Make Use of Learning Theory to Learn Efficient Machine Learning Models:

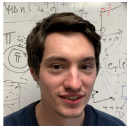
From PAC-Bayesian Generalization Bounds to (Self-Bounding) Learning Algorithms

Paul Viallard & Emilie Morvant
(paul.viallard@inria.fr – emilie.morvant@univ-st-etienne.fr)

June 30, 2025 — COLT 2025 Tutorial

https://paulviallard.github.io/colt25-pac-bayes-tutorial/

Paul Viallard — Researcher



- Inria, Rennes, France
- at IRISA in the MALT team

Emilie Morvant — Associate professor

- Jean Monnet University, Saint-Etienne, France
- at Hubert Curien Laboratory, Data Intelligence Team

Research area: **Statistical Machine Learning** (main topic: PAC-Bayesian Theory)

What's the point of theory?

What's the point of theory?

🧐 **Understand:** Formalize what learning means and identify its limits

📊 **Predict:** Anticipate performance and generalization to new data

🛡️ **Certify:** Provide guarantees—*not just hope*

Connect: Bridge the gap between understanding and deployment

🔧 **Design:** Build better algorithms by using theory as a guide

Theory is not the opposite of practice
it's what makes practice meaningful, trustworthy, and long-lasting

Our Main Objective

Giving insights on the **flexibility**, and the **practical impact** of PAC-Bayesian theory in modern machine learning

PART 1: From theory...

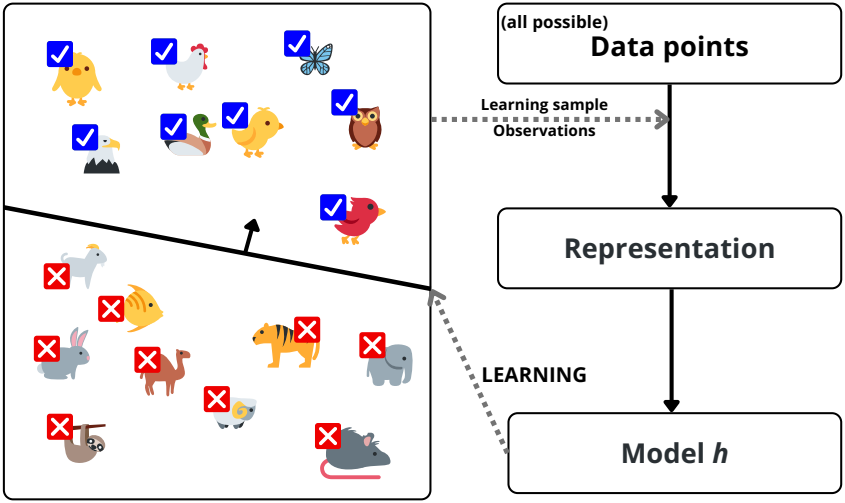
- Generalization bounds in general
- Foundations of the PAC-Bayesian theory
  - Seminal results
  - Main proof techniques
- Examples of other kinds of PAC-Bayesian bounds
  - Disintegrated PAC-Bayesian bounds
  - Bounds based on the Wasserstein distance

PART 2: ...to practices

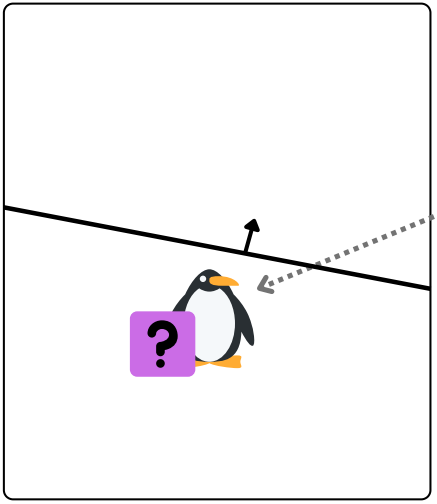
- On the majority vote
- On neural networks

Part I
FROM LEARNING THEORY

Classify animals into 2 categories: ☑ with wings or ☒ without wings



Classify animals into 2 categories: ✅ with wings or ❌ without wings



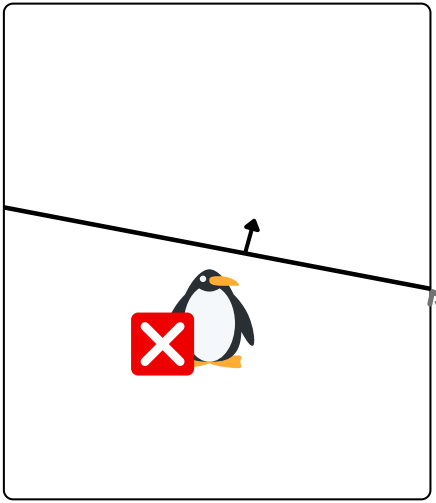
(all possible)
Data points

a new data point to classify

Representation

Model *h*

Classify animals into 2 categories: ✅ with wings or ❌ without wings



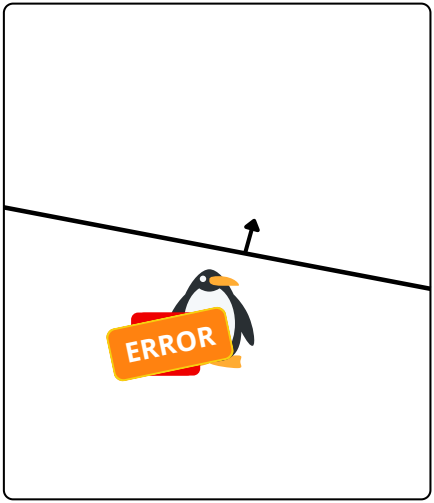
(all possible)
Data points

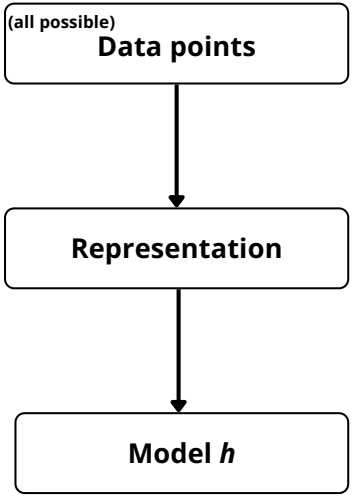
Representation

CLASSIFICATION

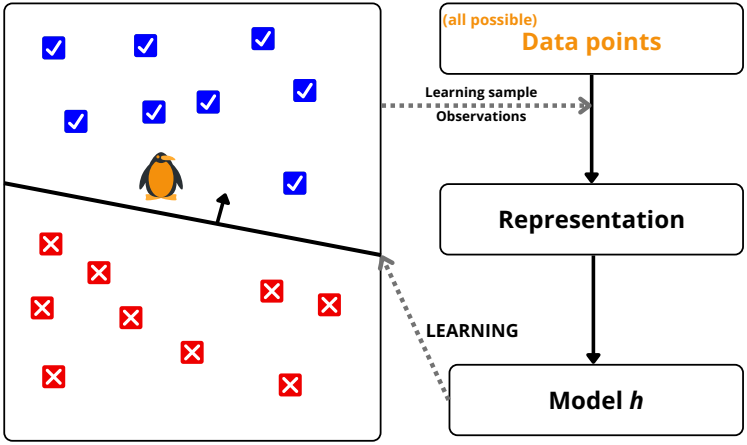
Model *h*

Classify animals into 2 categories: ✅ with wings or ❌ without wings





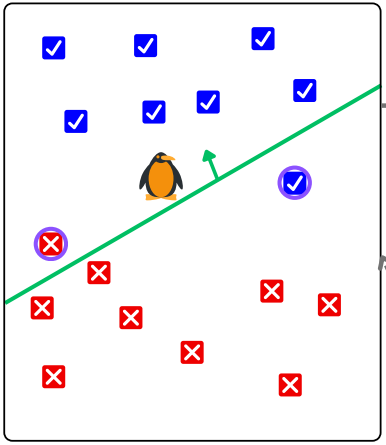
Generalization bounds



errors on all data $\leq$

NOT COMPUTABLE

Generalization bounds



(all possible)
Data points

Learning sample
Observations

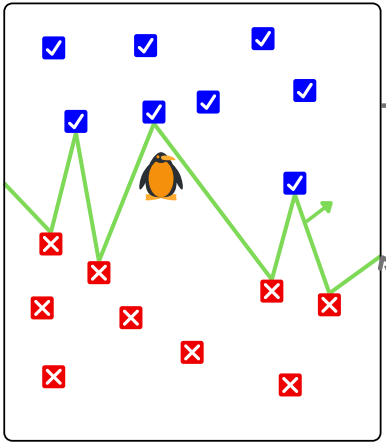
Representation

LEARNING

Model *h*

$$\underbrace{\text{errors on all data}}_{\text{NOT COMPUTABLE}} \leq \underbrace{\text{errors on the observed data} + f\left(\text{number of observed data}\right)}_{\text{COMPUTABLE (or estimable)}}$$

Generalization bounds



(all possible)
Data points

Learning sample
Observations

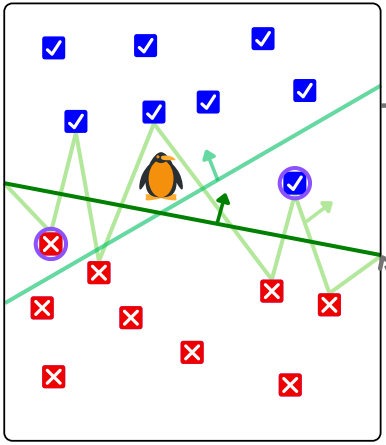
Representation

LEARNING

Model *h*

$$\underbrace{\text{errors on all data}}_{\text{NOT COMPUTABLE}} \leq \underbrace{+ f\left(\text{complexity of } h \;,\; \right)}_{\text{COMPUTABLE (or estimable)}}$$

Generalization bounds



(all possible)
Data points

Learning sample
Observations

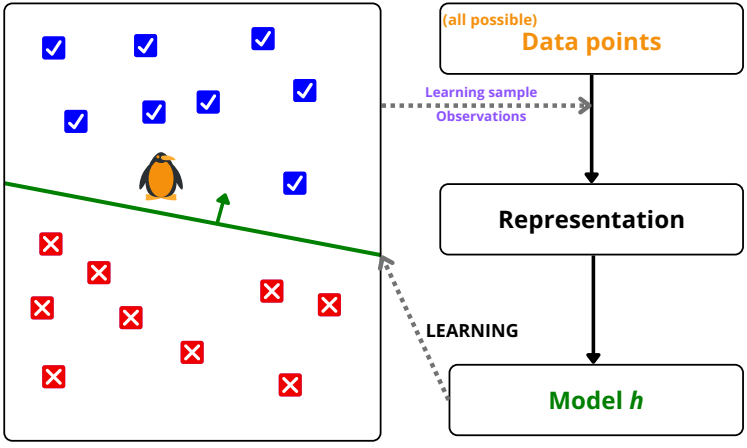
Representation

LEARNING

Model *h*

$$\underbrace{\text{errors on all data}}_{\text{NOT COMPUTABLE}} \leq \underbrace{\text{errors on the observed data} + f\left(\text{complexity of } h, \text{number of observed data}\right)}_{\text{COMPUTABLE (or estimable)}}$$

Generalization bounds



$$\text{errors on all data} \leq \text{errors on the observed data} + f\left(\text{complexity of } h, \text{number of observed data}\right)$$

To learn a model with good generalization guarantees, we can minimize the trade-off between errors on the observed data and complexity of the model

Formalization of supervised classification

$\mathbb{X} \subseteq \mathbb{R}^d$ input space

$\mathbb{Y} = \{-1, +1\}$ output/label space

$\mathbb{H}$ hypothesis space such that $\forall h \in \mathbb{H},\ h : \mathbb{X} \to \mathbb{Y}$

$\mathcal{D}$ unknown distribution on $\mathbb{X} \times \mathbb{Y}$

$\mathbb{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \overset{iid}{\sim} (\mathcal{D})^m$ learning set

$\ell : \mathbb{H} \times (\mathbb{X} \times \mathbb{Y}) \to \mathbb{R}$ **loss function**

$\ell\big(h, (\mathbf{x}, y)\big)$ measures the error of a hypothesis when predicting the label of an example

the choice depends on the task and the learning algorithm

$\mathbb{X} \subseteq \mathbb{R}^d$ input space

$\mathbb{Y} = \{-1, +1\}$ output/label space

$\mathbb{H}$ hypothesis space such that $\forall h \in \mathbb{H},\ h : \mathbb{X} \to \mathbb{Y}$

$\mathcal{D}$ unknown distribution on $\mathbb{X} \times \mathbb{Y}$

$\mathbb{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \overset{iid}{\sim} (\mathcal{D})^m$ learning set

Examples of common losses

$\ell(h,(\mathbf{x},y))$		$\ell(h,(\mathbf{x},y))$	
$\mathrm{I}[\, h(\mathbf{x}) \neq y \,]$	0–1 loss	$\frac{1}{2}\Big(1 - yh(\mathbf{x})\Big)$	linear loss
$\Big(y - h(\mathbf{x})\Big)^2$	square loss	$\max\Big\{0, 1 - yh(\mathbf{x})\Big\}$	Hinge loss
$\exp\Big(-yh(\mathbf{x})\Big)$	exponential loss	$\log\Big(1 + \exp(-yh(\mathbf{x}))\Big)$	logistic loss

Formalization of supervised classification

$\mathbb{X} \subseteq \mathbb{R}^d$ input space

$\mathbb{Y} = \{-1, +1\}$ output/label space

$\mathbb{H}$ hypothesis space such that $\forall h \in \mathbb{H},\ h : \mathbb{X} \to \mathbb{Y}$

$\mathcal{D}$ unknown distribution on $\mathbb{X} \times \mathbb{Y}$

$\mathbb{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \overset{iid}{\sim} (\mathcal{D})^m$ learning set

$\ell : \mathbb{H} \times (\mathbb{X} \times \mathbb{Y}) \to [0, 1]$ the 0-1 loss[1]

$\widehat{\mathsf{R}}_{\mathbb{S}}(h) = \dfrac{1}{m} \displaystyle\sum_{i=1}^{m} \mathrm{I}[\, h(\mathbf{x}_i) \neq y_i \,]$ associated empirical risk

Supervised classification

Goal: Find the **hypothesis** h in $\mathbb{H}$ that minimizes the true risk on all data

$$\underbrace{\mathsf{R}_{\mathcal{D}}(h) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathrm{I}[\, h(\mathbf{x}) \neq y \,]}_{\text{True risk}}$$

[1] In this tutorial, for pedagogical purposes, we mainly focuses on binary classification with 0-1 loss

Probably Approximately Correct generalization bound (Valiant, 1984)

Given a data distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for all $\delta\in(0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\underbrace{\mathrm{R}_{\mathcal{D}}(h)}_{\text{True risk}} \leq \underbrace{\widehat{\mathrm{R}}_{\mathbb{S}}(h)}_{\text{Empirical risk}} + \underbrace{\Phi(h,\mathbb{S},\delta)}_{\text{complexity}}\right] \geq 1-\delta$$

With high probability of at least $1-\delta$, the risk of h is lower than $\widehat{\mathrm{R}}_{\mathbb{S}}(h)+\Phi(h,\mathbb{S},\delta)$

We want a bound to be

- **Informative** $\iff \widehat{\mathrm{R}}_{\mathbb{S}}(h)+\Phi(h,\mathbb{S},\delta) \leq 1$
- **Tight** $\iff \left|\mathrm{R}_{\mathcal{D}}(h)-\widehat{\mathrm{R}}_{\mathbb{S}}(h)\right| \simeq \Phi(h,\mathbb{S},\delta)$
- **Trustworthy** $\iff$ a low δ increases its confidence, but decreases its tightness
- ***(easily)* Computable** $\iff \widehat{\mathrm{R}}_{\mathbb{S}}(h)+\Phi(h,\mathbb{S},\delta)$ is computable from $\mathbb{S}$

Probably Approximately Correct generalization bound (Valiant, 1984)

Given a data distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for all $\delta\in(0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\underbrace{\mathrm{R}_{\mathcal{D}}(h)}_{\text{True risk}} \leq \underbrace{\widehat{\mathrm{R}}_{\mathbb{S}}(h)}_{\text{Empirical risk}} + \underbrace{\Phi(h,\mathbb{S},\delta)}_{\text{complexity}}\right] \geq 1-\delta$$

With high probability of at least $1-\delta$, the risk of h is lower than $\widehat{\mathrm{R}}_{\mathbb{S}}(h)+\Phi(h,\mathbb{S},\delta)$

Main families of generalization bounds:

- Uniform convergence bounds (Vapnik *et al.*, 1971; Bartlett *et al.*, 2002)
- Algorithmic-dependent bounds (Bousquet *et al.*, 2002; Xu *et al.*, 2012)
- PAC-Bayesian bounds (Shawe-Taylor *et al.*, 1997; McAllester, 1998)

Key characteristics of the main families of generalization bounds

Uniform convergence bound for a hypothesis set $\mathbb{H}$

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\ \forall h \in \mathbb{H},\ \ \mathrm{R}_{\mathcal{D}}(h) \ \leq\ \widehat{\mathrm{R}}_{\mathbb{S}}(h) \ +\ \Phi(\mathbb{H}, \mathbb{S}, \delta)\ \right] \geq 1-\delta$$

worst-case bound

fixed for all $h \in \mathbb{H}$

- Valid for all hypotheses in $\mathbb{H}$

  $\hookrightarrow$ Worst-case bounds $\rightsquigarrow$ often too pessimistic

- $\Phi(h, \mathbb{S}, \delta)$ fixed for all hypotheses from $\mathbb{H}$, for example:

  - Vapnik-Chervonenkis (VC) dimension of $\mathbb{H}$
    - ✓ Can be simple to interpret (learning capacity of $\mathbb{H}$)
    - ✗ Can be very large or even infinite $\rightsquigarrow$ Non-informative bound
  - Rademacher complexity of $\mathbb{H}$
    - ✓ Tighter than VC-dim (measures how well a function can align with randomly perturbed labels)
    - ✗ Can be difficult to compute in practice

Key characteristics of the main families of generalization bounds

Algorithmic-dependent bound, with $\mathcal{A} : \mathbb{S} \mapsto h_{\mathbb{S}}$

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\mathsf{R}_{\mathcal{D}}(h_{\mathbb{S}}) \leq \widehat{\mathsf{R}}_{\mathbb{S}}(h_{\mathbb{S}}) + \Phi(\mathcal{A}, \mathbb{S}, \delta) \right] \geq 1 - \delta$$

valid only for $h_{\mathbb{S}}$ learned from $\mathcal{A}$ and $\mathbb{S}$

depends on properties of the algorithm $\mathcal{A}$

- Valid for a single hypothesis $h_{\mathbb{S}} \rightarrow$ the one learned with $\mathcal{A}$ and $\mathbb{S}$
  $\hookrightarrow$ More realistic but less general
- $\Phi(\mathcal{A}, \mathbb{S}, \delta)$ depends on algorithmic properties of $\mathcal{A}$, for example:
  - Algorithmic stability
    - ✓ Captures robustness to small variations in the data
    - ✗ May be hard to estimate (*e.g.*, if $\mathcal{A}$ constructs a sparse hypothesis)
  - Algorithmic robustness
    - ✓ Captures robustness to variations within the same region of the input space
    - ✗ Can be complicated to compute in practice

Key characteristics of the main families of generalization bounds

PAC-Bayesian bound

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall \rho \text{ on } \mathbb{H},\ \mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}_{\mathcal{D}}(h) \ \leq\ \mathop{\mathbb{E}}_{h\sim\rho} \widehat{\mathsf{R}}_{\mathbb{S}}(h) \ +\ \Phi(\Delta(\rho,\pi),\mathbb{S},\delta)\right] \geq 1-\delta$$

bound in expectation over $\mathbb{H}$

depends on a divergence Δ between ρ and an apriori π

- Expected bound over $\mathbb{H}$
  - $\hookrightarrow$ Usually tighter than worst-case bounds
  - $\hookrightarrow$ Bound on a stochastic risk over $\mathbb{H}$
- But the expectation over $\mathbb{H}$ is closely related to the majority vote over $\mathbb{H}$
- Can be easily computed (or upper-bounded)
  - $\hookrightarrow$ and minimized $\Longrightarrow$ self-bounding algorithm
- $\Phi(\Delta(\rho,\pi),\mathbb{S},\delta)$ depends on
  - the choice of a prior π over the hypothesis space
  - a divergence $\Delta(\rho,\pi)$ between ρ and π (*e.g.*, KL-divergence, Rényi divergence, Wassertein distance)

Key characteristics of the main families of generalization bounds

Uniform convergence bound for a hypothesis set $\mathbb{H}$

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\ \forall h \in \mathbb{H},\ \ \mathrm{R}_{\mathcal{D}}(h) \ \leq\ \widehat{\mathrm{R}}_{\mathbb{S}}(h) \ +\ \Phi(\mathbb{H}, \mathbb{S}, \delta)\ \right] \geq 1-\delta$$

worst-case bound

fixed for all $h \in \mathbb{H}$

Algorithmic-dependent bound with $\mathcal{A} : \mathbb{S} \mapsto h_{\mathbb{S}}$

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\ \mathrm{R}_{\mathcal{D}}(h_{\mathbb{S}}) \ \leq\ \widehat{\mathrm{R}}_{\mathbb{S}}(h_{\mathbb{S}}) \ +\ \Phi(\mathcal{A}, \mathbb{S}, \delta)\ \right] \geq 1-\delta$$

valid only for $h_{\mathbb{S}}$ learned from $\mathcal{A}$ and $\mathbb{S}$

depends on properties of the algorithm $\mathcal{A}$

PAC-Bayesian bound

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\ \forall \rho \text{ on } \mathbb{H},\ \ \mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}_{\mathcal{D}}(h) \ \leq\ \mathop{\mathbb{E}}_{h\sim\rho} \widehat{\mathrm{R}}_{\mathbb{S}}(h) \ +\ \Phi(\Delta(\rho, \pi), \mathbb{S}, \delta)\ \right] \geq 1-\delta$$

bound in expectation over $\mathbb{H}$

depends on a divergence Δ between ρ and an apriori π

The PAC-Bayesian Theory

Foundations & Basics

$\mathbb{X} \subseteq \mathbb{R}^d$ Input space

$\mathbb{Y} = \{-1, +1\}$ output space

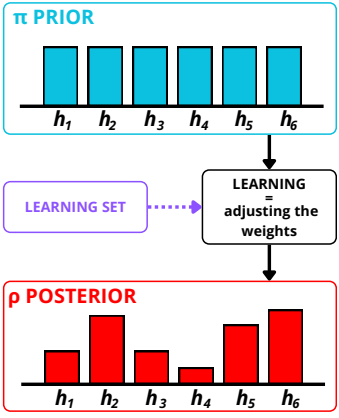
$\mathbb{H}$ hypothesis space *s.t.* $\forall h \in \mathbb{H},\ h : \mathbb{X} \to \mathbb{Y}$

$\mathcal{D}$ unknown distribution on $\mathbb{X} \times \mathbb{Y}$

$\mathbb{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \overset{iid}{\sim} \mathcal{D}^m$ learning set

π **prior distribution on** $\mathbb{H}$

ρ **posterior distribution on** $\mathbb{H}$

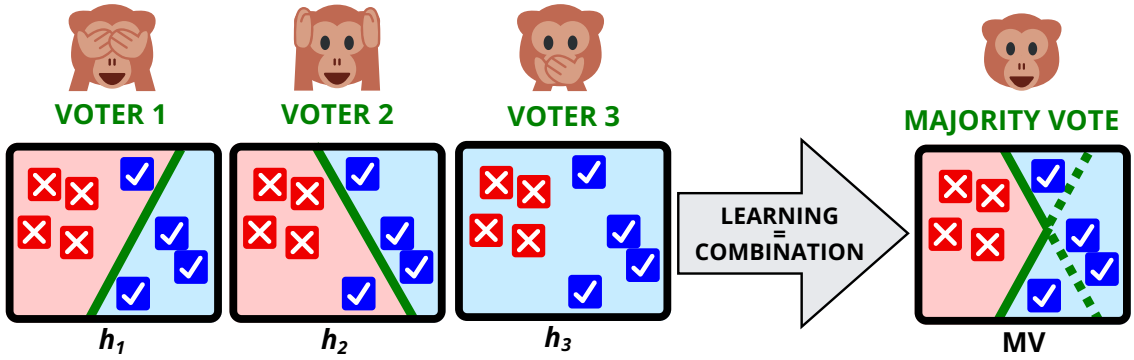


PAC-Bayesian supervised classification — Stochastic classifier

Goal : Find the posterior distribution ρ on $\mathbb{H}$ that minimizes the expected true risk on all data

$$\underbrace{\mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}_{\mathcal{D}}(h) = \mathop{\mathbb{E}}_{h\sim\rho} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathrm{I}[\,h(\mathbf{x}) \neq y\,]}_{\text{True Gibbs Risk}}$$

The ρ-weighted majority vote



PAC-Bayesian supervised classification — Majority vote classifier

Goal: Find the ρ-weighted majority vote MV on $\mathbb{H}$ that minimizes the true risk

$$\underbrace{\mathrm{R}_{\mathcal{D}}(\mathrm{MV}) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathrm{I}[\,\mathrm{MV}(\mathbf{x}) \neq y\,]}_{\text{True risk}}$$

where $\mathrm{MV}(\cdot) = \mathrm{sign}\left[\sum_{h\in\mathbb{H}} \overbrace{\rho(h)}^{\text{weight of } h}\, h(\cdot)\right]$

$$= \mathrm{sign}\left[\mathop{\mathbb{E}}_{h\sim\rho} h(\cdot)\right]$$

The Gibbs risk as a first surrogate of the risk of the majority vote

Recall: $\forall h \in \mathbb{H},\ h : \mathbb{X} \to \{-1, +1\}$ and $\mathbb{Y} = \{-1, +1\}$

The 2-factor bound

$$\mathsf{R}_{\mathcal{D}}(\mathsf{MV}) \ \leq \ 2 \times \mathop{\mathbb{E}}_{h \sim \rho} \mathsf{R}_{\mathcal{D}}(h)$$

A **bound in expectation** on a set of hypotheses $\mathbb{H}$ $\Longrightarrow$ a bound for a **deterministic** model

Recall: $\forall h \in \mathbb{H},\ h : \mathbb{X} \to \{-1, +1\}$ and $\mathbb{Y} = \{-1, +1\}$

The 2-factor bound

$$\mathrm{R}_{\mathcal{D}}(\mathrm{MV}) \ \leq \ 2 \times \mathop{\mathbb{E}}_{h \sim \rho} \mathrm{R}_{\mathcal{D}}(h)$$

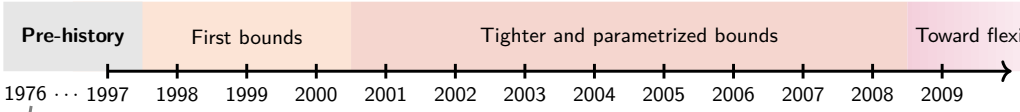
A **bound in expectation** on a set of hypotheses $\mathbb{H} \implies$ a bound for a **deterministic** model

Proof — From Markov's inequality: $\mathbb{P}\left(X \geq a\right) \leq \frac{\mathbb{E}\, X}{a}$

$$\begin{aligned}
\mathrm{R}_{\mathcal{D}}(\mathrm{MV}) &= \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathrm{I}[\,\mathrm{MV}(\mathbf{x}) \neq y\,] = \mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\mathop{\mathbb{E}}_{h\sim\rho} yh(\mathbf{x}) \leq 0 \right] \\
&= \mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\frac{1}{2}\left(1 - \mathop{\mathbb{E}}_{h\sim\rho} yh(\mathbf{x})\right) \geq \frac{1}{2} \right] \\
&\leq 2 \times \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \frac{1}{2}\left(1 - \mathop{\mathbb{E}}_{h\sim\rho} yh(\mathbf{x})\right) \\
&= 2 \times \mathop{\mathbb{E}}_{h\sim\rho} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \frac{1}{2}\Big(1 - yh(\mathbf{x})\Big) \ = \ 2 \times \mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}_{\mathcal{D}}(h)
\end{aligned}$$

Spoiler: There exist tighter surrogates of $\mathrm{R}_{\mathcal{D}}(\mathrm{MV})$ (*e.g.*, with Cantely-Chebychev's inequality)

History of PAC-Bayes



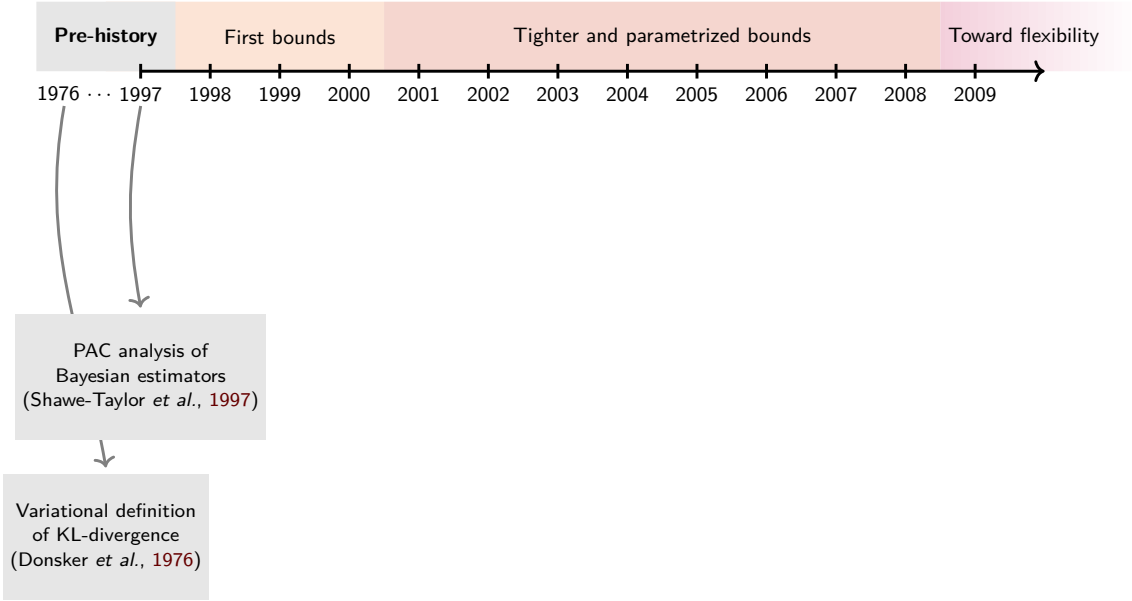
For any ρ and π on $\mathbb{H}$, for any measurable function $\phi : \mathbb{H} \to \mathbb{R}$, we have

$$\mathop{\mathbb{E}}_{h\sim\rho} \phi(h) \;\leq\; \mathrm{KL}(\rho\|\pi) + \ln\left[\mathop{\mathbb{E}}_{h\sim\pi} \exp\left(\phi(h)\right)\right]$$

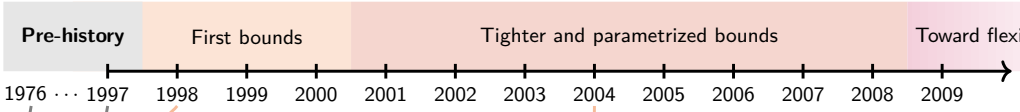
with $\mathrm{KL}(\rho\|\pi) = \mathop{\mathbb{E}}_{h\sim\rho} \ln \dfrac{\pi(h)}{\rho(h)}$ the Kullback-Leibler divergence

Variational definition of KL-divergence (Donsker *et al.*, 1976)

History of PAC-Bayes



History of PAC-Bayes



First PAC-Bayesian Theorem (McAllester, 1998) (Maurer, 2004)

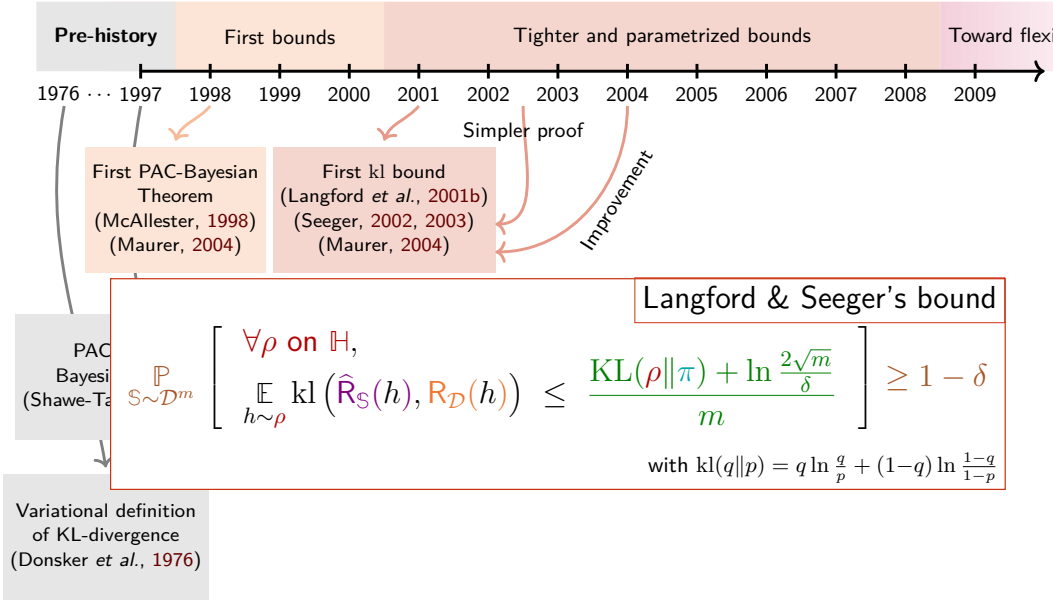
Improvement

McAllester's bound

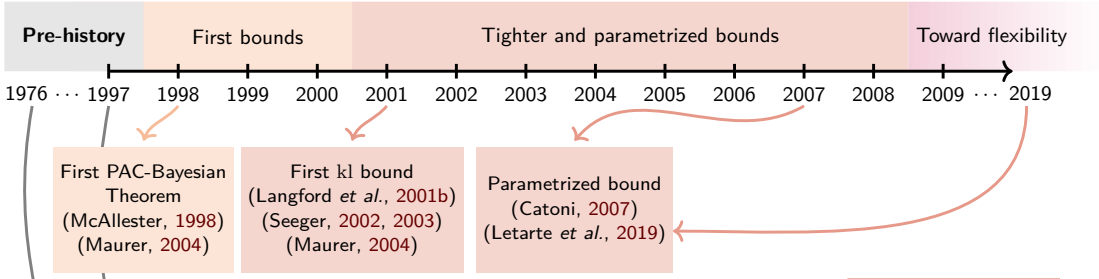
$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\ \forall\rho \text{ on } \mathbb{H},\quad \mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}_{\mathcal{D}}(h) \ \leq\ \mathop{\mathbb{E}}_{h\sim\rho} \widehat{\mathsf{R}}_{\mathbb{S}}(h) \ +\ \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{2\sqrt{m}}{\delta}}{m}}\ \right] \geq 1-\delta$$

Variational definition of KL-divergence (Donsker *et al.*, 1976)

History of PAC-Bayes



History of PAC-Bayes



For a fixed $c > 0$,

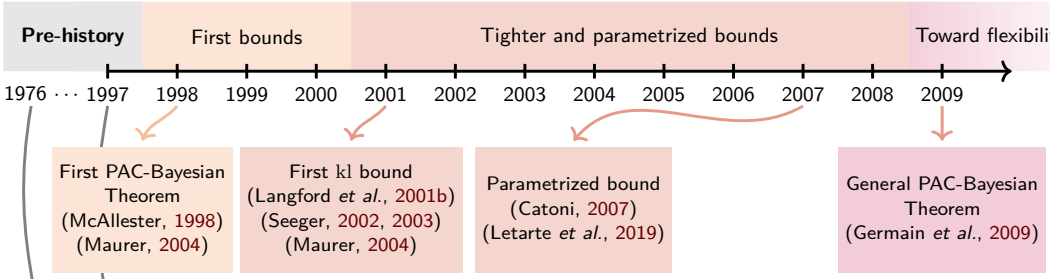
Catoni's bound

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\ \forall \rho \text{ on } \mathbb{H},\ \mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}_{\mathcal{D}}(h) \ \le\ \frac{1}{1-e^{-c}}\left(c \mathop{\mathbb{E}}_{h\sim\rho} \widehat{\mathsf{R}}_{\mathbb{S}}(h) + \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi) + \frac{1}{\delta}\right]\right)\ \right] \ge 1-\delta$$

Letarte's bound

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\ \forall \rho \text{ on } \mathbb{H},\ \forall c > 0,\ \mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}_{\mathcal{D}}(h) \ \le\ \frac{1}{1-e^{-c}}\left(c \mathop{\mathbb{E}}_{h\sim\rho} \widehat{\mathsf{R}}_{\mathbb{S}}(h) + \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi) + \frac{2\sqrt{m}}{\delta}\right]\right)\ \right] \ge 1-\delta$$

History of PAC-Bayes



PAC-Bayesian general bound

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\ \forall \rho \text{ on } \mathbb{H},\quad \mathrm{D}\left(\mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}_{\mathcal{D}}(h),\ \mathop{\mathbb{E}}_{h\sim\rho} \widehat{\mathrm{R}}_{\mathbb{S}}(h)\right) \ \leq\ \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{\mathcal{I}_{\mathrm{D}}(m)}{\delta}\right]\ \right] \geq 1-\delta$$

$\mathrm{D} : [0,1]\times[0,1]\to\mathbb{R}$ a convex function

$\Longrightarrow$ Measures the deviation between empirical and true risks

$\Longrightarrow$ Allows to retrieve the seminal bounds

A general PAC-Bayesian bound

General theorem (Germain *et al.*, 2009)

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall \rho \text{ on } \mathbb{H},\ \mathrm{D}\left(\mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}_{\mathcal{D}}(h), \mathop{\mathbb{E}}_{h\sim\rho} \widehat{\mathsf{R}}_{\mathbb{S}}(h)\right) \leq \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi) + \ln \frac{\mathcal{I}_{\mathrm{D}}(m)}{\delta}\right]\right] \geq 1-\delta$$

Rediscovering seminal bounds

- $\mathrm{D}(a,b) = 2(a-b)^2$
  $\hookrightarrow$ McAllester's bound (simple, interpretable)
- $\mathrm{D}(a,b) = \mathsf{kl}(a\|b)$
  $\hookrightarrow$ Seeger's bound (tighter when a, b close)
- $\mathrm{D}(a,b) = -\ln(1-b[1-e^{-c}]) - c\cdot a$
  $\hookrightarrow$ Catoni's bound (parametrized)

Key idea

Each choice of D corresponds to a way to capture the so-called generalization gap

$\Longrightarrow$ We can tailor a PAC-Bayes bound to fit the task's needs

General theorem (Germain *et al.*, 2009)

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall \rho \text{ on } \mathbb{H},\ \mathrm{D}\left(\mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}_{\mathcal{D}}(h), \mathop{\mathbb{E}}_{h\sim\rho} \widehat{\mathsf{R}}_{\mathbb{S}}(h)\right) \leq \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi) + \ln \frac{\mathcal{I}_{\mathrm{D}}(m)}{\delta}\right]\right] \geq 1-\delta$$

Proof main tools

- From posterior ρ to prior $\pi \rightsquigarrow$ Change of Measure Inequality (Donsker *et al.*, 1976)
  For any ρ and π on $\mathbb{H}$, for any measurable function $\phi : \mathbb{H} \to \mathbb{R}$, we have

$$\mathop{\mathbb{E}}_{h\sim\rho} \phi(h) \leq \mathrm{KL}(\rho\|\pi) + \ln\left[\mathop{\mathbb{E}}_{h\sim\pi} \exp\left(\phi(h)\right)\right]$$

- Markov's inequality

$$\mathbb{P}\left(X \leq \frac{\mathbb{E}\, X}{\delta}\right) \geq 1-\delta \qquad \Longleftrightarrow \qquad X \leq_{1-\delta} \frac{\mathbb{E}\, X}{\delta}$$

General theorem (Germain *et al.*, 2009)

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall \rho \text{ on } \mathbb{H},\ \mathrm{D}\left(\mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}_{\mathcal{D}}(h), \mathop{\mathbb{E}}_{h\sim\rho} \widehat{\mathsf{R}}_{\mathbb{S}}(h)\right) \le \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi) + \ln \frac{\mathcal{I}_{\mathrm{D}}(m)}{\delta}\right]\right] \ge 1-\delta$$

Proof

$$m\,\mathrm{D}\left(\mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}_{\mathcal{D}}(h), \mathop{\mathbb{E}}_{h\sim\rho} \widehat{\mathsf{R}}_{\mathbb{S}}(h)\right)$$

Jensen's inequality $\le$ $\mathop{\mathbb{E}}_{h\sim\rho} m\,\mathrm{D}\left(\mathsf{R}_{\mathcal{D}}(h), \widehat{\mathsf{R}}_{\mathbb{S}}(h)\right)$

Change of measure $\le$ $\mathrm{KL}(\rho\|\pi) + \ln \mathop{\mathbb{E}}_{h\sim\pi} \exp\left[m\,\mathrm{D}\left(\mathsf{R}_{\mathcal{D}}(h), \widehat{\mathsf{R}}_{\mathbb{S}}(h)\right)\right]$

Markov's inequality $\le_{1-\delta}$ $\mathrm{KL}(\rho\|\pi) + \ln \frac{1}{\delta} \mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{h\sim\pi} \exp\left[m\,\mathrm{D}\left(\mathsf{R}_{\mathcal{D}}(h), \widehat{\mathsf{R}}_{\mathbb{S}'}(h)\right)\right]$

Expectation swap $=$ $\mathrm{KL}(\rho\|\pi) + \ln \frac{1}{\delta} \mathop{\mathbb{E}}_{h\sim\pi} \mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m} \exp\left[m\,\mathrm{D}\left(\mathsf{R}_{\mathcal{D}}(h), \widehat{\mathsf{R}}_{\mathbb{S}'}(h)\right)\right]$

$= \mathrm{KL}(\rho\|\pi) + \ln \frac{\mathcal{I}_{\mathrm{D}}(m)}{\delta}$

Rooted in **concentration inequalities**

- Like many generalization bounds, PAC-Bayes relies on concentration inequalities

So what? Why is PAC-Bayes special?

🧮 Rooted in **concentration inequalities**

- Like many generalization bounds, PAC-Bayes relies on concentration inequalities

🧰 PAC-Bayes is a **flexible** framework

- The deviation term $\mathrm{D}(\text{true risk}, \text{empirical risk})$ can be user-defined
  $\Rightarrow$ One can adapt the bound to different tasks, measure of risks, etc

So what? Why is PAC-Bayes special?

🧮 Rooted in **concentration inequalities**

- Like many generalization bounds, PAC-Bayes relies on concentration inequalities

🧰 PAC-Bayes is a **flexible** framework

- The deviation term $\mathrm{D}(\text{true risk}, \text{empirical risk})$ can be user-defined
  $\Rightarrow$ One can adapt the bound to different tasks, measure of risks, etc

🔁 **But there's more**: a key step in the proof is the **change of measure**

- **Main trick**: compare the *apriori* belief to a data-dependent learned knowledge
- **Complexity control** via $\Delta(\text{posterior}, \text{prior})$

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall \rho,\ \mathrm{D}\left(\mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}_{\mathcal{D}}(h),\ \mathop{\mathbb{E}}_{h\sim\rho} \widehat{\mathrm{R}}_{\mathbb{S}}(h)\right) \leq \Phi(\Delta(\rho,\pi),\mathbb{S},\delta)\right] \geq 1-\delta$$

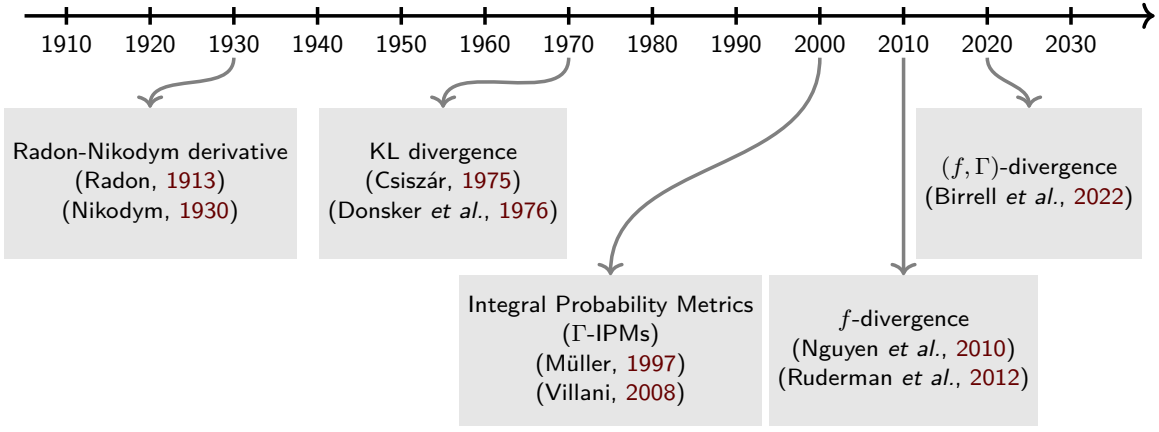
upper bound on the gap D between true and empirical risks

USER-DEFINED

depends on a divergence Δ between ρ and an apriori π

CHANGE OF MEASURE

History of the change of measure



Examples of some other PAC-Bayesian bounds

- with Rényi divergence (Bégin *et al.*, 2016; Alquier *et al.*, 2018)
- with general change of measure for f-divergence (Ohnishi *et al.*, 2021; Picard-Weibel *et al.*, 2022)
- with the Zhang-Cutkosky-Paschalidis (ZCP) divergence (Kuzborskij *et al.*, 2024)

Disintegrated PAC-Bayesian Bounds

Classical **PAC-Bayesian bound**

For any $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any prior π on $\mathbb{H}$, for any $\delta\in(0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\boxed{\forall\rho \text{ on } \mathbb{H}},\quad \mathop{\mathbb{E}}_{h\sim\rho}\mathrm{R}_{\mathcal{D}}(h)\leq \mathop{\mathbb{E}}_{h\sim\rho}\widehat{\mathrm{R}}_{\mathbb{S}}(h)+\sqrt{\frac{1}{2m}\left[\mathrm{KL}(\rho\|\pi)+\ln\frac{2\sqrt{m}}{\delta}\right]}\right]\geq 1-\delta$$

A drawback: **A PAC-Bayes bound is on a randomized/stochastic classifier**

❌ The randomized classifier is often not used in practice
↪ Most applications deploy a single (deterministic) model
↪ So, what about guarantees for the algorithm output?

Need for a **derandomization** of the bound!

⇓

Disintegrated PAC-Bayesian Bound

Classical PAC-Bayesian bound

For any $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any prior π on $\mathbb{H}$, for any $\delta\in(0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\boxed{\forall\rho \text{ on } \mathbb{H}},\quad \mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}_{\mathcal{D}}(h)\le\mathop{\mathbb{E}}_{h\sim\rho}\widehat{\mathsf{R}}_{\mathbb{S}}(h)+\sqrt{\frac{1}{2m}\left[\mathrm{KL}(\rho\|\pi)+\ln\frac{2\sqrt{m}}{\delta}\right]}\right]\ge 1-\delta$$

Disintegrated PAC-Bayesian bound (Blanchard *et al.*, 2007; Catoni, 2007; Rivasplata *et al.*, 2020)

For any $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any prior π on $\mathbb{H}$, for any $\delta\in(0,1]$, **for any algorithm** A, we have

$$\mathop{\mathbb{P}}_{\substack{\mathbb{S}\sim\mathcal{D}^m,\\ \boxed{h\sim\rho_{\mathbb{S}}}}}\left[\mathsf{R}_{\mathcal{D}}(h)\le\widehat{\mathsf{R}}_{\mathbb{S}}(h)+\sqrt{\frac{1}{2m}\left[\ln\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}+\ln\frac{2\sqrt{m}}{\delta}\right]_+}\right]\ge 1-\delta$$

where $\rho_{\mathbb{S}}=A(\mathbb{S},\pi)$ is output by the deterministic algorithm A and $[a]_+=\max(0,a)$

General disintegrated PAC-Bayesian bound

***Disintegrated* PAC-Bayesian bound** (Blanchard *et al.*, 2007; Catoni, 2007; Rivasplata *et al.*, 2020)
For any $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any prior π on $\mathbb{H}$, for any $\delta\in(0,1]$, **for any algorithm** A, we have

$$\mathop{\mathbb{P}}_{\substack{\mathbb{S}\sim\mathcal{D}^m,\\ \boxed{h\sim\rho_{\mathbb{S}}}}}\left[\mathrm{R}_{\mathcal{D}}(h)\leq \widehat{\mathrm{R}}_{\mathbb{S}}(h)+\sqrt{\frac{1}{2m}\left[\ln\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}+\ln\frac{2\sqrt{m}}{\delta}\right]_+}\right]\geq 1-\delta$$

where $\rho_{\mathbb{S}}=A(\mathbb{S},\pi)$ is output by the deterministic algorithm A and $[a]_+=\max(0,a)$

General disintegrated PAC-Bayesian bound (Rivasplata *et al.*, 2020)

$$\mathop{\mathbb{P}}_{\substack{\mathbb{S}\sim\mathcal{D}^m,\\ h\sim\rho_{\mathbb{S}}}}\left[\mathrm{D}\left(\mathrm{R}_{\mathcal{D}}(h),\widehat{\mathrm{R}}_{\mathbb{S}}(h)\right)\leq\frac{1}{m}\left[\ln\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}+\ln\frac{\mathcal{I}_{\mathrm{D}}(m)}{\delta}\right]\right]\geq 1-\delta$$

Proof of the disintegrated PAC-Bayesian bound of Rivasplata *et al.* (2020)

General disintegrated PAC-Bayesian bound (Rivasplata *et al.*, 2020)

$$\underset{\substack{\mathbb{S}\sim\mathcal{D}^m,\\ h\sim\rho_{\mathbb{S}}}}{\mathbb{P}}\left[\mathrm{D}\left(\mathrm{R}_{\mathcal{D}}(h),\widehat{\mathrm{R}}_{\mathbb{S}}(h)\right)\leq\frac{1}{m}\left[\ln\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}+\ln\frac{\mathcal{I}_{\mathrm{D}}(m)}{\delta}\right]\right]\geq 1-\delta$$

Proof

$$m\mathrm{D}\left(\mathrm{R}_{\mathcal{D}}(h),\widehat{\mathrm{R}}_{\mathbb{S}}(h)\right)-\ln\left(\frac{d\rho_{\mathbb{S}}}{d\pi}(h)\right)$$

Markov's inequality: $\leq_{1-\delta}$ $\ln\frac{1}{\delta}+\ln\left(\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\rho_{\mathbb{S}}}\exp\left(m\mathrm{D}\left(\mathrm{R}_{\mathcal{D}}(h),\widehat{\mathrm{R}}_{\mathbb{S}'}(h)\right)-\ln\frac{d\rho_{\mathbb{S}}}{d\pi}(h)\right)\right)$

Change of measure: $=$ $\ln\frac{1}{\delta}+\ln\left(\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\pi}\exp\left(m\mathrm{D}\left(\mathrm{R}_{\mathcal{D}}(h),\widehat{\mathrm{R}}_{\mathbb{S}'}(h)\right)\right)\right)$

Expectation swap: $=$ $\ln\frac{1}{\delta}+\ln\left(\mathop{\mathbb{E}}_{h\sim\pi}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\exp\left(m\mathrm{D}\left(\mathrm{R}_{\mathcal{D}}(h),\widehat{\mathrm{R}}_{\mathbb{S}'}(h)\right)\right)\right)$

$$=\ln\frac{\mathcal{I}_{\mathrm{D}}(m)}{\delta}$$

***Disintegrated* PAC-Bayesian bound** (example based on Rivasplata *et al.*, 2020)

For any $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any prior π on $\mathbb{H}$, for any $\delta\in(0,1]$, **for any algorithm** A, we have

$$\mathop{\mathbb{P}}_{\substack{\mathbb{S}\sim\mathcal{D}^m,\\ \boxed{h\sim\rho_{\mathbb{S}}}}}\left[\mathsf{R}_{\mathcal{D}}(h)\leq\widehat{\mathsf{R}}_{\mathbb{S}}(h)+\sqrt{\frac{1}{2m}\left[\ln\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}+\ln\frac{2\sqrt{m}}{\delta}\right]_+}\right]\geq 1-\delta$$

where $\rho_{\mathbb{S}}=A(\mathbb{S},\pi)$ is output by the deterministic algorithm A and $[a]_+=\max(0,a)$

A practical perspective (derandomization of the PAC-Bayesian bounds)

1. Sampling a learning sample $\mathbb{S}$ from $\mathcal{D}$
2. Sampling a hypothesis h from $\rho_{\mathbb{S}}=A(\mathbb{S},\pi)$
3. With high probability (at least $1-\delta$), a bound holds, *e.g.*, we have

$$\mathsf{R}_{\mathcal{D}}(h)\ \leq\ \widehat{\mathsf{R}}_{\mathbb{S}}(h)+\sqrt{\frac{1}{2m}\left[\ln\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}+\ln\frac{2\sqrt{m}}{\delta}\right]_+}$$

***Disintegrated* PAC-Bayesian bound** (example based on Rivasplata *et al.*, 2020)

For any $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any prior π on $\mathbb{H}$, for any $\delta\in(0,1]$, **for any algorithm** A, we have

$$\mathop{\mathbb{P}}_{\substack{\mathbb{S}\sim\mathcal{D}^m,\\ \boxed{h\sim\rho_{\mathbb{S}}}}}\left[\mathrm{R}_{\mathcal{D}}(h)\leq\widehat{\mathrm{R}}_{\mathbb{S}}(h)+\sqrt{\frac{1}{2m}\left[\ln\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}+\ln\frac{2\sqrt{m}}{\delta}\right]_+}\right]\geq 1-\delta$$

where $\rho_{\mathbb{S}}=A(\mathbb{S},\pi)$ is output by the deterministic algorithm A and $[a]_+=\max(0,a)$

1. **Advantage:** Bound for a *unique* hypothesis $h\sim\rho_{\mathbb{S}}$
2. **Drawback:** The term depends $\ln\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}$ on the hypothesis $h\sim\rho_{\mathbb{S}}$

   $\Rightarrow$ Difficult to control/minimize

A *practical* disintegrated PAC-Bayesian bound (Viallard *et al.*, 2024b)

For any $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any prior π on $\mathbb{H}$, for any $\lambda>1$ for any $\delta\in(0,1]$, **for any algorithm** A, we have

$$\mathop{\mathbb{P}}_{\substack{\mathbb{S}\sim\mathcal{D}^m,\\ h\sim\rho_{\mathbb{S}}}}\left[\mathrm{R}_{\mathcal{D}}(h)\leq\widehat{\mathrm{R}}_{\mathbb{S}}(h)+\sqrt{\frac{1}{2m}\left[\mathrm{D}_\lambda(\rho_{\mathbb{S}}\|\pi)+\frac{2\lambda-1}{\lambda-1}\ln\frac{2}{\delta}+\ln(2\sqrt{m})\right]}\right]\geq 1-\delta$$

with $\mathrm{D}_\lambda(\rho_{\mathbb{S}}\|\pi)=\frac{1}{\lambda-1}\ln\left[\mathop{\mathbb{E}}_{h\sim\pi}\left[\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}\right]^\lambda\right]$ and $\rho_{\mathbb{S}}=A(\mathbb{S},\pi)$ is output by the deterministic algorithm A

The main difference

The above bound depends on the Rényi divergence $\mathrm{D}_\lambda(\rho_{\mathbb{S}}\|\pi)$ instead of $\ln\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}$

$$\mathop{\mathbb{P}}_{\substack{\mathbb{S}\sim\mathcal{D}^m,\\ h\sim\rho_{\mathbb{S}}}}\left[\mathrm{R}_{\mathcal{D}}(h)\leq\widehat{\mathrm{R}}_{\mathbb{S}}(h)+\sqrt{\frac{1}{2m}\left[\ln\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}+\ln\frac{2\sqrt{m}}{\delta}\right]_+}\right]\geq 1-\delta$$

General disintegrated PAC-Bayesian bound of Viallard (2024b)

A *practical* disintegrated PAC-Bayesian bound (Viallard *et al.*, 2024b)

For any $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any prior π on $\mathbb{H}$, for any $\lambda>1$ for any $\delta\in(0,1]$, **for any algorithm** A, we have

$$\mathop{\mathbb{P}}_{\substack{\mathbb{S}\sim\mathcal{D}^m,\\ h\sim\rho_{\mathbb{S}}}}\left[\mathrm{R}_{\mathcal{D}}(h)\leq\widehat{\mathrm{R}}_{\mathbb{S}}(h)+\sqrt{\frac{1}{2m}\left[\mathrm{D}_\lambda(\rho_{\mathbb{S}}\|\pi)+\frac{2\lambda-1}{\lambda-1}\ln\frac{2}{\delta}+\ln(2\sqrt{m})\right]}\right]\geq 1-\delta$$

with $\mathrm{D}_\lambda(\rho_{\mathbb{S}}\|\pi)=\frac{1}{\lambda-1}\ln\left[\mathop{\mathbb{E}}_{h\sim\pi}\left[\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}\right]^\lambda\right]$ and $\rho_{\mathbb{S}}=A(\mathbb{S},\pi)$ is output by the deterministic algorithm A

General disintegrated PAC-Bayesian bound (Viallard *et al.*, 2024b)

For any $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any prior π on $\mathbb{H}$, for any $\lambda>1$ for any $\delta\in(0,1]$, **for any algorithm** A, we have

$$\mathop{\mathbb{P}}_{\substack{\mathbb{S}\sim\mathcal{D}^m,\\ h\sim\rho_{\mathbb{S}}}}\left[\mathrm{D}\left(\mathrm{R}_{\mathcal{D}}(h),\widehat{\mathrm{R}}_{\mathbb{S}}(h)\right)\leq\frac{1}{m}\left[\mathrm{D}_\lambda(\rho_{\mathbb{S}}\|\pi)+\frac{2\lambda-1}{\lambda-1}\ln\frac{2}{\delta}+\ln\mathcal{I}_{\mathrm{D}}(m)\right]\right]\geq 1-\delta$$

Proof of the disintegrated PAC-Bayesian bound of Viallard *et al.* (2024b)

General disintegrated PAC-Bayesian bound (Viallard *et al.*, 2024b)

$$\mathop{\mathbb{P}}_{\substack{\mathbb{S}\sim\mathcal{D}^m,\\ h\sim\rho_{\mathbb{S}}}}\left[\mathrm{D}\left(\mathrm{R}_{\mathcal{D}}(h),\widehat{\mathrm{R}}_{\mathbb{S}}(h)\right)\leq\frac{1}{m}\left[\mathrm{D}_\lambda(\rho_{\mathbb{S}}\|\pi)+\frac{2\lambda-1}{\lambda-1}\ln\frac{2}{\delta}+\ln\mathcal{I}_{\mathrm{D}}(m)\right]\right]\geq 1-\delta$$

Proof

$$m\,\mathrm{D}\left(\mathrm{R}_{\mathcal{D}}(h),\widehat{\mathrm{R}}_{\mathbb{S}}(h)\right)$$

Markov on h $\quad\leq_{1-\frac{\delta}{2}}\quad \frac{\lambda}{\lambda-1}\ln\left[\frac{2}{\delta}\mathop{\mathbb{E}}_{h'\sim\rho_{\mathbb{S}}}\exp\left(\frac{\lambda-1}{\lambda}m\mathrm{D}\left(\mathrm{R}_{\mathcal{D}}(h'),\widehat{\mathrm{R}}_{\mathbb{S}}(h')\right)\right)\right]$

Change of measure $\quad=\quad D_\lambda(\rho_{\mathbb{S}}\|\pi)+\frac{\lambda}{\lambda-1}\ln\frac{2}{\delta}+\ln\left[\mathop{\mathbb{E}}_{h'\sim\pi}\exp\left(m\,\mathrm{D}\left(\mathrm{R}_{\mathcal{D}}(h'),\widehat{\mathrm{R}}_{\mathbb{S}}(h')\right)\right)\right]$

Markov on $\mathbb{S}$ $\quad\leq_{1-\frac{\delta}{2}}\quad D_\lambda(\rho_{\mathbb{S}}\|\pi)+\frac{\lambda}{\lambda-1}\ln\frac{2}{\delta}+\ln\left[\frac{2}{\delta}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\exp\left(m\mathrm{D}\left(\mathrm{R}_{\mathcal{D}}(h'),\widehat{\mathrm{R}}_{\mathbb{S}'}(h')\right)\right)\right]$

Expectation swap $\quad=\quad D_\lambda(\rho_{\mathbb{S}}\|\pi)+\frac{\lambda}{\lambda-1}\ln\frac{2}{\delta}+\ln\left[\frac{2}{\delta}\mathop{\mathbb{E}}_{h'\sim\pi}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\exp\left(m\mathrm{D}\left(\mathrm{R}_{\mathcal{D}}(h'),\widehat{\mathrm{R}}_{\mathbb{S}'}(h')\right)\right)\right]$

$$=\quad D_\lambda(\rho_{\mathbb{S}}\|\pi)+\frac{2\lambda-1}{\lambda-1}\ln\frac{2}{\delta}+\ln\frac{\mathcal{I}_{\mathrm{D}}(m)}{\delta}$$

PAC-Bayesian Bounds with the Wasserstein Distance

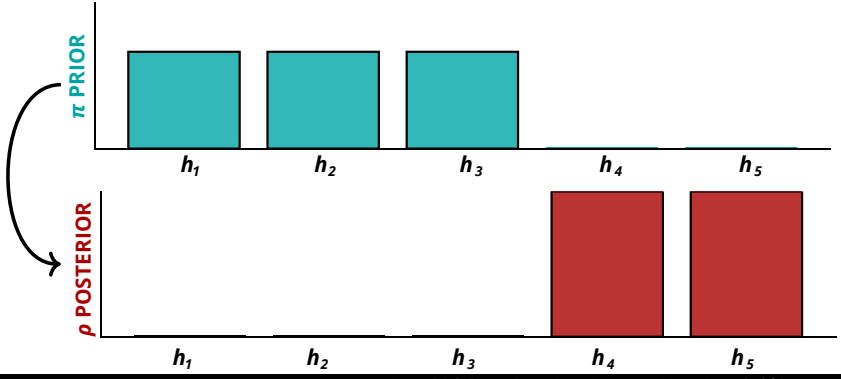
***Classical* PAC-Bayesian bound**

For any $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any prior π on $\mathbb{H}$, for any $\delta\in(0,1]$, we have

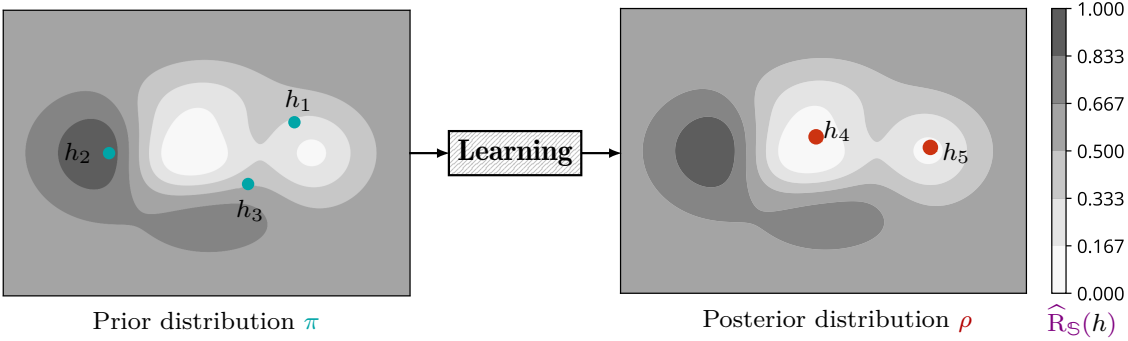
$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m}\left[\forall\rho \text{ on } \mathbb{H},\quad \mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}_{\mathcal{D}}(h) \le \mathop{\mathbb{E}}_{h\sim\rho} \widehat{\mathrm{R}}_{\mathcal{S}}(h) + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{2\sqrt{m}}{\delta}\right]}\right] \ge 1-\delta$$

Another drawback: **PAC-Bayesian bound is infinite if ρ and π have disjoint supports**

$\mathrm{KL}(\rho\|\pi) = +\infty$



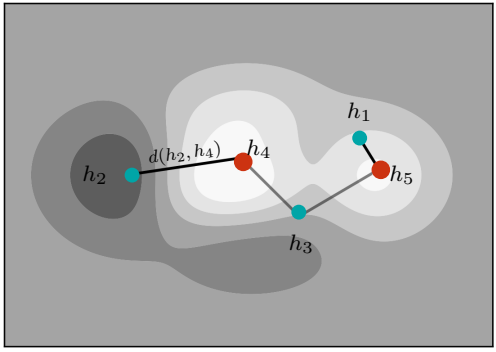
Problem: Having disjoint supports for ρ and π makes sense ...



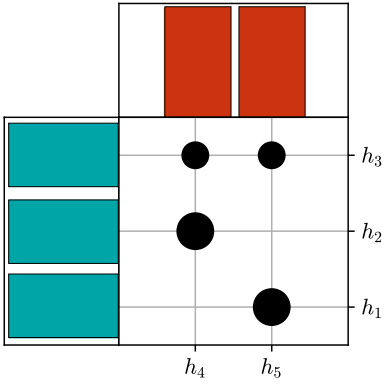
Prior distribution π Posterior distribution ρ

A solution

Using the **Wasserstein distance** to obtain PAC-Bayesian bounds (Amit *et al.*, 2022)



$$\mathrm{W}(\rho, \pi) := \inf_{\gamma \in \Gamma(\rho, \pi)} \left\{ \mathbb{E}_{(h,h') \sim \gamma} d(h, h') \right\}$$



Coupling γ of ρ and π

PAC-Bayesian bound of Amit *et al.* (2022)

For any $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, assume that for any $\delta' \in (0,1]$

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\left(\mathsf{R}^{\ell}_{\mathcal{D}}(h) - \widehat{\mathsf{R}}^{\ell}_{\mathcal{S}}(h)\right)^2 \text{ is } L_{\mathbb{S}}(\delta')\text{-Lipschitz } \textit{w.r.t.} \text{ the distance } d(\cdot,\cdot)\right] \geq 1-\delta'$$

for any prior π on $\mathbb{H}$, for any $\alpha > 1$, for any $\delta \in (0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall \rho \text{ on } \mathbb{H}, \quad \mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}^{\ell}_{\mathcal{D}}(h) \leq \mathop{\mathbb{E}}_{h\sim\rho} \widehat{\mathsf{R}}^{\ell}_{\mathbb{S}}(h) + \sqrt{L_{\mathbb{S}}\left(\tfrac{\delta}{2}\right) W(\rho,\pi) + \frac{\ln\frac{2m}{\delta}}{2(m-1)}}\right] \geq 1-\delta$$

with

- $\ell : \mathbb{H}\times(\mathbb{X}\times\mathbb{Y}) \to \mathbb{R}$ a loss function
- $\mathsf{R}^{\ell}_{\mathcal{D}}(h) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \ell(h,(\mathbf{x},y))$
- $\widehat{\mathsf{R}}^{\ell}_{\mathbb{S}}(h) = \dfrac{1}{m}\displaystyle\sum_{i=1}^{m} \ell(h,(\mathbf{x}_i,y_i))$

PAC-Bayesian bound of Amit *et al.* (2022)

For any $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, assume that for any $\delta' \in (0,1]$

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\left(\mathsf{R}^{\ell}_{\mathcal{D}}(h)-\widehat{\mathsf{R}}^{\ell}_{\mathbb{S}}(h)\right)^2 \text{ is } L_{\mathbb{S}}(\delta')\text{-Lipschitz } \textit{w.r.t.} \text{ the distance } d(\cdot,\cdot)\right] \geq 1-\delta'$$

for any prior π on $\mathbb{H}$, for any $\alpha > 1$, for any $\delta\in(0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall\rho \text{ on } \mathbb{H},\quad \mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^{\ell}_{\mathcal{D}}(h) \leq \mathop{\mathbb{E}}_{h\sim\rho}\widehat{\mathsf{R}}^{\ell}_{\mathbb{S}}(h) + \sqrt{L_{\mathbb{S}}\left(\tfrac{\delta}{2}\right)W(\rho,\pi) + \frac{\ln\frac{2m}{\delta}}{2(m-1)}}\right] \geq 1-\delta$$

Example: Bound for finite hypothesis sets

For any $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$,

for any loss $\ell : \mathbb{H}\times(\mathbb{X}\times\mathbb{Y}) \to [0,1]$ that is L-Lipschitz *w.r.t.* $d(\cdot,\cdot)$ for all $(x,y)\in\mathbb{X}\times\mathbb{Y}$,

for any prior π on $\mathbb{H}$, for any $\delta\in(0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall\rho \text{ on } \mathbb{H},\quad \mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^{\ell}_{\mathcal{D}}(h) \leq \mathop{\mathbb{E}}_{h\sim\rho}\widehat{\mathsf{R}}^{\ell}_{\mathbb{S}}(h) + \sqrt{\frac{8L\ln\frac{4|\mathbb{H}|}{\delta}}{m}W(\rho,\pi) + \frac{\ln\frac{2m}{\delta}}{2(m-1)}}\right] \geq 1-\delta$$

PAC-Bayesian bound of Amit *et al.* (2022)

For any $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, assume that for any $\delta' \in (0,1]$

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\left(\mathrm{R}^{\ell}_{\mathcal{D}}(h) - \widehat{\mathrm{R}}^{\ell}_{\mathbb{S}}(h)\right)^2 \text{ is } L_{\mathbb{S}}(\delta')\text{-Lipschitz } \textit{w.r.t.} \text{ the distance } d(\cdot,\cdot)\right] \geq 1-\delta'$$

for any prior π on $\mathbb{H}$, for any $\alpha > 1$, for any $\delta \in (0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall \rho \text{ on } \mathbb{H}, \quad \mathop{\mathbb{E}}_{h\sim\rho}\mathrm{R}^{\ell}_{\mathcal{D}}(h) \leq \mathop{\mathbb{E}}_{h\sim\rho}\widehat{\mathrm{R}}^{\ell}_{\mathbb{S}}(h) + \sqrt{L_{\mathbb{S}}\left(\tfrac{\delta}{2}\right)W(\rho,\pi) + \frac{\ln\frac{2m}{\delta}}{2(m-1)}}\right] \geq 1-\delta$$

Example: Uniform convergence bound for finite hypothesis sets

For any $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$,

for any loss $\ell : \mathbb{H}\times(\mathbb{X}\times\mathbb{Y}) \to [0,1]$ that is L-Lipschitz *w.r.t.* $d(\cdot,\cdot)$ for all $(x,y)\in\mathbb{X}\times\mathbb{Y}$,

for any prior π on $\mathbb{H}$, for any $\delta\in(0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall h \text{ on } \mathbb{H}, \quad \mathrm{R}^{\ell}_{\mathcal{D}}(h) \leq \widehat{\mathrm{R}}^{\ell}_{\mathbb{S}}(h) + \sqrt{\frac{8L\ln\frac{4|\mathbb{H}|}{\delta}}{m}\mathop{\mathbb{E}}_{h'\sim\pi}d(h,h') + \frac{\ln\frac{2m}{\delta}}{2(m-1)}}\right] \geq 1-\delta$$

PAC-Bayesian bound of Amit *et al.* (2022)

For any $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, assume that for any $\delta' \in (0,1]$

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\left(\mathsf{R}^{\ell}_{\mathcal{D}}(h)-\widehat{\mathsf{R}}^{\ell}_{\mathcal{S}}(h)\right)^2 \text{ is } L_{\mathbb{S}}(\delta')\text{-Lipschitz } \textit{w.r.t.} \text{ the distance } d(\cdot,\cdot)\right] \geq 1-\delta'$$

for any prior π on $\mathbb{H}$, for any $\alpha > 1$, for any $\delta \in (0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall \rho \text{ on } \mathbb{H}, \quad \mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}^{\ell}_{\mathcal{D}}(h) \leq \mathop{\mathbb{E}}_{h\sim\rho} \widehat{\mathsf{R}}^{\ell}_{\mathbb{S}}(h) + \sqrt{L_{\mathbb{S}}\left(\tfrac{\delta}{2}\right) W(\rho,\pi) + \frac{\ln\frac{2m}{\delta}}{2(m-1)}}\right] \geq 1-\delta$$

Advantage: The bound is both "PAC-Bayesian" and "Uniform Convergence"
Drawback: In the general case, the Wasserstein distance is not divided by m

We can have both a KL divergence and a Wasserstein distance!

PAC-Bayesian bound of Viallard *et al.* (2024c)

For any $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, assume that for any $\delta' \in (0,1]$

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\left(\mathrm{R}_{\mathcal{D}}^{\ell}(h) - \widehat{\mathrm{R}}_{\mathbb{S}}^{\ell}(h)\right)^2 \text{ is } L_{\mathbb{S}}(\delta')\text{-Lipschitz } \textit{w.r.t.} \text{ the distance } d(\cdot,\cdot)\right] \geq 1-\delta'$$

for any prior π si on $\mathbb{H}$, for any $\alpha > 1$, for any $\delta \in (0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall \rho,\eta \text{ on } \mathbb{H}, \mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}_{\mathcal{D}}^{\ell}(h) \leq \mathop{\mathbb{E}}_{h\sim\rho} \widehat{\mathrm{R}}_{\mathbb{S}}^{\ell}(h) + \sqrt{L_{\mathbb{S}}\left(\tfrac{\delta}{2}\right) W(\rho,\eta) + \frac{\mathrm{KL}(\eta,\pi) + \ln\frac{4\sqrt{m}}{\delta}}{m}}\right] \geq 1-\delta$$

with $\ell : \mathbb{H}\times(\mathbb{X}\times\mathbb{Y}) \to \mathbb{R}$ a loss function, $\mathrm{R}_{\mathcal{D}}^{\ell}(h) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\,\ell(h,(\mathbf{x},y))$ and $\widehat{\mathrm{R}}_{\mathbb{S}}^{\ell}(h) = \frac{1}{m}\sum_{i=1}^{m}\ell(h,(\mathbf{x}_i,y_i))$

PAC-Bayesian bound of Viallard *et al.* (2024c)

For any $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, assume that for any $\delta' \in (0,1]$

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\left(\mathrm{R}^{\ell}_{\mathcal{D}}(h) - \widehat{\mathrm{R}}^{\ell}_{\mathbb{S}}(h)\right)^2 \text{ is } L_{\mathbb{S}}(\delta')\text{-Lipschitz } \textit{w.r.t.} \text{ the distance } d(\cdot,\cdot)\right] \geq 1-\delta'$$

for any prior π si on $\mathbb{H}$, for any $\alpha > 1$, for any $\delta \in (0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall \rho,\eta \text{ on } \mathbb{H}, \mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}^{\ell}_{\mathcal{D}}(h) \leq \mathop{\mathbb{E}}_{h\sim\rho} \widehat{\mathrm{R}}^{\ell}_{\mathbb{S}}(h) + \sqrt{L_{\mathbb{S}}\left(\tfrac{\delta}{2}\right) W(\rho,\eta) + \frac{\mathrm{KL}(\eta,\pi) + \ln\frac{4\sqrt{m}}{\delta}}{m}}\right] \geq 1-\delta$$

Advantages:

- ✓ The impact of the Wasserstein distance can be reduced with the KL divergence
- ✓ It is not restricted to the KL divergence

We can have a bound on $L_{\mathbb{S}}(\delta')$ in the general case!

Estimation of $L_{\mathbb{S}}(\delta')$ for a general hypothesis set (Viallard *et al.*, 2024c)

For any hypothesis set $\mathbb{H}$, for any L-Lipschitz loss $\ell : \mathbb{H} \times \mathcal{Z} \to \mathbb{R}$ *w.r.t.* the distance $d(\cdot, \cdot)$, we have

$$\mathop{\mathbb{P}}_{\substack{\mathbb{S}\sim\mathcal{D}^m \\ \boldsymbol{\varepsilon}\sim\mathcal{E}^m}} \left[\begin{array}{l} h \mapsto \left| \mathsf{R}_{\mathcal{D}}^{\ell}(h) - \widehat{\mathsf{R}}_{\mathbb{S}}^{\ell}(h) \right| \text{ is } L_{\mathbb{S}}(\delta') = \left(2\mathscr{R}_{\mathcal{S}}^{\boldsymbol{\varepsilon}}(h) + 3L\sqrt{\frac{2\ln\frac{4}{\delta'}}{m}} \right)\text{-Lipschitz} \\ \text{where} \quad \mathscr{R}_{\mathcal{S}}^{\boldsymbol{\varepsilon}}(h) = \sup_{h \neq h' \in \mathbb{H}} \frac{1}{m} \sum_{i=1}^{m} \varepsilon_i \frac{[\ell(h', \mathbf{z}_i) - \ell(h, \mathbf{z}_i)]}{d(h, h')} \end{array} \right] \geq 1 - \delta'$$

Key properties of the PAC-Bayesian bounds

- **Flexible** and adaptable to various scenarios
- **Tight** and informative
- **Intermediate** between uniform-convergence and algorithm-dependent bounds

Bound in expectation over $\mathbb{H}$

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall\rho,\ \mathrm{D}\left(\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}_{\mathcal{D}}(h),\mathop{\mathbb{E}}_{h\sim\rho}\widehat{\mathsf{R}}_{\mathbb{S}}(h)\right)\ \leq\ \Phi(\Delta(\rho,\pi),\mathbb{S},\delta)\right]\geq 1-\delta$$

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m,\,h\sim\rho_{\mathbb{S}}}\left[\mathrm{D}\left(\mathsf{R}_{\mathcal{D}}(h)\ ,\ \widehat{\mathsf{R}}_{\mathbb{S}}(h)\right)\ \leq\ \Phi(\Delta(\rho_{\mathbb{S}},\pi),\mathbb{S},\delta)\right]\geq 1-\delta$$

Bound for only one $h\sim\rho_{\mathbb{S}}$

Questions ? :)

Part II
TO LEARNING PRACTICE

Doing learning theory

Not just to analyze

But to learn with and from confidence

BUT WHY ? 🤔

What do practitioners usually do?

A typical machine learning objective

$$\underset{\text{model}}{\text{minimize}} \left\{ \widehat{R}_{\mathbb{S}}(\text{model}) \; + \; \text{regularization}(\text{model}) \right\}$$

From observed data $\mathbb{S}$, practitioners learn a model

🤞 And they "**hope**" the model will generalize well to new data

🤔 *But…where is the confidence?*

After learning *(not always)*

Confidence is *sometimes* analyzed with generalization bounds

What do practitioners usually do?

A typical machine learning objective

$$\underset{\text{model}}{\text{minimize}} \left\{ \widehat{R}_{\mathbb{S}}(\text{model}) \; + \; \text{regularization}(\text{model}) \right\}$$

From observed data $\mathbb{S}$, practitioners learn a model

🤞 And they "**hope**" the model will generalize well to new data

🤔 *But…where is the confidence?*

After learning *(not always)*

Confidence is *sometimes* analyzed with generalization bounds

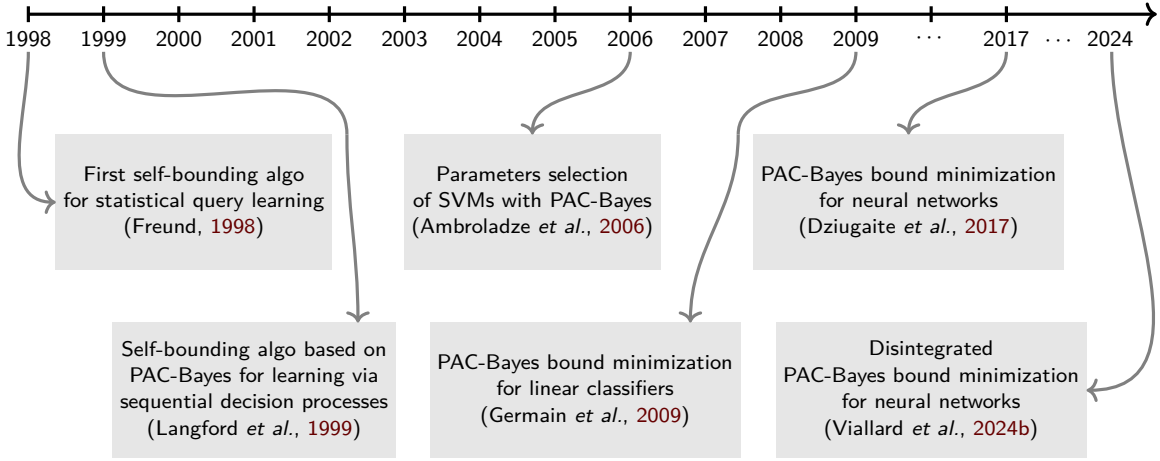
But the learning process ignores these guarantees!

A solution from statistical learning theory

Directly minimizing a generalization bound

$\Longrightarrow$ Self-certified / Self-bounding algorithm (Freund, 1998)

Key dates for PAC-Bayesian self-bounding algorithms



A core feature of PAC-Bayesian bounds

PAC-Bayesian guarantees are tighter when the **prior** is **well-aligned with the task**

↪ This is directly due to the divergence term $\mathrm{KL}(\rho\|\pi)$ (or any $\Delta(\rho\|\pi)$)

↪ **Smaller KL** $\Longrightarrow$ **Tighter bound** (for fixed empirical risk)

Algorithmic perspective

Learning becomes: refining a prior knowledge into a posterior distribution

A core feature of PAC-Bayesian bounds

PAC-Bayesian guarantees are tighter when the **prior** is **well-aligned with the task**

↪ This is directly due to the divergence term $\mathrm{KL}(\rho\|\pi)$ (or any $\Delta(\rho\|\pi)$)

↪ **Smaller KL** $\Longrightarrow$ **Tighter bound** (for fixed empirical risk)

Algorithmic perspective

Learning becomes: refining a prior knowledge into a posterior distribution

Can we learn a good *prior*? $\rightarrow$ YES!

- **Data-dependent priors** (learned on a portion of the data — 50/50 is a good solution)

(Ambroladze *et al.*, 2006; Germain *et al.*, 2009; Mhammedi *et al.*, 2019; Dziugaite *et al.*, 2021; Pérez-Ortiz *et al.*, 2021)

- **Distribution-dependent priors** (learned from $\mathcal{D}$ or a proxy)

(Lever *et al.*, 2013; Dziugaite *et al.*, 2018; Rivasplata *et al.*, 2018)

A core feature of PAC-Bayesian bounds

PAC-Bayesian guarantees are tighter when the **prior** is **well-aligned with the task**

↪ This is directly due to the divergence term $\mathrm{KL}(\rho\|\pi)$ (or any $\Delta(\rho\|\pi)$)

↪ **Smaller KL** $\Longrightarrow$ **Tighter bound** (for fixed empirical risk)

Algorithmic perspective

Learning becomes: refining a prior knowledge into a posterior distribution

Can we use constrained *prior*? $\rightarrow$ YES!

Example: symmetric prior on auto-complemented voters (*e.g.*, Roy *et al.*, 2011; Bellet *et al.*, 2014; Germain *et al.*, 2015)

If for every $h \in \mathbb{H}$, we also have $-h \in \mathbb{H}$, we can define a prior π that is symmetric over $\{h, -h\}$

If the posterior respects this symmetry, then the posterior and the prior are **aligned**, and we have

$\boxed{\mathrm{KL}(\rho\|\pi) \text{ disappears}} \quad \Rightarrow \quad$ **The bound gets tighter!**

A core feature of PAC-Bayesian bounds

PAC-Bayesian guarantees are tighter when the **prior** is **well-aligned with the task**

↪ This is directly due to the divergence term $\mathrm{KL}(\rho\|\pi)$ (or any $\Delta(\rho\|\pi)$)

↪ **Smaller KL** $\Longrightarrow$ **Tighter bound** (for fixed empirical risk)

Algorithmic perspective

Learning becomes: refining a prior knowledge into a posterior distribution

Take-home message

Designing smart priors = better guarantees

Not just a theoretical tool — but a key design element in self-certified learning

PAC-Bayesian generalization bound

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\ \forall \rho \text{ on } \mathbb{H},\ \mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}_{\mathcal{D}}(h) \ \leq\ \underbrace{\mathop{\mathbb{E}}_{h\sim\rho} \widehat{\mathsf{R}}_{\mathbb{S}}(h) \ +\ \Phi(\Delta(\rho,\pi),\mathbb{S},\delta)}\ \right] \geq 1-\delta$$

PAC-Bayesian-inspired learning algorithm

$$\mathop{\text{minimize}}_{\rho}\ \left\{\ \overbrace{\mathop{\mathbb{E}}_{h\sim\rho} \widehat{\mathsf{R}}_{\mathbb{S}}(h) \ +\ \Phi(\Delta(\rho,\pi),\mathbb{S},\delta)}\ \right\}$$

This leads to theoretically-founded strategies to learn a model defined by ρ

How to classify with a learned distribution ρ on $\mathbb{H}$?

How to classify with a learned distribution ρ on $\mathbb{H}$?

Stochastic

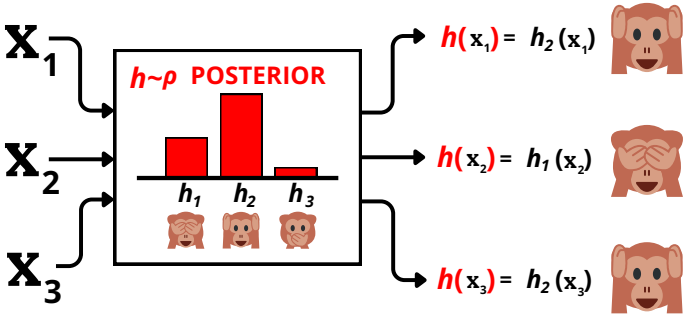
For each example $\mathbf{x}$

1. Randomly pick $h \sim \rho$
2. Predict $h(\mathbf{x})$

$\Rightarrow$

Using a classical PAC-Bayesian bound

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\Big[\forall \rho,\ \mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}_{\mathcal{D}}(h) \ \leq\ \mathop{\mathbb{E}}_{h\sim\rho} \widehat{\mathrm{R}}_{\mathbb{S}}(h) + \Phi(\Delta(\rho,\pi),\mathbb{S},\delta)\Big] \geq 1-\delta$$



How to classify with a learned distribution ρ on $\mathbb{H}$?

Stochastic

For each example $\mathbf{x}$

1. Randomly pick $h \sim \rho$
2. Predict $h(\mathbf{x})$

$\Rightarrow$ Using a classical PAC-Bayesian bound

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\Big[\forall \rho,\ \mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}_{\mathcal{D}}(h) \ \leq\ \mathop{\mathbb{E}}_{h\sim\rho} \widehat{\mathrm{R}}_{\mathbb{S}}(h) + \Phi(\Delta(\rho,\pi),\mathbb{S},\delta)\Big] \geq 1-\delta$$

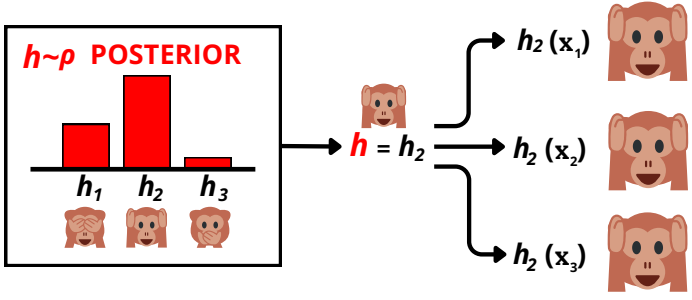
Deterministic

Pick $h \sim \rho$

Predict with $h(\cdot)$

$\Rightarrow$ Using a PAC-Bayesian disintegrated bound

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m,\ h\sim\rho}\Big[\ \mathrm{R}_{\mathcal{D}}(h) \ \leq\ \widehat{\mathrm{R}}_{\mathbb{S}}(h) \ +\ \Phi(\Delta(\rho,\pi),\mathbb{S},\delta)\Big] \geq 1-\delta$$



What kind of models do we learn with PAC-Bayes?

How to classify with a learned distribution ρ on $\mathbb{H}$?

Stochastic

For each example $\mathbf{x}$

1. Randomly pick $h \sim \rho$

2. Predict $h(\mathbf{x})$

$\Rightarrow$ Using a classical PAC-Bayesian bound

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\Big[\forall \rho,\ \mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}_{\mathcal{D}}(h) \ \le\ \mathop{\mathbb{E}}_{h\sim\rho} \widehat{\mathsf{R}}_{\mathbb{S}}(h) + \Phi(\Delta(\rho,\pi),\mathbb{S},\delta)\Big] \ge 1-\delta$$

Deterministic

Pick $h \sim \rho$

Predict with $h(\cdot)$

$\Rightarrow$ Using a PAC-Bayesian disintegrated bound

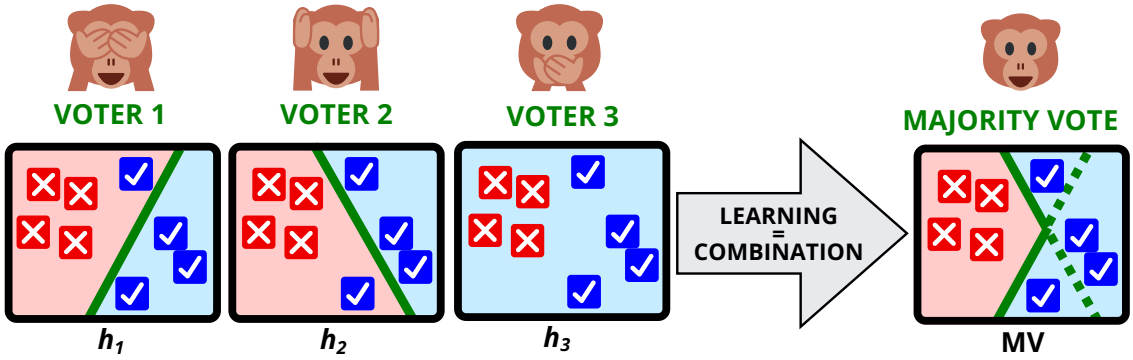
$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m,\ h\sim\rho}\Big[\ \mathsf{R}_{\mathcal{D}}(h) \ \le\ \widehat{\mathsf{R}}_{\mathbb{S}}(h) \ +\ \Phi(\Delta(\rho,\pi),\mathbb{S},\delta)\Big] \ge 1-\delta$$

Deterministic

Predict with $\mathop{\mathbb{E}}_{h\sim\rho} h(\cdot)$

$\Rightarrow$ Using a PAC-Bayesian bound on a surrogate of the risk of the ρ-weighted majority vote

Recall: The ρ-weighted majority vote



PAC-Bayesian supervised classification — Majority vote classifier

Goal: Find the ρ-weighted majority vote MV on $\mathbb{H}$ that minimizes the true risk

$$\underbrace{\mathrm{R}_{\mathcal{D}}(\mathrm{MV}) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathrm{I}[\,\mathrm{MV}(\mathbf{x}) \neq y\,]}_{\text{True risk}}$$

where $\mathrm{MV}(\cdot) = \mathrm{sign}\left[\sum_{h\in\mathbb{H}} \overbrace{\rho(h)}^{\text{weight of } h}\, h(\cdot)\right]$

$$= \mathrm{sign}\left[\mathop{\mathbb{E}}_{h\sim\rho} h(\cdot)\right]$$

The Majority Vote Case

Why the ρ-weighted majority vote is important in machine learning?

Majority vote combines multiple predictors/functions to make a final decision

$\hookrightarrow$ Common in ensemble methods

Classical examples: SVM, Bagging, Boosting, Random Forest, Neural Networks

Challenge

$R_{\mathcal{D}}(\mathrm{MV})$ is hard to bound and minimize directly

In PAC-Bayes

We can use a surrogate/upper bound on $R_{\mathcal{D}}(\mathrm{MV})$

Upper bound's minimization on the majority vote risk

The classical factor-2 bound

$$\mathrm{R}_{\mathcal{D}}(\mathrm{MV}) \;\leq\; 2 \times \mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}_{\mathcal{D}}(h)$$

$\hookrightarrow$ The factor 2 makes this bound quite loose $\rightsquigarrow$ even if we have tight bounds on $\mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}_{\mathcal{D}}(h)$

🤔 **BUT** there's nothing to stop you minimizing a bound !

The classical factor-2 bound

$$\mathrm{R}_{\mathcal{D}}(\mathrm{MV}) \;\leq\; 2 \times \mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}_{\mathcal{D}}(h)$$

$\hookrightarrow$ The factor 2 makes this bound quite loose $\rightsquigarrow$ even if we have tight bounds on $\mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}_{\mathcal{D}}(h)$

🤔 **BUT** there's nothing to stop you minimizing a bound !

Let us consider the Catoni's bound on $\mathrm{R}_{\mathcal{D}}(\mathrm{MV})$

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m}\left[\; \begin{array}{l} \forall \rho \text{ on } \mathbb{H}, \\ \mathrm{R}_{\mathcal{D}}(\mathrm{MV}) \end{array} \leq \frac{2}{1-e^{-c}}\left(c \mathop{\mathbb{E}}_{h\sim\rho} \widehat{\mathrm{R}}_{\mathcal{S}}(h) + \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi) + \frac{1}{\delta}\right]\right) \;\right] \geq 1-\delta$$

Why Catoni's bound ? $\rightarrow$ Because of the parameter c

- it controls the trade-off between $\mathbb{E}_{h\sim\rho} \widehat{\mathrm{R}}_{\mathcal{S}}(h)$ and $\mathrm{KL}(\rho\|\pi)$
- it can be tuned
- or there exists a closed-formed solution for c

An example — The algorithm PBGD in a nutshell (Germain *et al.*, 2009, 2020)

- Specialization to a set of linear classifiers $\mathbb{H} = \left\{ h_{\mathbf{v}} = \text{sign}\left[\mathbf{v}\cdot\mathbf{x}\right] \;\middle|\; \mathbf{v}\in\mathbb{R}^d \right\}$ (Langford, 2005)

$\Rightarrow$ then $\text{MV}(\mathbf{x}) = \text{sign}\left[\mathop{\mathbb{E}}_{\mathbf{v}\sim\rho_{\mathbf{w}}} \text{sign}\left[\mathbf{v}\cdot\mathbf{x}\right]\right] = \text{sign}\left[\mathbf{w}\cdot\mathbf{x}\right] = h_{\mathbf{w}}(\mathbf{x})$

Prior $\pi_{\mathbf{0}}$: isotropic Gaussian centered on $\mathbf{0}$
Posterior $\rho_{\mathbf{w}}$: isotropic Gaussian centered on $\mathbf{w}$

$\Rightarrow\ \text{KL}(\rho_{\mathbf{w}}\|\pi_{\mathbf{0}}) = \frac{1}{2}\|\mathbf{w}\|^2$

Given $\mathbb{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ and a hyper-parameter $C > 0$, we minimize the Catoni's bound with

$$\underset{\mathbf{w}}{\text{argmin}}\ C \underbrace{\sum_{i=1}^m \Phi\left(y_i \frac{\mathbf{w}\cdot\mathbf{x}_i}{\|\mathbf{x}_i\|}\right)}_{\mathbb{E}_{\mathbf{v}\sim\rho_{\mathbf{w}}} \widehat{\text{R}}_{\mathbb{S}}(h_{\mathbf{v}})} + \underbrace{\frac{1}{2}\|\mathbf{w}\|^2}_{\text{KL}(\rho_{\mathbf{w}}\|\pi_{\mathbf{0}})}$$

where $\phi(a) = \frac{1}{2}\left[1 - \text{Erf}\left(\frac{a}{\sqrt{2}}\right)\right]$, with Erf() the Gauss error function (a smooth surrogate of the $0-1$ loss)

Remarks: Can be used with a kernel function, or a data-dependent prior

The classical factor-2 bound — A key decomposition

$$\mathrm{R}_{\mathcal{D}}(\mathrm{MV}) \;\leq\; 2 \times \mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}_{\mathcal{D}}(h) \;=\; 2 \times \Big[\mathrm{e}_{\mathcal{D}}(\rho) + \tfrac{1}{2}\mathrm{d}_{\mathcal{D}_{\mathcal{X}}}(\rho)\Big]$$

where $\mathrm{d}_{\mathcal{D}_{\mathcal{X}}}(\rho) = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}} \mathop{\mathbb{E}}_{(h,h')\sim\rho^2} \mathrm{I}[\, h(\mathbf{x}) \neq h'(\mathbf{x}) \,]$ is the disagreement

↪ captures the **diversity** between the voters ⇝ important in ensemble methods

↪ if all the voters agree, we do not need to combine them 🗳️

and $\mathrm{e}_{\mathcal{D}}(\rho) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathop{\mathbb{E}}_{(h,h')\sim\rho^2} \mathrm{I}[\, h(\mathbf{x}) \neq y \,\wedge\, h'(\mathbf{x}) \neq y \,]$ is the joint error

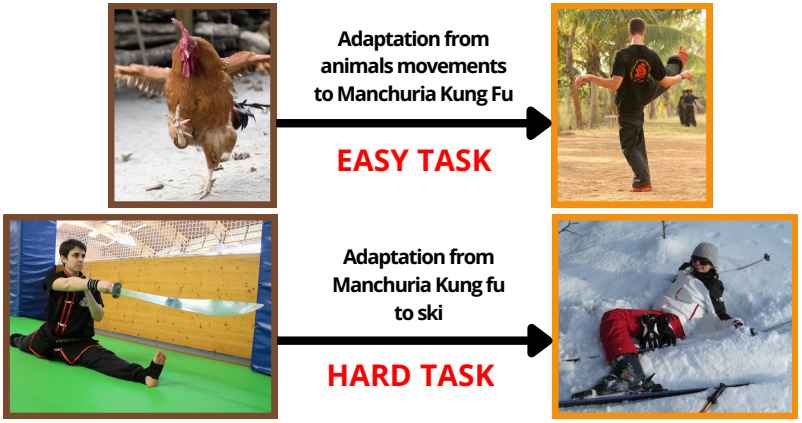
↪ we want to combine diverse voters that are **not too bad** !

↪ $\mathrm{e}_{\mathcal{D}}(\rho)$ captures how often voters **err together** ❌

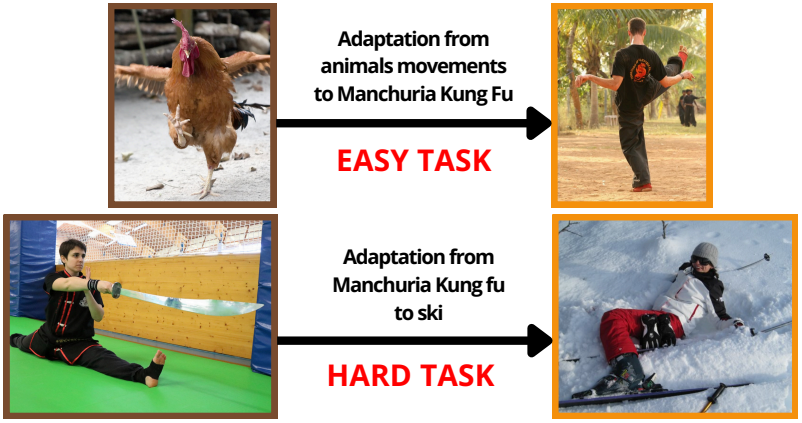
⚠️ Limitation

As itself, $\mathrm{e}_{\mathcal{D}}(\rho) + \frac{1}{2}\mathrm{d}_{\mathcal{D}_{\mathcal{X}}}(\rho)$ is limited since a 0 error MV implies $\mathrm{d}_{\mathcal{D}_{\mathcal{X}}}(\rho) = \mathrm{e}_{\mathcal{D}}(\rho) = 0$

BUT can be useful for specific settings such as for domain adaptation



Humans can adapt to a new task by reusing previously acquired knowledge



In **machine learning**, we need a measure of adaptation difficulty between tasks

Upper bound's minimization on the majority vote risk — *in domain adaptation*

Adaptation from animals movements to Manchuria Kung Fu

EASY TASK

Adaptation from Manchuria Kung fu to ski

HARD TASK

Ingredients in (unsupervised) domain adaptation

Two data distributions on $\mathbb{X}\times\mathbb{Y}$
$\mathcal{S}$ source domain and $\mathcal{T}$ target domain

Two learning samples
$\mathbb{S}\sim\mathcal{S}$ labeled and $\mathbb{T}\sim\mathcal{T}_{\mathbb{X}}$ unlabeled

$$\beta = \sup_{(\mathbf{x},y)\in\text{SUPP}(\mathcal{S})} \frac{\mathcal{T}(\mathbf{x},y)}{\mathcal{S}(\mathbf{x},y)} \text{ : Difficulty between domains}$$

The classical factor-2 bound — A key decomposition for DA (Germain *et al.*, 2016b, 2020)

$$\mathrm{R}_{\mathcal{T}}(\mathrm{MV}) \;\leq\; 2 \times \Big[\beta\, \mathrm{e}_{\mathcal{S}}(\rho) + \tfrac{1}{2}\mathrm{d}_{\mathcal{T}_{\mathbb{X}}}(\rho) + \eta\Big]$$

with $\eta = \underset{(\mathbf{x},y)\sim\mathcal{T}}{\mathbb{P}}((\mathbf{x},y) \notin \mathrm{SUPP}(\mathcal{S}))\ \sup_{h\in\mathbb{H}} \mathrm{R}_{\mathcal{T}\setminus\mathcal{S}}(h)$ the worst risk on $\mathcal{T}$ for which $\mathcal{S}$ is not informative

↪ η is not computable in practice and is **assumed to be low**

i.e. the domains are sufficiently similar *(a very strong but common assumption in DA)*

↪ β is not computable but **can be tuned**

to control the trade-off between source and target information

Ingredients in (unsupervised) domain adaptation

Two data distributions on $\mathbb{X}\times\mathbb{Y}$

$\mathcal{S}$ source domain and $\mathcal{T}$ target domain

Two learning samples

$\mathbb{S}\sim\mathcal{S}$ labeled and $\mathbb{T}\sim\mathcal{T}_{\mathbb{X}}$ unlabeled

$$\beta = \sup_{(\mathbf{x},y)\in\mathrm{SUPP}(\mathcal{S})} \frac{\mathcal{T}(\mathbf{x},y)}{\mathcal{S}(\mathbf{x},y)} : \text{Difficulty between domains}$$

The classical factor-2 bound — A key decomposition for DA (Germain *et al.*, 2016b, 2020)

$$\mathrm{R}_{\mathcal{T}}(\mathrm{MV}) \;\leq\; 2 \times \Big[\beta\, \mathrm{e}_{\mathcal{S}}(\rho) + \tfrac{1}{2}\mathrm{d}_{\mathcal{T}_{\mathbb{X}}}(\rho) + \eta\Big]$$

A Catoni's style PAC-Bayesian bound for DA

$$\underset{\substack{\mathbb{S}\sim(\mathcal{S})^{m_s}\\ \mathbb{T}\sim(\mathcal{T}_{\mathbb{X}})^{m_t}}}{\mathbb{P}}\left[\begin{array}{l}\forall \rho \text{ on } \mathbb{H},\\ \mathrm{R}_{\mathcal{T}}(\mathrm{MV}) \;\leq\; 2\,\beta\,\widehat{\mathrm{e}}_{\mathbb{S}}(\rho) \;+\; c\,\widehat{\mathrm{d}}_{\mathbb{T}}(\rho) \;+\; 2O\!\left(\dfrac{2\mathrm{KL}(\rho\|\pi)}{\beta m_s}+\dfrac{2\mathrm{KL}(\rho\|\pi)}{c m_t}\right)+2\,\eta\end{array}\right] \geq 1-\delta$$

DALC — A PBGD like algorithm for DA with Linear Classifiers

Given $\mathbb{S}=\{(\mathbf{x}_s,y_s)\}_{s=1}^{m_s}$, $\mathbb{T}=\{\mathbf{x}_t\}_{t=1}^{m_t}$, $C>0$, $B>0$, we minimize the computables terms

$$\underset{\mathbf{w}}{\operatorname{argmin}}\; B \underbrace{\sum_{s=1}^{m_s}\Phi_e\left(y_s\frac{\mathbf{w}\cdot\mathbf{x}_s}{\|\mathbf{x}_s\|}\right)}_{\widehat{\mathrm{e}}_{\mathbb{S}}(h_{\mathbf{w}})} + C \underbrace{\sum_{t=1}^{m_t}\Phi_d\left(\frac{\mathbf{w}\cdot\mathbf{x}_t}{\|\mathbf{x}_t\|}\right)}_{\widehat{\mathrm{d}}_{\mathbb{T}}(h_{\mathbf{w}})} + \underbrace{\|\mathbf{w}\|^2}_{2\mathrm{KL}(\rho_{\mathbf{w}}\|\pi_{\mathbf{0}})}$$

with $\phi_e(a)=[\phi(a)]^2$ and $\phi_d(a)=2\phi(a)\phi(-a)$ (Can be used with a kernel function, or a data-dependent prior)

Upper bound's minimization on the majority vote risk

The classical factor-2 bound

$$\mathrm{R}_{\mathcal{D}}(\mathrm{MV}) \;\leq\; 2 \times \mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}_{\mathcal{D}}(h) = 2 \times \left[\mathrm{e}_{\mathcal{D}}(\rho) + \tfrac{1}{2}\mathrm{d}_{\mathcal{D}_\mathcal{X}}(\rho)\right]$$

The C-bound (Breiman, 2001; Lacasse *et al.*, 2006) *comes from Cantelli-Chebitchev inequality*

$$\text{If } \mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}_{\mathcal{D}}(h) \leq \frac{1}{2}, \text{ then } \mathrm{R}_{\mathcal{D}}(\mathrm{MV}) \leq 1 - \frac{\left(1 - 2 \mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}_{\mathcal{D}}(h)\right)^2}{1 - 2\,\mathrm{d}_{\mathcal{D}_\mathcal{X}}(\rho)} = C_{\mathcal{D}}(\rho)$$

⇝ It is often significantly **tighter than the factor-2 bound**

⚖️ It better captures the key trade-off in ensemble methods

- $\mathbb{E}_{h\sim\rho}\,\mathrm{R}_{\mathcal{D}}(h)$: the average risk of voters
- $\mathrm{d}_{\mathcal{D}_\mathcal{X}}(\rho)$: the disagreement/diversity between voters

🤔 **Intuition**: A performing majority vote needs both accurate and diverse voters

The classical factor-2 bound

$$\mathrm{R}_{\mathcal{D}}(\mathrm{MV}) \;\leq\; 2 \times \mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}_{\mathcal{D}}(h) = 2 \times \left[\mathrm{e}_{\mathcal{D}}(\rho) + \tfrac{1}{2}\mathrm{d}_{\mathcal{D}_{\mathcal{X}}}(\rho)\right]$$

The C-bound (Breiman, 2001; Lacasse *et al.*, 2006) *comes from Cantelli-Chebitchev inequality*

$$\text{If } \mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}_{\mathcal{D}}(h) \leq \frac{1}{2}, \text{ then } \mathrm{R}_{\mathcal{D}}(\mathrm{MV}) \leq 1 - \frac{\left(1 - 2 \mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}_{\mathcal{D}}(h)\right)^2}{1 - 2\,\mathrm{d}_{\mathcal{D}_{\mathcal{X}}}(\rho)} = C_{\mathcal{D}}(\rho)$$

The joint error (Masegosa *et al.*, 2020) *comes from 2nd order Markov's inequality* $\mathbb{P}(X \geq \delta) \leq \frac{1}{\delta^2}\,\mathbb{E}[X^2]$

$$\mathrm{R}_{\mathcal{D}}(\mathrm{MV}) \leq 4 \times \mathrm{e}_{\mathcal{D}}(\rho)$$

if $\mathbb{E}_{h\sim\rho}\,\mathrm{R}_{\mathcal{D}}(h) < \frac{1}{2}$, we have

(i) $\mathrm{R}_{\mathcal{D}}(\mathrm{MV}) \;\leq\; C_{\mathcal{D}}(\rho) \;\leq\; 4\,\mathrm{e}_{\mathcal{D}}(\rho) \;\leq\; 2 \mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}_{\mathcal{D}}(h)$, if $\mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}_{\mathcal{D}}(h) \leq \mathrm{d}_{\mathcal{D}_{\mathcal{X}}}(\rho)$

(ii) $\mathrm{R}_{\mathcal{D}}(\mathrm{MV}) \;\leq\; 2 \mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}_{\mathcal{D}}(h) \;\leq\; C_{\mathcal{D}}(\rho) \;\leq\; 4\,\mathrm{e}_{\mathcal{D}}(\rho)$, otherwise

Upper bound's minimization on the majority vote risk

The classical factor-2 bound

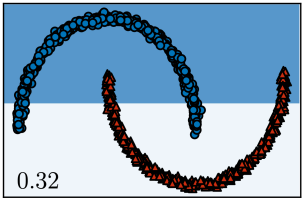
$$\mathrm{R}_{\mathcal{D}}(\mathrm{MV}) \;\leq\; 2 \times \mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}_{\mathcal{D}}(h) = 2 \times \left[\mathrm{e}_{\mathcal{D}}(\rho) + \tfrac{1}{2}\mathrm{d}_{\mathcal{D}_X}(\rho)\right]$$

The C-bound (Breiman, 2001; Lacasse *et al.*, 2006) *comes from Cantelli-Chebitchev inequality*

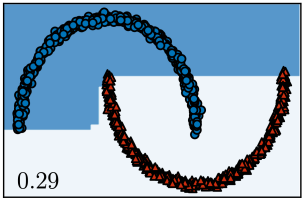
$$\text{If } \mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}_{\mathcal{D}}(h) \leq \frac{1}{2}, \text{ then } \mathrm{R}_{\mathcal{D}}(\mathrm{MV}) \leq 1 - \frac{\left(1 - 2 \mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}_{\mathcal{D}}(h)\right)^2}{1 - 2\,\mathrm{d}_{\mathcal{D}_X}(\rho)} = C_{\mathcal{D}}(\rho)$$

The joint error (Masegosa *et al.*, 2020) *comes from 2nd order Markov's inequality* $\mathbb{P}(X \geq \delta) \leq \frac{1}{\delta^2}\,\mathbb{E}[X^2]$

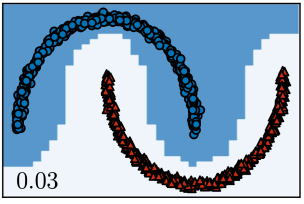
$$\mathrm{R}_{\mathcal{D}}(\mathrm{MV}) \leq 4 \times \mathrm{e}_{\mathcal{D}}(\rho)$$



Minimization of $2\,\mathbb{E}_{h\sim\rho}\,\widehat{\mathrm{R}}_{\mathbb{S}}(h)$



Minimization of $4\,\widehat{\mathrm{e}}_{\mathbb{S}}(\rho)$



Minimization of $\widehat{C}_{\mathbb{S}}(\rho)$

From a first PAC-Bayesian bound for the C-bound (Roy *et al.*, 2016)...

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m}\left[\forall\rho \text{ on } \mathcal{H},\ \mathrm{R}_{\mathcal{D}}(\mathrm{MV}) \le 1 - \frac{\left[1-2\,\min\left\{\frac{1}{2},\ \mathop{\mathbb{E}}_{h\sim\rho}\widehat{\mathrm{R}}_{\mathcal{S}}(h) + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{4\sqrt{m}}{\delta}\right]}\right\}\right]^2}{1-2\,\max\left\{0,\ \widehat{\mathrm{d}}_{\mathcal{S}}(\rho) - \sqrt{\frac{1}{2m}\left[2\,\mathrm{KL}(\rho\|\pi) + \ln\frac{4\sqrt{m}}{\delta}\right]}\right\}}\right] \ge 1-\delta$$

- Numerator = minimizing the PAC-Bayes bound on individual risks of the voters
  under the constraint that $\mathop{\mathbb{E}}_{h\sim\rho}\widehat{\mathrm{R}}_{\mathcal{S}}(h) + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{4\sqrt{m}}{\delta}\right]} \le \frac{1}{2}$
- Denominator = while maximizing PAC-Bayes bound on the diversity between voters

For pedagogical purposes we present here the self-bounding algorithm based on the PAC-Bayesian bound of Roy *et al.* (2016), but there are others versions that lead to slightly different algorithms (see Viallard *et al.* (2021a) for details)

...To a first algorithm (Viallard *et al.*, 2021a)

$$\min_{\rho} \overbrace{\left\{ 1 - \frac{\left[1 - 2\min\left\{\frac{1}{2}, \mathbb{E}_{h\sim\rho}\,\widehat{\mathsf{R}}_{\mathcal{S}}(h) + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{4\sqrt{m}}{\delta}\right]}\right\}\right]^2}{1 - 2\max\left\{0, \widehat{\mathrm{d}}_{\mathcal{S}}(\rho) - \sqrt{\frac{1}{2m}\left[2\,\mathrm{KL}(\rho\|\pi) + \ln\frac{4\sqrt{m}}{\delta}\right]}\right\}} \right\}}^{\widehat{C}^{\mathrm{PB}}_{\mathcal{S}}(\rho)}$$

such that $\quad \mathop{\mathbb{E}}_{h\sim\rho} \widehat{\mathsf{R}}_{\mathcal{S}}(h) + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{4\sqrt{m}}{\delta}\right]} \leq \frac{1}{2}$

- $\widehat{C}^{\mathrm{PB}}_{\mathcal{S}}(\rho) =$ Objective
  $\hookrightarrow$ PAC-Bayesian upper bound on the risk of the majority vote
- Constraint ensures non 0 gradient (numerator term $< \frac{1}{2}$)

...To a first algorithm (Viallard *et al.*, 2021a)

$$\min_{\rho} \overbrace{\left\{ 1 - \frac{\left[1 - 2\min\left\{\frac{1}{2}, \mathbb{E}_{h\sim\rho}\,\widehat{\mathsf{R}}_{\mathbb{S}}(h) + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{4\sqrt{m}}{\delta}\right]}\right\}\right]^2}{1 - 2\max\left\{0, \widehat{\mathrm{d}}_{\mathbb{S}}(\rho) - \sqrt{\frac{1}{2m}\left[2\,\mathrm{KL}(\rho\|\pi) + \ln\frac{4\sqrt{m}}{\delta}\right]}\right\}} \right\}}^{\widehat{C}^{\mathrm{PB}}_{\mathbb{S}}(\rho)}$$

such that $\mathop{\mathbb{E}}_{h\sim\rho} \widehat{\mathsf{R}}_{\mathbb{S}}(h) + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{4\sqrt{m}}{\delta}\right]} \leq \frac{1}{2}$

Constrained problem → Unconstrained problem

$$\min_{\rho} \left\{ \widehat{C}^{\mathrm{PB}}_{\mathbb{S}}(\rho) + \mathbf{B}_{\lambda}\left(\mathop{\mathbb{E}}_{h\sim\rho} \widehat{\mathsf{R}}_{\mathbb{S}}(h) + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{4\sqrt{m}}{\delta}\right]} - \frac{1}{2} \right) \right\}$$

(can be minimized with mini-batch)

with $\mathbf{B}_{\lambda}(a) = \begin{cases} -\frac{1}{\lambda}\ln(-a) & \text{if } a \leq -\frac{1}{\lambda^2} \\ \lambda a - \frac{1}{\lambda}\ln(\frac{1}{\lambda^2}) + \frac{1}{\lambda} & \text{otherwise} \end{cases}$ is a log-barrier function (Kervadec *et al.*, 2019)

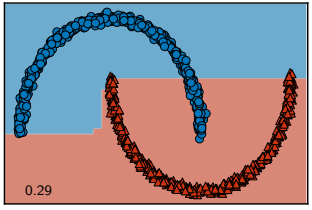
Drawbacks of the surrogates of the majority vote

- PAC-Bayesian generalization bounds are not necessarily tight
- Self-bounding algorithms do not take full advantage of the combination of voters

Example

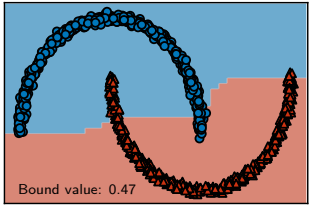
1. Surrogates are not necessarily precise for minimizing $\widehat{\mathrm{R}}_{\mathcal{S}}(\mathrm{MV})$

$$\min_{\rho} \left\{ 4\, \widehat{\mathrm{e}}_{\mathcal{S}}(\rho) \right\}$$



2. Even less precise when minimizing a generalization bound

$$\min_{\rho} \left\{ 4\, \widehat{\mathrm{e}}_{\mathcal{S}}(\rho) + \sqrt{\frac{8}{m}\left[\mathrm{KL}(\rho \| \pi) + \ln \frac{2\sqrt{m}}{\delta} \right]} \right\}$$



Upper bound's minimization on the majority vote risk—A drawback

Drawbacks of the surrogates of the majority vote

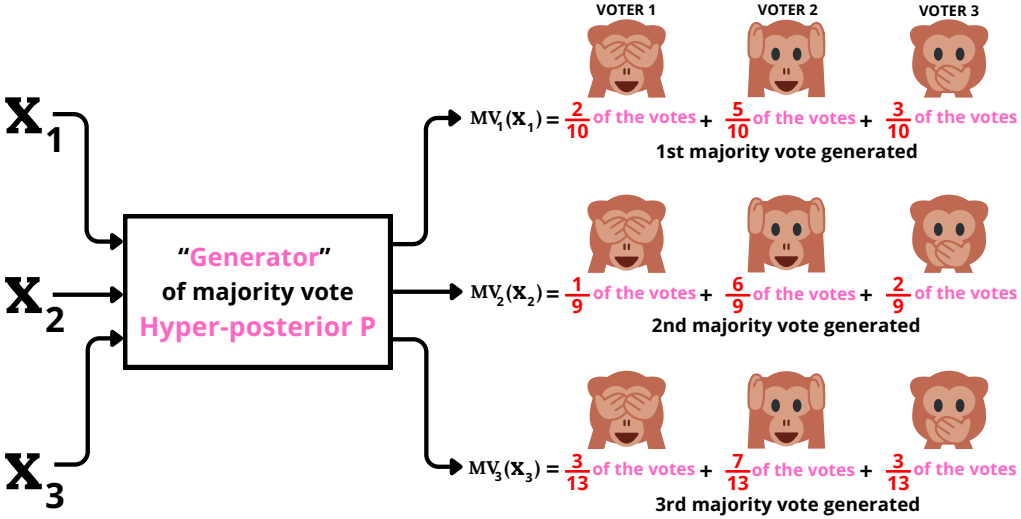
- PAC-Bayesian generalization bounds are not necessarily tight
- Self-bounding algorithms do not take full advantage of the combination of voters

One solution

The **Stochastic** Majority Vote (Zantedeschi *et al.*, 2021)

- Randomness for each input
- There is a closed-form solution for the empirical risk
- One can derive a self-bounding algorithm

The stochastic majority vote



Objective

Find a *hyper-posterior* P minimizing $\mathop{\mathbb{E}}_{\rho\sim\mathrm{P}} \mathrm{R}_{\mathcal{D}}(\mathrm{MV}_\rho)$

PAC-Bayesian bound for the stochastic majority vote

For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any finite hypothesis set $\mathbb{H}$, for any hyper-prior distribution Π, for any $\delta \in (0,1]$,

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\begin{array}{l}\forall\, \mathrm{P} \text{ on } \mathbb{H},\\ \mathop{\mathbb{E}}_{\rho\sim\mathrm{P}} \mathrm{R}_{\mathcal{D}}(\mathrm{MV}_\rho) \le \mathop{\mathbb{E}}_{\rho\sim\mathrm{P}} \widehat{\mathrm{R}}_{\mathbb{S}}(\mathrm{MV}_\rho) + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(\mathrm{P}\|\Pi) + \ln\frac{2\sqrt{m}}{\delta}\right]}\end{array}\right] \ge 1-\delta$$

Self-bounding algorithm

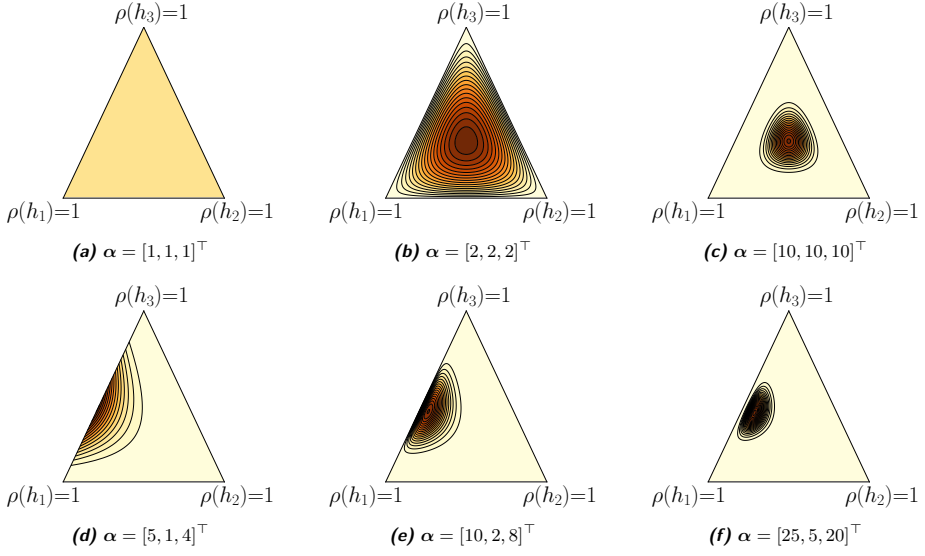
$$\min_{\mathrm{P}}\left\{\mathop{\mathbb{E}}_{\rho\sim\mathrm{P}} \widehat{\mathrm{R}}_{\mathbb{S}}(\mathrm{MV}_\rho) + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(\mathrm{P}\|\Pi) + \ln\frac{2\sqrt{m}}{\delta}\right]}\right\}$$

Issue

The right-hand side of the inequality is not computable (in general)

$\Rightarrow$ Special case of hyper-priors Π and hyper-posteriors P with a computable bound

Hyper-posterior P = Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\alpha})$



(a) $\boldsymbol{\alpha} = [1, 1, 1]^\top$ ***(b)*** $\boldsymbol{\alpha} = [2, 2, 2]^\top$ ***(c)*** $\boldsymbol{\alpha} = [10, 10, 10]^\top$

(d) $\boldsymbol{\alpha} = [5, 1, 4]^\top$ ***(e)*** $\boldsymbol{\alpha} = [10, 2, 8]^\top$ ***(f)*** $\boldsymbol{\alpha} = [25, 5, 20]^\top$

For each example $(\mathbf{x}_i, y_i)$, the parameters $\boldsymbol{\alpha}$ are separated into two subsets:

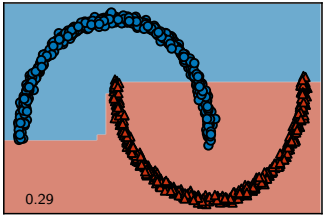
- $\mathbb{T}(\mathbf{x}_i, y_i) =$ Indices of the Dirichlet parameters of the voters that **correctly classify** $(\mathbf{x}_i, y_i)$
- $\mathbb{F}(\mathbf{x}_i, y_i) =$ Indices of the Dirichlet parameters of the voters that **misclassify** $(\mathbf{x}_i, y_i)$

Closed-form solution of the empirical risk (in binary classification)

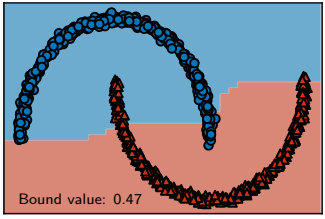
$$\mathop{\mathbb{E}}_{\rho\sim\mathrm{P}} \widehat{\mathsf{R}}_{\mathbb{S}}(\mathsf{MV}_\rho) = \frac{1}{m}\sum_{i=1}^{m} I_{0.5}\left(\sum_{j\in\mathbb{T}(\mathbf{x}_i,y_i)} \alpha_j, \sum_{j\in\mathbb{F}(\mathbf{x}_i,y_i)} \alpha_j\right)$$

with $I_{0.5}()$ the regularized incomplete beta function evaluated at 0.5

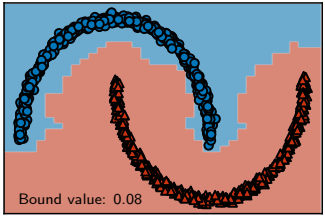
Advantage and limitations of the stochastic majority vote



(a) **Majority Vote**
Risk minimization



(b) **Majority Vote**
Self-bounding algorithm



(c) **Stochastic Majority Vote**
Self-bounding algorithm

Advantages

+ The PAC-Bayesian generalization bounds obtained are tight
+ A self-bounding algorithm

Limitations

− The majority vote is stochastic

⇒ The PAC-Bayesian bound does not consider a unique majority vote

The Neural Network Case

Objective function

Consider a PAC-Bayesian bound (e.g., McAllester's bound)

$$\underset{\rho}{\text{minimize}} \left\{ \underset{h\sim\rho}{\mathbb{E}} \widehat{\mathsf{R}}_{\mathcal{S}}(h) \; + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{2\sqrt{m}}{\delta}\right]} \right\}$$

Prior and posterior distribution

(see e.g., Langford *et al.*, 2001a; Dziugaite *et al.*, 2017; Pérez-Ortiz *et al.*, 2021; Viallard *et al.*, 2024b)

- $\mathbf{w}_\pi \in \mathbb{R}^d$ is the prior neural network's weights
- Prior $\pi = \mathcal{N}(\mathbf{w}_\pi, \sigma^2 I_d)$ (where σ^2 is the variance parameter)
- $\mathbf{w}$ is the neural network's weights
- Posterior $\rho_{\mathbf{w}} = \mathcal{N}(\mathbf{w}, \sigma^2 I_d)$

Prior and posterior distribution

(see e.g., Langford *et al.*, 2001a; Dziugaite *et al.*, 2017; Pérez-Ortiz *et al.*, 2021; Viallard *et al.*, 2024b)

- $\mathbf{w}_\pi \in \mathbb{R}^d$ is the prior neural network's weights
- Prior $\pi = \mathcal{N}(\mathbf{w}_\pi, \sigma^2 I_d)$ (where σ^2 is the variance parameter)
- $\mathbf{w}$ is the neural network's weights
- Posterior $\rho_{\mathbf{w}} = \mathcal{N}(\mathbf{w}, \sigma^2 I_d)$

Objective function

Consider a PAC-Bayesian bound (*e.g.*, McAllester's bound)

$$\underset{\mathbf{w}}{\text{minimize}} \left\{ \underset{h\sim\rho_{\mathbf{w}}}{\mathbb{E}} \widehat{\mathsf{R}}_{\mathbb{S}}(h) \; + \sqrt{\frac{1}{2m}\left[\frac{\|\mathbf{w}-\mathbf{w}_\pi\|_2^2}{2\sigma^2} + \ln\frac{2\sqrt{m}}{\delta}\right]} \right\}$$

How to minimize such an objective function? How to evaluate the final bound?

Problem 1: Unlike for PBGD we have no closed-form for the risk $\mathbb{E}_{h\sim\rho_{\mathbf{w}}}\widehat{\mathsf{R}}_{\mathbb{S}}(h)$

Solution: Learning algorithm with estimation of the risk $\mathbb{E}_{h\sim\rho_{\mathbf{w}}}\widehat{\mathsf{R}}_{\mathbb{S}}(h)$

Input: Datasets $\mathbb{S}_\pi$ and $\mathbb{S}_\rho$
Learn a prior π with the data $\mathbb{S}_\pi$
Initialize $\rho_{\mathbf{w}} = \mathcal{N}(\mathbf{w}, \sigma^2 I_d)$
while not converged **do**
  Sample a mini-batch U from $\mathbb{S}_\rho$
  Sample $h_{\mathbf{w}} \sim \rho_{\mathbf{w}}$
  Compute $R_U(h_{\mathbf{w}})$ and $\mathrm{KL}(\rho_{\mathbf{w}}\|\pi) = \frac{1}{2\sigma^2}\|\mathbf{w}-\mathbf{w}_\pi\|_2^2$
  Compute the gradient $\nabla_{\mathbf{w}}\left(R_U(h_{\mathbf{w}}) + \sqrt{\frac{1}{m}\left[\frac{1}{2\sigma^2}\|\mathbf{w}-\mathbf{w}_\pi\|_2^2 + \ln\frac{2\sqrt{m}}{\delta}\right]}\right)$
  Make a gradient step with, *e.g.*, SGD to update $\rho_{\mathbf{w}}$
end while
return posterior $\rho_{\mathbf{w}}$

Problem 2: To evaluate the PAC-Bayes bound, we need to estimate $\mathbb{E}_{h\sim\rho_{\mathbf{w}}} \widehat{\mathsf{R}}_{\mathbb{S}}(h)$ again

Solution: Sample some neural networks (by Monte Carlo)

for t=1…T **do**
  Sample a neural network $h_t \sim \rho_{\mathbf{w}}$
  Compute the empirical risk $\widehat{\mathsf{R}}_{\mathbb{S}}(h_t)$
end for
return $\frac{1}{T}\sum_{t=1}^{T} \widehat{\mathsf{R}}_{\mathbb{S}}(h_t)$

We have to integrate the Monte Carlo sampling into the bound!

Problem 3: We need a bound to take into account the estimation $\frac{1}{T}\sum_{t=1}^{T}\widehat{\mathsf{R}}_{\mathbb{S}}(h_t)$

Solution: Hoeffding's inequality (or the sample convergence bound of Langford *et al.* (2001a))

$$\mathop{\mathbb{P}}_{h_1\sim\rho_{\mathbf{w}},\ldots,h_T\sim\rho_{\mathbf{w}}}\left[\mathop{\mathbb{E}}_{h\sim\rho_{\mathbf{w}}}\widehat{\mathsf{R}}_{\mathbb{S}}(h)\leq\frac{1}{T}\sum_{t=1}^{T}\widehat{\mathsf{R}}_{\mathbb{S}}(h_t)+\sqrt{\frac{\ln\frac{1}{\delta}}{2T}}\right]\geq 1-\delta$$

We are now able to compute the (final) bound!

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D},h_1\sim\rho_{\mathbf{w}},\ldots,h_T\sim\rho_{\mathbf{w}}}\left[\mathop{\mathbb{E}}_{h\sim\rho_{\mathbf{w}}}\mathsf{R}_{\mathcal{D}}(h)\leq\frac{1}{T}\sum_{t=1}^{T}\widehat{\mathsf{R}}_{\mathbb{S}}(h_t)+\sqrt{\frac{1}{2m}\left[\frac{1}{2\sigma^2}\|\mathbf{w}-\mathbf{w}_\pi\|_2^2+\ln\frac{4\sqrt{m}}{\delta}\right]}+\sqrt{\frac{\ln\frac{2}{\delta}}{2T}}\right]\geq 1-\delta$$

Problem 4: It is computationally heavy...

Problem 4: It is computationally heavy...
Solution: Use a disintegrated PAC-Bayesian bound to sample only one hypothesis!

***Disintegrated* PAC-Bayesian bound** (example based on Rivasplata *et al.*, 2020)

For any $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any prior π on $\mathbb{H}$, for any $\delta\in(0,1]$, **for any algorithm** A, we have

$$\mathop{\mathbb{P}}_{\substack{\mathbb{S}\sim\mathcal{D}^m,\\ \boxed{h\sim\rho_{\mathbb{S}}}}}\left[\mathsf{R}_{\mathcal{D}}(h)\leq\widehat{\mathsf{R}}_{\mathbb{S}}(h)+\sqrt{\frac{1}{2m}\left[\ln\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}+\ln\frac{2\sqrt{m}}{\delta}\right]_+}\right]\geq 1-\delta$$

After learning...

1. Sampling a hypothesis h from $\rho_{\mathbb{S}}$
2. With high probability (at least $1-\delta$), a bound holds, *e.g.*, we have

$$\mathsf{R}_{\mathcal{D}}(h)\leq\widehat{\mathsf{R}}_{\mathbb{S}}(h)+\sqrt{\frac{1}{2m}\left[\ln\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}+\ln\frac{2\sqrt{m}}{\delta}\right]_+}$$

***Disintegrated* PAC-Bayesian bound** (example based on Rivasplata *et al.*, 2020)

For any $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any prior π on $\mathbb{H}$, for any $\delta\in(0,1]$, **for any algorithm** A, we have

$$\mathop{\mathbb{P}}_{\substack{\mathbb{S}\sim\mathcal{D}^m,\\ h\sim\rho_{\mathbb{S}}}}\left[\mathsf{R}_{\mathcal{D}}(h)\le\widehat{\mathsf{R}}_{\mathbb{S}}(h)+\sqrt{\frac{1}{2m}\left[\ln\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}+\ln\frac{2\sqrt{m}}{\delta}\right]_+}\right]\ge 1-\delta$$

After learning...

1. Sampling a hypothesis h from $\rho_{\mathbb{S}}$
2. With high probability (at least $1-\delta$), a bound holds, *e.g.*, we have

$$\mathsf{R}_{\mathcal{D}}(h)\ \le\ \widehat{\mathsf{R}}_{\mathbb{S}}(h)+\sqrt{\frac{1}{2m}\left[\ln\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}+\ln\frac{2\sqrt{m}}{\delta}\right]_+}$$

Problem 5: Sampling a model is not realistic in practice...
Solution: Use a PAC-Bayesian bound with a Wasserstein distance!

Problem 5: Sampling a model is not realistic in practice...
Solution: Use a PAC-Bayesian bound with a Wasserstein distance!

Bound with the Wasserstein / KL on deterministic hypotheses (Viallard *et al.*, 2024c)

For any $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, assume that for any $\delta' \in (0,1]$, $(\mathrm{R}_{\mathcal{D}}(h) - \widehat{\mathrm{R}}_{\mathbb{S}}(h))^2$ is $L_{\mathbb{S}}(\delta')$-Lipschitz *w.r.t.* the distance $d(\cdot,\cdot)$, for any prior π on $\mathbb{H}$, for any $\alpha > 1$, for any $\delta \in (0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall h \in \mathbb{H},\ \eta \text{ on } \mathbb{H},\ \mathrm{R}_{\mathcal{D}}(h) \le \widehat{\mathrm{R}}_{\mathbb{S}}(h) + \sqrt{L_{\mathbb{S}}\left(\tfrac{\delta}{2}\right) \mathop{\mathbb{E}}_{h'\sim\eta} d(h,h') + \frac{\mathrm{KL}(\eta,\pi) + \ln\frac{4\sqrt{m}}{\delta}}{m}}\right] \ge 1-\delta$$

Self-bounding algorithm

$$\min_{h\in\mathbb{H},\eta}\left\{\widehat{\mathrm{R}}_{\mathbb{S}}(h) + \sqrt{L_{\mathbb{S}}\left(\tfrac{\delta}{2}\right) \mathop{\mathbb{E}}_{h'\sim\eta} d(h,h') + \frac{\mathrm{KL}(\eta,\pi) + \ln\frac{4\sqrt{m}}{\delta}}{m}}\right\}$$

Bound with the Wasserstein / KL on deterministic hypotheses (Viallard *et al.*, 2024c)

For any $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, assume that for any $\delta' \in (0,1]$, $(\mathrm{R}_{\mathcal{D}}(h) - \widehat{\mathrm{R}}_{\mathbb{S}}(h))^2$ is $L_{\mathbb{S}}(\delta')$-Lipschitz *w.r.t.* the distance $d(\cdot,\cdot)$, for any prior π on $\mathbb{H}$, for any $\alpha > 1$, for any $\delta \in (0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall h \in \mathbb{H},\ \eta \text{ on } \mathbb{H},\ \mathrm{R}_{\mathcal{D}}(h) \le \widehat{\mathrm{R}}_{\mathbb{S}}(h) + \sqrt{L_{\mathbb{S}}\left(\tfrac{\delta}{2}\right) \mathop{\mathbb{E}}_{h'\sim\eta} d(h,h') + \frac{\mathrm{KL}(\eta,\pi) + \ln\frac{4\sqrt{m}}{\delta}}{m}}\right] \ge 1-\delta$$

Self-bounding algorithm

$$\min_{h\in\mathbb{H},\eta}\left\{\widehat{\mathrm{R}}_{\mathbb{S}}(h) + \sqrt{L_{\mathbb{S}}\left(\tfrac{\delta}{2}\right) \mathop{\mathbb{E}}_{h'\sim\eta} d(h,h') + \frac{\mathrm{KL}(\eta,\pi) + \ln\frac{4\sqrt{m}}{\delta}}{m}}\right\}$$

Problems

- How to choose the distributions η, π and the distance $d(\cdot,\cdot)$?
- How to approximate/compute the Lipschitz constant $L_{\mathbb{S}}\left(\frac{\delta}{2}\right)$?

Choosing the distributions and the distance

How to choose the distributions η, π and the distance $d(\cdot, \cdot)$?

Distributions:

- Prior distribution $\pi = \mathcal{N}(\mathbf{w}_\pi, \sigma_\pi^2 I_d)$
- η distribution: $\eta = \mathcal{N}(\mathbf{w}_\eta, \sigma_\eta^2 I_d)$
- Posterior distribution: Dirac distribution $\rho = \delta_h$ where $h \in \mathbb{H}$

Distance / divergence:

- $\mathrm{KL}(\eta \| \pi) = \frac{1}{2}\left[\frac{\sigma_\eta^2}{\sigma_\pi^2} d - d + \frac{1}{\sigma_\pi^2}\|\mathbf{w}_\eta - \mathbf{w}_\pi\|_2^2 + d \ln\left(\frac{\sigma_\pi^2}{\sigma_\eta^2}\right)\right]$
- $\mathop{\mathbb{E}}_{h' \sim \eta} d(h, h') \leq \|\mathbf{w} - \mathbf{w}_\eta\|_2 + \sigma_\eta \sqrt{2} \frac{\Gamma((d+1)/2)}{\Gamma(d/2)}$

How to approximate/compute the Lipschitz constant $L_{\mathbb{S}}(\delta/2)$?
$\hookrightarrow$ Computation with SGD (and approximation from McDiarmid's inequality)

Our contribution: Estimation of $L_{\mathbb{S}}(\delta')$ for a general hypothesis set (Viallard *et al.*, 2024c)

For any hypothesis set $\mathbb{H}$, for any L-Lipschitz loss $\ell : \mathbb{H} \times \mathcal{Z} \to \mathbb{R}$ *w.r.t.* the distance $d(\cdot, \cdot)$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, \boldsymbol{\varepsilon}\sim\boldsymbol{\mathcal{E}}^m} \left[\begin{array}{l} h \mapsto |\mathsf{R}_{\mathcal{D}}(h) - \widehat{\mathsf{R}}_{\mathbb{S}}(h)| \text{ is } L_{\mathbb{S}}(\delta') = \left(2\mathscr{R}^{\boldsymbol{\varepsilon}}_{\mathcal{S}}(h) + 3L\sqrt{\frac{2\ln\frac{4}{\delta'}}{m}} \right)\text{-Lipschitz,} \\ \text{where} \quad \mathscr{R}^{\boldsymbol{\varepsilon}}_{\mathcal{S}}(h) = \sup\limits_{h\neq h'\in\mathbb{H}} \dfrac{1}{m}\displaystyle\sum_{i=1}^{m} \varepsilon_i \dfrac{[\ell(h', \mathbf{z}_i) - \ell(h, \mathbf{z}_i)]}{d(h, h')} \end{array} \right] \geq 1 - \delta',$$

Estimating $\mathscr{R}^{\boldsymbol{\varepsilon}}_{\mathcal{S}}(h)$

$$\text{Learning by SGD:} \quad \max_{h\neq h'\in\mathbb{H}} \left\{ \frac{1}{m}\sum_{i=1}^{m} \varepsilon_i \frac{[\ell(h', \mathbf{z}_i) - \ell(h, \mathbf{z}_i)]}{d(h, h')} \right\}$$

Key take-aways

Why PAC-Bayes for learning algorithms?

- **Bound as Objective:** Guarantee can *guide* training (*self-bounding*)
- **Prior Matters:** Data-dependent priors tighten guarantees at no extra cost
- **Flexibility:** KL, Rényi, Wasserstein... choose the divergence that fits your setting

What we present in this part?

- Linear classifiers $\rightarrow$ PBGD, DALC for domain adaptation.
- Majority votes $\rightarrow$ C-bound, joint-error, stochastic majority votes.
- Neural networks $\rightarrow$ Stochastic and deterministic neural networks

(Practical and Theoretical)
Flexibility of PAC-Bayes

PAC-Bayes bounds can depend on a function defined by the user
(Rivasplata *et al.*, 2020; Viallard *et al.*, 2024a; Maurer, 2025)

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m,\, h'\sim\pi,\, h\sim\rho_{\mathbb{S}}} \Big[\mathsf{R}_{\mathcal{D}}(h) \;\leq\; \mathsf{R}_{\mathbb{S}}(h) \;+\; \mathsf{Complexity}_{\mu}(h, h', \mathbb{S}, \delta) \Big] \;\geq\; 1-\delta,$$

$$\text{where} \quad \mathsf{Complexity}_{\mu}(h, h', \mathbb{S}, \delta) \;=\; \sqrt{\frac{1}{2m}\Big[\mu(h', \mathbb{S}) \;-\; \mu(h, \mathbb{S}) \;+\; \ln \frac{8\sqrt{m}}{\delta^2}\Big]_+}$$

Applications:

- Learning with a custom complexity measure
- Compare different complexity measures

Classical PAC-Bayesian theory (McAllester, 1998): Hypothesis h sampled from ρ
PAC-Bayes for random sets (Dupuis *et al.*, 2024): Hypothesis set $\mathbb{H}$ sampled from ρ

$$\mathop{\mathbb{E}}_{\mathbb{H}\sim\rho}\left[\mathrm{D}(h)\right] \leq \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{\mathcal{I}_{\mathrm{D}}(m)}{\delta}\right]$$

Applications:

- Generalization bound based on data-dependent Rademacher complexity
- Generalization bound based on fractal dimensions
  (Simsekli *et al.*, 2020; Birdal *et al.*, 2021; Camuto *et al.*, 2021; Hodgkinson *et al.*, 2022; Andreeva *et al.*, 2023; Dupuis *et al.*, 2023)

PAC-Bayesian theory has been applied to several types of models
(on supervised classification)

- Linear classifiers / Majority votes:

  (Langford *et al.*, 2002; McAllester, 2003; Ambroladze *et al.*, 2006; Laviolette *et al.*, 2007; Germain *et al.*, 2009; Morvant *et al.*, 2012; Parrado-Hernández *et al.*, 2012; Germain *et al.*, 2015; Lorenzen *et al.*, 2019; Masegosa *et al.*, 2020; Viallard *et al.*, 2021a; Wu *et al.*, 2021; Zantedeschi *et al.*, 2021; Wu *et al.*, 2022)

- Neural networks:

  (Langford *et al.*, 2001a; Dziugaite *et al.*, 2017; Neyshabur *et al.*, 2017, 2018; Letarte *et al.*, 2019; Nagarajan *et al.*, 2019; Zhou *et al.*, 2019; Dziugaite *et al.*, 2021; Pérez-Ortiz *et al.*, 2021; Viallard *et al.*, 2023, 2024b,c)

PAC-Bayesian theory has been applied to several settings

- Transductive learning:

  (Derbeko *et al.*, 2004; Catoni, 2006; Bégin *et al.*, 2014)

- Domain adaptation / transfer learning:

  (Germain *et al.*, 2013, 2016b; McNamara *et al.*, 2017; Germain *et al.*, 2020; Sicilia *et al.*, 2022)

- Multiview / Multimodal Learning:

  (Morvant *et al.*, 2014; Goyal *et al.*, 2017; Sun *et al.*, 2017, 2022)

- Reinforcement learning:

  (Fard *et al.*, 2010, 2011; Seldin *et al.*, 2011a,b, 2012; Sakhi *et al.*, 2023)

- Non-i.i.d / heavy-tailed / unbounded losses

  (Ralaivola *et al.*, 2010; Alquier *et al.*, 2016; Germain *et al.*, 2016a; Alquier *et al.*, 2018; Holland, 2019; Haddouche *et al.*, 2022; Chugg *et al.*, 2023; Haddouche *et al.*, 2023)

PAC-Bayesian theory has been applied to several settings

- Algorithmic stability & PAC-Bayes

  (London, 2017; Mou *et al.*, 2018; Li *et al.*, 2020; Zhou *et al.*, 2023)

- Generative models:

  (Chérief-Abdellatif *et al.*, 2022; Mbacke *et al.*, 2023a,b)

- Adversarial Robustness:

  (Farnia *et al.*, 2019; Viallard *et al.*, 2021b; Wang *et al.*, 2023; Xiao *et al.*, 2023; Mustafa *et al.*, 2024)

- Meta learning / Lifelong learning / continual learning:

  (Pentina *et al.*, 2014, 2015; Amit *et al.*, 2018; Zakerinia *et al.*, 2024)

Open Questions and Discussion

Having tight PAC-Bayesian bounds is not sufficient to understand generalization!

☹ Problems:

- To be tight, PAC-Bayes bounds rely on learning a good prior
- The prior is learned without any guarantees
- Without prior learning, the bound is too high

😀 (Possible?) solutions:

- Derive better complexity measures that avoid norms
- Control the prior construction process

Towards New Roles for Generalization Bounds?

If these bounds do not depend on learning the prior to be tight ...A bound might

- (Really) explain why a model generalizes (by having non-vacuous bounds)
- Certify the performance of deployed models
  - Accuracy
  - Robustness
  - Fairness

Generalization bounds can support trustworthy AI!

A New Machine Learning Paradigm?

☹ Training is mostly empirical: more data, larger models, etc

😀 But theory may help us *redefine* our learning strategies:

- Learn a model with performance guarantees
- Select models based on bounds, not only validation loss
- Learn smaller models that still generalize well

Can we shift from empirical risk minimization to bound-driven learning?

Thank you for your attention!

Questions? 🤔

Special thanks to

François Laviolette

Pascal Germain — Amaury Habrard

Valentina Zantedeschi — Mario Marchand

Benjamin Dupuis — Maxime Haddouche

Benjamin Guedj — Umut Şimşekli

Rémi Emonet — Christoph Lampert

Liva Ralaivola

who initiated EM to PAC-Bayes (Morvant et al., 2012)

📣 We are hiring!

emilie.morvant@univ-st-etienne.fr | https://emorvant.github.io

paul.viallard@inria.fr | https://paulviallard.github.io

References I

Johann Radon. Theorie und Anwendungen der absolut additiven Mengenfunktionen. *Hölder*. (1913).

Otton Nikodym. Sur une généralisation des intégrales de M. J. Radon. *Fundamenta Mathematicae*. (1930).

Vladimir Vapnik and Alexey Chervonenkis. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theory of Probability and its Applications*. (1971).

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The annals of probability*. (1975).

Monroe Donsker and Srinivasa Varadhan. Asymptotic evaluation of certain Markov process expectations for large time - III. *Communications on pure and applied Mathematics*. (1976).

Leslie Valiant. A Theory of the Learnable. *Communications of the ACM*. (1984).

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in applied probability*. (1997).

John Shawe-Taylor and Robert Williamson. A PAC Analysis of a Bayesian Estimator. *Annual Conference on Learning Theory*. (1997).

Yoav Freund. Self Bounding Learning Algorithms. *Conference on Learning Theory*. (1998).

David McAllester. Some PAC-Bayesian Theorems. *Annual Conference on Learning Theory*. (1998).

References II

John Langford and Avrim Blum. Microchoice Bounds and Self Bounding Learning Algorithms. *Conference on Computational Learning Theory*. (1999).

Leo Breiman. Random Forests. *Machine Learning*. (2001).

John Langford and Rich Caruana. (Not) Bounding the True Error. *Advances in Neural Information Processing Systems*. (2001a).

John Langford and Matthias Seeger. Bounds for Averaging Classifiers. Tech. rep. CMU-CS-01-102. Carnegie Mellon University, (2001b).

Peter Bartlett and Shahar Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research*. (2002).

Olivier Bousquet and André Elisseeff. Stability and Generalization. *Journal of Machine Learning Research*. (2002).

John Langford and John Shawe-Taylor. PAC-Bayes & Margins. *Advances in Neural Information Processing Systems 15 [Neural Information Processing Systems*. (2002).

Matthias W. Seeger. PAC-Bayesian Generalisation Error Bounds for Gaussian Process Classification. *Journal of Machine Learning Research*. (2002).

David A. McAllester. Simplified PAC-Bayesian Margin Bounds. *Annual Conference on Computational Learning Theory*. (2003).

References III

Matthias W. Seeger. Bayesian Gaussian process models : PAC-Bayesian generalisation error bounds and sparse approximations. PhD thesis. University of Edinburgh, UK, (2003).

Philip Derbeko, Ran El-Yaniv, and Ron Meir. Explicit Learning Curves for Transduction and Application to Clustering and Compression Algorithms. *Journal of Artificial Intelligence Research*. (2004).

Andreas Maurer. A Note on the PAC Bayesian Theorem. *arXiv*. cs.LG/0411099. (2004).

John Langford. Tutorial on practical prediction theory for classification. *Journal of machine learning research*. (2005).

Amiran Ambroladze, Emilio Parrado-Hernández, and John Shawe-Taylor. Tighter PAC-Bayes Bounds. *Advances in Neural Information Processing Systems*. (2006).

Olivier Catoni. PAC-Bayesian inductive and transductive learning. *arXiv*. math/0605793. (2006).

Alexandre Lacasse, François Laviolette, Mario Marchand, Pascal Germain, and Nicolas Usunier. PAC-Bayes Bounds for the Risk of the Majority Vote and the Variance of the Gibbs Classifier. *Advances in Neural Information Systems*. (2006).

Gilles Blanchard and François Fleuret. Occam's Hammer. *COLT*. (2007).

Olivier Catoni. PAC-Bayesian supervised classification: the thermodynamics of statistical learning. *arXiv preprint arXiv:0712.0248*. (2007).

References IV

François Laviolette and Mario Marchand. PAC-Bayes Risk Bounds for Stochastic Averages and Majority Votes of Sample-Compressed Classifiers. *Journal of Machine Learning Research*. (2007).

Cédric Villani. Optimal transport: old and new. *Springer*. (2008).

Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. PAC-Bayesian learning of linear classifiers. *International Conference on Machine Learning*. (2009).

Mahdi Milani Fard and Joelle Pineau. PAC-Bayesian Model Selection for Reinforcement Learning. *Advances in Neural Information Processing Systems*. (2010).

XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*. (2010).

Liva Ralaivola, Marie Szafranski, and Guillaume Stempfel. Chromatic PAC-Bayes Bounds for Non-IID Data: Applications to Ranking and Stationary β-Mixing Processes. *Journal of Machine Learning Research*. (2010).

Mahdi Milani Fard, Joelle Pineau, and Csaba Szepesvári. PAC-Bayesian Policy Evaluation for Reinforcement Learning. *Conference on Uncertainty in Artificial Intelligence*. (2011).

Jean-Francis Roy, François Laviolette, and Mario Marchand. From PAC-Bayes Bounds to Quadratic Programs for Majority Votes. *International Conference on Machine Learning*, (2011).

References V

Yevgeny Seldin, Peter Auer, John Shawe-Taylor, Ronald Ortner, and François Laviolette. PAC-Bayesian analysis of contextual bandits. *NeurIPS*. 24. (2011).

Yevgeny Seldin, François Laviolette, John Shawe-Taylor, Jan Peters, and Peter Auer. PAC-Bayesian Analysis of Martingales and Multiarmed Bandits. *arXiv preprint arXiv:1105.2416*. (2011).

Emilie Morvant, Sokol Koço, and Liva Ralaivola. PAC-Bayesian Generalization Bound on Confusion Matrix for Multi-Class Classification. *International Conference on Machine Learning*. (2012).

Emilio Parrado-Hernández, Amiran Ambroladze, John Shawe-Taylor, and Shiliang Sun. PAC-bayes bounds with data dependent priors. *Journal of Machine Learning Research*. (2012).

Avraham Ruderman, Mark D. Reid, Dario García-García, and James Petterson. Tighter Variational Representations of f-Divergences via Restriction to Probability Measures. *International Conference on Machine Learning*. (2012).

Yevgeny Seldin, Nicolò Cesa-Bianchi, Peter Auer, François Laviolette, and John Shawe-Taylor. PAC-Bayes-Bernstein Inequality for Martingales and its Application to Multiarmed Bandits. *Workshop on On-line Trading of Exploration and Exploitation 2*. (2012).

Huan Xu and Shie Mannor. Robustness and Generalization. *Machine Learning*. (2012).

Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. A PAC-Bayesian Approach for Domain Adaptation with Specialization to Linear Classifiers. *International Conference on Machine Learning*. (2013).

References VI

Guy Lever, François Laviolette, and John Shawe-Taylor. Tighter PAC-Bayes bounds through distribution-dependent priors. *Theoretical Computer Science*. (2013).

Luc Bégin, Pascal Germain, François Laviolette, and Jean-Francis Roy. PAC-Bayesian Theory for Transductive Learning. *International Conference on Artificial Intelligence and Statistics*. (2014).

Aurélien Bellet, Amaury Habrard, Emilie Morvant, and Marc Sebban. Learning A Priori Constrained Weighted Majority Votes. *Machine Learning*. (2014).

Emilie Morvant, Amaury Habrard, and Stéphane Ayache. Majority vote of diverse classifiers for late fusion. *S+SSPR 2014*. (2014).

Anastasia Pentina and Christoph H. Lampert. A PAC-Bayesian bound for Lifelong Learning. *International Conference on Machine Learning*. (2014).

Pascal Germain, Alexandre Lacasse, François Laviolette, Mario Marchand, and Jean-Francis Roy. Risk bounds for the majority vote: from a PAC-Bayesian analysis to a learning algorithm. *Journal of Machine Learning Research*. (2015).

Anastasia Pentina and Christoph H. Lampert. Lifelong Learning with Non-i.i.d. Tasks. *Advances in Neural Information Processing Systems*. (2015).

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research*. (2016).

References VII

Luc Bégin, Pascal Germain, François Laviolette, and Jean-Francis Roy. PAC-Bayesian Bounds based on the Rényi Divergence. *International Conference on Artificial Intelligence and Statistics*. (2016).

Pascal Germain, Francis R. Bach, Alexandre Lacoste, and Simon Lacoste-Julien. PAC-Bayesian Theory Meets Bayesian Inference. *Advances in Neural Information Processing Systems*. (2016a).

Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. A New PAC-Bayesian Perspective on Domain Adaptation. *International Conference on Machine Learning*. (2016b).

Jean-Francis Roy, Mario Marchand, and François Laviolette. A Column Generation Bound Minimization Approach with PAC-Bayesian Generalization Guarantees. *International Conference on Artificial Intelligence and Statistics*. (2016).

Gintare Karolina Dziugaite and Daniel M. Roy. Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data. *Conference on Uncertainty in Artificial Intelligence. AUAI Press*. (2017).

Anil Goyal, Emilie Morvant, Pascal Germain, and Massih-Reza Amini. Pac-bayesian analysis for a two-step hierarchical multiview learning approach. *ECML-PKDD*. (2017).

Ben London. A PAC-Bayesian Analysis of Randomized Learning with Application to Stochastic Gradient Descent. *Advances in Neural Information Processing Systems*. (2017).

Daniel McNamara and Maria-Florina Balcan. Risk Bounds for Transferring Representations With and Without Fine-Tuning. *International Conference on Machine Learning*. (2017).

References VIII

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring Generalization in Deep Learning. *Advances in Neural Information Processing Systems*. (2017).

Shiliang Sun, John Shawe-Taylor, and Liang Mao. PAC-Bayes analysis of multi-view learning. *Information Fusion*. 35. (2017).

Pierre Alquier and Benjamin Guedj. Simpler PAC-Bayesian bounds for hostile data. *Machine Learning*. (2018).

Ron Amit and Ron Meir. Meta-Learning by Adjusting Priors Based on Extended PAC-Bayes Theory. *International Conference on Machine Learning*. (2018).

Gintare Karolina Dziugaite and Daniel M. Roy. Data-dependent PAC-Bayes priors via differential privacy. *Advances in Neural Information Processing Systems*. (2018).

Wenlong Mou, Liwei Wang, Xiyu Zhai, and Kai Zheng. Generalization Bounds of SGLD for Non-convex Learning: Two Theoretical Viewpoints. *Conference On Learning Theory*. (2018).

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-Bayesian Approach to Spectrally-Normalized Margin Bounds for Neural Networks. *International Conference on Learning Representations*. (2018).

Omar Rivasplata, Csaba Szepesvári, John Shawe-Taylor, Emilio Parrado-Hernández, and Shiliang Sun. PAC-Bayes bounds for stable algorithms with instance-dependent priors. *Advances in Neural Information Processing Systems*. (2018).

References IX

Farzan Farnia, Jesse M. Zhang, and David Tse. Generalizable Adversarial Training via Spectral Normalization. *International Conference on Learning Representations*. (2019).

Matthew J. Holland. PAC-Bayes under potentially heavy tails. *Advances in Neural Information Processing Systems*. (2019).

H. Kervadec, J. Dolz, J. Yuan, C. Desrosiers, E. Granger, and I. Ayed. Constrained Deep Networks: Lagrangian Optimization via Log-Barrier Extensions. *arXiv preprint arXiv:1904.04205*. (2019).

Gaël Letarte, Pascal Germain, Benjamin Guedj, and François Laviolette. Dichotomize and Generalize: PAC-Bayesian Binary Activated Deep Neural Networks. *Advances in Neural Information Processing Systems*. (2019).

Stephan Sloth Lorenzen, Christian Igel, and Yevgeny Seldin. On PAC-Bayesian bounds for random forests. *Machine Learning*. (2019).

Zakaria Mhammedi, Peter Grünwald, and Benjamin Guedj. PAC-Bayes Un-Expected Bernstein Inequality. *Advances in Neural Information Processing Systems*. (2019).

Vaishnavh Nagarajan and J. Zico Kolter. Deterministic PAC-Bayesian generalization bounds for deep networks via generalizing noise-resilience. *International Conference on Learning Representations*. (2019).

Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P. Adams, and Peter Orbanz. Non-vacuous Generalization Bounds at the ImageNet Scale: a PAC-Bayesian Compression Approach. *International Conference on Learning Representations*. (2019).

References X

Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. PAC-Bayes and domain adaptation. *Neurocomputing*. (2020).

Jian Li, Xuanyuan Luo, and Mingda Qiao. On Generalization Error Bounds of Noisy Gradient Methods for Non-Convex Learning. *International Conference on Learning Representations*. (2020).

Andrés R. Masegosa, Stephan Sloth Lorenzen, Christian Igel, and Yevgeny Seldin. Second Order PAC-Bayesian Bounds for the Weighted Majority Vote. *Advances in Neural Information Processing Systems*. (2020).

Omar Rivasplata, Ilja Kuzborskij, Csaba Szepesvári, and John Shawe-Taylor. PAC-Bayes Analysis Beyond the Usual Bounds. *Advances in Neural Information Processing System (NeurIPS)*. (2020).

Umut Simsekli, Ozan Sener, George Deligiannidis, and Murat A. Erdogdu. Hausdorff Dimension, Heavy Tails, and Generalization in Neural Networks. *Advances in Neural Information Processing Systems*. (2020).

Tolga Birdal, Aaron Lou, Leonidas J. Guibas, and Umut Simsekli. Intrinsic Dimension, Persistent Homology and Generalization in Neural Networks. *Advances in Neural Information Processing Systems*. (2021).

Alexander Camuto, George Deligiannidis, Murat A. Erdogdu, Mert Gürbüzbalaban, Umut Simsekli, and Lingjiong Zhu. Fractal Structure and Generalization Properties of Stochastic Optimization Algorithms. *Advances in Neural Information Processing Systems*. (2021).

Gintare Karolina Dziugaite, Kyle Hsu, Waseem Gharbieh, Gabriel Arpino, and Daniel M. Roy. On the role of data in PAC-Bayes. *International Conference on Artificial Intelligence and Statistics*. (2021).

References XI

Yuki Ohnishi and Jean Honorio. Novel Change of Measure Inequalities with Applications to PAC-Bayesian Bounds and Monte Carlo Estimation. *International Conference on Artificial Intelligence and Statistics*. (2021).

María Pérez-Ortiz, Omar Rivasplata, John Shawe-Taylor, and Csaba Szepesvári. Tighter Risk Certificates for Neural Networks. *Journal of Machine Learning Research*. (2021).

Paul Viallard, Pascal Germain, Amaury Habrard, and Emilie Morvant. Self-bounding Majority Vote Learning Algorithms by the Direct Minimization of a Tight PAC-Bayesian C-Bound. *European Conference on Machine Learning and Knowledge Discovery in Databases*. (2021a).

Paul Viallard, Guillaume Vidot, Amaury Habrard, and Emilie Morvant. A PAC-Bayes Analysis of Adversarial Robustness. *Advances in Neural Information Processing Systems*. (2021b).

Yi-Shan Wu, Andrés R. Masegosa, Stephan Sloth Lorenzen, Christian Igel, and Yevgeny Seldin. Chebyshev-Cantelli PAC-Bayes-Bennett Inequality for the Weighted Majority Vote. *Advances in Neural Information Processing Systems*. (2021).

Valentina Zantedeschi, Paul Viallard, Emilie Morvant, Rémi Emonet, Amaury Habrard, Pascal Germain, and Benjamin Guedj. Learning Stochastic Majority Votes by Minimizing a PAC-Bayes Generalization Bound. *Advances in Neural Information Processing Systems*. (2021).

Ron Amit, Baruch Epstein, Shay Moran, and Ron Meir. Integral Probability Metrics PAC-Bayes Bounds. *Advances in Neural Information Processing Systems (NeurIPS)*. (2022).

References XII

Jeremiah Birrell, Paul Dupuis, Markos A. Katsoulakis, Yannis Pantazis, and Luc Rey-Bellet. (f, Gamma)-Divergences: Interpolating between f-Divergences and Integral Probability Metrics. *Journal of Machine Learning Research*. (2022).

Badr-Eddine Chérief-Abdellatif, Yuyang Shi, Arnaud Doucet, and Benjamin Guedj. On PAC-Bayesian reconstruction guarantees for VAEs. *International Conference on Artificial Intelligence and Statistics*. (2022).

Maxime Haddouche and Benjamin Guedj. Online PAC-Bayes Learning. *Advances in Neural Information Processing Systems*. (2022).

Liam Hodgkinson, Umut Simsekli, Rajiv Khanna, and Michael W. Mahoney. Generalization Bounds using Lower Tail Exponents in Stochastic Optimizers. *International Conference on Machine Learning*. (2022).

Antoine Picard-Weibel and Benjamin Guedj. On change of measure inequalities for f-divergences. *CoRR*. abs/2202.05568. (2022). arXiv: 2202.05568. URL: https://arxiv.org/abs/2202.05568.

Anthony Sicilia, Katherine Atwell, Malihe Alikhani, and Seong Jae Hwang. PAC-Bayesian domain adaptation bounds for multiclass learners. *Conference on Uncertainty in Artificial Intelligence*. (2022).

Shiliang Sun, Mengran Yu, John Shawe-Taylor, and Liang Mao. Stability-based PAC-Bayes analysis for multi-view learning algorithms. *Information Fusion*. 86. (2022).

Yi-Shan Wu and Yevgeny Seldin. Split-kl and PAC-Bayes-split-kl Inequalities for Ternary Random Variables. *Advances in Neural Information Processing Systems*. (2022).

References XIII

Rayna Andreeva, Katharina Limbeck, Bastian Rieck, and Rik Sarkar. Metric Space Magnitude and Generalisation in Neural Networks. *Topological, Algebraic and Geometric Learning Workshops*. (2023).

Ben Chugg, Hongjian Wang, and Aaditya Ramdas. A Unified Recipe for Deriving (Time-Uniform) PAC-Bayes Bounds. *Journal of Machine Learning Research*. (2023).

Benjamin Dupuis and Paul Viallard. From Mutual Information to Expected Dynamics: New Generalization Bounds for Heavy-Tailed SGD. *arXiv*. abs/2312.00427. (2023).

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

Sokhna Diarra Mbacke, Florence Clerc, and Pascal Germain. PAC-Bayesian Generalization Bounds for Adversarial Generative Models. *International Conference on Machine Learning*. (2023a).

Sokhna Diarra Mbacke, Florence Clerc, and Pascal Germain. Statistical Guarantees for Variational Autoencoders using PAC-Bayesian Theory. *Advances in Neural Information Processing Systems*. (2023b).

Otmane Sakhi, Pierre Alquier, and Nicolas Chopin. PAC-Bayesian Offline Contextual Bandits With Guarantees. *International Conference on Machine Learning*. (2023).

Paul Viallard, Maxime Haddouche, Umut Simsekli, and Benjamin Guedj. Learning via Wasserstein-Based High Probability Generalisation Bounds. *Conference on Neural Information Processing Systems (NeurIPS)*. (2023).

References XIV

Zifan Wang, Nan Ding, Tomer Levinboim, Xi Chen, and Radu Soricut. Improving Robust Generalization by Direct PAC-Bayesian Bound Minimization. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2023).

Jiancong Xiao, Ruoyu Sun, and Zhi-Quan Luo. PAC-Bayesian Spectrally-Normalized Bounds for Adversarially Robust Generalization. *Advances in Neural Information Processing Systems*. (2023).

Sijia Zhou, Yunwen Lei, and Ata Kabán. Toward Better PAC-Bayes Bounds for Uniformly Stable Algorithms. *Advances in Neural Information Processing Systems*. (2023).

Benjamin Dupuis, Paul Viallard, George Deligiannidis, and Umut Simsekli. Uniform Generalization Bounds on Data-Dependent Hypothesis Sets via PAC-Bayesian Theory on Random Sets. *Journal of Machine Learning Research*. (2024).

Ilja Kuzborskij, Kwang-Sung Jun, Yulian Wu, Kyoungseok Jang, and Francesco Orabona. Better-than-KL PAC-Bayes Bounds. *Conference on Learning Theory*. (2024).

Waleed Mustafa, Philipp Liznerski, Antoine Ledent, Dennis Wagner, Puyu Wang, and Marius Kloft. Non-vacuous Generalization Bounds for Adversarial Risk in Stochastic Neural Networks. *International Conference on Artificial Intelligence and Statistics*. (2024).

Paul Viallard, Rémi Emonet, Amaury Habrard, Emilie Morvant, and Valentina Zantedeschi. Leveraging PAC-Bayes Theory and Gibbs Distributions for Generalization Bounds with Complexity Measures. *International Conference on Artificial Intelligence and Statistics*. (2024a).

References XV

Paul Viallard, Pascal Germain, Amaury Habrard, and Emilie Morvant. A general framework for the practical disintegration of PAC-Bayesian bounds. *Machine Learning*. (2024).

Paul Viallard, Maxime Haddouche, Umut Simsekli, and Benjamin Guedj. Tighter Generalisation Bounds via Interpolation. *arXiv*. abs/2402.05101. (2024).

Hossein Zakerinia, Amin Behjati, and Christoph H. Lampert. More Flexible PAC-Bayesian Meta-Learning by Learning Learning Algorithms. *International Conference on Machine Learning*. (2024).

Andreas Maurer. Generalization of the Gibbs algorithm with high probability at low temperatures. *arXiv*. abs/2502.11071. (2025).