A Framework for Bounding Deterministic Risk with PAC-Bayes: Applications to Majority Votes

Benjamin Leblanc
https://benjaminlblanc.github.io/
Université Laval, département d'informatique et de génie logiciel

November 20th, 2025

Plan

1. Preamble
   - Definitions
   - Generalization guarantees / PAC-Bayes
2. Derandomizing PAC-Bayes bounds - Majority votes
   - Categorical distribution
   - Dirichlet distribution
   - Gaussian distribution
   - Empirical results

Definitions

An **observation** $\mathbf{z} := (\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y} = \mathcal{Z}$ is a pair **features** - **labels**.

Type of task

$\mathcal{X} \subseteq \mathbb{R}^d$, $\mathcal{Y} \in \{\mathbf{y} \in \{0,1\}^c \mid \sum_{i=1}^c y_i = 1\}$

Data generating distribution

Examples comes from a **distribution** D over $\mathcal{Z}$.

Training set

$$S \; := \; \{\, \mathbf{z}_1,\; \mathbf{z}_2,\; \ldots,\; \mathbf{z}_m \,\} \; \sim \; D^m$$

Classifier (hypothesis), hypothesis space

$$h : \mathcal{X} \to \mathcal{Y}', \quad \mathcal{Y}' \supseteq \mathcal{Y}$$

Hypothesis space

$$h \in \mathcal{H}$$

Loss function

$$\ell : \mathcal{Y}' \times \mathcal{Y} \to \{0,1\}$$

Definitions

Training loss

$$\widehat{\mathcal{L}}_S(h) := \frac{1}{n}\sum_{i=1}^{m} \ell(h(\mathbf{x}_i), y_i)$$

Goal : minimize the average loss on unseen examples (generalization loss, risk)

$$\mathcal{L}_D(h) := \mathop{\mathbb{E}}_{z\sim D} \ell(h(\mathbf{x}), y)$$

$\rightarrow$ Throughout the presentation, for any quantity K_D that depends on D :

- $\widehat{K_S}$ represents the empirical estimation of K_D from S ;
- $\widetilde{K_S}$ represents a bound on K_D using S.

The generalization problem

PAC (Probably Approximately Correct) bounds

"With probability at least "$1-\delta$", the generalization loss of h will be at most $\varepsilon(\cdot,\dots)$"

$$\mathbb{P}_{S\sim D^m}\left(\mathcal{L}_D(h) \leq \varepsilon(\widehat{\mathcal{L}}_S(h), m, \delta, \dots)\right) \geq 1-\delta$$

Why PAC-Bayes ?

PAC (Probably Approximately Correct)

With probability at least "$1-\delta$", the generalization loss of h will be at most "$\varepsilon(\cdot,\ldots,\cdot)$"

$$\mathbb{P}_{S\sim D^m}\Big(\mathcal{L}_D(h) \;\leq\; \varepsilon(\widehat{\mathcal{L}}_S(h), m, \delta, \ldots)\Big) \;\geq\; 1-\delta$$

- A single hypothesis ($\mathcal{H} = \{h\}$) :

$$\mathcal{L}_D(h) \leq \widehat{\mathcal{L}}_S(h) + \sqrt{\tfrac{1}{2m}\log\left(\tfrac{1}{\delta}\right)}.$$

- Finite hypothesis space ($|\mathcal{H}| < \infty$) :

$$\forall h \in \mathcal{H}, \quad \mathcal{L}_D(h) \leq \widehat{\mathcal{L}}_S(h) + \sqrt{\tfrac{1}{2m}\log\left(\tfrac{|\mathcal{H}|}{\delta}\right)}$$

- Countable hypothesis space with probability *a priori* $p(h_i)$:

$$\forall h \in \mathcal{H}, \quad \mathcal{L}_D(h) \leq \widehat{\mathcal{L}}_S(h) + \sqrt{\tfrac{1}{2m}\log\left(\tfrac{1}{p(h)\delta}\right)}$$

- Uncountable hypothesis space : VC dimension, Rademacher complexity, PAC-Bayes...

Typical PAC-Bayes

PAC-Bayesian considerations

We will define our bounds on average over the random draw of a hypothesis following an *a posteriori* distribution Q over $\mathcal{H}$.

$$\underset{S\sim D^m}{\mathbb{P}}\left(\underset{h'\sim Q}{\mathbb{E}}\mathcal{L}_D(h') \leq \ldots\right) \geq 1-\delta \quad \underset{S\sim D^m}{\mathbb{P}}\left(\underbrace{\underset{h'\sim Q}{\mathbb{E}}\mathcal{L}_D(h')}_{:=\mathcal{L}_D(Q),\text{"Gibbs risk"}} \leq \ldots\right) \geq 1-\delta$$

$\rightarrow$ Throughout the presentation :

- h' is always drawn from Q, whereas h is chosen.

Major con of such an approach ?

We do not have any information on any single predictor ! Not usable in many applications...

Plan

The question of interest : Is it possible to bound a *single, chosen* classifier of an *uncountable hypothesis space* using a PAC-Bayes bound ?

i.e. : $\mathcal{L}_D(h) \leq f\left(\mathcal{L}_D(Q)\right) \leq f\left(\widetilde{\mathcal{L}_D}(Q)\right) \mathcal{L}_D(h) \leq \underbrace{f\left(\mathcal{L}_D(Q)\right) \leq f\left(\widetilde{\mathcal{L}_D}(Q)\right)}_{\text{This is what PAC-Bayes does}} \overbrace{\mathcal{L}_D(h) \leq f\left(\mathcal{L}_D(Q)\right)}^{\text{This is what we week!}} \leq$

This is what P

A first step

Proposition 1.

For any data distribution D, distribution Q over $\mathcal{H}$, classifier $h \in \mathcal{H}$: if $c_Q(h) \neq b_Q(h)$, then

$$\mathcal{L}_D(h) = \color{red}{\mathcal{L}_D(h)} \color{black} = \frac{\mathcal{L}_D(Q)\color{red}{\mathcal{L}_D(Q)}\color{black} - b_Q(h)}{c_Q(h) - b_Q(h)},$$

where

$$c_Q(h)\,\underbrace{c_Q(h)} = [\ldots],\ \ b_Q(h)\,\underbrace{b_Q(h)} = [\ldots].$$

Involve both Q and h.

A first step

Proposition 1. *proof sketch.*

$$
\begin{aligned}
\mathcal{L}_D(Q) &= \underset{(\mathbf{x},y)\sim D}{\mathbb{E}} \; \underset{h'\sim Q}{\mathbb{E}} \ell(h'(\mathbf{x}), y) \\
&= \underset{(\mathbf{x},y)\sim D}{\mathbb{P}} \Big(\ell(h(\mathbf{x}), y) = 1\Big) \underset{(\mathbf{x},y)\sim D}{\mathbb{E}} \left[\underset{h'\sim Q}{\mathbb{E}} \ell(h'(\mathbf{x}), y) \;\middle|\; \ell(h(\mathbf{x}), y) = 1 \right] + \\
&\quad \underset{(\mathbf{x},y)\sim D}{\mathbb{P}} \Big(\ell(h(\mathbf{x}), y) = 0\Big) \underset{(\mathbf{x},y)\sim D}{\mathbb{E}} \left[\underset{h'\sim Q}{\mathbb{E}} \ell(h'(\mathbf{x}), y) \;\middle|\; \ell(h(\mathbf{x}), y) = 0 \right] \\
&= \underbrace{\underset{(\mathbf{x},y)\sim D}{\mathbb{P}} \Big(\ell(h(\mathbf{x}), y) = 1\Big)}_{\mathcal{L}_D(h)} \underbrace{\underset{(\mathbf{x},y)\sim D}{\mathbb{E}} \left[\underset{h'\sim Q}{\mathbb{E}} \ell(h'(\mathbf{x}), y) \;\middle|\; \ell(h(\mathbf{x}), y) = 1 \right]}_{c_Q(h)} + \\
&\quad \underbrace{\underset{(\mathbf{x},y)\sim D}{\mathbb{P}} \Big(\ell(h(\mathbf{x}), y) = 0\Big)}_{1-\mathcal{L}_D(h)} \underbrace{\underset{(\mathbf{x},y)\sim D}{\mathbb{E}} \left[\underset{h'\sim Q}{\mathbb{E}} \ell(h'(\mathbf{x}), y) \;\middle|\; \ell(h(\mathbf{x}), y) = 0 \right]}_{b_Q(h)}
\end{aligned}
$$

A first step

Proposition 1.

For any data distribution D, distribution Q over $\mathcal{H}$, classifier $h \in \mathcal{H}$: if $c_Q(h) \neq b_Q(h)$, then

$$\mathcal{L}_D(h) = \frac{\mathcal{L}_D(Q) - b_Q(h)}{c_Q(h) - b_Q(h)},$$

where

$$c_Q(h) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}\left[\mathop{\mathbb{E}}_{h'\sim Q} \ell(h'(\mathbf{x}), y) \,\middle|\, \ell(h(\mathbf{x}), y) = 1\right], \; b_Q(h) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}\left[\mathop{\mathbb{E}}_{h'\sim Q} \ell(h'(\mathbf{x}), y) \,\middle|\, \ell(h(\mathbf{x}), y) = 0\right].$$

Corollary 1.

Let $\mathcal{L}_D(Q) \leq \widetilde{\mathcal{L}_S}(Q)$ ✓, $c_Q(h) \geq \widetilde{c_S}(h)$?, and $b_Q(h) \geq \widetilde{b_S}(h)$? : if $\widetilde{c_S}(h) > \widetilde{b_S}(h)$?, then

$$\mathcal{L}_D(h) \leq \frac{\widetilde{\mathcal{L}_S}(Q) - \widetilde{b_S}(h)}{\widetilde{c_S}(h) - \widetilde{b_S}(h)}.$$

A first step

Why not simply use regular PAC-Bayesian results to bound $c_Q(h)$ and $b_Q(h)$?

Lemma 1.

Let $\mathrm{kl}(q,p) = q\ln(\frac{q}{p}) + (1-q)\ln(\frac{1-q}{1-p})$. For any distribution D, hypothesis set $\mathcal{H}$, prior distribution P over $\mathcal{H}$, $\delta \in (0,1]$, we have with probability at least $1-\delta$ over the random choice $S \sim D^m$ that for every Q over $\mathcal{H}$:

$$\mathrm{kl}\left(\widehat{\mathcal{L}_S}(Q) \middle\| \mathcal{L}_D(Q)\right) \leq \frac{1}{m}\left[\mathrm{KL}(Q||P) + \ln\left(\frac{2\sqrt{m}}{\delta}\right)\right].$$

$$\mathrm{kl}\left(\widehat{b_S}(h) \middle\| b_Q(h)\right) \leq \frac{1}{m}\left[\mathrm{KL}(Q||P) + \ln\left(\frac{2\sqrt{m}}{\delta}\right)\right]?$$

$$\rightarrow b_Q(h) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}\left[\mathop{\mathbb{E}}_{h'\sim Q} \ell(h'(\mathbf{x}), y) \;\middle|\; \ell(h(\mathbf{x}), y) = 0\right] :($$

A first step

PAC-Bayes won't be of help here... We need the bound to hold simultaneously for all $h \in \mathcal{H}$!

We need a way to :

- Bound $b_Q(h) \geq \widetilde{b_S}(h)$ and $c_Q(h) \geq \widetilde{c_S}(h)$ with probability 1 ;
- Ensure that $\widetilde{c_S}(h) > \widetilde{b_S}(h)$.

Possible solution : bound $b_Q(h)$ and $c_Q(h)$ by their worst-case values ! In there current form $\left(\text{i.e. } \mathbb{E}_{(\mathbf{x},y)\sim D}\left[\mathbb{E}_{h'\sim Q}\ell(h'(\mathbf{x}),y) \;\middle|\; \ell(h(\mathbf{x}),y) = 0.5 \pm 0.5\right]\right)$, hard to grasp what their worst value could be ; missing link between h and Q...

Solution : we focus on specific $\mathcal{H}$ and Q, so that simplification might occur !

A specialization to majority votes

Base classifiers

$$\mathcal{F} = \{f_1, \ldots, f_n\},\ f_i\ :\ \mathcal{X} \to \mathcal{Y}\ \forall i \in \{1, \ldots, n\}$$

Hypotesis space

$$\mathcal{H} = \{\textstyle\sum_{i=1}^{n} \alpha_i f_i \mid \boldsymbol{\alpha} \in \mathcal{A}\},\ \mathcal{A} \subseteq \mathbb{R}^n$$

Majority vote

$$h_{\boldsymbol{\alpha}}(\mathbf{x}) = \textstyle\sum_{i=1}^{n} \alpha_i f_i(\mathbf{x}) = \boldsymbol{\alpha} \cdot \mathbf{f}(\mathbf{x}) \in \mathcal{Y}'$$

$\rightarrow$ For simplicity, we now define Q over the weight space A : each vector of weights $\boldsymbol{\alpha}$ corresponds to a unique hypothesis $h_{\boldsymbol{\alpha}}$.

Gibbs risk (for any Q over $\mathcal{A}$)

$$\mathcal{L}_D(Q) := \mathop{\mathbb{E}}_{z \sim D} \mathop{\mathbb{E}}_{\boldsymbol{\alpha} \sim Q} \ell\left(h_{\boldsymbol{\alpha}}(\mathbf{x}), y\right) = \mathop{\mathbb{E}}_{\boldsymbol{\alpha} \sim Q} \mathcal{L}_D(h_{\boldsymbol{\alpha}})$$

Theory under the Categorical assumption

We first explore the most restrictive support for Q : $\{\boldsymbol{\alpha} \in \{0,1\}^n \mid \sum_{i=1}^n \alpha_i = 1\}$.
A natural choice for Q is the Categorical distribution, and a natural choice for ℓ is

$$\ell(h_{\boldsymbol{\alpha}}(\mathbf{x}), y) = \mathbb{1}\left\{ \textstyle\sum_{i=1}^n \alpha_i \mathbb{1}\left\{f_i(\mathbf{x}) \neq y\right\} \underbrace{\displaystyle\sum_{i=1}^n \alpha_i \mathbb{1}\left\{f_i(\mathbf{x}) \neq y\right\}}_{\boldsymbol{\alpha}(\mathbf{x},y)} \geq 0.5 \right\}.$$

Theory under the Categorical assumption

Let us recall that $Q \sim \mathcal{C}(\mathbf{p})$. Given $(\mathbf{x}, y)$, we have

$$c_Q(h_\cdot) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}\left[\mathop{\mathbb{E}}_{\boldsymbol{\alpha}\sim Q}\ell(h_{\boldsymbol{\alpha}}(\mathbf{x}), y)\,{\color{red}\mathop{\mathbb{E}}_{\boldsymbol{\alpha}\sim Q}\ell(h_{\boldsymbol{\alpha}}(\mathbf{x}), y)}\;\middle|\;\ell(h_\cdot(\mathbf{x}), y) = 1\right].$$

We first focus on the inner content of $c_Q(h)$ and $b_Q(h)$: ${\color{red}\mathop{\mathbb{E}}_{\boldsymbol{\alpha}\sim Q}\ell(h_{\boldsymbol{\alpha}}(\mathbf{x}), y)}$.

Theory under the Categorical assumption

Remember that $Q \sim \mathcal{C}(\mathbf{pp})$.

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{\alpha}\sim Q} \ell(h_{\boldsymbol{\alpha}}(\mathbf{x}), y) &= \mathop{\mathbb{E}}_{\boldsymbol{\alpha}\sim Q} \mathbb{1}\left\{\sum_{j=1}^{n} \alpha_j \mathbb{1}\left\{f_j(\mathbf{x}) \neq y\right\} \geq 0.5\right\} \\
&= \sum_{i=1}^{n} p_i \mathbb{1}\{\mathbb{1}\left\{f_i(\mathbf{x}) \neq y\right\} \geq 0.5\} \\
&= \sum_{i=1}^{n} p_i \mathbb{1}\{f_i(\mathbf{x}) \neq y\} \\
&= \mathbf{pp}(\mathbf{x}, y)
\end{aligned}
$$

Theory under the Categorical assumption

$$
\begin{aligned}
c_Q(h_\cdot) &= \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}\left[\mathop{\mathbb{E}}_{\boldsymbol{\alpha}\sim Q}\ell(h_{\boldsymbol{\alpha}}(\mathbf{x}),y) \,\middle|\, \ell(h_\cdot(\mathbf{x}),y)=1\right] \\
&= \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}\left[\mathbf{p}(\mathbf{x},y) \mid \ell(h_\cdot(\mathbf{x}),y)=1\right] \\
&= \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}\left[\mathbf{p}(\mathbf{x},y) \mid \ell(h_{\mathbf{p}}(\mathbf{x}),y)=1\right] && \langle\text{By choosing } h_\cdot := h_{\mathbf{p}}\rangle \\
&= \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}\left[\mathbf{p}(\mathbf{x},y) \mid \mathbf{p}(\mathbf{x},y)\geq 0.5\right]
\end{aligned}
$$

Theory under the Categorical assumption

Proposition 1.1.

Let $Q \sim \mathcal{C}(\mathbf{p})$ be a Categorical distribution with parameters $\mathbf{p}$. For any data distribution D, $\mathbf{p} \in [0,1]^n$ s.t. $\sum_{i=1}^n p_i = 1$:

$$c_Q(h_{\mathbf{p}}) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}\left[\mathbf{p}(\mathbf{x},y) \;\middle|\; \mathbf{p}(\mathbf{x},y) \geq 0.5\right], \;\; b_Q(h_{\mathbf{p}}) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}\left[\mathbf{p}(\mathbf{x},y) \;\middle|\; \mathbf{p}(\mathbf{x},y) < 0.5\right],$$

where $\mathbf{p}(\mathbf{x},y) = \sum_{i=1}^n p_i \mathbb{1}\{f_i(\mathbf{x}) \neq y\}$.

We need a way to :

- Bound $b_Q(h) \geq \widetilde{b_S}(h)$ and $c_Q(h) \geq \widetilde{c_S}(h)$ with probability 1 ;
- Ensure that $\widetilde{c_S}(h) > \widetilde{b_S}(h)$. ✓

Theory under the Categorical assumption

We seek upper bounds on $b_Q(h_{\mathbf{p}})$ and $c_Q(h_{\mathbf{p}})$ that hold with probability 1.

Given $\mathbf{p}$:

$$\begin{aligned} b_Q(h_{\mathbf{p}}) &= \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}\left[p(\mathbf{x},y) \;\middle|\; p(\mathbf{x},y) < 0.5\right] \\ &\geq \min_{(\mathbf{x},y)\in\mathcal{X}\times\mathcal{Y}}\left[p(\mathbf{x},y) \;\middle|\; p(\mathbf{x},y) < 0.5\right] \\ &= \min_{(\mathbf{x},y)\in\mathcal{X}\times\mathcal{Y}}\left[p(\mathbf{x},y)\right] \end{aligned}$$

$\rightarrow$ Easy to compute for many hypotheses.

Theory under the Categorical assumption

Similarly :

$$\begin{aligned} c_Q(h_{\mathbf{p}}) &= \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}\left[p(\mathbf{x},y) \;\middle|\; p(\mathbf{x},y) \geq 0.5\right] \\ &\geq \min_{(\mathbf{x},y)\in\mathcal{X}\times\mathcal{Y}}\left[p(\mathbf{x},y) \;\middle|\; p(\mathbf{x},y) \geq 0.5\right] \\ &\geq \min_{\boldsymbol{\beta}\in\{0,1\}^n}\left[\boldsymbol{\beta}\cdot\mathbf{p} \;\middle|\; \boldsymbol{\beta}\cdot\mathbf{p} \geq 0.5\right]. \end{aligned}$$

Let's make an example : suppose $\mathbf{p} = [0.33, 0.33, 0.33]$. What is $\min_{\boldsymbol{\beta}\in\{0,1\}^n}\left[\boldsymbol{\beta}\cdot\mathbf{p} \;\middle|\; \boldsymbol{\beta}\cdot\mathbf{p} \geq 0.5\right]$?

Now suppose $|\mathbf{p}| > 1000$! There are $O(2^n)$ possible values for $\boldsymbol{\beta}\cdot\mathbf{p}$...

Knapsack problem representation non efficient : $\mathbf{p}$ is a real-valued vector...

Solution : partition problem representation !

Theory under the Categorical assumption

The partition problem

Given a set of number $\mathbf{p}$, the partition problem consists in the following :

$$\text{Part}(\mathbf{p}) = \operatorname*{argmin}_{\mathbf{p}_1, \mathbf{p}_2} \left\{ \left| \text{sum}(\mathbf{p}_1) - \text{sum}(\mathbf{p}_2) \right| \; s.t. \; \mathbf{p}_1 \cup \mathbf{p}_2 = \mathbf{p} \; \wedge \; \mathbf{p}_1 \cap \mathbf{p}_2 = \emptyset \right\}.$$

Proposition 1.1.1.

Given a set of number $\mathbf{p} \in [0,1]^n$, where $\text{sum}(\mathbf{p}) = 1$, let $\text{Part}(\mathbf{p}) = \mathbf{p}_1, \mathbf{p}_2$. Then :

$$\min_{\boldsymbol{\beta} \in \{0,1\}^n} \left[\boldsymbol{\beta} \cdot \mathbf{p} \;\middle|\; \boldsymbol{\beta} \cdot \mathbf{p} \geq 0.5 \right] = \max(\text{sum}(\mathbf{p}_1), \text{sum}(\mathbf{p}_2)).$$

Indeed : $\text{Part}([0.33, 0.33, 0.33]) = [0.33], [0.33, 0.33]$.

$\rightarrow$ There exist complete greedy algorithms for the partitioning problem !

Theory under the Categorical assumption

Proposition 1.1.1.

Given a set of number $\mathbf{p} \in [0,1]^n$, where $\mathrm{sum}(\mathbf{p}) = 1$, let $\mathrm{Part}(\mathbf{p}) = \mathbf{p}_1, \mathbf{p}_2$. Then :

$$\min_{\beta \in \{0,1\}^n} \left[\beta \cdot \mathbf{p} \;\middle|\; \beta \cdot \mathbf{p} \geq 0.5 \right] = \max(\mathrm{sum}(\mathbf{p}_1), \mathrm{sum}(\mathbf{p}_2)).$$

We need a way to :

- Bound $b_Q(h) \geq \widetilde{b_S}(h)$ and $c_Q(h) \geq \widetilde{c_S}(h)$ with probability 1 ;✓
- Ensure that $\widetilde{c_S}(h) > \widetilde{b_S}(h)$.✓

Corollary 1.

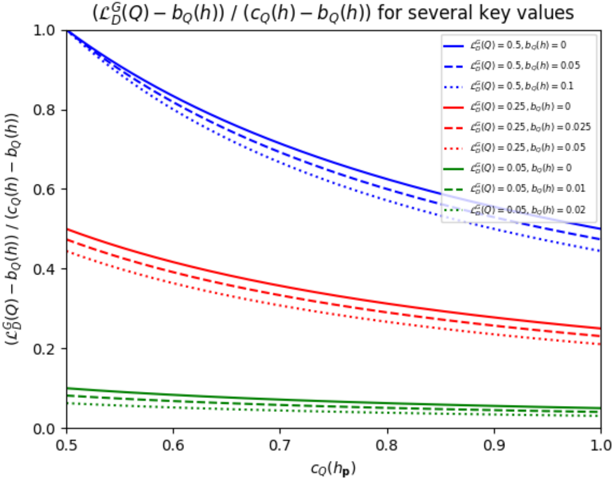
Let $\mathcal{L}_D(Q) \leq \widetilde{\mathcal{L}_S}(Q)$, $c_Q(h_\mathbf{p}) \geq \widetilde{c_S}(h_\mathbf{p})$, and $b_Q(h_\mathbf{p}) \geq \widetilde{b_S}(h_\mathbf{p})$: if $\widetilde{c_S}(h_\mathbf{p}) > \widetilde{b_S}(h_\mathbf{p})$, then

$$\mathcal{L}_D(h_\mathbf{p}) \leq \frac{\widetilde{\mathcal{L}_S}(Q) - \widetilde{b_S}(h_\mathbf{p})}{\widetilde{c_S}(h_\mathbf{p}) - \widetilde{b_S}(h_\mathbf{p})}.$$

Theory under the Categorical assumption

Worst-case scenario :

- $\widetilde{b_S}(h_\mathbf{p}) = 0$,
- $\widetilde{c_S}(h_\mathbf{p}) = 0.5$:
- $\frac{\mathcal{L}_D(Q) - \widetilde{b_S}(h_\mathbf{p})}{\widetilde{c_S}(h_\mathbf{p}) - \widetilde{b_S}(h_\mathbf{p})} = 2\mathcal{L}_D(Q)$.



Theory under the Dirichlet assumption

We now explore a lesser restrictive support for Q : $\{\boldsymbol{\alpha} \in [0,1]^n \mid \sum_{i=1}^n \alpha_i = 1\}$.
A natural choice for Q is the Dirichlet distribution.

Proposition 1.2.

Let $\mathcal{Y}' = \mathcal{Y}$, $\ell(h_{\boldsymbol{\alpha}}(\mathbf{x}), y) = \mathbb{1}\left\{\sum_{i=1}^n \alpha_i \mathbb{1}\left\{f_i(\mathbf{x}) \neq y\right\} \geq 0.5\right\}$ and $Q \sim \mathcal{D}(\mathbf{p})$ be a Dirichlet distribution with parameters $\mathbf{p}$. For any data distribution D, $\mathbf{p} \in \mathbb{R}^n_{>0}$:

$$c_Q(h_{\mathbf{p}}) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}\left[I_{0.5}\left(\overline{\mathbf{p}} - \mathbf{p}(\mathbf{x},y), \mathbf{p}(\mathbf{x},y)\right) \;\middle|\; \frac{\mathbf{p}(\mathbf{x},y)}{\overline{\mathbf{p}}} \geq 0.5 \right],$$

$$b_Q(h_{\mathbf{p}}) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D}\left[I_{0.5}\left(\overline{\mathbf{p}} - \mathbf{p}(\mathbf{x},y), \mathbf{p}(\mathbf{x},y)\right) \;\middle|\; \frac{\mathbf{p}(\mathbf{x},y)}{\overline{\mathbf{p}}} < 0.5 \right],$$

where $I_x(\cdot,\cdot)$ is the regularized incomplete beta function evaluated at x, and $\overline{\mathbf{p}} = \sum_{i=1}^n a_i$.

$\rightarrow$ $c_Q(h_{\mathbf{p}})$ and $b_Q(h_{\mathbf{p}})$ are proportional to $\mathbf{p}(\mathbf{x},y)$!

Theory under the Dirichlet assumption

Proposition 1.2. *proof sketch*

Based on ZANTEDESCHI et al. 2021, we use the following result :

$$\mathop{\mathbb{E}}_{\boldsymbol{\alpha}\sim\mathcal{D}(\mathbf{p})} \mathbb{1}\left\{\sum_{i=1}^{n} \alpha_i \mathbb{1}\left\{f_i(\mathbf{x}) \neq y\right\} \geq 0.5\right\} = I_{0.5}\left(\overline{\mathbf{p}} - \mathbf{p}(\mathbf{x}, y), \mathbf{p}(\mathbf{x}, y)\right).$$

Theory under the Gaussian assumption

We now explore the most general support for Q : $\mathbb{R}^n$.

A natural choice for Q is the Gaussian distribution. What loss to use ? Remember : we must have an analytic form for $\underset{\boldsymbol{\alpha}\sim Q}{\mathbb{E}} \ell(h_{\boldsymbol{\alpha}}(\mathbf{x}), y)$...

Such a result only exists for $\mathcal{Y} = \{-1, +1\}$ and $\ell(h_{\boldsymbol{\alpha}}(\mathbf{x}), y) = \mathbb{1}\left\{y \sum_{i=1}^n \alpha_i f_i(\mathbf{x}) \leq 0\right\}$.

Proposition 1.3.1

Let $\mathcal{Y} = \{-1, +1\}$ and $\mathcal{Y}' \subseteq \mathbb{R}$. Let $\ell(h_{\boldsymbol{\alpha}}(\mathbf{x}), y) = \mathbb{1}\left\{y \sum_{i=1}^n \alpha_i f_i(\mathbf{x}) \leq 0\right\}$ and $Q \sim \mathcal{N}(\boldsymbol{\mu}, I)$ be an isotropic Gaussian distribution with mean $\boldsymbol{\mu}$. For any data distribution D, $\boldsymbol{\mu} \in \mathbb{R}^n$:

$$c_Q(h_{\boldsymbol{\mu}}) = \underset{(\mathbf{x},\cdot)\sim D}{\mathbb{E}} \left[1 - \Phi\left(\frac{|\boldsymbol{\mu} \cdot \mathbf{f}(\mathbf{x})|}{||\mathbf{f}(\mathbf{x})||}\right)\right],$$

$$b_Q(h_{\boldsymbol{\mu}}) = \underset{(\mathbf{x},\cdot)\sim D}{\mathbb{E}} \left[\Phi\left(\frac{|\boldsymbol{\mu} \cdot \mathbf{f}(\mathbf{x})|}{||\mathbf{f}(\mathbf{x})||}\right)\right],$$

where $\Phi(k) = \frac{1}{2}\left(1 - \frac{2}{\sqrt{\pi}} \int_0^{k/\sqrt{2}} e^{t^2} dt\right)$.

Empirical results

- Tested on 11 binary/multiclass datasets
- Using Seeger's bound to bound the Gibbs risk
- Base classifier are :
    - Uniformly spaced decision stumps (binary classification), or
    - Decision trees learned on half the examples (multiclass classification)
- Five benchmarks from the literature

Empirical results

Table – Comparison of our proposed approach and a few baselines for different types of predictors. **Bolded** values correspond to the best bound values. Average and standard deviations computed over 5 random seeds.

Task	FO		SO		Bin		CBnd		VC-dim		**Part**	
	Bound	Test error	Bound	Test error	Bound	Test error	Bound	Test error	Bound	Test error	Bound	Test error
ADULT	46.9 ± 0.0	22.2 ± 0.1	52.9 ± 0.1	15.8 ± 0.2	51.4 ± 0.0	24.1 ± 0.0	74.5 ± 0.4	21.0 ± 1.6	78.9 ± 0.0	17.2 ± 0.3	**23.2** ± 2.4	21.7 ± 2.5
CODRNA	40.2 ± 0.1	12.0 ± 0.1	47.4 ± 0.1	11.7 ± 0.1	42.8 ± 5.2	15.2 ± 4.5	63.7 ± 1.0	23.1 ± 1.3	38.8 ± 1.6	23.1 ± 1.5	**24.7** ± 0.3	23.6 ± 0.3
HABER	75.2 ± 0.9	26.5 ± 1.8	110.5 ± 1.1	26.1 ± 2.1	84.7 ± 0.8	26.1 ± 0.7	100.0 ± 0.0	27.1 ± 1.8	112.0 ± 1.0	28.4 ± 1.8	**37.8** ± 0.5	26.5 ± 1.8
MUSH	12.2 ± 0.3	4.8 ± 0.5	20.6 ± 1.2	0.5 ± 0.3	12.6 ± 0.4	1.1 ± 0.1	22.0 ± 0.5	4.8 ± 0.5	59.1 ± 0.1	4.8 ± 0.5	**7.9** ± 2.4	4.8 ± 0.5
PHIS	**25.8** ± 0.3	11.1 ± 0.6	33.6 ± 0.2	6.4 ± 0.5	27.5 ± 0.2	6.8 ± 0.5	37.1 ± 0.5	10.4 ± 1.3	85.2 ± 0.2	11.1 ± 0.6	**25.8** ± 0.4	11.1 ± 0.6
SVMG	17.4 ± 0.6	7.0 ± 1.0	28.2 ± 0.7	7.0 ± 1.0	22.0 ± 0.6	7.0 ± 1.0	29.2 ± 0.7	7.0 ± 1.0	37.6 ± 0.3	7.0 ± 1.0	**8.7** ± 0.3	7.0 ± 1.0
TTT	75.9 ± 1.9	30.4 ± 3.8	100.0 ± 0.7	30.2 ± 2.7	87.2 ± 1.2	29.5 ± 4.3	94.8 ± 0.7	30.4 ± 3.8	117.0 ± 1.8	32.2 ± 2.9	**69.8** ± 6.5	30.4 ± 3.8
FASHION	40.7 ± 0.2	17.1 ± 0.2	58.6 ± 0.2	13.0 ± 0.3	49.9 ± 0.3	13.0 ± 0.3	N/A	N/A	40.3 ± 0.3	18.3 ± 0.3	**26.6** ± 0.1	25.4 ± 0.2
MNIST	26.9 ± 0.1	8.0 ± 0.4	44.7 ± 0.1	4.4 ± 0.1	35.8 ± 0.2	4.3 ± 0.2	N/A	N/A	37.4 ± 2.6	15.8 ± 2.7	**23.6** ± 0.2	22.6 ± 0.3
PEND	12.0 ± 0.2	2.5 ± 0.3	17.1 ± 0.1	1.5 ± 0.2	15.5 ± 0.2	1.5 ± 0.2	N/A	N/A	52.7 ± 0.6	5.8 ± 0.7	**10.6** ± 0.2	8.3 ± 0.7
PROTEIN	84.9 ± 0.6	34.9 ± 0.3	138.3 ± 0.3	36.6 ± 0.5	120.1 ± 1.3	55.0 ± 0.5	N/A	N/A	73.4 ± 0.3	39.7 ± 0.5	**54.1** ± 0.2	52.0 ± 0.6
SENSOR	16.8 ± 0.4	4.1 ± 0.3	21.2 ± 1.0	3.9 ± 0.4	15.3 ± 1.0	3.0 ± 0.2	N/A	N/A	32.9 ± 2.0	9.6 ± 2.0	**11.4** ± 1.0	10.6 ± 0.6

Empirical results

Table – Proposed approach's bound and test error for every distribution. **Bolded** values are the best bounds, underlined values are the best error. Average and standard deviations over 5 random seeds.

Task	Categorical - Part		Dirichlet - Part		Gaussian - Part	
	Bound	Test error	Bound	Test error	Bound	Test error
ADULT	46.9 $\pm$ 0.0	22.2 $\pm$ 0.1	**23.2** $\pm$ 2.4	<u>21.6</u> $\pm$ 2.5	48.7 $\pm$ 4.1	<u>21.6</u> $\pm$ 4.3
CODRNA	**24.7** $\pm$ 0.3	23.6 $\pm$ 0.3	29.4 $\pm$ 4.8	28.4 $\pm$ 4.6	39.7 $\pm$ 1.4	<u>12.0</u> $\pm$ 0.1
HABER	**37.8** $\pm$ 0.5	<u>26.5</u> $\pm$ 1.8	84.5 $\pm$ 21.2	29.7 $\pm$ 5.9	82.3 $\pm$ 1.0	<u>26.5</u> $\pm$ 0.9
MUSH	**7.9** $\pm$ 2.4	4.8 $\pm$ 0.5	**7.9** $\pm$ 1.0	4.1 $\pm$ 1.8	15.7 $\pm$ 1.3	<u>0.6</u> $\pm$ 0.4
PHIS	**25.8** $\pm$ 0.4	11.1 $\pm$ 0.6	26.8 $\pm$ 0.3	<u>6.5</u> $\pm$ 0.4	27.2 $\pm$ 0.2	<u>6.5</u> $\pm$ 0.4
SVMG	**8.7** $\pm$ 0.3	7.0 $\pm$ 1.0	9.6 $\pm$ 1.8	<u>5.3</u> $\pm$ 1.4	13.0 $\pm$ 0.3	6.9 $\pm$ 1.5
TTT	**69.8** $\pm$ 6.5	30.4 $\pm$ 3.8	84.2 $\pm$ 1.3	29.8 $\pm$ 4.2	75.9 $\pm$ 2.2	<u>29.3</u> $\pm$ 2.3
FASHION	**26.6** $\pm$ 0.1	25.4 $\pm$ 0.2	45.9 $\pm$ 0.2	<u>17.0</u> $\pm$ 0.2	N/A	N/A
MNIST	**23.6** $\pm$ 0.2	22.6 $\pm$ 0.3	43.7 $\pm$ 0.2	<u>8.0</u> $\pm$ 0.4	N/A	N/A
PEND	**10.6** $\pm$ 0.2	8.3 $\pm$ 0.7	12.9 $\pm$ 0.3	<u>2.4</u> $\pm$ 0.2	N/A	N/A
PROTEIN	**54.1** $\pm$ 0.2	<u>52.0</u> $\pm$ 0.6	65.5 $\pm$ 29.2	52.8 $\pm$ 0.6	N/A	N/A
SENSOR	**11.4** $\pm$ 1.0	10.6 $\pm$ 0.6	12.3 $\pm$ 0.8	<u>3.0</u> $\pm$ 0.3	N/A	N/A

Thank you for attending :) !

References I

ZANTEDESCHI, Valentina, Paul VIALLARD, Emilie MORVANT, Rémi EMONET, Amaury HABRARD, Pascal GERMAIN et Benjamin GUEDJ (2021). "Learning Stochastic Majority Votes by Minimizing a PAC-Bayes Generalization Bound". In : *NeurIPS*, p. 455-467.