PAC-Bayes & Variational Inference

Badr-Eddine Chérief-Abdellatif
Chargé de Recherche CNRS

badr.eddine.cherief.abdellatif@gmail.com

Master 2, Sorbonne Université
Paris, Spring 2023

Reminder of the setting

Training dataset : $\mathcal{S} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ i.i.d $\sim \mathbb{P}$, $X_i \in \mathcal{X} \subset \mathbb{R}^d$, $Y_i \in \mathcal{Y}$.

- $\mathcal{Y} = \{\text{lion}, \text{gazelle}\}$: Binary Classification.
- $\mathcal{Y} = \mathbb{R}$: Regression.

Reminder of the setting

Training dataset : $\mathcal{S} = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ i.i.d $\sim \mathbb{P}$, $X_i \in \mathcal{X} \subset \mathbb{R}^d$, $Y_i \in \mathcal{Y}$.

- $\mathcal{Y} = \{\text{lion}, \text{gazelle}\}$: Binary Classification.
- $\mathcal{Y} = \mathbb{R}$: Regression.

Loss $\ell(y', y)$ quantifies the price to predict y' instead of y.

- $\mathcal{Y} = \{\text{cat}, \text{dog}\}$: $\ell(y', y) = \mathbb{1}(y' \neq y)$.
- $\mathcal{Y} = \mathbb{R}$: $\ell(y', y) = (y' - y)^2$.

Reminder of the setting

Training dataset : $\mathcal{S} = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ i.i.d $\sim \mathbb{P}$, $X_i \in \mathcal{X} \subset \mathbb{R}^d$, $Y_i \in \mathcal{Y}$.

- $\mathcal{Y} = \{\text{lion}, \text{gazelle}\}$: Binary Classification.
- $\mathcal{Y} = \mathbb{R}$: Regression.

Loss $\ell(y', y)$ quantifies the price to predict y' instead of y.

- $\mathcal{Y} = \{\text{cat}, \text{dog}\}$: $\ell(y', y) = \mathbb{1}(y' \neq y)$.
- $\mathcal{Y} = \mathbb{R}$: $\ell(y', y) = (y' - y)^2$.

We then define the (theoretical) risk of a predictor $f_\theta, \theta \in \Theta$:

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

and the empirical risk $\hat{R}_{\mathcal{S}}(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(f_\theta(X_i), Y_i)$.

Reminder of the setting

Training dataset : $\mathcal{S} = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ i.i.d $\sim \mathbb{P}$, $X_i \in \mathcal{X} \subset \mathbb{R}^d$, $Y_i \in \mathcal{Y}$.

- $\mathcal{Y} = \{\text{lion}, \text{gazelle}\}$: Binary Classification.
- $\mathcal{Y} = \mathbb{R}$: Regression.

Loss $\ell(y', y)$ quantifies the price to predict y' instead of y.

- $\mathcal{Y} = \{\text{cat}, \text{dog}\}$: $\ell(y', y) = \mathbb{1}(y' \neq y)$.
- $\mathcal{Y} = \mathbb{R}$: $\ell(y', y) = (y' - y)^2$.

We then define the (theoretical) risk of a predictor $f_\theta, \theta \in \Theta$:

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

and the empirical risk $\hat{R}_{\mathcal{S}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(f_\theta(X_i), Y_i)$.

General goal : Learn using the data a predictor $\hat{\theta}$ with small risk

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

Reminder of the previous lecture

In this course, we want to derive Probably Approximately Correct generalization bounds for predictors $\hat{\theta}$ or randomized predictors $\hat{\rho}$.

Reminder of the previous lecture

In this course, we want to derive Probably Approximately Correct generalization bounds for predictors $\hat{\theta}$ or randomized predictors $\hat{\rho}$.

Typical PAC bounds : with high probability, the generalization gap of θ is at most something we can control & compute. For any δ,

$$\mathbb{P}_{\mathcal{S}}\left[\forall \theta \in \Theta,\ \left|R(\theta) - \hat{R}_{\mathcal{S}}(\theta)\right| \lesssim \sqrt{\frac{\mathrm{comp}(\Theta) + \log(\frac{1}{\delta})}{n}}\right] \geq 1 - \delta.$$

Reminder of the previous lecture

In this course, we want to derive Probably Approximately Correct generalization bounds for predictors $\hat{\theta}$ or randomized predictors $\hat{\rho}$.

Typical PAC bounds : with high probability, the generalization gap of θ is at most something we can control & compute. For any δ,

$$\mathbb{P}_{\mathcal{S}}\left[\forall \theta \in \Theta,\ \left|R(\theta) - \hat{R}_{\mathcal{S}}(\theta)\right| \lesssim \sqrt{\frac{\mathrm{comp}(\Theta) + \log(\frac{1}{\delta})}{n}}\right] \geq 1 - \delta.$$

Some limits to this classical statistical learning theory :

Reminder of the previous lecture

In this course, we want to derive Probably Approximately Correct generalization bounds for predictors $\hat{\theta}$ or randomized predictors $\hat{\rho}$.

Typical PAC bounds : with high probability, the generalization gap of θ is at most something we can control & compute. For any δ,

$$\mathbb{P}_{\mathcal{S}}\left[\forall \theta \in \Theta, \ \left|R(\theta) - \hat{R}_{\mathcal{S}}(\theta)\right| \lesssim \sqrt{\frac{\mathrm{comp}(\Theta) + \log(\frac{1}{\delta})}{n}}\right] \geq 1 - \delta.$$

Some limits to this classical statistical learning theory :

- relies on restricting the complexity of Θ,

Reminder of the previous lecture

In this course, we want to derive Probably Approximately Correct generalization bounds for predictors $\hat{\theta}$ or randomized predictors $\hat{\rho}$.

Typical PAC bounds : with high probability, the generalization gap of θ is at most something we can control & compute. For any δ,

$$\mathbb{P}_{\mathcal{S}}\left[\forall\theta\in\Theta,\ \left|R(\theta)-\hat{R}_{\mathcal{S}}(\theta)\right|\lesssim\sqrt{\frac{\mathrm{comp}(\Theta)+\log(\frac{1}{\delta})}{n}}\right]\geq 1-\delta.$$

Some limits to this classical statistical learning theory :

- relies on restricting the complexity of Θ,
- too conservative as Θ is rarely entirely explored by the algorithm $\{\hat{\theta}_t\}_t$,

Reminder of the previous lecture

In this course, we want to derive Probably Approximately Correct generalization bounds for predictors $\hat{\theta}$ or randomized predictors $\hat{\rho}$.

Typical PAC bounds : with high probability, the generalization gap of θ is at most something we can control & compute. For any δ,

$$\mathbb{P}_{\mathcal{S}}\left[\forall \theta \in \Theta,\ \left|R(\theta) - \hat{R}_{\mathcal{S}}(\theta)\right| \lesssim \sqrt{\frac{\mathrm{comp}(\Theta) + \log(\frac{1}{\delta})}{n}}\right] \geq 1 - \delta.$$

Some limits to this classical statistical learning theory :

- relies on restricting the complexity of Θ,
- too conservative as Θ is rarely entirely explored by the algorithm $\{\hat{\theta}_t\}_t$,
- ignores the interaction between the dataset $\mathcal{S}$ and the algorithm $\hat{\theta}$.

An information-theoretic notion of stability

An information-theoretic notion of stability

- Key idea : a (possibly randomized) learning algorithm $\theta \sim \hat{\rho}_{\mathcal{S}}$ is stable if it doesn't reveal too much information about the dataset $\mathcal{S}$ it was trained on.

An information-theoretic notion of stability

- Key idea : a (possibly randomized) learning algorithm $\theta \sim \hat{\rho}_{\mathcal{S}}$ is stable if it doesn't reveal too much information about the dataset $\mathcal{S}$ it was trained on.
- Question : how do we formalize this dependence ?

An information-theoretic notion of stability

- Key idea : a (possibly randomized) learning algorithm $\theta \sim \hat{\rho}_{\mathcal{S}}$ is stable if it doesn't reveal too much information about the dataset $\mathcal{S}$ it was trained on.
- Question : how do we formalize this dependence ?
- Answer : the mutual information !

$$\mathcal{I}(\hat{\rho}_{\mathcal{S}}; \mathcal{S}) = \mathbb{E}_{\mathcal{S}} \left[\mathrm{KL} \left(\hat{\rho}_{\mathcal{S}} \| \mathbb{E}_{\mathcal{S}}[\hat{\rho}_{\mathcal{S}}] \right) \right] = \mathrm{KL}(P_{\theta,\mathcal{S}} \| P_{\theta} \otimes P_{\mathcal{S}}).$$

An information-theoretic notion of stability

- Key idea : a (possibly randomized) learning algorithm $\theta \sim \hat{\rho}_{\mathcal{S}}$ is stable if it doesn't reveal too much information about the dataset $\mathcal{S}$ it was trained on.
- Question : how do we formalize this dependence ?
- Answer : the mutual information !

$$\mathcal{I}(\hat{\rho}_{\mathcal{S}}; \mathcal{S}) = \mathbb{E}_{\mathcal{S}} \left[\mathrm{KL} \left(\hat{\rho}_{\mathcal{S}} \| \mathbb{E}_{\mathcal{S}}[\hat{\rho}_{\mathcal{S}}] \right) \right] = \mathrm{KL}(P_{\theta,\mathcal{S}} \| P_{\theta} \otimes P_{\mathcal{S}}).$$

- MI quantifies the number of bits of information the algorithm leaks about the training data into the parameters it learns.

An information-theoretic notion of stability

- Key idea : a (possibly randomized) learning algorithm $\theta \sim \hat{\rho}_{\mathcal{S}}$ is stable if it doesn't reveal too much information about the dataset $\mathcal{S}$ it was trained on.
- Question : how do we formalize this dependence ?
- Answer : the mutual information !

$$\mathcal{I}(\hat{\rho}_{\mathcal{S}};\mathcal{S}) = \mathbb{E}_{\mathcal{S}}\left[\mathsf{KL}\left(\hat{\rho}_{\mathcal{S}}\|\mathbb{E}_{\mathcal{S}}[\hat{\rho}_{\mathcal{S}}]\right)\right] = \mathsf{KL}(P_{\theta,\mathcal{S}}\|P_{\theta}\otimes P_{\mathcal{S}}).$$

- MI quantifies the number of bits of information the algorithm leaks about the training data into the parameters it learns.
- The higher the MI, the worse the stability of the algorithm, the worse the generalization ability.

An information-theoretic notion of stability

- Key idea : a (possibly randomized) learning algorithm $\theta \sim \hat{\rho}_{\mathcal{S}}$ is stable if it doesn't reveal too much information about the dataset $\mathcal{S}$ it was trained on.
- Question : how do we formalize this dependence ?
- Answer : the mutual information !

$$\mathcal{I}(\hat{\rho}_{\mathcal{S}}; \mathcal{S}) = \mathbb{E}_{\mathcal{S}}\left[\mathsf{KL}\left(\hat{\rho}_{\mathcal{S}} \| \mathbb{E}_{\mathcal{S}}[\hat{\rho}_{\mathcal{S}}]\right)\right] = \mathsf{KL}(P_{\theta,\mathcal{S}} \| P_{\theta} \otimes P_{\mathcal{S}}).$$

- MI quantifies the number of bits of information the algorithm leaks about the training data into the parameters it learns.
- The higher the MI, the worse the stability of the algorithm, the worse the generalization ability.

Theorem (Russo & Zhou, '16 / Xu & Raginsky, '17 / Catoni, '07) : if $\ell(\cdot,\cdot) \leq 1$,

$$\mathbb{E}_{\mathcal{S}}\left[R(\hat{\rho}_{\mathcal{S}}) - \hat{R}_{\mathcal{S}}(\hat{\rho}_{\mathcal{S}})\right] \leq \sqrt{\frac{2 \cdot \mathcal{I}(\hat{\rho}_{\mathcal{S}}; \mathcal{S})}{n}}.$$

Information-theory vs PAC-Bayes

Information theoretic generalization bound ($\ell(\cdot,\cdot) \leq 1$) :

$$\mathbb{E}_{\mathcal{S}}\left[R(\hat{\rho}_{\mathcal{S}}) - \hat{R}_{\mathcal{S}}(\hat{\rho}_{\mathcal{S}})\right] \leq \sqrt{\frac{2 \cdot \mathbb{E}_{\mathcal{S}}\left[\mathrm{KL}\left(\hat{\rho}_{\mathcal{S}} \| \mathbb{E}_{\mathcal{S}}[\hat{\rho}_{\mathcal{S}}]\right)\right]}{n}}.$$

Information-theory vs PAC-Bayes

Information theoretic generalization bound ($\ell(\cdot,\cdot) \leq 1$) :

$$\mathbb{E}_{\mathcal{S}} \left[R(\hat{\rho}_{\mathcal{S}}) - \hat{R}_{\mathcal{S}}(\hat{\rho}_{\mathcal{S}}) \right] \leq \sqrt{\frac{2 \cdot \mathbb{E}_{\mathcal{S}} \left[\mathrm{KL} \left(\hat{\rho}_{\mathcal{S}} \| \mathbb{E}_{\mathcal{S}}[\hat{\rho}_{\mathcal{S}}] \right) \right]}{n}}.$$

vs PAC-Bayes generalization bound ($\ell(\cdot,\cdot) \leq 1$) :

$$\mathbb{E}_{\mathcal{S}} \left[R(\hat{\rho}_{\mathcal{S}}) - \hat{R}_{\mathcal{S}}(\hat{\rho}_{\mathcal{S}}) \right] \leq \sqrt{\frac{2 \cdot \mathbb{E}_{\mathcal{S}} \left[\mathrm{KL} \left(\hat{\rho}_{\mathcal{S}} \| \pi \right) \right]}{n}}.$$

Information-theory vs PAC-Bayes

Information theoretic generalization bound ($\ell(\cdot,\cdot) \leq 1$) :

$$\mathbb{E}_{\mathcal{S}}\left[R(\hat{\rho}_{\mathcal{S}}) - \hat{R}_{\mathcal{S}}(\hat{\rho}_{\mathcal{S}})\right] \leq \sqrt{\frac{2 \cdot \mathbb{E}_{\mathcal{S}}\left[\mathrm{KL}\left(\hat{\rho}_{\mathcal{S}} \| \mathbb{E}_{\mathcal{S}}[\hat{\rho}_{\mathcal{S}}]\right)\right]}{n}}.$$

vs PAC-Bayes generalization bound ($\ell(\cdot,\cdot) \leq 1$) :

$$\mathbb{E}_{\mathcal{S}}\left[R(\hat{\rho}_{\mathcal{S}}) - \hat{R}_{\mathcal{S}}(\hat{\rho}_{\mathcal{S}})\right] \leq \sqrt{\frac{2 \cdot \mathbb{E}_{\mathcal{S}}\left[\mathrm{KL}\left(\hat{\rho}_{\mathcal{S}} \| \pi\right)\right]}{n}}.$$

The prior minimizing the PAC-Bayes bound is $\pi := \mathbb{E}_{\mathcal{S}}[\hat{\rho}_{\mathcal{S}}]$.

Information-theory vs PAC-Bayes

Information theoretic generalization bound ($\ell(\cdot,\cdot) \leq 1$) :

$$\mathbb{E}_{\mathcal{S}}\left[R(\hat{\rho}_{\mathcal{S}}) - \hat{R}_{\mathcal{S}}(\hat{\rho}_{\mathcal{S}})\right] \leq \sqrt{\frac{2 \cdot \mathbb{E}_{\mathcal{S}}\left[\mathrm{KL}\left(\hat{\rho}_{\mathcal{S}} \| \mathbb{E}_{\mathcal{S}}[\hat{\rho}_{\mathcal{S}}]\right)\right]}{n}}.$$

vs PAC-Bayes generalization bound ($\ell(\cdot,\cdot) \leq 1$) :

$$\mathbb{E}_{\mathcal{S}}\left[R(\hat{\rho}_{\mathcal{S}}) - \hat{R}_{\mathcal{S}}(\hat{\rho}_{\mathcal{S}})\right] \leq \sqrt{\frac{2 \cdot \mathbb{E}_{\mathcal{S}}\left[\mathrm{KL}\left(\hat{\rho}_{\mathcal{S}} \| \pi\right)\right]}{n}}.$$

The prior minimizing the PAC-Bayes bound is $\pi := \mathbb{E}_{\mathcal{S}}[\hat{\rho}_{\mathcal{S}}]$.

Typical PAC-Bayes bound : with high probability, the generalization gap of ρ is at most something we can control & compute. If we assume that $\ell(\cdot,\cdot) \leq 1$, then $\forall \delta \in (0,1)$,

$$\mathbb{P}_{\mathcal{S}}\left[\forall \rho \in \mathcal{P}(\Theta),\ \left|R(\rho) - \hat{R}_{\mathcal{S}}(\rho)\right| \lesssim \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \log(\frac{1}{\delta})}{n}}\right] \geq 1 - \delta.$$

Overview of the course

The course will be divided in 5 lectures :

- Lecture 1 : Introduction & Motivation
- Lecture 2 : Basics of PAC-Bayes Theory
- Lecture 3 : Advances in PAC-Bayes Theory
- Lecture 4 : Basics of Variational Inference
- Lecture 5 : Advances in Variational Inference

Lecture 3 : Advances in PAC-Bayes Theory

Outline of the lecture

- PAC-Bayes bounds robust to heavy-tails.
- PAC-Bayes bounds achieving fast rates.
- Towards tight certificates in Deep Learning.
- Generalization bounds for SGD using information bounds.

PAC-Bayes bounds robust to heavy-tails

A generic PAC-Bayes bound

A generic PAC-Bayes bound

Germain, Lacasse, Laviolette and Marchand [2009]

For any convex function $\mathcal{D} : [0,1]^2 \to \mathbb{R}$, with proba $\geq 1-\delta$:

$$\forall \rho \in \mathcal{P}(\Theta), \quad \mathcal{D}\left(\hat{R}_{\mathcal{S}}(\rho), R(\rho)\right) \leq \frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\pi}[e^{n\mathcal{D}\left(\hat{R}_{\mathcal{S}}(\theta), R(\theta)\right)}]}{\delta}\right)}{n}.$$

A generic PAC-Bayes bound

Germain, Lacasse, Laviolette and Marchand [2009]

For any convex function $\mathcal{D} : [0,1]^2 \to \mathbb{R}$, with proba $\geq 1-\delta$:

$$\forall \rho \in \mathcal{P}(\Theta), \quad \mathcal{D}\left(\hat{R}_{\mathcal{S}}(\rho), R(\rho)\right) \leq \frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\pi}[e^{n\mathcal{D}\left(\hat{R}_{\mathcal{S}}(\theta), R(\theta)\right)}]}{\delta}\right)}{n}.$$

$$\begin{aligned} n\mathcal{D}\left(\mathbb{E}_{\theta\sim\rho}\left[\hat{R}_{\mathcal{S}}(\theta)\right], \mathbb{E}_{\theta\sim\rho}[R(\theta)]\right) &\leq n \cdot \mathbb{E}_{\theta\sim\rho}\left[\mathcal{D}\left(\hat{R}_{\mathcal{S}}(\theta), R(\theta)\right)\right] \\ &\leq \mathrm{KL}(\rho\|\pi) + \log\left(\mathbb{E}_{\theta\sim\pi}[e^{n\mathcal{D}\left(\hat{R}_{\mathcal{S}}(\theta), R(\theta)\right)}]\right) \\ &\leq \mathrm{KL}(\rho\|\pi) + \log\left(\frac{\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\pi}[e^{n\mathcal{D}\left(\hat{R}_{\mathcal{S}}(\theta), R(\theta)\right)}]}{\delta}\right) \end{aligned}$$

A generic PAC-Bayes bound

Germain, Lacasse, Laviolette and Marchand [2009]

For any convex function $\mathcal{D} : [0,1]^2 \to \mathbb{R}$, with proba $\geq 1-\delta$:

$$\forall \rho \in \mathcal{P}(\Theta), \quad \mathcal{D}\left(\hat{R}_{\mathcal{S}}(\rho), R(\rho)\right) \leq \frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\pi}[e^{n\mathcal{D}\left(\hat{R}_{\mathcal{S}}(\theta), R(\theta)\right)}]}{\delta}\right)}{n}.$$

$$\begin{aligned}
n\mathcal{D}\left(\mathbb{E}_{\theta\sim\rho}\left[\hat{R}_{\mathcal{S}}(\theta)\right], \mathbb{E}_{\theta\sim\rho}[R(\theta)]\right) &\leq n \cdot \mathbb{E}_{\theta\sim\rho}\left[\mathcal{D}\left(\hat{R}_{\mathcal{S}}(\theta), R(\theta)\right)\right] \\
&\leq \mathrm{KL}(\rho\|\pi) + \log\left(\mathbb{E}_{\theta\sim\pi}[e^{n\mathcal{D}\left(\hat{R}_{\mathcal{S}}(\theta), R(\theta)\right)}]\right) \\
&\leq \mathrm{KL}(\rho\|\pi) + \log\left(\frac{\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\pi}[e^{n\mathcal{D}\left(\hat{R}_{\mathcal{S}}(\theta), R(\theta)\right)}]}{\delta}\right)
\end{aligned}$$

Germain's bound is a generalization of both McAllester's and Catoni's bounds (and many others) : if $\ell(\cdot,\cdot) \leq 1$,

McAllester [1999] : $\quad R(\rho) \leq \hat{R}_{\mathcal{S}}(\rho) + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{2n}}.$

Catoni [2003] : $\quad R(\rho) \leq \hat{R}_{\mathcal{S}}(\rho) + \frac{\mathrm{KL}(\rho\|\pi)}{\lambda} + \frac{\lambda}{8n} + \frac{\log\left(\frac{1}{\delta}\right)}{\lambda}.$

Towards a robust PAC-Bayes bound ?

Towards a robust PAC-Bayes bound ?

Catoni's PAC-Bayes bound [2003]

For any λ, with probability $\geq 1-\delta : \forall \rho \in \mathcal{P}(\Theta)$,

$$\left|\hat{R}_{\mathcal{S}}(\rho) - R(\rho)\right| \leq \frac{\mathrm{KL}(\rho\|\pi) + \log\left(\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\pi}[e^{\lambda\left|\hat{R}_{\mathcal{S}}(\theta)-R(\theta)\right|}]\right) + \log\left(\frac{1}{\delta}\right)}{\lambda}.$$

Towards a robust PAC-Bayes bound ?

Catoni's PAC-Bayes bound [2003]

For any λ, with probability $\geq 1 - \delta$: $\forall \rho \in \mathcal{P}(\Theta)$,

$$\left|\hat{R}_{\mathcal{S}}(\rho) - R(\rho)\right| \leq \frac{\mathrm{KL}(\rho\|\pi) + \log\left(\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\pi}[e^{\lambda\left|\hat{R}_{\mathcal{S}}(\theta)-R(\theta)\right|}]\right) + \log\left(\frac{1}{\delta}\right)}{\lambda}.$$

What if the exponential moment does not exist ?

Towards a robust PAC-Bayes bound ?

Catoni's PAC-Bayes bound [2003]

For any λ, with probability $\geq 1-\delta$: $\forall \rho \in \mathcal{P}(\Theta)$,

$$\left|\hat{R}_{\mathcal{S}}(\rho) - R(\rho)\right| \leq \frac{\mathrm{KL}(\rho\|\pi) + \log\left(\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\pi}[e^{\lambda\left|\hat{R}_{\mathcal{S}}(\theta)-R(\theta)\right|}]\right) + \log\left(\frac{1}{\delta}\right)}{\lambda}.$$

What if the exponential moment does not exist ?

We assume instead a much weaker assumption : for some integer q,

$$\mathcal{M}_q := \mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\pi}\left[\left|\hat{R}_{\mathcal{S}}(\theta) - R(\theta)\right|^q\right] < +\infty.$$

Towards a robust PAC-Bayes bound ?

Catoni's PAC-Bayes bound [2003]

For any λ, with probability $\geq 1-\delta$: $\forall \rho \in \mathcal{P}(\Theta)$,

$$\left|\hat{R}_{\mathcal{S}}(\rho) - R(\rho)\right| \leq \frac{\mathrm{KL}(\rho\|\pi) + \log\left(\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\pi}[e^{\lambda\left|\hat{R}_{\mathcal{S}}(\theta)-R(\theta)\right|}]\right) + \log\left(\frac{1}{\delta}\right)}{\lambda}.$$

What if the exponential moment does not exist ?

We assume instead a much weaker assumption : for some integer q,

$$\mathcal{M}_q := \mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\pi}\left[\left|\hat{R}_{\mathcal{S}}(\theta) - R(\theta)\right|^q\right] < +\infty.$$

To get a PAC-Bayes bound, we need to consider Csiszàr ϕ-divergences :
let ϕ be a convex function with $\phi(1) = 0$,

$$D_\phi(\rho\|\pi) := \mathbb{E}_\pi\left[\phi\left(\frac{d\rho}{d\pi}\right)\right],$$

when $\rho << \pi$ and $D_\phi(\rho\|\pi) = +\infty$ otherwise.

Towards a robust PAC-Bayes bound ?

Catoni's PAC-Bayes bound [2003]

For any λ, with probability $\geq 1-\delta$: $\forall \rho \in \mathcal{P}(\Theta)$,

$$\left|\hat{R}_{\mathcal{S}}(\rho) - R(\rho)\right| \leq \frac{\mathrm{KL}(\rho\|\pi) + \log\left(\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\pi}[e^{\lambda\left|\hat{R}_{\mathcal{S}}(\theta)-R(\theta)\right|}]\right) + \log\left(\frac{1}{\delta}\right)}{\lambda}.$$

What if the exponential moment does not exist ?

We assume instead a much weaker assumption : for some integer q,

$$\mathcal{M}_q := \mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\pi}\left[\left|\hat{R}_{\mathcal{S}}(\theta) - R(\theta)\right|^q\right] < +\infty.$$

To get a PAC-Bayes bound, we need to consider Csiszàr ϕ-divergences :
let ϕ be a convex function with $\phi(1) = 0$,

$$D_\phi(\rho\|\pi) := \mathbb{E}_\pi\left[\phi\left(\frac{d\rho}{d\pi}\right)\right],$$

when $\rho << \pi$ and $D_\phi(\rho\|\pi) = +\infty$ otherwise.
The KL is given by the special case $\mathrm{KL}(\rho\|\pi) = D_{x\log(x)}(\rho\|\pi)$.

A robust PAC-Bayes bound for heavy tails

Fix $p > 1$, $q = p/(p-1)$, $\delta \in (0,1)$ and let $\phi_p : x \mapsto x^p$.

A robust PAC-Bayes bound for heavy tails

Fix $p > 1$, $q = p/(p-1)$, $\delta \in (0,1)$ and let $\phi_p : x \mapsto x^p$.

Alquier & Guedj PAC-Bayes bound [2018]

With probability $\geq 1-\delta$: $\forall \rho \in \mathcal{P}(\Theta)$,

$$\left|\hat{R}_{\mathcal{S}}(\rho) - R(\rho)\right| \leq \left(\frac{\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\pi}\left[\left|\hat{R}_{\mathcal{S}}(\theta) - R(\theta)\right|^q\right]}{\delta} \right)^{\frac{1}{q}} \cdot \left(\mathbb{E}_{\pi}\left[\left(\frac{d\rho}{d\pi}\right)^p\right]\right)^{\frac{1}{p}}$$

A robust PAC-Bayes bound for heavy tails

Fix $p > 1$, $q = p/(p-1)$, $\delta \in (0,1)$ and let $\phi_p : x \mapsto x^p$.

Alquier & Guedj PAC-Bayes bound [2018]

With probability $\geq 1 - \delta$: $\forall \rho \in \mathcal{P}(\Theta)$,

$$\left|\hat{R}_{\mathcal{S}}(\rho) - R(\rho)\right| \leq \left(\frac{\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\pi}\left[\left|\hat{R}_{\mathcal{S}}(\theta) - R(\theta)\right|^q\right]}{\delta}\right)^{\frac{1}{q}} \cdot \left(\mathbb{E}_\pi\left[\left(\frac{d\rho}{d\pi}\right)^p\right]\right)^{\frac{1}{p}}$$

The bound decouples :

- The moment $\mathcal{M}_q$ (depending on the distribution of the data).
- The divergence $D_{\phi_p - 1}(\rho\|\pi) + 1$ (measure of complexity).

A robust PAC-Bayes bound for heavy tails

Fix $p > 1$, $q = p/(p-1)$, $\delta \in (0,1)$ and let $\phi_p : x \mapsto x^p$.

Alquier & Guedj PAC-Bayes bound [2018]

With probability $\geq 1 - \delta$: $\forall \rho \in \mathcal{P}(\Theta)$,

$$\left|\hat{R}_{\mathcal{S}}(\rho) - R(\rho)\right| \leq \left(\frac{\mathbb{E}_{\mathcal{S}} \mathbb{E}_{\theta \sim \pi} \left[\left| \hat{R}_{\mathcal{S}}(\theta) - R(\theta) \right|^q \right]}{\delta} \right)^{\frac{1}{q}} \cdot \left(\mathbb{E}_{\pi} \left[\left(\frac{d\rho}{d\pi} \right)^p \right] \right)^{\frac{1}{p}}$$

The bound decouples :

- The moment $\mathcal{M}_q$ (depending on the distribution of the data).
- The divergence $D_{\phi_p - 1}(\rho \| \pi) + 1$ (measure of complexity).

Note the weak dependence $\delta^{-1/q}$ vs $\sqrt{\log(1/\delta)}$ (there's no free lunch)...

A robust PAC-Bayes bound for heavy tails

Fix $p > 1$, $q = p/(p-1)$, $\delta \in (0,1)$ and let $\phi_p : x \mapsto x^p$.

Alquier & Guedj PAC-Bayes bound [2018]

With probability $\geq 1 - \delta$: $\forall \rho \in \mathcal{P}(\Theta)$,

$$\left|\hat{R}_{\mathcal{S}}(\rho) - R(\rho)\right| \leq \left(\frac{\mathbb{E}_{\mathcal{S}} \mathbb{E}_{\theta \sim \pi} \left[\left| \hat{R}_{\mathcal{S}}(\theta) - R(\theta) \right|^q \right]}{\delta} \right)^{\frac{1}{q}} \cdot \left(\mathbb{E}_{\pi} \left[\left(\frac{d\rho}{d\pi} \right)^p \right] \right)^{\frac{1}{p}}$$

The bound decouples :

- The moment $\mathcal{M}_q$ (depending on the distribution of the data).
- The divergence $D_{\phi_p - 1}(\rho \| \pi) + 1$ (measure of complexity).

Note the weak dependence $\delta^{-1/q}$ vs $\sqrt{\log(1/\delta)}$ (there's no free lunch)...

For $p = q = 2$, for $\mathcal{V} := \mathbb{E}_{\theta \sim \pi} \mathbb{V}_{(X,Y) \sim \mathbb{P}}[\ell(f_\theta(x), y)] < +\infty$,

A robust PAC-Bayes bound for heavy tails

Fix $p > 1$, $q = p/(p-1)$, $\delta \in (0,1)$ and let $\phi_p : x \mapsto x^p$.

Alquier & Guedj PAC-Bayes bound [2018]

With probability $\geq 1-\delta$: $\forall \rho \in \mathcal{P}(\Theta)$,

$$\left|\hat{R}_{\mathcal{S}}(\rho) - R(\rho)\right| \leq \left(\frac{\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\pi}\left[\left|\hat{R}_{\mathcal{S}}(\theta) - R(\theta)\right|^q\right]}{\delta}\right)^{\frac{1}{q}} \cdot \left(\mathbb{E}_{\pi}\left[\left(\frac{d\rho}{d\pi}\right)^p\right]\right)^{\frac{1}{p}}$$

The bound decouples :

- The moment $\mathcal{M}_q$ (depending on the distribution of the data).
- The divergence $D_{\phi_p-1}(\rho\|\pi)+1$ (measure of complexity).

Note the weak dependence $\delta^{-1/q}$ vs $\sqrt{\log(1/\delta)}$ (there's no free lunch)...

For $p = q = 2$, for $\mathcal{V} := \mathbb{E}_{\theta\sim\pi}\mathbb{V}_{(X,Y)\sim\mathbb{P}}[\ell(f_\theta(x), y)] < +\infty$, w.p. $\geq 1-\delta$,

$$\forall \rho \in \mathcal{P}(\Theta), \quad R(\rho) \leq \hat{R}_{\mathcal{S}}(\rho) + \sqrt{\frac{\mathcal{V}\left(1+\chi^2(\rho\|\pi)\right)}{n\delta}}.$$

Proof of Alquier & Guedj's bound

Alquier & Guedj PAC-Bayes bound [2018]

With probability $\geq 1-\delta$: $\forall \rho \in \mathcal{P}(\Theta)$,

$$\left|\hat{R}_{\mathcal{S}}(\rho) - R(\rho)\right| \leq \left(\frac{\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\pi}\left[\left|\hat{R}_{\mathcal{S}}(\theta) - R(\theta)\right|^{q}\right]}{\delta}\right)^{\frac{1}{q}} \cdot \left(\mathbb{E}_{\pi}\left[\left(\frac{d\rho}{d\pi}\right)^{p}\right]\right)^{\frac{1}{p}}.$$

Proof of Alquier & Guedj's bound

Alquier & Guedj PAC-Bayes bound [2018]

With probability $\geq 1-\delta$: $\forall \rho \in \mathcal{P}(\Theta)$,

$$\left|\hat{R}_{\mathcal{S}}(\rho) - R(\rho)\right| \leq \left(\frac{\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\pi}\left[\left|\hat{R}_{\mathcal{S}}(\theta) - R(\theta)\right|^q\right]}{\delta}\right)^{\frac{1}{q}} \cdot \left(\mathbb{E}_{\pi}\left[\left(\frac{d\rho}{d\pi}\right)^p\right]\right)^{\frac{1}{p}}.$$

$$\begin{aligned}
\left|\mathbb{E}_{\theta\sim\rho}\left[\hat{R}_{\mathcal{S}}(\theta)\right] - \mathbb{E}_{\theta\sim\rho}\left[R(\theta)\right]\right| &\leq \mathbb{E}_{\theta\sim\rho}\left[\left|\hat{R}_{\mathcal{S}}(\theta) - R(\theta)\right|\right] \\
&= \mathbb{E}_{\theta\sim\pi}\left[\left|\hat{R}_{\mathcal{S}}(\theta) - R(\theta)\right| \frac{d\rho}{d\pi}(\theta)\right] \\
&\leq \mathbb{E}_{\theta\sim\pi}\left[\left|\hat{R}_{\mathcal{S}}(\theta) - R(\theta)\right|^q\right]^{\frac{1}{q}} \cdot \mathbb{E}_{\pi}\left[\left(\frac{d\rho}{d\pi}\right)^p\right]^{\frac{1}{p}} \\
&\leq \frac{\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\pi}\left[\left|\hat{R}_{\mathcal{S}}(\theta) - R(\theta)\right|^q\right]^{\frac{1}{q}}}{\delta^{\frac{1}{q}}} \cdot \mathbb{E}_{\pi}\left[\left(\frac{d\rho}{d\pi}\right)^p\right]^{\frac{1}{p}}
\end{aligned}$$

PAC-Bayes bounds achieving fast rates

Reminder on Tolstikhin's bound

Reminder on Tolstikhin's bound

Tolstikhin's bound

With proba $\geq 1-\delta$ ($\ell(\cdot,\cdot)\leq 1$),

$$\forall \rho \in \mathcal{P}(\Theta), \;\; R(\rho) \leq \hat{R}_{\mathcal{S}}(\rho) + \sqrt{2\hat{R}_{\mathcal{S}}(\rho)\frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{n}} + 2\frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{n}.$$

Reminder on Tolstikhin's bound

Tolstikhin's bound

With proba $\geq 1-\delta$ $(\ell(\cdot,\cdot)\leq 1)$,

$$\forall \rho \in \mathcal{P}(\Theta),\;\; R(\rho) \leq \hat{R}_{\mathcal{S}}(\rho) + \sqrt{2\hat{R}_{\mathcal{S}}(\rho)\frac{\mathrm{KL}(\rho\|\pi)+\log\left(\frac{2\sqrt{n}}{\delta}\right)}{n}} + 2\frac{\mathrm{KL}(\rho\|\pi)+\log\left(\frac{2\sqrt{n}}{\delta}\right)}{n}.$$

We get this bound from Seeger's one via a refinement of Pinsker's inequality $\mathrm{kl}(p\|q)\geq (p-q)^2/2q$ i.e. $\mathrm{kl}^{-1}(q\|b)\leq q+\sqrt{2qb}+2b$, hence improving over McAllester's bound.

Reminder on Tolstikhin's bound

Tolstikhin's bound

With proba $\geq 1-\delta$ ($\ell(\cdot,\cdot) \leq 1$),

$$\forall \rho \in \mathcal{P}(\Theta), \quad R(\rho) \leq \hat{R}_{\mathcal{S}}(\rho) + \sqrt{2\hat{R}_{\mathcal{S}}(\rho) \frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{n}} + 2\frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{n}.$$

We get this bound from Seeger's one via a refinement of Pinsker's inequality $\mathrm{kl}(p\|q) \geq (p-q)^2/2q$ i.e. $\mathrm{kl}^{-1}(q\|b) \leq q + \sqrt{2qb} + 2b$, hence improving over McAllester's bound.

An amazing characteristic of this bound : while its order is in general in $1/\sqrt{n}$, as all the PAC-Bayes bounds seen so far, the dependence drops to $1/n$ in the noiseless case where $\hat{R}_{\mathcal{S}}(\rho) = 0$.

Reminder on Tolstikhin's bound

Tolstikhin's bound

With proba $\geq 1-\delta$ ($\ell(\cdot,\cdot) \leq 1$),

$$\forall \rho \in \mathcal{P}(\Theta), \quad R(\rho) \leq \hat{R}_{\mathcal{S}}(\rho) + \sqrt{2\hat{R}_{\mathcal{S}}(\rho)\frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{n}} + 2\frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{n}.$$

We get this bound from Seeger's one via a refinement of Pinsker's inequality $\mathrm{kl}(p\|q) \geq (p-q)^2/2q$ i.e. $\mathrm{kl}^{-1}(q\|b) \leq q + \sqrt{2qb} + 2b$, hence improving over McAllester's bound.

An amazing characteristic of this bound : while its order is in general in $1/\sqrt{n}$, as all the PAC-Bayes bounds seen so far, the dependence drops to $1/n$ in the noiseless case where $\hat{R}_{\mathcal{S}}(\rho) = 0$.

Question : is it possible to achieve fast rates more generally ?

Reminder on Tolstikhin's bound

Tolstikhin's bound

With proba $\geq 1-\delta$ ($\ell(\cdot,\cdot) \leq 1$),

$$\forall \rho \in \mathcal{P}(\Theta), \quad R(\rho) \leq \hat{R}_{\mathcal{S}}(\rho) + \sqrt{2\hat{R}_{\mathcal{S}}(\rho)\frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{n}} + 2\frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{n}.$$

We get this bound from Seeger's one via a refinement of Pinsker's inequality $\mathrm{kl}(p\|q) \geq (p-q)^2/2q$ i.e. $\mathrm{kl}^{-1}(q\|b) \leq q + \sqrt{2qb} + 2b$, hence improving over McAllester's bound.

An amazing characteristic of this bound : while its order is in general in $1/\sqrt{n}$, as all the PAC-Bayes bounds seen so far, the dependence drops to $1/n$ in the noiseless case where $\hat{R}_{\mathcal{S}}(\rho) = 0$.

Question : is it possible to achieve fast rates more generally ? Yes ! Under some specific required assumptions.

Oracle inequalities : controlling the excess risk

Let's start in a deterministic setting where we want to quantify the performance of $f_{\hat{\theta}}$ via ***excess risk bounds/oracle inequalities***.

Oracle inequalities : controlling the excess risk

Let's start in a deterministic setting where we want to quantify the performance of $f_{\hat{\theta}}$ via ***excess risk bounds/oracle inequalities***.

Given $\mathcal{S}$, how well does $f_{\hat{\theta}}$ predict on unseen data ?

$$R(\hat{\theta}) = \mathbb{E}\left[\ell(f_{\hat{\theta}}(X), Y) \middle| \mathcal{S}\right].$$

Oracle inequalities : controlling the excess risk

Let's start in a deterministic setting where we want to quantify the performance of $f_{\hat{\theta}}$ via ***excess risk bounds/oracle inequalities***.

Given $\mathcal{S}$, how well does $f_{\hat{\theta}}$ predict on unseen data ?

$$R(\hat{\theta}) = \mathbb{E}\left[\ell(f_{\hat{\theta}}(X), Y) \middle| \mathcal{S}\right].$$

What is the best predictor I could (should ?) have chosen ?

$$R(\theta^*) = \inf_{\theta \in \Theta} \mathbb{E}\Big[\ell\left(f_\theta(X), Y\right)\Big].$$

Oracle inequalities : controlling the excess risk

Let's start in a deterministic setting where we want to quantify the performance of $f_{\hat{\theta}}$ via ***excess risk bounds/oracle inequalities***.

Given $\mathcal{S}$, how well does $f_{\hat{\theta}}$ predict on unseen data ?

$$R(\hat{\theta}) = \mathbb{E}\left[\ell(f_{\hat{\theta}}(X), Y) \middle| \mathcal{S}\right].$$

What is the best predictor I could (should ?) have chosen ?

$$R(\theta^*) = \inf_{\theta \in \Theta} \mathbb{E}\Big[\ell\left(f_\theta(X), Y\right)\Big].$$

With high probability, the excess risk of $f_{\hat{\theta}}$ within Θ is expected to be (for bounded losses), for any δ, of order,

$$\mathbb{P}_{\mathcal{S}}\left[R(\hat{\theta}) - \inf_{\theta\in\Theta} R(\theta) \lesssim \sqrt{\frac{\text{comp}(\Theta)}{n}} \times \sqrt{\log\left(\frac{1}{\delta}\right)}\right] \geq 1 - \delta.$$

Oracle inequalities : controlling the excess risk

Let's start in a deterministic setting where we want to quantify the performance of $f_{\hat{\theta}}$ via ***excess risk bounds/oracle inequalities***.

Given $\mathcal{S}$, how well does $f_{\hat{\theta}}$ predict on unseen data ?

$$R(\hat{\theta}) = \mathbb{E}\left[\ell(f_{\hat{\theta}}(X), Y) \middle| \mathcal{S}\right].$$

What is the best predictor I could (should ?) have chosen ?

$$R(\theta^*) = \inf_{\theta \in \Theta} \mathbb{E}\Big[\ell\left(f_\theta(X), Y\right)\Big].$$

With high probability, the excess risk of $f_{\hat{\theta}}$ within Θ is expected to be (for bounded losses), for any δ, of order,

$$\mathbb{P}_{\mathcal{S}}\left[R(\hat{\theta}) - \inf_{\theta \in \Theta} R(\theta) \lesssim \sqrt{\frac{\mathrm{comp}(\Theta)}{n}} \times \sqrt{\log\left(\frac{1}{\delta}\right)}\right] \geq 1 - \delta.$$

Is it possible to achieve faster rates for bounded losses ?

Oracle inequalities : controlling the excess risk

Let's start in a deterministic setting where we want to quantify the performance of $f_{\hat{\theta}}$ via ***excess risk bounds/oracle inequalities***.

Given $\mathcal{S}$, how well does $f_{\hat{\theta}}$ predict on unseen data ?

$$R(\hat{\theta}) = \mathbb{E}\left[\ell(f_{\hat{\theta}}(X), Y) \middle| \mathcal{S}\right].$$

What is the best predictor I could (should ?) have chosen ?

$$R(\theta^*) = \inf_{\theta \in \Theta} \mathbb{E}\left[\ell\left(f_\theta(X), Y\right)\right].$$

With high probability, the excess risk of $f_{\hat{\theta}}$ within Θ is expected to be (for bounded losses), for any δ, of order,

$$\mathbb{P}_{\mathcal{S}}\left[R(\hat{\theta}) - \inf_{\theta \in \Theta} R(\theta) \lesssim \sqrt{\frac{\mathrm{comp}(\Theta)}{n}} \times \sqrt{\log\left(\frac{1}{\delta}\right)}\right] \geq 1 - \delta.$$

Is it possible to achieve faster rates for bounded losses ? *Yes ! Under further assumptions on the "easiness" of the problem.*

Faster rates of convergence

Actually, the optimal excess risk of a rule $f_{\hat{\theta}}$ is usually of order

$$R(\hat{\theta}) - \inf_{\theta \in \Theta} R(\theta) \lesssim \left(\frac{\mathrm{comp}(\Theta)}{n} \right)^{\gamma}$$

where $\gamma \in \left[\frac{1}{2}, 1\right]$ reflects the easiness of the problem $(\mathbb{P}, \ell, \Theta)$.

Faster rates of convergence

Actually, the optimal excess risk of a rule $f_{\hat{\theta}}$ is usually of order

$$R(\hat{\theta}) - \inf_{\theta \in \Theta} R(\theta) \lesssim \left(\frac{\text{comp}(\Theta)}{n} \right)^{\gamma}$$

where $\gamma \in \left[\frac{1}{2}, 1\right]$ reflects the easiness of the problem $(\mathbb{P}, \ell, \Theta)$.

- If the problem is "hard" (e.g. no specific assumption), $\gamma = 1/2$ and :

$$R(\hat{\theta}) - \inf_{\theta \in \Theta} R(\theta) \lesssim \sqrt{\frac{\text{comp}(\Theta)}{n}}.$$

Faster rates of convergence

Actually, the optimal excess risk of a rule $f_{\hat{\theta}}$ is usually of order

$$R(\hat{\theta}) - \inf_{\theta \in \Theta} R(\theta) \lesssim \left(\frac{\text{comp}(\Theta)}{n} \right)^{\gamma}$$

where $\gamma \in \left[\frac{1}{2}, 1\right]$ reflects the easiness of the problem $(\mathbb{P}, \ell, \Theta)$.

- If the problem is "hard" (e.g. no specific assumption), $\gamma = 1/2$ and :

$$R(\hat{\theta}) - \inf_{\theta \in \Theta} R(\theta) \lesssim \sqrt{\frac{\text{comp}(\Theta)}{n}}.$$

- If the problem is "easy" (e.g. noiseless/low-noise settings), $\gamma = 1$ and :

$$R(\hat{\theta}) - \inf_{\theta \in \Theta} R(\theta) \lesssim \frac{\text{comp}(\Theta)}{n}.$$

Fast rates in classification

Consider the following binary classification problem $(\mathbb{P}, \ell, \Theta)$:

Fast rates in classification

Consider the following binary classification problem $(\mathbb{P}, \ell, \Theta)$:

- $\ell(y', y) = \mathbb{1}(y' \neq y)$ so that $R(\theta) = \mathbb{P}\left(f_\theta(X) \neq Y\right)$

Fast rates in classification

Consider the following binary classification problem $(\mathbb{P}, \ell, \Theta)$:

- $\ell(y', y) = \mathbb{1}(y' \neq y)$ so that $R(\theta) = \mathbb{P}\left(f_\theta(X) \neq Y\right)$
- $f^*(x) = \arg\max_y \mathbb{P}(Y = y | X = x) \in \{f_\theta; \theta \in \Theta\}$ *(assumption)*

Fast rates in classification

Consider the following binary classification problem $(\mathbb{P}, \ell, \Theta)$:

- $\ell(y', y) = \mathbb{1}(y' \neq y)$ so that $R(\theta) = \mathbb{P}\left(f_\theta(X) \neq Y\right)$
- $f^*(x) = \arg\max_y \mathbb{P}(Y = y | X = x) \in \{f_\theta; \theta \in \Theta\}$ *(assumption)*

Question : when is the problem difficult ?

Fast rates in classification

Consider the following binary classification problem $(\mathbb{P}, \ell, \Theta)$:

- $\ell(y', y) = \mathbb{1}(y' \neq y)$ so that $R(\theta) = \mathbb{P}\left(f_\theta(X) \neq Y\right)$
- $f^*(x) = \arg\max_y \mathbb{P}(Y = y | X = x) \in \{f_\theta; \theta \in \Theta\}$ *(assumption)*

Question : when is the problem difficult ?

- Answer : when $\mathbb{P}(Y = 1|X)$ very close to $1/2$! But then learning is (almost) useless...

Fast rates in classification

Consider the following binary classification problem $(\mathbb{P}, \ell, \Theta)$:

- $\ell(y', y) = \mathbb{1}(y' \neq y)$ so that $R(\theta) = \mathbb{P}\left(f_\theta(X) \neq Y\right)$
- $f^*(x) = \arg\max_y \mathbb{P}(Y = y|X = x) \in \{f_\theta; \theta \in \Theta\}$ *(assumption)*

Question : when is the problem difficult ?

- Answer : when $\mathbb{P}(Y = 1|X)$ very close to $1/2$! But then learning is (almost) useless...
- Solution : assume that $\mathbb{P}(Y = 1|X)$ not too close to $1/2$!

Fast rates in classification

Consider the following binary classification problem $(\mathbb{P}, \ell, \Theta)$:

- $\ell(y', y) = \mathbb{1}(y' \neq y)$ so that $R(\theta) = \mathbb{P}\left(f_\theta(X) \neq Y\right)$
- $f^*(x) = \arg\max_y \mathbb{P}(Y = y | X = x) \in \{f_\theta; \theta \in \Theta\}$ *(assumption)*

Question : when is the problem difficult ?

- Answer : when $\mathbb{P}(Y = 1|X)$ very close to $1/2$! But then learning is (almost) useless...
- Solution : assume that $\mathbb{P}(Y = 1|X)$ not too close to $1/2$!

$$\text{w.h.p. over } X \quad , \quad \left|\mathbb{P}(Y = 1|X) - \frac{1}{2}\right| \text{ is large.}$$

The margin condition in classification

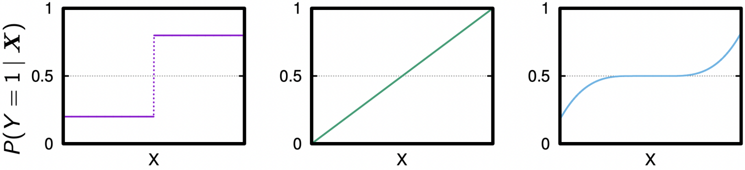
Tsybakov's α-margin condition [Tsybakov, AoS 2004]

$$\mathbb{P}_X\left[\left|\mathbb{P}(Y=1|X)-\frac{1}{2}\right|\leq t\right]\leq ct^{\alpha}.$$

The margin condition in classification

Tsybakov's α-margin condition [Tsybakov, AoS 2004]

$$\mathbb{P}_X\left[\left|\mathbb{P}(Y=1|X)-\frac{1}{2}\right|\leq t\right]\leq ct^{\alpha}.$$

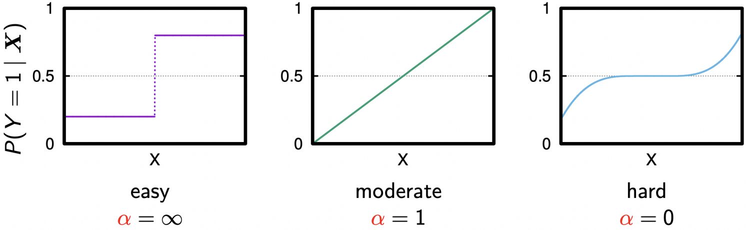


easy $\alpha=\infty$ | moderate $\alpha=1$ | hard $\alpha=0$

The margin condition in classification

Tsybakov's α-margin condition [Tsybakov, AoS 2004]

$$\mathbb{P}_X\left[\left|\mathbb{P}(Y=1|X)-\frac{1}{2}\right|\leq t\right]\leq ct^{\alpha}.$$



$$R(\hat{\theta})-\inf_{\theta\in\Theta}R(\theta)\lesssim\left(\frac{\mathrm{comp}(\Theta)}{n}\right)^{\frac{1+\alpha}{2+\alpha}}.$$

The general setting : Bernstein condition

Bernstein's condition [Bartlett & Mendelson, PTFR 2006]

For some $\beta \in [0, 1]$ and $B > 0$, with the notation $\ell_\theta = \ell(f_\theta(X), Y)$,

$$\mathbb{E}\left[(\ell_\theta - \ell_{\theta^*})^2\right] \leq B\mathbb{E}\left[\ell_\theta - \ell_{\theta^*}\right]^\beta.$$

The general setting : Bernstein condition

Bernstein's condition [Bartlett & Mendelson, PTFR 2006]

For some $\beta \in [0, 1]$ and $B > 0$, with the notation $\ell_\theta = \ell\left(f_\theta(X), Y\right)$,

$$\mathbb{E}\left[\left(\ell_\theta - \ell_{\theta^*}\right)^2\right] \leq B\mathbb{E}\left[\ell_\theta - \ell_{\theta^*}\right]^\beta .$$

Under β-Bernstein condition,

$$R(\hat{\theta}) - \inf_{\theta \in \Theta} R(\theta) \lesssim \left(\frac{\text{comp}(\Theta)}{n}\right)^{\frac{1}{2-\beta}} .$$

The general setting : Bernstein condition

Bernstein's condition [Bartlett & Mendelson, PTFR 2006]

For some $\beta \in [0,1]$ and $B > 0$, with the notation $\ell_\theta = \ell\left(f_\theta(X), Y\right)$,

$$\mathbb{E}\left[\left(\ell_\theta - \ell_{\theta^*}\right)^2\right] \leq B\mathbb{E}\left[\ell_\theta - \ell_{\theta^*}\right]^\beta .$$

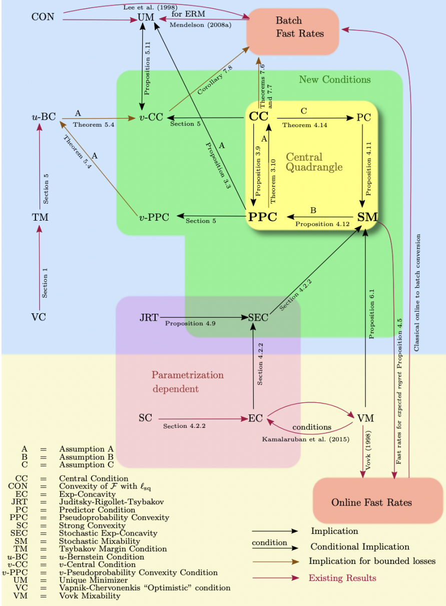
Under β-Bernstein condition,

$$R(\hat{\theta}) - \inf_{\theta \in \Theta} R(\theta) \lesssim \left(\frac{\text{comp}(\Theta)}{n}\right)^{\frac{1}{2-\beta}} .$$

Bernstein condition is satisfied in the following settings :

- In noiseless classification $R(\theta^*) = 0$, with $\beta = 1$.
- Under Tsybakov's α-margin assumption, with $\beta = \frac{\alpha}{1+\alpha}$.
- For Lipschitz and strongly-convex loss functions, with $\beta = 1$.

Welcome to the zoo [Van Erven et al., JMLR 2015]



How about fast PAC-Bayes bounds ?

Reminder on the Gibbs posterior :

$$\hat{\rho}_\lambda(d\theta) := \arg\min_{\rho \in \mathcal{P}(\Theta)} \left\{ \hat{R}_{\mathcal{S}}(\rho) + \frac{\mathrm{KL}(\rho \| \pi)}{\lambda} \right\} \propto \exp\left(-\lambda \sum_{i=1}^{n} \ell(f_\theta(X_i), Y_i) \right) \pi(d\theta).$$

How about fast PAC-Bayes bounds ?

Reminder on the Gibbs posterior :

$$\hat{\rho}_\lambda(d\theta) := \arg\min_{\rho \in \mathcal{P}(\Theta)} \left\{ \hat{R}_{\mathcal{S}}(\rho) + \frac{\mathrm{KL}(\rho\|\pi)}{\lambda} \right\} \propto \exp\left(-\lambda \sum_{i=1}^{n} \ell(f_\theta(X_i), Y_i) \right) \pi(d\theta).$$

We denote $\mathcal{R}(\theta) = R(\theta) - \inf_\Theta R$ the excess risk, and assume $\ell(\cdot,\cdot) \leq 1$.

How about fast PAC-Bayes bounds ?

Reminder on the Gibbs posterior :

$$\hat{\rho}_\lambda(d\theta) := \arg\min_{\rho\in\mathcal{P}(\Theta)} \left\{ \hat{R}_{\mathcal{S}}(\rho) + \frac{\mathrm{KL}(\rho\|\pi)}{\lambda} \right\} \propto \exp\left(-\lambda \sum_{i=1}^{n} \ell(f_\theta(X_i), Y_i) \right) \pi(d\theta).$$

We denote $\mathcal{R}(\theta) = R(\theta) - \inf_\Theta R$ the excess risk, and assume $\ell(\cdot,\cdot) \leq 1$.

Slow excess risk bound (in expectation) for the Gibbs posterior :

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

How about fast PAC-Bayes bounds ?

Reminder on the Gibbs posterior :

$$\hat{\rho}_\lambda(d\theta) := \arg\min_{\rho\in\mathcal{P}(\Theta)} \left\{ \hat{R}_{\mathcal{S}}(\rho) + \frac{\mathrm{KL}(\rho\|\pi)}{\lambda} \right\} \propto \exp\left(-\lambda \sum_{i=1}^{n} \ell(f_\theta(X_i), Y_i) \right) \pi(d\theta).$$

We denote $\mathcal{R}(\theta) = R(\theta) - \inf_\Theta R$ the excess risk, and assume $\ell(\cdot,\cdot) \leq 1$.

Slow excess risk bound (in expectation) for the Gibbs posterior :

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

Fast excess risk bound (in expectation) for the Gibbs posterior under Bernstein's condition ($\beta = 1$) :

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

How about fast PAC-Bayes bounds ?

Reminder on the Gibbs posterior :

$$\hat{\rho}_\lambda(d\theta) := \arg\min_{\rho\in\mathcal{P}(\Theta)} \left\{ \hat{R}_{\mathcal{S}}(\rho) + \frac{\mathrm{KL}(\rho\|\pi)}{\lambda} \right\} \propto \exp\left(-\lambda \sum_{i=1}^{n} \ell(f_\theta(X_i), Y_i) \right) \pi(d\theta).$$

We denote $\mathcal{R}(\theta) = R(\theta) - \inf_\Theta R$ the excess risk, and assume $\ell(\cdot,\cdot) \le 1$.

Slow excess risk bound (in expectation) for the Gibbs posterior :

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

Fast excess risk bound (in expectation) for the Gibbs posterior under Bernstein's condition ($\beta = 1$) :

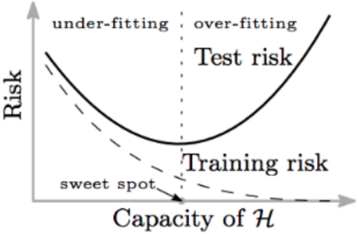
$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\hat{\rho}_\lambda}\left[\mathcal{R}(\theta)\right] \le \frac{1}{1 - \frac{B\lambda/n}{2(1-\lambda/n)}} \mathbb{E}_{\mathcal{S}}\left[\inf_{\rho\in\mathcal{P}(\Theta)} \left\{ \mathbb{E}_{\theta\sim\rho}\left[\mathcal{R}(\theta)\right] + \frac{\mathrm{KL}(\rho\|\pi)}{\lambda} \right\} \right].$$

For $\lambda = n/(1+B)$, under Bernstein's condition ($\beta = 1$) :

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\hat{\rho}_\lambda}\left[\mathcal{R}(\theta)\right] \le 2\cdot \mathbb{E}_{\mathcal{S}}\left[\inf_{\rho\in\mathcal{P}(\Theta)} \left\{ \mathbb{E}_{\theta\sim\rho}\left[\mathcal{R}(\theta)\right] + \frac{(1+B)\mathrm{KL}(\rho\|\pi)}{n} \right\} \right].$$

Towards Tight Certificates in Deep Learning

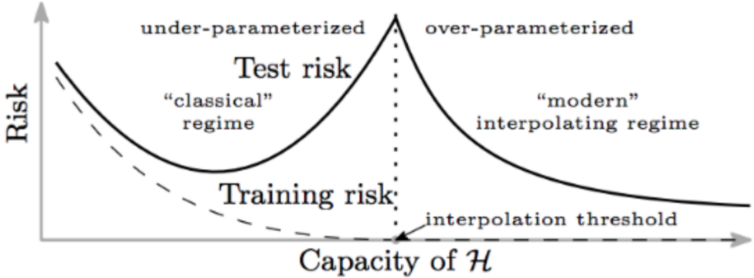
Rethinking generalization with DL



(a)

- Many parameters !
- NNs trained with SGD achieve 0 training error.
- NNs can overfit but in practice don't : **why** ?
- **Hypothesis** : complexity « number of parameters.

Rethinking generalization with DL



(b)

A breakthrough : [Dziugaite and Roy, 2017]

[Dziugaite and Roy, 2017] achieved non-vacuous bounds on binary MNIST using PAC-Bayes bounds (≈ 0.2 vs 0.03 test error).

A breakthrough : [Dziugaite and Roy, 2017]

[Dziugaite and Roy, 2017] achieved non-vacuous bounds on binary MNIST using PAC-Bayes bounds (≈ 0.2 vs 0.03 test error).

- Choose a Gaussian posterior $\rho_{w,s^2} = \mathcal{N}(w, s^2 I_p)$ and minimize McAllester's PAC-Bayes bound wrt (w, s^2).
- Upper bound the 0-1 loss by a convex, Lipschitz upper bound in order to make the bound easier to minimize $\mathbb{1}(f_\theta(x) \neq y) \leq \log(1 + e^{-yf_\theta(x)}) / \log(2)$.
- Use SGD to solve the optimization problem (importance of achieving flat minima).
- Important : use a data-dependent prior ! Optimize the prior variance (union bound argument), mean equal to 0 or randomly chosen.
- Do not use $\mathrm{kl}\left(\hat{R}_{\mathcal{S}}(\rho), R(\rho)\right) \geq 2\left(R(\rho) - \hat{R}_{\mathcal{S}}(\rho)\right)^2$ but (right) invert the kl directly $\rightarrow$ evaluate the subsequent bound at $\hat{\rho}_{\mathcal{S}}$.

Two different bounds

Langford & Seeger's PAC-Bayes bound

With probability $\geq 1 - \delta$,

$$\forall \rho \in \mathcal{P}(\mathcal{F}), \quad R(\rho) \leq \mathrm{kl}^{-1}\left(\hat{R}_{\mathcal{S}}(\rho) \middle\| \frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{n} \right).$$

Two different bounds

Langford & Seeger's PAC-Bayes bound

With probability $\geq 1-\delta$,

$$\forall \rho \in \mathcal{P}(\mathcal{F}), \quad R(\rho) \leq \mathrm{kl}^{-1}\left(\hat{R}_{\mathcal{S}}(\rho) \middle\| \frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{n} \right).$$

This leads to two different PAC-Bayes bounds :

- A bound for the training stage (not tight) : wp $\geq 1-\delta$ over data samples, uniformly over $\rho \in \mathcal{P}(\mathcal{F})$,

$$R^{x}(\rho) \leq \hat{R}^{x}_{\mathcal{S}}(\rho) + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{2n}}.$$

- A bound for the evaluation stage (not practical) : wp $\geq 1-\delta-\delta'$ over data + MC samples, uniformly over $\rho \in \mathcal{P}(\mathcal{F})$,

$$R^{01}(\rho) \leq \mathrm{kl}^{-1}\left(\mathrm{kl}^{-1}\left(\hat{R}^{01}_{\mathcal{S}}(\tilde{\rho}_m), \frac{\log\left(\frac{2\sqrt{m}}{\delta'}\right)}{m} \right), \frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{2n} \right).$$

Algorithm 1 PAC-Bayes with Backprop (PBB)

Input:

μ_0 ▷ Prior center parameters (random init.)

ρ_0 ▷ Prior scale hyper-parameter

$Z_{1:n}$ ▷ Training examples (inputs + labels)

$\delta \in (0,1)$ ▷ Confidence parameter

$\alpha \in (0,1), T$ ▷ Learning rate; # of iterations

Output: Optimal μ, ρ ▷ Centers, scales

procedure PB_QUAD_GAUSS

  $\mu \leftarrow \mu_0$ ▷ Set posterior centers to init. of prior

  $\rho \leftarrow \rho_0$ ▷ Set posterior scale to ρ_0 hyperparam.

  for $t \leftarrow 1 : T$ **do** ▷ Run SGD for T iterations.

    Sample $V \sim \mathcal{N}(0, I)$

    $W = \mu + \log(1 + \exp(\rho)) \odot V$

    $f(\mu, \rho) = f_{\text{quad}}(Z_{1:n}, W, \mu, \rho, \mu_0, \rho_0, \delta)$

    SGD gradient step using $\begin{bmatrix} \nabla_\mu f \\ \nabla_\rho f \end{bmatrix}$

  return μ, ρ

PAC-Bayes workflow [Pérez-Ortiz et al., 2021]

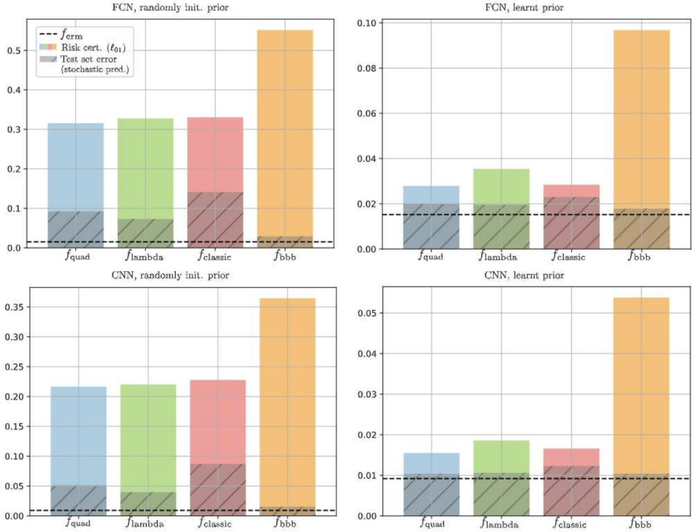
- Split the dataset in two separate subsets $\mathcal{S} = \mathcal{S}_{\text{prior}} \cup \mathcal{S}_{\text{eval}}$.
- Learn the prior using $\mathcal{S}_{\text{prior}}$ by ERM with dropout.
- Learn the posterior using the whole dataset $\mathcal{S}$,

$$\min_{\rho} \left\{ \hat{R}^{x}_{\mathcal{S}}(\rho) + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{2\sqrt{n}}{\delta}\right)}{2n}} \right\}.$$

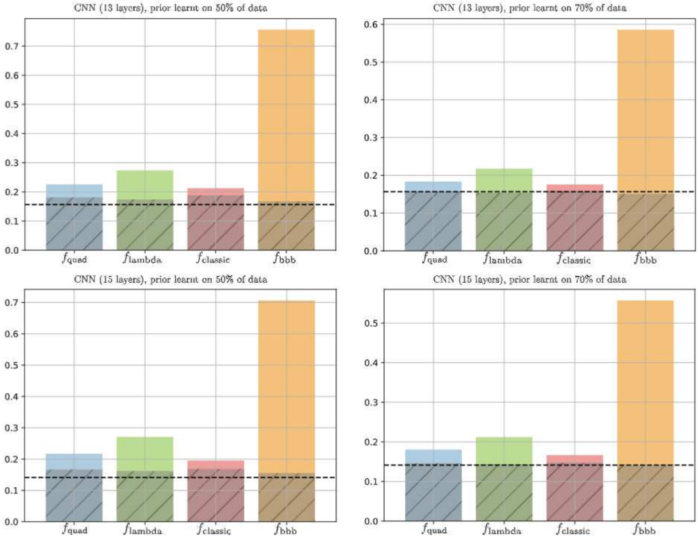
- Evaluate the bound at the learned posterior ρ using $\mathcal{S}_{\text{eval}}$,

$$\mathrm{kl}^{-1}\left(\mathrm{kl}^{-1}\left(\hat{R}^{01}_{\mathcal{S}}(\tilde{\rho}_m), \frac{\log\left(\frac{2}{\delta'}\right)}{m} \right), \frac{\mathrm{KL}(\rho\|\pi) + \log\left(\frac{2\sqrt{|\mathcal{S}_{\text{eval}}|}}{\delta}\right)}{2|\mathcal{S}_{\text{eval}}|} \right).$$

MNIST experiments [Pérez-Ortiz et al., 2021]



CIFAR 10 experiments [Pérez-Ortiz et al., 2021]



Conclusions of [Pérez-Ortiz et al., 2021]

- Model selection feasible without data splitting.
- Non-vacuous and tight bounds achievable.
- Choosing a prior centered at the ERM is key.
- Different trade-offs between test error and risk certificate.
- Extensive experiments for FCNs and CNNs.
- How about specific models ?
- How about different learning strategies ?

Generalization bounds for SGD using information bounds

Stochastic Gradient Descent

SGD algorithm :

$$\theta_{t+1} = \theta_t - \eta_t g(\theta_t; X_{I_t}, X_{I_t})$$

where η_t is the learning rate, I_t is the index set of minibatch of datapoints (ind. of $\mathcal{S}$) at step t, and $g(\theta; x, y) = \nabla_\theta \ell(f_\theta(x), y)$ is the gradient (averaged over the minibatch). The stepsize and sampling rule are fixed but arbitrary.

Stochastic Gradient Descent

SGD algorithm :

$$\theta_{t+1} = \theta_t - \eta_t g(\theta_t; X_{I_t}, X_{I_t})$$

where η_t is the learning rate, I_t is the index set of minibatch of datapoints (ind. of $\mathcal{S}$) at step t, and $g(\theta; x, y) = \nabla_\theta \ell(f_\theta(x), y)$ is the gradient (averaged over the minibatch). The stepsize and sampling rule are fixed but arbitrary.

(Expected) generalization gap of SGD

$$\mathbb{E}_{\mathcal{S}}\left[R(\theta_T) - \hat{R}_{\mathcal{S}}(\theta_T)\right] \leq \ ?$$

Stochastic Gradient Descent

SGD algorithm :

$$\theta_{t+1} = \theta_t - \eta_t g(\theta_t; X_{I_t}, X_{I_t})$$

where η_t is the learning rate, I_t is the index set of minibatch of datapoints (ind. of $\mathcal{S}$) at step t, and $g(\theta; x, y) = \nabla_\theta \ell(f_\theta(x), y)$ is the gradient (averaged over the minibatch). The stepsize and sampling rule are fixed but arbitrary.

(Expected) generalization gap of SGD

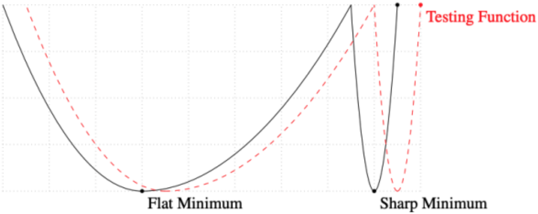
$$\mathbb{E}_{\mathcal{S}}\left[R(\theta_T) - \hat{R}_{\mathcal{S}}(\theta_T)\right] \leq \ ?$$

Question : when does SGD generalize ?
Attempt by Neu, Dziugaite, Haghifam & Roy (COLT 2021) via Information bounds !

When does a predictor generalize ?

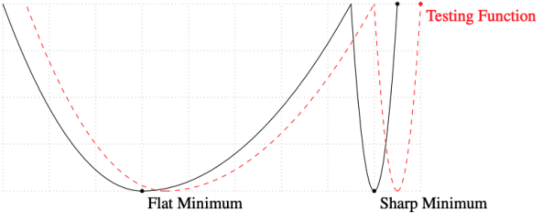
When does a predictor generalize ?

- Flatness (Hochreiter & Schmidhuber, Neural computation 1997, Keskar et al., ICLR 2017)
  - belief that algorithms that find wide optima of the loss landscape generalize well to test data
  - some flaws (difficult to define, parameterization,...)



When does a predictor generalize ?

- Flatness (Hochreiter & Schmidhuber, Neural computation 1997, Keskar et al., ICLR 2017)
  - belief that algorithms that find wide optima of the loss landscape generalize well to test data
  - some flaws (difficult to define, parameterization,...)



- Stability (Hardt, Recht & Singer, ICML 2016)
  - SGD has strong stability conditions
  - stability improves as assumptions get stronger

Fixing the mutual information bounds

The problem with mutual information bounds ?

Fixing the mutual information bounds

The problem with mutual information bounds ? The MI between deterministic quantities is equal to $+\infty$!

Fixing the mutual information bounds

The problem with mutual information bounds ? The MI between deterministic quantities is equal to $+\infty$!

Idea : *artificially* perturb the iterates !

Fixing the mutual information bounds

The problem with mutual information bounds ? The MI between deterministic quantities is equal to $+\infty$!

Idea : *artificially* perturb the iterates ! We can define a randomly perturbed version of SGD :

$$\tilde{\theta}_{t+1} = \theta_t + \zeta_t \quad \text{with} \quad \zeta_t = \sum_{k=1}^{t-1} \varepsilon_k \sim \mathcal{N}(0, \sigma_{1:t}^2),$$

$$\tilde{\theta}_{t+1} = \tilde{\theta}_t - \eta_t g(\theta_t; X_{I_t}, Y_{I_t}) + \varepsilon_t \quad \text{where} \quad \varepsilon_t \sim \mathcal{N}(0, \sigma_t^2 I).$$

Fixing the mutual information bounds

The problem with mutual information bounds ? The MI between deterministic quantities is equal to $+\infty$!

Idea : *artificially* perturb the iterates ! We can define a randomly perturbed version of SGD :

$$\tilde{\theta}_{t+1} = \theta_t + \zeta_t \quad \text{with} \quad \zeta_t = \sum_{k=1}^{t-1} \varepsilon_k \sim \mathcal{N}(0, \sigma_{1:t}^2),$$

$$\tilde{\theta}_{t+1} = \tilde{\theta}_t - \eta_t g(\theta_t; X_{I_t}, Y_{I_t}) + \varepsilon_t \quad \text{where} \quad \varepsilon_t \sim \mathcal{N}(0, \sigma_t^2 I).$$

We get :

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}}\left[R(\theta_T) - \hat{R}_{\mathcal{S}}(\theta_T)\right] &= \mathbb{E}_{\zeta_T,\mathcal{S}}\left[R(\tilde{\theta}_T) - \hat{R}_{\mathcal{S}}(\tilde{\theta}_T)\right] \\
&+ \mathbb{E}_{\zeta_T,\mathcal{S},\mathcal{S}'}\left[\hat{R}_{\mathcal{S}'}(\theta_T) - \hat{R}_{\mathcal{S}'}(\tilde{\theta}_T)\right] + \mathbb{E}_{\zeta_T,\mathcal{S},\mathcal{S}'}\left[\hat{R}_{\mathcal{S}}(\tilde{\theta}_T) - \hat{R}_{\mathcal{S}}(\theta_T)\right] \\
&\leq \sqrt{\frac{2 \cdot \mathcal{I}(\tilde{\theta}_T; \mathcal{S})}{n}} + \mathbb{E}_{\mathcal{S},\mathcal{S}'}[\Delta_{\sigma_{1:T}}(\theta_T, \mathcal{S}') - \Delta_{\sigma_{1:T}}(\theta_T, \mathcal{S})].
\end{aligned}$$

Main result

A generalization bound depending on the :

Main result

A generalization bound depending on the :

- variance of the gradients along the SGD path,

Main result

A generalization bound depending on the :

- variance of the gradients along the SGD path,
- perturbation-sensitivity of the gradients along the SGD path,

Main result

A generalization bound depending on the :

- variance of the gradients along the SGD path,
- perturbation-sensitivity of the gradients along the SGD path,
- perturbation-sensitivity of the loss at the final output.

Main result

A generalization bound depending on the :

- variance of the gradients along the SGD path,
- perturbation-sensitivity of the gradients along the SGD path,
- perturbation-sensitivity of the loss at the final output.

Thm [Neu, Dziugaite, Haghifam & Roy, COLT 2021] : assume that $\ell(\cdot,\cdot) \leq 1$, then for any $(\sigma_1, \cdots, \sigma_T)$, $\sigma_{1:T} = \sqrt{\sum_{k=1}^{T-1} \sigma_k^2}$,

$$\left|\mathbb{E}_{\mathcal{S}}\left[R(\theta_T) - \hat{R}_{\mathcal{S}}(\theta_T)\right]\right| \leq \sqrt{\frac{4}{n}\sum_{t=1}^{T}\frac{\eta_t^2}{\sigma_t^2}\mathop{\mathbb{E}}_{\mathcal{S}\sim P_{\mathcal{S}}}[\Gamma_{\sigma_{1:t}}(\theta_t) + V_t(\theta_t)]}$$
$$+ \left|\mathbb{E}_{\mathcal{S},\mathcal{S}'}[\Delta_{\sigma_{1:T}}(\theta_T,\mathcal{S}') - \Delta_{\sigma_{1:T}}(\theta_T,\mathcal{S})]\right| .$$

Variance of the gradients

The gradient variance V_t measures the variability of the gradients with respect to the randomness of the data :

$$V_t(\theta) = \mathbb{E}_{\mathcal{S}}\left[\left\| g(\theta; X_{I_t}, Y_{I_t}) - \bar{g}(\theta) \right\|_2^2 \middle| \theta_t = \theta \right]$$

where $\bar{g}(\theta) = \mathbb{E}_{(X,Y)\sim\mathbb{P}}[g(\theta; X, Y)]$.

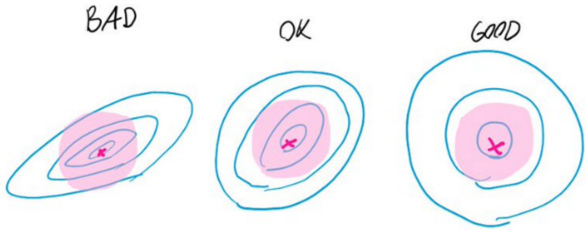
The gradient sensitivity Γ_σ measures the variability of the gradients to small perturbations in the parameter space.

$$\Gamma_\sigma(\theta) = \mathop{\mathbb{E}}_{(X,Y)\sim\mathbb{P},\zeta\sim\mathcal{N}(0,\sigma^2 I)} \left[\left\| g(\theta, Z) - g(\theta+\zeta, Z) \right\|_2^2 \right].$$

Value sensitivity

The value sensitivity Δ_σ measures the variability of the loss function to small perturbations in the parameter space.

$$\Delta_\sigma(\theta, s) = \mathop{\mathbb{E}}_{\zeta \sim \mathcal{N}(0, \sigma^2 I)} \left[\left\| \hat{R}_s(\theta) - \hat{R}_s(\theta + \zeta) \right\|_2^2 \right].$$



Summary of the quantities

Thm : for any $(\sigma_1, \cdots, \sigma_T)$, $\sigma_{1:T} = \sqrt{\sum_{k=1}^{T-1} \sigma_k^2}$, for losses ≤ 1,

$$\left|\mathbb{E}_{\mathcal{S}}\left[R(\theta_T) - \hat{R}_{\mathcal{S}}(\theta_T)\right]\right| \leq \sqrt{\frac{4}{n}\sum_{t=1}^{T}\frac{\eta_t^2}{\sigma_t^2}\mathbb{E}_{\mathcal{S}}[\Gamma_{\sigma_{\mathbf{1}:t}}(\theta_t) + V_t(\theta_t)]}$$
$$+ \left|\mathbb{E}_{\mathcal{S},\mathcal{S}'}[\Delta_{\sigma_{\mathbf{1}:T}}(\theta_T, \mathcal{S}') - \Delta_{\sigma_{\mathbf{1}:T}}(\theta_T, \mathcal{S})]\right|,$$

with the variance of the gradients along the SGD path

$$V_t(\theta) = \mathbb{E}_{\mathcal{S}}\left[\left\|g(\theta; X_{I_t}, Y_{I_t}) - \bar{g}(\theta)\right\|_2^2 \middle| \theta_t = \theta\right],$$

the sensitivity of the gradients along the SGD path

$$\Gamma_{\sigma_{\mathbf{1}:t}}(\theta) = \mathbb{E}_{(X,Y)\sim\mathbb{P},\zeta\sim\mathcal{N}(0,\sigma_{\mathbf{1}:t}^2 I)}\left[\left\|g(\theta; X, Y) - g(\theta+\zeta; X, Y)\right\|_2^2\right],$$

and the sensitivity of the loss at the final output :

$$\Delta_{\sigma_{\mathbf{1}:t}}(\theta, s) = \mathbb{E}_{\zeta\sim\mathcal{N}(0,\sigma_{\mathbf{1}:t}^2 I)}\left[\left\|\hat{R}_s(\theta) - \hat{R}_s(\theta+\zeta)\right\|_2^2\right].$$

Result for smooth functions

Assume that :

- $\eta_t = \eta$ and minibatches of size b,
- for each $i = 1, \cdots, n$, there is exactly one index t such that $i \in I_t$,
- $\mathbb{E}_{(X,Y)\sim\mathbb{P}}[\|g(\theta; X, Y) - \bar{g}(\theta)\|_2^2] \leq v$ for all θ,
- ℓ is globally μ-smooth i.e.

$$\|g(\theta; x, y) - g(\theta + u; x, y)\|_2 \leq \mu \|u\|_2$$

for all θ, u and all x, y.

Result for smooth functions

Assume that :

- $\eta_t = \eta$ and minibatches of size b,
- for each $i = 1, \cdots, n$, there is exactly one index t such that $i \in I_t$,
- $\mathbb{E}_{(X,Y)\sim\mathbb{P}}[\|g(\theta; X, Y) - \bar{g}(\theta)\|_2^2] \leq v$ for all θ,
- ℓ is globally μ-smooth i.e.

$$\|g(\theta; x, y) - g(\theta + u; x, y)\|_2 \leq \mu \|u\|_2$$

for all θ, u and all x, y.

Proposition : for any σ,

Result for smooth functions

Assume that :

- $\eta_t = \eta$ and minibatches of size b,
- for each $i = 1, \cdots, n$, there is exactly one index t such that $i \in I_t$,
- $\mathbb{E}_{(X,Y)\sim\mathbb{P}}[\|g(\theta; X, Y) - \bar{g}(\theta)\|_2^2] \leq v$ for all θ,
- ℓ is globally μ-smooth i.e.

$$\|g(\theta; x, y) - g(\theta + u; x, y)\|_2 \leq \mu\|u\|_2$$

for all θ, u and all x, y.

Proposition : for any σ,

$$\left|\mathbb{E}_{\mathcal{S}}\left[R(\theta_T) - \hat{R}_{\mathcal{S}}(\theta_T)\right]\right| = \mathcal{O}\left(\sqrt{\frac{R^2\eta^2 T}{\sigma^2 n}\left(\mu^2\sigma^2 dT + \frac{v}{b}\right)} + \mu\sigma^2 dT\right)$$

Different rates and regimes

Proposition : for any σ,

$$\left|\mathbb{E}_{\mathcal{S}}\left[R(\theta_T) - \hat{R}_{\mathcal{S}}(\theta_T)\right]\right| = \mathcal{O}\left(\sqrt{\frac{R^2\eta^2 T}{\sigma^2 n}\left(\mu^2\sigma^2 dT + \frac{v}{b}\right)} + \mu\sigma^2 dT\right)$$

Different rates and regimes

Proposition : for any σ,

$$\left|\mathbb{E}_{\mathcal{S}}\left[R(\theta_T) - \hat{R}_{\mathcal{S}}(\theta_T)\right]\right| = \mathcal{O}\left(\sqrt{\frac{R^2\eta^2 T}{\sigma^2 n}\left(\mu^2\sigma^2 dT + \frac{v}{b}\right)} + \mu\sigma^2 dT\right)$$

- Small-batch SGD : $T = \mathcal{O}(n)$ and $b = \mathcal{O}(1)$.

Different rates and regimes

Proposition : for any σ,

$$\left|\mathbb{E}_{\mathcal{S}}\left[R(\theta_T) - \hat{R}_{\mathcal{S}}(\theta_T)\right]\right| = \mathcal{O}\left(\sqrt{\frac{R^2\eta^2 T}{\sigma^2 n}\left(\mu^2\sigma^2 dT + \frac{v}{b}\right)} + \mu\sigma^2 dT\right)$$

- Small-batch SGD : $T = \mathcal{O}(n)$ and $b = \mathcal{O}(1)$. With $\eta = \mathcal{O}(1/n)$ & $\sigma = \Theta(n^{-4/3})$:

$$\left|\mathbb{E}_{\mathcal{S}}\left[R(\theta_T) - \hat{R}_{\mathcal{S}}(\theta_T)\right]\right| = n^{-1/3}.$$

Different rates and regimes

Proposition : for any σ,

$$\left|\mathbb{E}_{\mathcal{S}}\left[R(\theta_T) - \hat{R}_{\mathcal{S}}(\theta_T)\right]\right| = \mathcal{O}\left(\sqrt{\frac{R^2\eta^2 T}{\sigma^2 n}\left(\mu^2\sigma^2 dT + \frac{v}{b}\right)} + \mu\sigma^2 dT\right)$$

- Small-batch SGD : $T = \mathcal{O}(n)$ and $b = \mathcal{O}(1)$. With $\eta = \mathcal{O}(1/n)$ & $\sigma = \Theta(n^{-4/3})$:

$$\left|\mathbb{E}_{\mathcal{S}}\left[R(\theta_T) - \hat{R}_{\mathcal{S}}(\theta_T)\right]\right| = n^{-1/3}.$$

- Large-batch SGD : $T = \mathcal{O}(\sqrt{n})$ & $b = \Omega(\sqrt{n})$.

Different rates and regimes

Proposition : for any σ,

$$\left|\mathbb{E}_{\mathcal{S}}\left[R(\theta_T) - \hat{R}_{\mathcal{S}}(\theta_T)\right]\right| = \mathcal{O}\left(\sqrt{\frac{R^2\eta^2 T}{\sigma^2 n}\left(\mu^2\sigma^2 dT + \frac{v}{b}\right)} + \mu\sigma^2 dT\right)$$

- Small-batch SGD : $T = \mathcal{O}(n)$ and $b = \mathcal{O}(1)$. With $\eta = \mathcal{O}(1/n)$ & $\sigma = \Theta(n^{-4/3})$:

$$\left|\mathbb{E}_{\mathcal{S}}\left[R(\theta_T) - \hat{R}_{\mathcal{S}}(\theta_T)\right]\right| = n^{-1/3}.$$

- Large-batch SGD : $T = \mathcal{O}(\sqrt{n})$ & $b = \Omega(\sqrt{n})$. With $\eta = \mathcal{O}(1/T) = \mathcal{O}(1/\sqrt{n})$ and $\sigma = \Theta(1/\sqrt{n})$:

$$\left|\mathbb{E}_{\mathcal{S}}\left[R(\theta_T) - \hat{R}_{\mathcal{S}}(\theta_T)\right]\right| = 1/\sqrt{n}.$$

Many points to discuss

- Guarantees obtained for non-randomized predictors
- Small values of Γ, V and Δ imply good generalization
- How do we measure them ?
- Why would SGD make them small ?
- How do we adjust SGD so that they become smaller ?
- Is it necessary ?
- Limitations of the geometry
- Choice of the surrogate in the proof
- How about the subGaussian assumption ?

Next lecture : Variational inference !