An Introduction to PAC-Bayes Bounds

Luis A. Ortega

April 3, 2025

Universidad Autónoma de Madrid

In a **supervised learning** problem, we are given a data set, and

1. Fix a family of predictors.
2. Find a good predictor in this set.

In a **supervised learning** problem, we are given a data set, and

1. Fix a family of predictors.
2. Find a good predictor in this set.

For example, for **linear regression**, you 1) choose to consider only **linear predictors** and 2) use the **least-square method** to choose your linear predictor.

Notation I

The objective is to ***learn from examples to assign labels to objects***.

Notation I

The objective is to ***learn from examples to assign labels to objects***.

- The set of all possible objects will be denoted by $\mathcal{X} \subset \mathbb{R}^d$.

Notation I

The objective is to ***learn from examples to assign labels to objects***.

- The set of all possible objects will be denoted by $\mathcal{X} \subset \mathbb{R}^d$.
- The set of labels will be denoted by $\mathcal{Y}$.

Notation I

The objective is to ***learn from examples to assign labels to objects***.

- The set of all possible objects will be denoted by $\mathcal{X} \subset \mathbb{R}^d$.
- The set of labels will be denoted by $\mathcal{Y}$.
- A predictor is a function $f : \mathcal{X} \to \mathcal{Y}$. We are usually interested in parametric sets of predictors. That is, we consider $\{f_\theta, \theta \in \Theta\}$.

Notation I

The objective is to ***learn from examples to assign labels to objects***.

- The set of all possible objects will be denoted by $\mathcal{X} \subset \mathbb{R}^d$.
- The set of labels will be denoted by $\mathcal{Y}$.
- A predictor is a function $f : \mathcal{X} \to \mathcal{Y}$. We are usually interested in parametric sets of predictors. That is, we consider $\{f_\theta, \theta \in \Theta\}$.
- A loss function $\ell : \mathcal{Y}^2 \to [0, +\infty)$; where $\ell(y, y) = 0$. The $0 - 1$ loss for classification:

$$\ell(y, y') = \begin{cases} 0 & \text{if} \quad y = y' \, , \\ 1 & \text{if} \quad y \neq y' \, . \end{cases}$$

Notation II

We want to **predict the label of objects** in the future.

Notation II

We want to **predict the label of objects** in the future.

- Let P denote the probability distribution over $\mathcal{X} \times \mathcal{Y}$.

Notation II

We want to **predict the label of objects** in the future.

- Let P denote the probability distribution over $\mathcal{X} \times \mathcal{Y}$.
- The **expected error** (generalization risk) is

$$R(f) := \mathbb{E}_{(X,Y)\sim P}[\ell(f(X), Y)] \,, \quad R(\theta) := R(f_\theta) \,.$$

Notation II

We want to **predict the label of objects** in the future.

- Let P denote the probability distribution over $\mathcal{X} \times \mathcal{Y}$.
- The **expected error** (generalization risk) is

$$R(f) := \mathbb{E}_{(X,Y)\sim P}[\ell(f(X), Y)]\,, \quad R(\theta) := R(f_\theta)\,.$$

- Observations, $S := [(X_1, Y_1), \ldots, (X_n, Y_n)]$ are i.i.d from P.

Notation II

We want to **predict the label of objects** in the future.

- Let P denote the probability distribution over $\mathcal{X} \times \mathcal{Y}$.
- The **expected error** (generalization risk) is

$$R(f) := \mathbb{E}_{(X,Y)\sim P}[\ell(f(X), Y)]\,, \quad R(\theta) := R(f_\theta)\,.$$

- Observations, $S := [(X_1, Y_1), \ldots, (X_n, Y_n)]$ are i.i.d from P.
- The **empirical risk**:

$$r(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(f_\theta(X_i), Y_i), \quad \mathbb{E}_S[r(\theta)] = R(\theta)\,.$$

Notation III

An **estimator** takes observations and returns a predictor:

$$\hat{\theta} : \bigcup_{n=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^n \to \Theta \,.$$

Notation III

An **estimator** takes observations and returns a predictor:

$$\hat{\theta} : \bigcup_{n=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^n \to \Theta \,.$$

For example, the **empirical risk minimizer (ERM)**:

$$\hat{\theta}_{ERM} = \underset{\theta \in \Theta}{\arg\min}\, r(\theta) = \underset{\theta \in \Theta}{\arg\min}\, \tfrac{1}{n} \sum_{i=1}^{n} \ell(f_\theta(X_i), Y_i) \,.$$

PAC Bounds

$$\hat{\theta}_{ERM} = \arg\min_{\theta \in \Theta} r(\theta) \;\not\Longrightarrow\; \hat{\theta}_{ERM} = \arg\min_{\theta \in \Theta} R(\theta)\,.$$

ERM relies on the *hope* that "*these two functions are not so different*".

PAC Bounds

$$\hat{\theta}_{ERM} = \arg\min_{\theta \in \Theta} r(\theta) \;\not\Rightarrow\; \hat{\theta}_{ERM} = \arg\min_{\theta \in \Theta} R(\theta)\,.$$

ERM relies on the *hope* that "*these two functions are not so different*".

Proposition 1. If $\ell(\cdot,\cdot)$ is **bounded** in [0, C]; for any $\theta \in \Theta$ and $\delta \in (0,1)$,

$$\mathbb{P}_S\left[R(\theta) > r(\theta) + C\sqrt{\frac{\log\frac{1}{\delta}}{2n}}\right] \leq \delta$$

$$\hat{\theta}_{ERM} = \arg\min_{\theta \in \Theta} r(\theta) \;\not\Rightarrow\; \hat{\theta}_{ERM} = \arg\min_{\theta \in \Theta} R(\theta)\,.$$

ERM relies on the *hope* that "*these two functions are not so different*".

Proposition 1. If $\ell(\cdot,\cdot)$ is **bounded** in [0, C]; for any $\theta \in \Theta$ and $\delta \in (0,1)$,

$$\mathbb{P}_S\left[R(\theta) > r(\theta) + C\sqrt{\frac{\log\frac{1}{\delta}}{2n}}\right] \leq \delta \iff \mathbb{P}_S\left[R(\theta) \leq r(\theta) + C\sqrt{\frac{\log\frac{1}{\delta}}{2n}}\right] \geq 1 - \delta\,.$$

Proof. Hoeffding's inequality to $U_i = \mathbb{E}[\ell_i(\theta)] - \ell_i(\theta)$.

Proposition 1. If $\ell(\cdot, \cdot)$ is **bounded** in [0, C]; for any $\theta \in \Theta$ and $\delta \in (0, 1)$,

$$\mathbb{P}_S \left[R(\theta) > r(\theta) + C\sqrt{\frac{\log \frac{1}{\delta}}{2n}} \right] \leq \delta$$

Proposition 1. If $\ell(\cdot,\cdot)$ is **bounded** in [0, C]; for any $\theta \in \Theta$ and $\delta \in (0,1)$,

$$\mathbb{P}_S\left[R(\theta) > r(\theta) + C\sqrt{\frac{\log\frac{1}{\delta}}{2n}}\right] \leq \delta \iff \mathbb{P}_S\left[R(\theta) \leq r(\theta) + C\sqrt{\frac{\log\frac{1}{\delta}}{2n}}\right] \geq 1-\delta\,.$$

Proposition 1. If $\ell(\cdot,\cdot)$ is **bounded** in [0, C]; for any $\theta \in \Theta$ and $\delta \in (0,1)$,

$$\mathbb{P}_S\left[R(\theta) > r(\theta) + C\sqrt{\frac{\log\frac{1}{\delta}}{2n}}\right] \leq \delta \iff \mathbb{P}_S\left[R(\theta) \leq r(\theta) + C\sqrt{\frac{\log\frac{1}{\delta}}{2n}}\right] \geq 1-\delta\,.$$

1. Proposition 1 states that $R(\theta)$ will “usually” not exceed $r(\theta)$ by more than a term in $1/\sqrt{n}$.

Proposition 1. If $\ell(\cdot,\cdot)$ is **bounded** in [0, C]; for any $\theta \in \Theta$ and $\delta \in (0,1)$,

$$\mathbb{P}_S\left[R(\theta) > r(\theta) + C\sqrt{\frac{\log\frac{1}{\delta}}{2n}}\right] \leq \delta \iff \mathbb{P}_S\left[R(\theta) \leq r(\theta) + C\sqrt{\frac{\log\frac{1}{\delta}}{2n}}\right] \geq 1-\delta\,.$$

1. Proposition 1 states that $R(\theta)$ will "usually" not exceed $r(\theta)$ by more than a term in $1/\sqrt{n}$.
2. This is **not enough**, to justify the use of the ERM.

Proposition 1. If $\ell(\cdot,\cdot)$ is **bounded** in [0, C]; for any $\theta \in \Theta$ and $\delta \in (0,1)$,

$$\mathbb{P}_S\left[R(\theta) > r(\theta) + C\sqrt{\frac{\log\frac{1}{\delta}}{2n}}\right] \leq \delta \iff \mathbb{P}_S\left[R(\theta) \leq r(\theta) + C\sqrt{\frac{\log\frac{1}{\delta}}{2n}}\right] \geq 1-\delta\,.$$

1. Proposition 1 states that $R(\theta)$ will "usually" not exceed $r(\theta)$ by more than a term in $1/\sqrt{n}$.
2. This is **not enough**, to justify the use of the ERM.
3. The result is only true for a **fixed** θ, and we cannot apply it to $\hat{\theta}_{ERM}$ that is a function of the data.

PAC Bound on ERM

The usual approach to control $R(\hat{\theta}_{ERM})$ is to use:

$$R(\hat{\theta}_{ERM}) - r(\hat{\theta}_{ERM}) \leq \sup_{\theta \in \Theta} R(\theta) - r(\theta)\,.$$

Theorem 2. Assume that $\Theta = \{\theta_1, \ldots, \theta_M\}$. Then, for any $\delta \in (0, 1)$,

$$\mathbb{P}_S \left[R(\hat{\theta}_{ERM}) \leq \inf_{\theta \in \Theta} r(\theta) + C \sqrt{\frac{\log \frac{M}{\delta}}{2n}} \right] \geq 1 - \delta\,.$$

PAC Bound on ERM

The usual approach to control $R(\hat{\theta}_{ERM})$ is to use:

$$R(\hat{\theta}_{ERM}) - r(\hat{\theta}_{ERM}) \leq \sup_{\theta \in \Theta} R(\theta) - r(\theta)\,.$$

Theorem 2. Assume that $\Theta = \{\theta_1, \ldots, \theta_M\}$. Then, for any $\delta \in (0, 1)$,

$$\mathbb{P}_S \left[R(\hat{\theta}_{ERM}) \leq \inf_{\theta \in \Theta} r(\theta) + C \sqrt{\frac{\log \frac{M}{\delta}}{2n}} \right] \geq 1 - \delta\,.$$

These are called **Probably Approximately Correct (PAC) Bounds**.

$r(\hat{\theta}_{ERM}) = \inf_{\theta \in \Theta} r(\theta)$ approximates $R(\hat{\theta}_{ERM})$ within $C \sqrt{\frac{\log \frac{M}{\delta}}{2n}}$ with prob. $1 - \delta\,.$

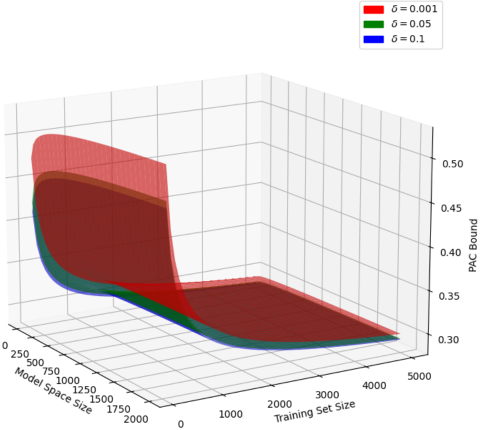
PAC Bound Example

$$\mathbb{P}_S\left[R(\hat{\theta}_{ERM}) \leq \inf_{\theta\in\Theta} r(\theta) + C\sqrt{\frac{\log\frac{M}{\delta}}{2n}}\right] \geq 1-\delta\,.$$

Let $\min_{\theta\in\Theta} r(\theta) = 0.26$, $C = 1$, $M = 100$, $n = 1000$ and $\delta = 0.05$

$$\mathbb{P}_S\left(R(\hat{\theta}_{ERM}) \leq 0.26 + 1 \times \sqrt{\frac{\log\frac{100}{0.05}}{2\times 1000}}\right)$$

$$\mathbb{P}_S\left(R(\hat{\theta}_{ERM}) \leq 0.26 + 0.06165\right) \geq 0.95\,.$$



PAC Bound Proof Elements

The proof is based on:

1. **Chernoff's Inequality**: for any $t > 0$,

$$\mathbb{P}[U > s] = \mathbb{P}\left[e^{tU} > e^{ts}\right] \leq \frac{\mathbb{E}\left[e^{tU}\right]}{e^{ts}} .$$

2. **The Union bound**:

$$\mathbb{P}\left[\sup_{1\leq i\leq M} U_i \ > s\right] = \mathbb{P}\left[\bigcup_{1\leq i\leq M} \{U_i > s\}\right] \leq \sum_{i=1}^{M} \mathbb{P}\left[U_i > s\right] .$$

PAC Bound Proof Elements

The proof is based on:

1. **Chernoff's Inequality**: for any $t > 0$,

$$\mathbb{P}[U > s] = \mathbb{P}\left[e^{tU} > e^{ts}\right] \leq \frac{\mathbb{E}\left[e^{tU}\right]}{e^{ts}}.$$

2. **The Union bound**:

$$\mathbb{P}\left[\sup_{1 \leq i \leq M} U_i > s\right] = \mathbb{P}\left[\bigcup_{1 \leq i \leq M} \{U_i > s\}\right] \leq \sum_{i=1}^{M} \mathbb{P}\left[U_i > s\right].$$

PAC-Bayes bounds are a generalization of the union bound argument that will allow us to deal with any parameter set Θ.

What are PAC-Bayes Bounds?

A **data-dependent probability measure** is a function:

$$\hat{\rho} : \bigcup_{n=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{P}(\Theta) \, .$$

To get a **predictor**:

1. Draw a parameter $\tilde{\theta} \sim \hat{\rho}$, **randomized estimator**.
2. **Average** predictors

$$f_{\hat{\rho}}(\cdot) := \mathbb{E}_{\theta \sim \hat{\rho}}[f_\theta(\cdot)]$$

With PAC-Bayes Bounds, we can obtain bounds related to

1. The risk of a randomized estimator, $R(\tilde{\theta})$.
2. The average risk of randomized estimators, $\mathbb{E}_{\theta\sim\hat{\rho}}[R(\theta)]$.
3. The risk of the aggregated estimator, $R(f_{\hat{\rho}})$.

A first PAC-Bayes Bound

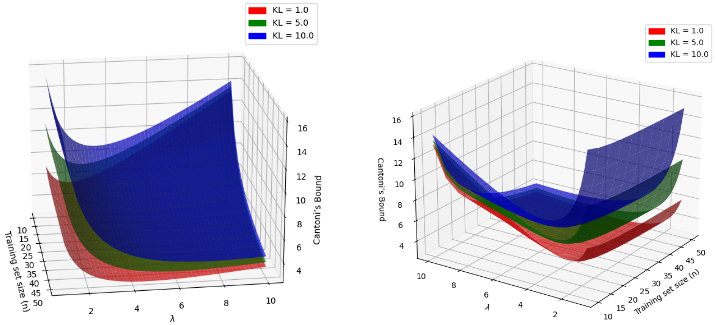
Let $\pi \in \mathcal{P}(\Theta)$ be a **fixed** prob. measure (the prior), and $\ell(\cdot, \cdot)$ be **bounded** in $[0, C]$.

A first PAC-Bayes Bound

Let $\pi \in \mathcal{P}(\Theta)$ be a **fixed** prob. measure (the prior), and $\ell(\cdot,\cdot)$ be **bounded** in $[0, C]$.

Cantoni's Bound, 2003. For any $\lambda > 0$, and any $\delta \in (0,1)$,

$$\mathbb{P}_S\left[\forall \rho \in \mathcal{P}(\Theta),\ \mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \mathbb{E}_{\theta\sim\rho}[r(\theta)] + \frac{\lambda C^2}{8n} + \frac{\mathrm{KL}(\rho|\pi) + \log\frac{1}{\delta}}{\lambda}\right] \geq 1-\delta\,.$$



Gibbs Posterior

$$\hat{\rho}_\lambda := \operatorname*{arg\,min}_{\rho \in \mathcal{P}(\Theta)} \left\{ \mathbb{E}_{\theta \sim \rho}[r(\theta)] + \frac{\mathsf{KL}(\rho|\pi)}{\lambda} \right\} .$$

Due to Donsker and Varadhan's variational formula:

$$\hat{\rho}_\lambda \propto e^{-\lambda r(\theta)} \pi(\theta) .$$

Gibbs Posterior

$$\hat{\rho}_\lambda := \operatorname*{arg\,min}_{\rho \in \mathcal{P}(\Theta)} \left\{ \mathbb{E}_{\theta\sim\rho}[r(\theta)] + \frac{\mathsf{KL}(\rho|\pi)}{\lambda} \right\} .$$

Due to Donsker and Varadhan's variational formula:

$$\hat{\rho}_\lambda \propto e^{-\lambda r(\theta)} \pi(\theta) .$$

$$\text{If} \quad r(\theta) := -\ln P(S|\theta) , \quad \text{then,} \quad \hat{\rho}_\lambda \propto P(S|\theta)^\lambda \pi(\theta) .$$

Related to **generalized** Bayesian framework and **tempered posteriors**.

Gibbs Posterior

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

Due to Donsker and Varadhan's variational formula:

$$\hat{\rho}_\lambda \propto e^{-\lambda r(\theta)} \pi(\theta) .$$

$$\text{If} \quad r(\theta) := -\ln P(S|\theta) , \quad \text{then,} \quad \hat{\rho}_\lambda \propto P(S|\theta)^\lambda \pi(\theta) .$$

Related to **generalized** Bayesian framework and **tempered posteriors**.

$$\mathbb{P}_S \left(\mathbb{E}_{\theta\sim\hat{\rho}_\lambda}[R(\theta)] \leq \inf_{\rho \in \mathcal{P}(\theta)} \left[\mathbb{E}_{\theta\sim\rho}[r(\theta)] + \frac{\lambda C^2}{8n} + \frac{\mathsf{KL}(\rho|\pi) + \log\frac{1}{\delta}}{\lambda} \right] \right) \geq 1 - \delta .$$

Order of Magnitude

Finite case $\Theta = \{\theta_1, \dots, \theta_M\}$.

$$\mathbb{E}_{\theta \sim \hat{\rho}_\lambda}[R(\theta)] \leq \inf_{\rho \in \mathcal{P}(\theta)} \left[\mathbb{E}_{\theta \sim \rho}[r(\theta)] + \frac{\lambda C^2}{8n} + \frac{\mathsf{KL}(\rho|\pi) + \log \frac{1}{\delta}}{\lambda} \right]$$

Order of Magnitude

Finite case $\Theta = \{\theta_1, \ldots, \theta_M\}$.

$$\begin{aligned}
\mathbb{E}_{\theta\sim\hat{\rho}_\lambda}[R(\theta)] &\leq \inf_{\rho\in\mathcal{P}(\theta)} \left[\mathbb{E}_{\theta\sim\rho}[r(\theta)] + \frac{\lambda C^2}{8n} + \frac{\mathsf{KL}(\rho|\pi) + \log\frac{1}{\delta}}{\lambda}\right] \\
\left[\text{D.V. formula}\right] \quad &\leq -\frac{1}{\lambda}\log\sum_{\theta\in\Theta}\pi(\theta)e^{-\lambda r(\theta)} + \frac{\lambda C^2}{8n} + \frac{\log\frac{1}{\delta}}{\lambda}
\end{aligned}$$

Order of Magnitude

Finite case $\Theta = \{\theta_1, \dots, \theta_M\}$.

$$\begin{aligned}
\mathbb{E}_{\theta\sim\hat{\rho}_\lambda}[R(\theta)] &\leq \inf_{\rho\in\mathcal{P}(\theta)} \left[\mathbb{E}_{\theta\sim\rho}[r(\theta)] + \frac{\lambda C^2}{8n} + \frac{\mathsf{KL}(\rho|\pi) + \log\frac{1}{\delta}}{\lambda}\right] \\
\left[\text{D.V. formula}\right] \quad &\leq -\frac{1}{\lambda}\log\sum_{\theta\in\Theta}\pi(\theta)e^{-\lambda r(\theta)} + \frac{\lambda C^2}{8n} + \frac{\log\frac{1}{\delta}}{\lambda} \\
\left[e^{-\lambda r(\theta)} \leq e^{-\lambda \inf_{\eta\in\Theta} r(\eta)}\right] \quad &\leq \inf_{\theta\in\Theta}\left[r(\theta) + \frac{\log\frac{1}{\pi(\theta)}}{\lambda}\right] + \frac{\lambda C^2}{8n} + \frac{\log\frac{1}{\delta}}{\lambda}
\end{aligned}$$

Order of Magnitude

Finite case $\Theta = \{\theta_1, \ldots, \theta_M\}$.

$$\begin{aligned}
\mathbb{E}_{\theta\sim\hat{\rho}_\lambda}[R(\theta)] &\leq \inf_{\rho\in\mathcal{P}(\theta)} \left[\mathbb{E}_{\theta\sim\rho}[r(\theta)] + \frac{\lambda C^2}{8n} + \frac{\mathsf{KL}(\rho|\pi) + \log\frac{1}{\delta}}{\lambda}\right] \\
\left[\text{D.V. formula}\right] \quad &\leq -\frac{1}{\lambda}\log\sum_{\theta\in\Theta}\pi(\theta)e^{-\lambda r(\theta)} + \frac{\lambda C^2}{8n} + \frac{\log\frac{1}{\delta}}{\lambda} \\
\left[e^{-\lambda r(\theta)} \leq e^{-\lambda \inf_{\eta\in\Theta} r(\eta)}\right] \quad &\leq \inf_{\theta\in\Theta}\left[r(\theta) + \frac{\log\frac{1}{\pi(\theta)}}{\lambda}\right] + \frac{\lambda C^2}{8n} + \frac{\log\frac{1}{\delta}}{\lambda}
\end{aligned}$$

Tight if $r(\theta)$ and $1/\pi(\theta)$ are **small simultaneously**; π cannot be large everywhere. The larger Θ, the more “spread” π is.

$$\mathbb{E}_{\theta\sim\hat{\rho}_\lambda}[R(\theta)] \leq \inf_{\theta\in\Theta}\left\{r(\theta) + \frac{\log\frac{1}{\pi(\theta)\delta}}{\lambda} + \frac{\lambda C^2}{8n}\right\}$$

If we choose an uniform prior $\pi(\theta) = 1/M$, the optimal $\lambda = \sqrt{8n\log(M/\delta)/C^2}$

$$\mathbb{P}_S\left(\mathbb{E}_{\theta\sim\hat{\rho}_\lambda}[R(\theta)] \leq \inf_{\theta\in\Theta}\{r(\theta)\} + C\sqrt{\frac{\log\frac{M}{\delta}}{2n}}\right) \geq 1-\delta\,.$$

$$\mathbb{E}_{\theta\sim\hat{\rho}_\lambda}[R(\theta)] \leq \inf_{\theta\in\Theta}\left\{r(\theta) + \frac{\log\frac{1}{\pi(\theta)\delta}}{\lambda} + \frac{\lambda C^2}{8n}\right\}$$

If we choose an uniform prior $\pi(\theta) = 1/M$, the optimal $\lambda = \sqrt{8n\log(M/\delta)/C^2}$

$$\mathbb{P}_S\left(\mathbb{E}_{\theta\sim\hat{\rho}_\lambda}[R(\theta)] \leq \inf_{\theta\in\Theta}\{r(\theta)\} + C\sqrt{\frac{\log\frac{M}{\delta}}{2n}}\right) \geq 1-\delta\,.$$

1. The Gibbs posterior $\hat{\rho}_\lambda$ satisfies the **same bound as the ERM**.
2. However $\hat{\rho}_\lambda$ and $\hat{\theta}_{ERM}$ are **not** equivalent!
3. The PAC-Bayes bound **can be tighter**.

Dirac Delta Posteriors

Cantoni's Bound, 2003. For any $\lambda > 0$, and any $\delta \in (0,1)$,

$$\mathbb{P}_S\left(\forall \rho \in \mathcal{P}(\Theta),\ \mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \mathbb{E}_{\theta\sim\rho}[r(\theta)] + \frac{\lambda C^2}{8n} + \frac{\mathsf{KL}(\rho|\pi) + \log\frac{1}{\delta}}{\lambda}\right) \geq 1-\delta\,.$$

It holds for every $\rho \in \mathcal{P}(\Theta)$. Then, consider a fixed parameter θ and $\delta_\theta \in \mathcal{P}(\Theta)$.

Dirac Delta Posteriors

Cantoni's Bound, 2003. For any $\lambda > 0$, and any $\delta \in (0,1)$,

$$\mathbb{P}_S\left(\forall \rho \in \mathcal{P}(\Theta),\ \mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \mathbb{E}_{\theta\sim\rho}[r(\theta)] + \frac{\lambda C^2}{8n} + \frac{\mathsf{KL}(\rho|\pi) + \log\frac{1}{\delta}}{\lambda}\right) \geq 1-\delta\,.$$

It holds for every $\rho \in \mathcal{P}(\Theta)$. Then, consider a fixed parameter θ and $\delta_\theta \in \mathcal{P}(\Theta)$.

1. $\mathbb{E}_{\eta\sim\delta_\theta}[r(\eta)] = r(\theta)$.
2. $\mathsf{KL}(\delta_\theta|\pi) = -\log \pi(\theta)$.

$$\mathbb{P}_S\left(\forall \theta \in \Theta,\ R(\theta) \leq r(\theta) + \frac{\lambda C^2}{8n} + \frac{\log\frac{1}{\delta} + \log\frac{1}{\pi(\theta)}}{\lambda}\right) \geq 1-\delta\,.$$

$$\mathbb{P}_S\left(\forall\theta\in\Theta,\ R(\theta)\leq r(\theta)+\frac{\lambda C^2}{8n}+\frac{\log\frac{1}{\delta}+\log\frac{1}{\pi(\theta)}}{\lambda}\right)\geq 1-\delta\,.$$

Taking the infimum over θ with $\Theta=\{\theta_1,\ldots,\theta_M\}$:

$$R(\hat{\theta}_{ERM})\leq\inf_{\theta\in\Theta}\{r(\theta)\}+\frac{\lambda C^2}{8n}+\frac{\log\frac{M}{\delta}}{\lambda}\,.$$

Taking again $\lambda=\sqrt{8n\log(M/\delta)/C^2}$

$$R(\hat{\theta}_{ERM})\leq\inf_{\theta\in\Theta}\{r(\theta)\}+C\sqrt{\frac{\log\frac{M}{\delta}}{2n}}\,.$$

Remarks

1. PAC-Bayes can be used to prove **generalization bounds for Gibbs posteriors**.

Remarks

1. PAC-Bayes can be used to prove **generalization bounds for Gibbs posteriors**.
2. Recent papers study **non-Bayesian robust estimators** of the mean and covariance matrix of **heavy-tailed** random vectors.

Remarks

1. PAC-Bayes can be used to prove **generalization bounds for Gibbs posteriors**.
2. Recent papers study **non-Bayesian robust estimators** of the mean and covariance matrix of **heavy-tailed** random vectors.
3. The choice of λ has a different status:

Remarks

1. PAC-Bayes can be used to prove **generalization bounds for Gibbs posteriors**.
2. Recent papers study **non-Bayesian robust estimators** of the mean and covariance matrix of **heavy-tailed** random vectors.
3. The choice of λ has a different status:
    3.1 Bound on the ERM: λ is chosen to **minimize the bound**, but the estimation procedure is not affected by λ.

Remarks

1. PAC-Bayes can be used to prove **generalization bounds for Gibbs posteriors**.
2. Recent papers study **non-Bayesian robust estimators** of the mean and covariance matrix of **heavy-tailed** random vectors.
3. The choice of λ has a different status:
    3.1 Bound on the ERM: λ is chosen to **minimize the bound**, but the estimation procedure is not affected by λ.
    3.2 Bound for the Gibbs posterior is also minimized with respect to λ, but $\hat{\rho}_\lambda$ **depends on** λ.

Example: Lipschitz loss and Gaussian prior

Assumptions:

1. $\Theta = \mathbb{R}^d$.
2. $\theta \mapsto \ell(f_\theta(x), y)$ is L-Lipschitz for any (x, y).
3. $\pi(\theta) = \mathcal{N}(0, \sigma^2 I_d)$.

Example: Lipschitz loss and Gaussian prior

Assumptions:

1. $\Theta = \mathbb{R}^d$.
2. $\theta \mapsto \ell(f_\theta(x), y)$ is L-Lipschitz for any (x, y).
3. $\pi(\theta) = \mathcal{N}(0, \sigma^2 I_d)$.

Starting point:

$$\forall \rho \in \mathcal{P}(\Theta),\ \mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \mathbb{E}_{\theta\sim\rho}[r(\theta)] + \frac{\lambda C^2}{8n} + \frac{\mathrm{KL}(\rho|\pi) + \log\frac{1}{\delta}}{\lambda}.$$

Example: Lipschitz loss and Gaussian prior

Assumptions:

1. $\Theta = \mathbb{R}^d$.
2. $\theta \mapsto \ell(f_\theta(x), y)$ is L-Lipschitz for any (x, y).
3. $\pi(\theta) = \mathcal{N}(0, \sigma^2 I_d)$.

Starting point:

$$\forall \rho \in \mathcal{P}(\Theta),\ \mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \mathbb{E}_{\theta\sim\rho}[r(\theta)] + \frac{\lambda C^2}{8n} + \frac{\mathsf{KL}(\rho|\pi) + \log\frac{1}{\delta}}{\lambda}\,.$$

Simplifications:

$$\mathsf{KL}(\rho|\pi) = \frac{\|m\|^2}{2\sigma^2} + \frac{d}{2}\left[\frac{s^2}{\sigma^2} + \log\frac{\sigma^2}{s^2} - 1\right]\,.$$

$$r(\theta) \text{ is } L\text{-Lipschitz} \implies \mathbb{E}_{\theta\sim\rho}[r(\theta)] \leq r(m) + Ls\sqrt{d}\,.$$

$$(\tilde{m}, \tilde{s}) = \operatorname*{arg\,min}_{m \in \mathbb{R}^d,\ s>0} \left\{ r(m) + \frac{\lambda C^2}{8n} + \frac{\frac{\|m\|^2}{2\sigma^2} + \frac{d}{2}\left[\frac{s^2}{\sigma^2} + \log\frac{\sigma^2}{s^2} - 1\right] + \log\frac{1}{\delta}}{\lambda} \right\} .$$

$\tilde{\rho}_\lambda := \mathcal{N}(\tilde{m}, \tilde{s}^2 I_d)$ is a variational approximation of $\hat{\rho}_\lambda$.

The choice of λ

In general, is **not possible** to optimize the right-hand side of the PAC-Bayes equality **with respect to** λ.

The choice of λ

In general, is **not possible** to optimize the right-hand side of the PAC-Bayes equality **with respect to** λ.

The **optimal value** of λ could depend on ρ.

The choice of λ

In general, is **not possible** to optimize the right-hand side of the PAC-Bayes equality **with respect to** λ.

The **optimal value** of λ could depend on ρ.

A natural idea is to propose a **finite grid** $\Lambda \subset (0, +\infty)$ and to minimize over this grid, which can be justified by a **union bound argument**:

$$\mathbb{P}_S\left[\forall \rho \in \mathcal{P}(\Theta),\ \mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \mathbb{E}_{\theta\sim\rho}[r(\theta)] + \frac{\lambda C^2}{8n} + \frac{\mathsf{KL}(\rho|\pi) + \log\frac{\mathsf{card}(\Lambda)}{\delta}}{\lambda}\right] \geq 1-\delta\,.$$

Final Remarks

1. Optimizing ρ and λ is an **open-problem**.
2. "There is no PAC-Bound tight for **all data-generating distributions**" — Gastpar et al., *Fantastic generalization measures are nowhere to be found*, ICLR (2024).

Final Remarks

1. Optimizing ρ and λ is an **open-problem**.
2. "There is no PAC-Bound tight for **all data-generating distributions**" — Gastpar et al., *Fantastic generalization measures are nowhere to be found*, ICLR (2024).

$$\downarrow$$

Data-distribution dependent or Algorithm dependent bounds

3. PAC-Bayes Bounds for **unbounded** losses are an open problem.

Thank you for your **attention**!

Alquier, Pierre (2024). **"User-friendly Introduction to PAC-Bayes Bounds".** In: *Foundations and Trends® in Machine Learning* 17.2, pp. 174–303. ISSN: 1935-8245. URL: `https://arxiv.org/pdf/2110.11216`.

Casado, Ioar et al. (2024). **"PAC-Bayes-Chernoff Bounds for Unbounded Losses".** In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. URL: `https://openreview.net/pdf?id=CyzZeND3LB`.

Gastpar, Michael et al. (2024). **"Fantastic Generalization Measures are Nowhere to be Found".** In: *The Twelfth International Conference on Learning Representations*. URL: `https://openreview.net/forum?id=NkmJotfL42`.

Jiang, Yiding et al. (2020). **"Fantastic Generalization Measures and Where to Find Them".** In: *International Conference on Learning Representations*. URL: `https://openreview.net/pdf?id=SJgIPJBFvH`.

Kullback-Leibler Divergence

Given two probability measures μ and ν in $\mathcal{P}(\Theta)$, the Kullback-Leibler (or simply KL) divergence between μ and ν is defined as

$$\mathsf{KL}(\mu|\nu) = \int \log\left(\frac{d\mu}{d\nu}(\theta)\right)\mu d(\theta)$$

Under absolutely continuity assumptions:

$$\mathsf{KL}(\mu|\nu) = \int \mu(\theta)\log\left(\frac{\mu(\theta)}{\nu(\theta)}\right)\, d(\theta)\,.$$

Hoeffding's Inequality

Let $X_1, X_2, \ldots, X_n$ be independent random variables such that $a_i \leq X_i \leq b_i$ almost surely. Then, consider

$$S_n = X_1 + \cdots + X_n \,.$$

It verifies that

$$P(S_n - \mathbb{E}[S_n]) \geq t) \leq \exp\left(\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \,.$$

Donsker and Varadhan's Variational Formula

For any measurable, bounded function $h : \Theta \to \mathbb{R}$ we have:

$$\log \mathbb{E}_{\theta\sim\pi}[e^{h(\theta)}] = \sup_{\rho\in\mathcal{P}(\Theta)} \left[\mathbb{E}_{\theta\sim\rho}[h(\theta)] - \mathsf{KL}(\rho|\pi)\right] .$$

It verifies that

$$P(S_n - \mathbb{E}[S_n]) \geq t) \leq \exp\left(\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) .$$