A Gentle Introduction to PAC-Bayes Bounds

PAC-Bayes Mini-tutorial

Erfan Mirzaei

Istituto Italiano di Tecnologia, University of Genoa & École Polytechnique

erfunmirzaei@gmail.com

First session:2026-07-04

Roadmap

1. Supervised Learning Problems
2. Probability Ideas
3. PAC Bounds
4. Occam's Razor: Min Description Length Principle
5. Beyond PAC Bounds

Supervised Learning Problems

We are given a dataset, and we should:

1. fix a set of *predictors*
2. find a good predictor in this set.

Goal of Supervised Learning

Goal: Building a machine to predict the output of the objects that it will encounter in the future. [Alquier, 2024]

Goal of Supervised Learning

Goal: Building a machine to predict the output of the objects that it will encounter in the future. [Alquier, 2024]

Mathematical Translation: Minimizing the expected prediction mistake, aka the ***true risk*** of f:

$$R(f) := \mathbb{E}_{(X,Y)\sim P}[l(f(X), Y)]$$

Here, P denotes the probability distribution of $\mathcal{X} \times \mathcal{Y}$.

Goal of Supervised Learning

Goal: Building a machine to predict the output of the objects that it will encounter in the future. [Alquier, 2024]

Mathematical Translation: Minimizing the expected prediction mistake, aka the ***true risk*** of f:

$$R(f) := \mathbb{E}_{(X,Y)\sim P}[l(f(X), Y)]$$

Here, P denotes the probability distribution of $\mathcal{X} \times \mathcal{Y}$.

Since we don't know P in practice, we can not minimize it directly.
Instead, we will train our machine based on examples.

Learning from Examples

We define the *empirical risk*:

$$r(\theta) = r(f_\theta) = \frac{1}{n}\sum_{i=1}^{n} l(f_\theta(X_i), Y_i),$$

where $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d from P. Note that it satisfies:

$$\mathbb{E}_{\mathcal{S}}[r(\theta)] := \mathbb{E}_{(X_1,Y_1),\ldots,(X_n,Y_n)}[r(\theta)] = R(\theta).$$

Learning from Examples

We define the *empirical risk*:

$$r(\theta) = r(f_\theta) = \frac{1}{n}\sum_{i=1}^{n} l(f_\theta(X_i), Y_i),$$

where $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d from P. Note that it satisfies:

$$\mathbb{E}_{\mathcal{S}}[r(\theta)] := \mathbb{E}_{(X_1,Y_1),\ldots,(X_n,Y_n)}[r(\theta)] = R(\theta).$$

Finally, an *estimator* is a function that takes a sample and returns a guess for the parameter θ. Formally,

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

Learning from Examples

We define the *empirical risk*:

$$r(\theta) = r(f_\theta) = \frac{1}{n}\sum_{i=1}^{n} l(f_\theta(X_i), Y_i),$$

where $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d from P. Note that it satisfies:

$$\mathbb{E}_{\mathcal{S}}[r(\theta)] := \mathbb{E}_{(X_1,Y_1),\ldots,(X_n,Y_n)}[r(\theta)] = R(\theta).$$

Finally, an *estimator* is a function that takes a sample and returns a guess for the parameter θ. Formally,

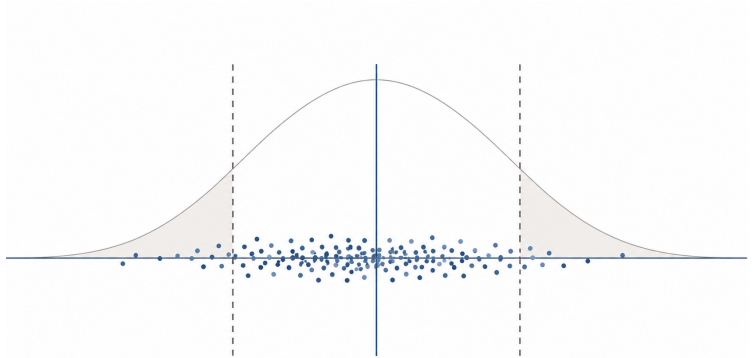
$$\widehat{\theta} : \bigcup_{n=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^n \to \Theta.$$

Question: when does small training error imply small true error?

Roadmap

1. Supervised Learning Problems
2. **Probability Ideas**
3. PAC Bounds
4. Occam's Razor: Min Description Length Principle
5. Beyond PAC Bounds

Concentration Inequalities: the Basic Idea



Message: A random average is usually close to expectation. Large deviations are rare.

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - Z\right| > \varepsilon\right) \quad \text{is small.}$$

Why Concentration Matters for Learning

Empirical risk is an average:

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

True risk is its expectation:

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

Concentration intuition

For a fixed predictor θ, $r(\theta) \approx R(\theta)$ with high probability.

This is the probabilistic engine behind PAC bounds.

Three Probability Ideas We Will Use

1. **i.i.d. samples:**The examples are independent draws from the same distribution.

2. **High probability:** A statement holds with probability at least $1 - \delta$.

3. **Union bound:** Controlling many bad events costs an extra complexity term.

Roadmap

1. Supervised Learning Problems
2. Probability Ideas
3. **PAC Bounds**
4. Occam's Razor: Min Description Length Principle
5. Beyond PAC Bounds

PAC Bounds

Of course, our objective is to **minimize R, not r**.

PAC Bounds

Of course, our objective is to **minimize** R**, not** r.

So the ERM strategy is motivated by the **hope** that these two functions are not so different, so that the minimizer of r almost minimizes R.

PAC Bounds

Of course, our objective is to **minimize** R**, not** r.

So the ERM strategy is motivated by the **hope** that these two functions are not so different, so that the minimizer of r almost minimizes R.

PAC question

Can we guarantee that $R(\theta) \approx r(\theta)$ with high probability?

PAC Bounds

Of course, our objective is to **minimize** R**, not** r.

So the ERM strategy is motivated by the **hope** that these two functions are not so different, so that the minimizer of r almost minimizes R.

PAC question

Can we guarantee that $R(\theta) \approx r(\theta)$ with high probability?

Proposition

For any $\theta \in \Theta$*, for any* $\delta \in (0, 1)$*,*

$$\mathbb{P}_{\mathcal{S}}\big(R(\theta) > r(\theta) + C\sqrt{\frac{\log(\frac{1}{\delta})}{2n}}\big) \leq \delta.$$

Proof: Hoeffding's + Chernoff's Bounding Technique

Lemma (Hoeffding's Lemma: See Ch 2 of [Boucheron et al., 2013])

Let $U_1, \ldots, U_n$ be independent random variables taking values in an interval $[a, b]$. Then, for any $t > 0$,

$$\mathbb{E}\left[e^{t\sum_{i=1}^{n}[U_i - \mathbb{E}[U_i]]}\right] \leq e^{\frac{nt^2(b-a)^2}{8}}$$

PAC Bounds

This is **not enough**, though, to justify the use of the ERM. Indeed, the proposition is only true for the θ that was fixed above, and we cannot apply it to $\hat{\theta}_{\text{ERM}}$, which is a function of the data.

PAC Bounds

This is **not enough**, though, to justify the use of the ERM. Indeed, the proposition is only true for the θ that was fixed above, and we cannot apply it to $\hat{\theta}_{\text{ERM}}$, which is a function of the data.

To control $R(\hat{\theta}_{\text{ERM}})$, we need an inequality,

$$R(\hat{\theta}_{\text{ERM}}) - r(\hat{\theta}_{\text{ERM}}) < \sup_{\theta \in \Theta} [R(\theta) - r(\theta)]. \tag{1}$$

Theorem (A simple PAC Bound)

Assume the card$(\Theta) = M < +\infty$. *For any* $\delta \in (0, 1)$,

$$\mathbb{P}_{\mathcal{S}}\Big(R(\hat{\theta}_{ERM}) < \inf_{\theta \in \Theta} r(\theta) + C\sqrt{\frac{\log(\frac{M}{\delta})}{2n}} \quad \Big) \geq 1 - \delta.$$

Proof: Last Preposition + Union Bound

Roadmap

1. Supervised Learning Problems
2. Probability Ideas
3. PAC Bounds
4. **Occam's Razor: Min Description Length Principle**
5. Beyond PAC Bounds

From Uniform Bounds to Weighted Bounds

In PAC bounds, we assumed a **uniform** "uncertainty/complexity" for all predictors (log of cardinality M). [Fleuret, 2011]

From Uniform Bounds to Weighted Bounds

In PAC bounds, we assumed a **uniform** "uncertainty/complexity" for all predictors (log of cardinality M). [Fleuret, 2011]

- What if we believe some θ are more likely to be the true solution than others *before* seeing the data?

From Uniform Bounds to Weighted Bounds

In PAC bounds, we assumed a **uniform** "uncertainty/complexity" for all predictors (log of cardinality M). [Fleuret, 2011]

- What if we believe some θ are more likely to be the true solution than others *before* seeing the data?
- Let us assign a "prior" weight $\pi(\theta)$ to each $\theta \in \Theta$ such that:

$$\pi(\theta) \in (0, 1) \quad \text{and} \quad \sum_{\theta \in \Theta} \pi(\theta) = 1$$

From Uniform Bounds to Weighted Bounds

In PAC bounds, we assumed a **uniform** "uncertainty/complexity" for all predictors (log of cardinality M). [Fleuret, 2011]

- What if we believe some θ are more likely to be the true solution than others *before* seeing the data?
- Let us assign a "prior" weight $\pi(\theta)$ to each $\theta \in \Theta$ such that:

$$\pi(\theta) \in (0, 1) \quad \text{and} \quad \sum_{\theta \in \Theta} \pi(\theta) = 1$$

Modified Union Bound: Instead of allocating the failure probability δ uniformly (δ/M per hypothesis), we allocate it proportionally to $\pi(\theta)$.

$$\mathbb{P}_{\mathcal{S}}\big(\exists \theta \in \Theta : R(\theta) - r(\theta) > \epsilon(\theta)\big) \leq \sum_{\theta \in \Theta} \delta \cdot \pi(\theta) = \delta$$

Occam's Bound

By solving for $\epsilon(\theta)$ using the weighted allocation, we get a θ-dependent bound.

Occam's Bound

By solving for $\epsilon(\theta)$ using the weighted allocation, we get a θ-dependent bound.

Theorem (Occam's Bound)

For any discrete set Θ, any prior distribution π on Θ, and any $\delta \in (0, 1)$, with probability at least $1 - \delta$:

$$\forall \theta \in \Theta, \quad R(\theta) \leq r(\theta) + \sqrt{\frac{\log \frac{1}{\pi(\theta)} + \log \frac{1}{\delta}}{2n}}$$

Occam's Bound

By solving for $\epsilon(\theta)$ using the weighted allocation, we get a θ-dependent bound.

Theorem (Occam's Bound)

For any discrete set Θ, any prior distribution π on Θ, and any $\delta \in (0, 1)$, with probability at least $1 - \delta$:

$$\forall \theta \in \Theta, \quad R(\theta) \leq r(\theta) + \sqrt{\frac{\log \frac{1}{\pi(\theta)} + \log \frac{1}{\delta}}{2n}}$$

Proof: By Hoeffding's inequality, for a fixed θ, this probability is bounded by $e^{-2n\epsilon(\theta)^2}$ that equals $\pi(\theta)$ due to the definition of the last slide.

Occam's Bound

By solving for $\epsilon(\theta)$ using the weighted allocation, we get a θ-dependent bound.

Theorem (Occam's Bound)

For any discrete set Θ, any prior distribution π on Θ, and any $\delta \in (0,1)$, with probability at least $1-\delta$:

$$\forall \theta \in \Theta, \quad R(\theta) \leq r(\theta) + \sqrt{\frac{\log \frac{1}{\pi(\theta)} + \log \frac{1}{\delta}}{2n}}$$

Proof: By Hoeffding's inequality, for a fixed θ, this probability is bounded by $e^{-2n\epsilon(\theta)^2}$ that equals $\pi(\theta)$ due to the definition of the last slide.

Interpretation: The term $\log \frac{1}{\pi(\theta)}$ can be viewed as the **description length** of θ (in bits) using an optimal code (Shannon/Huffman coding).

The Principle of Parsimony

This result offers a mathematical justification for **Occam's Razor** ("Entities should not be multiplied unnecessarily").

The Principle of Parsimony

This result offers a mathematical justification for **Occam's Razor** ("Entities should not be multiplied unnecessarily").

- **High $\pi(\theta)$ (Simple Model):** Short description length $\implies$ Small $\log \frac{1}{\pi(\theta)}$ $\implies$ **Tighter generalization gap**.

The Principle of Parsimony

This result offers a mathematical justification for **Occam's Razor** ("Entities should not be multiplied unnecessarily").

- **High $\pi(\theta)$ (Simple Model):** Short description length $\implies$ Small $\log \frac{1}{\pi(\theta)}$ $\implies$ **Tighter generalization gap**.
- **Low $\pi(\theta)$ (Complex Model):** Long description length $\implies$ Large $\log \frac{1}{\pi(\theta)}$ $\implies$ **Looser generalization gap**.

The Principle of Parsimony

This result offers a mathematical justification for **Occam's Razor** ("Entities should not be multiplied unnecessarily").

- **High $\pi(\theta)$ (Simple Model):** Short description length $\implies$ Small $\log \frac{1}{\pi(\theta)}$ $\implies$ **Tighter generalization gap**.
- **Low $\pi(\theta)$ (Complex Model):** Long description length $\implies$ Large $\log \frac{1}{\pi(\theta)}$ $\implies$ **Looser generalization gap**.

Occam's Bound: For a fixed prior π , the complexity of a single hypothesis θ is the code length:

$$\text{Complexity}(\theta) = \log \frac{1}{\pi(\theta)} \quad \text{(bits to encode } \theta\text{)}$$

The Principle of Parsimony

This result offers a mathematical justification for **Occam's Razor** ("Entities should not be multiplied unnecessarily").

- **High $\pi(\theta)$ (Simple Model):** Short description length $\implies$ Small $\log \frac{1}{\pi(\theta)}$ $\implies$ **Tighter generalization gap**.
- **Low $\pi(\theta)$ (Complex Model):** Long description length $\implies$ Large $\log \frac{1}{\pi(\theta)}$ $\implies$ **Looser generalization gap**.

Occam's Bound: For a fixed prior π , the complexity of a single hypothesis θ is the code length:

$$\text{Complexity}(\theta) = \log \frac{1}{\pi(\theta)} \quad \text{(bits to encode } \theta\text{)}$$

Conclusion: Complex hypotheses (those with long descriptions) require more data (n) to ensure the empirical risk $r(\theta)$ is close to the true risk $R(\theta)$.

Roadmap

1. Supervised Learning Problems
2. Probability Ideas
3. PAC Bounds
4. Occam's Razor: Min Description Length Principle
5. **Beyond PAC Bounds**

Analyzing ERM Beyond Finite Parameter Set

To go beyond the finite parameter set, the union-bound argument has to be modified, and things become a little more complicated. To do so, we have to follow one of the following ways:

Analyzing ERM Beyond Finite Parameter Set

To go beyond the finite parameter set, the union-bound argument has to be modified, and things become a little more complicated. To do so, we have to follow one of the following ways:

- Vapnik-Chervonenkis dimension

Analyzing ERM Beyond Finite Parameter Set

To go beyond the finite parameter set, the union-bound argument has to be modified, and things become a little more complicated. To do so, we have to follow one of the following ways:

- Vapnik-Chervonenkis dimension
- Rademacher complexity

Analyzing ERM Beyond Finite Parameter Set

To go beyond the finite parameter set, the union-bound argument has to be modified, and things become a little more complicated. To do so, we have to follow one of the following ways:

- Vapnik-Chervonenkis dimension
- Rademacher complexity
- PAC-Bayes framework

Analyzing ERM Beyond Finite Parameter Set

To go beyond the finite parameter set, the union-bound argument has to be modified, and things become a little more complicated. To do so, we have to follow one of the following ways:

- Vapnik-Chervonenkis dimension
- Rademacher complexity
- PAC-Bayes framework

Analyzing ERM Beyond Finite Parameter Set

To go beyond the finite parameter set, the union-bound argument has to be modified, and things become a little more complicated. To do so, we have to follow one of the following ways:

- Vapnik-Chervonenkis dimension
- Rademacher complexity
- PAC-Bayes framework

Using the first two methods leads to technical difficulties or strong restrictions, such as the compactness of Θ.

References I

Alquier, P. (2024).
User-friendly introduction to PAC-Bayes bounds.
Foundations and Trends® in Machine Learning, 17(2):174–303.

Boucheron, S., Lugosi, G., and Massart, P. (2013).
Concentration Inequalities.
Oxford University Press.

Fleuret, F. (2011).
Machine learning: Pac-learning.
Slides, May 11, 2011.