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

Second Session: 2026-07-05

Roadmap

1. Review: Learning Theory
2. Occam's Razor: Min Description Length Principle
3. Beyond PAC Bounds
4. Intro to PAC-Bayes Bounds

The Basic Problem

We want to minimize the **true risk**

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

But P is unknown, so we only observe data:

$$S = \{(X_i, Y_i)\}_{i=1}^n.$$

Question: When does a small training error imply a small true error?

From Conditional Risk to Bayes Risk

By the law of total expectation,

$$R(f) = \mathbb{E}[\ell(y, f(x))] = \mathbb{E}_{x'}\left[\mathbb{E}[\ell(y, f(x')) \mid x = x']\right].$$

From Conditional Risk to Bayes Risk

By the law of total expectation,

$$R(f) = \mathbb{E}[\ell(y, f(x))] = \mathbb{E}_{x'}\left[\mathbb{E}[\ell(y, f(x')) \mid x = x']\right].$$

For each fixed input x', define the conditional risk

$$r(z \mid x') = \mathbb{E}[\ell(y, z) \mid x = x'].$$

From Conditional Risk to Bayes Risk

By the law of total expectation,

$$R(f) = \mathbb{E}[\ell(y, f(x))] = \mathbb{E}_{x'}\left[\mathbb{E}[\ell(y, f(x')) \mid x = x']\right].$$

For each fixed input x', define the conditional risk

$$r(z \mid x') = \mathbb{E}[\ell(y, z) \mid x = x'].$$

The best possible prediction is the Bayes predictor:

$$f_*(x') \in \arg\min_{z \in \mathcal{Y}} r(z \mid x').$$

The Bayes risk is

$$R_* = R(f_*).$$

Excess Risk

The Bayes risk R_* is the best achievable risk among all prediction functions.

Excess Risk

The Bayes risk R_* is the best achievable risk among all prediction functions.

For any predictor f, its **excess risk** is

$$R(f) - R_*.$$

Excess Risk

The Bayes risk R_* is the best achievable risk among all prediction functions.

For any predictor f, its **excess risk** is

$$R(f) - R_*.$$

Interpretation

Excess risk measures how far f is from the best possible predictor.

Excess Risk

The Bayes risk R_* is the best achievable risk among all prediction functions.

For any predictor f, its **excess risk** is

$$R(f) - R_*.$$

Interpretation

Excess risk measures how far f is from the best possible predictor.

In learning theory, the goal is not only to make $R(f)$ small, but to understand

$$R(f) - R_*.$$

Approximation and Estimation Error

Consider a hypothesis set

$$\mathcal{F} = \{f_\theta : \theta \in \Theta\}.$$

Approximation and Estimation Error

Consider a hypothesis set

$$\mathcal{F} = \{f_\theta : \theta \in \Theta\}.$$

Let θ' be a minimizer of the expected risk inside Θ:

$$\theta' \in \arg\min_{\theta \in \Theta} R(f_\theta).$$

Approximation and Estimation Error

Consider a hypothesis set

$$\mathcal{F} = \{f_\theta : \theta \in \Theta\}.$$

Let θ' be a minimizer of the expected risk inside Θ:

$$\theta' \in \arg\min_{\theta \in \Theta} R(f_\theta).$$

Then

$$R(f_{\hat{\theta}}) - R_* = \underbrace{R(f_{\hat{\theta}}) - R(f_{\theta'})}_{\text{estimation error}} + \underbrace{R(f_{\theta'}) - R_*}_{\text{approximation error}}.$$

What Do These Errors Mean?

Approximation error

$$R(f_{\theta'}) - R_*$$

comes from restricting ourselves to the class $\mathcal{F}$.

What Do These Errors Mean?

Approximation error

$$R(f_{\theta'}) - R_*$$

comes from restricting ourselves to the class $\mathcal{F}$.

Estimation error

$$R(f_{\hat{\theta}}) - R(f_{\theta'})$$

comes from learning from finitely many samples.

What Do These Errors Mean?

Approximation error

$$R(f_{\theta'}) - R_*$$

comes from restricting ourselves to the class $\mathcal{F}$.

Estimation error

$$R(f_{\hat{\theta}}) - R(f_{\theta'})$$

comes from learning from finitely many samples.

Trade-off

A larger class $\mathcal{F}$ may reduce approximation error, but can increase estimation error.

Estimation Error: Uniform Deviation + Optimization

Let $\theta' \in \arg\min_{\theta \in \Theta} R(f_\theta)$. Then we can decompose the estimation error:

Estimation Error: Uniform Deviation + Optimization

Let $\theta' \in \arg\min_{\theta \in \Theta} R(f_\theta)$. Then we can decompose the estimation error:

$$R(f_{\hat{\theta}}) - R(f_{\theta'}) = \underbrace{R(f_{\hat{\theta}}) - r(f_{\hat{\theta}})}_{\text{generalization gap}} + \underbrace{r(f_{\hat{\theta}}) - r(f_{\theta'})}_{\text{empirical optimization}} + \underbrace{r(f_{\theta'}) - R(f_{\theta'})}_{\text{generalization gap}}.$$

Estimation Error: Uniform Deviation + Optimization

Let $\theta' \in \arg\min_{\theta\in\Theta} R(f_\theta)$. Then we can decompose the estimation error:

$$R(f_{\hat{\theta}}) - R(f_{\theta'}) = \underbrace{R(f_{\hat{\theta}}) - r(f_{\hat{\theta}})}_{\text{generalization gap}} + \underbrace{r(f_{\hat{\theta}}) - r(f_{\theta'})}_{\text{empirical optimization}} + \underbrace{r(f_{\theta'}) - R(f_{\theta'})}_{\text{generalization gap}}.$$

Therefore,

$$R(f_{\hat{\theta}}) - R(f_{\theta'}) \leq 2 \sup_{\theta\in\Theta} |r(f_\theta) - R(f_\theta)| + \left[r(f_{\hat{\theta}}) - \inf_{\theta\in\Theta} r(f_\theta) \right].$$

What Controls the Estimation Error?

The estimation error is controlled by two quantities:

$$R(f_{\hat{\theta}}) - R(f_{\theta'}) \leq 2 \sup_{\theta \in \Theta} |r(f_\theta) - R(f_\theta)| + \text{optimization error.}$$

What Controls the Estimation Error?

The estimation error is controlled by two quantities:

$$R(f_{\hat{\theta}}) - R(f_{\theta'}) \leq 2 \sup_{\theta \in \Theta} |r(f_\theta) - R(f_\theta)| + \text{optimization error}.$$

$$\text{optimization error} = r(f_{\hat{\theta}}) - \inf_{\theta \in \Theta} r(f_\theta).$$

What Controls the Estimation Error?

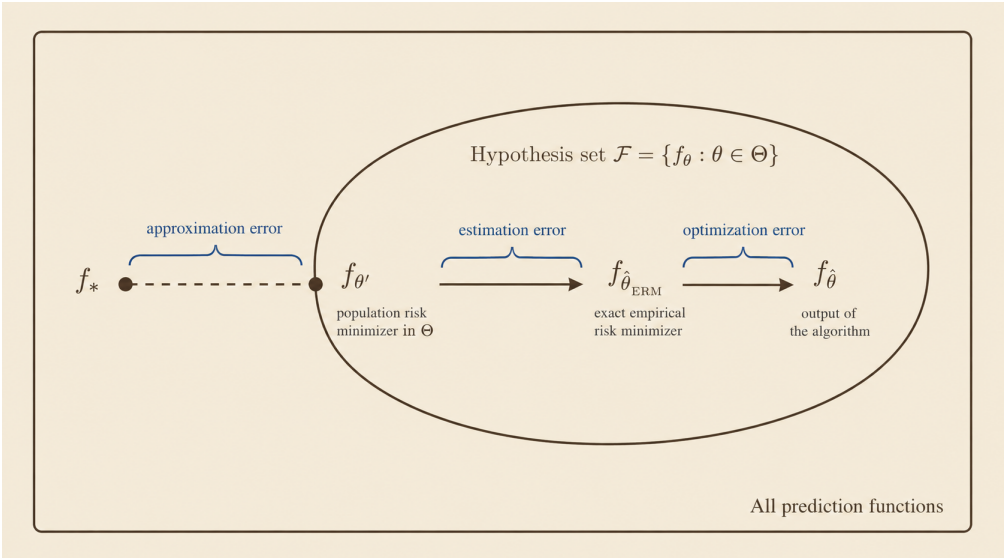
The estimation error is controlled by two quantities:

$$R(f_{\hat{\theta}}) - R(f_{\theta'}) \leq 2 \sup_{\theta \in \Theta} |r(f_\theta) - R(f_\theta)| + \text{optimization error}.$$

$$\text{optimization error} = r(f_{\hat{\theta}}) - \inf_{\theta \in \Theta} r(f_\theta).$$

Key point: concentration inequalities control the uniform deviation term. [Bach, 2024]

Error Decomposition in One Picture



From Empirical Risk to True Risk

$$r(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(f_\theta(X_i), Y_i) \qquad \text{vs.} \qquad R(\theta) = \mathbb{E}\,\ell(f_\theta(X), Y)$$

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* $\delta \in (0, 1)$*,*

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

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* $\delta \in (0, 1)$,

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

Roadmap

1. Review: Learning Theory
2. **Occam's Razor: Min Description Length Principle**
3. Beyond PAC Bounds
4. Intro to PAC-Bayes 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. Review: Learning Theory
2. Occam's Razor: Min Description Length Principle
3. **Beyond PAC Bounds**
4. Intro to PAC-Bayes 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.

Roadmap

1. Review: Learning Theory
2. Occam's Razor: Min Description Length Principle
3. Beyond PAC Bounds
4. **Intro to PAC-Bayes Bounds**

Bayesian Supervised Learning

In **classical** supervised learning problem, We are given a dataset, and we should:

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

Bayesian Supervised Learning

In **classical** supervised learning problem, We are given a dataset, and we should:

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

Bayesian Supervised Learning

In **classical** supervised learning problem, We are given a dataset, and we should:

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

While in **Bayesian** supervised learning problem, we are given a dataset, and after the first step, we should:

- draw a predictor according to some prescribed probability distribution

What are PAC-Bayes Bounds?

- An extension of the union-bound argument (and maybe more!)
- Finite, infinite, or even continuous Parameter set
- But a different estimator!

What are PAC-Bayes Bounds?

- An extension of the union-bound argument (and maybe more!)
- Finite, infinite, or even continuous Parameter set
- But a different estimator!

Definition

Let $P(\Theta)$ be the set of all probability distributions on Θ equipped with a σ-algebra $\mathcal{T}$. A **data-dependent probability measure** is a function:

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

What are PAC-Bayes Bounds?

- An extension of the union-bound argument (and maybe more!)
- Finite, infinite, or even continuous Parameter set
- But a different estimator!

Definition

Let $P(\Theta)$ be the set of all probability distributions on Θ equipped with a σ-algebra $\mathcal{T}$. A **data-dependent probability measure** is a function:

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

Now, to bound data-depe quantities the supremum in Eq. 1 needs to be replaced by:

$$\mathbb{E}_{\theta \sim \hat{\rho}}[R(\theta) - r(\theta)] \leq \sup_{\rho \in P(\Theta)} \mathbb{E}_{\theta \sim \rho}[R(\theta) - r(\theta)]$$

Why PAC-Bayes Bounds?

The last equation might look scary(!) and unnecessarily complicated at first sight, but it is more convenient for many reasons:

Why PAC-Bayes Bounds?

The last equation might look scary(!) and unnecessarily complicated at first sight, but it is more convenient for many reasons:

- PAC-Bayes bounds could handle **infinite** parameter set easily

Why PAC-Bayes Bounds?

The last equation might look scary(!) and unnecessarily complicated at first sight, but it is more convenient for many reasons:

- PAC-Bayes bounds could handle **infinite** parameter set easily
- **Randomized estimators** are fairly common in ML

Why PAC-Bayes Bounds?

The last equation might look scary(!) and unnecessarily complicated at first sight, but it is more convenient for many reasons:

- PAC-Bayes bounds could handle **infinite** parameter set easily
- **Randomized estimators** are fairly common in ML
- Bayesian estimators incorporate **prior knowledge** through a prior distribution π on Θ.

Why PAC-Bayes Bounds?

The last equation might look scary(!) and unnecessarily complicated at first sight, but it is more convenient for many reasons:

- PAC-Bayes bounds could handle **infinite** parameter set easily
- **Randomized estimators** are fairly common in ML
- Bayesian estimators incorporate **prior knowledge** through a prior distribution π on Θ.

Q: If Θ is infinite, what will the $log(M)$ term be replaced with?

A Simple PAC-Bayes Bound

Let us fix a probability measure $\pi \in \mathcal{P}(\Theta)$, which will be called the *prior*.

A Simple PAC-Bayes Bound

Let us fix a probability measure $\pi \in \mathcal{P}(\Theta)$, which will be called the *prior*.

Theorem (A simple PAC-Bayes Bound ([Alquier, 2024]))

For any $\lambda > 0$, for any $\delta \in (0,1)$,

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

From Occam's Razor to PAC-Bayes

PAC-Bayes bounds are fundamentally an expression of **Occam's Razor**: simpler descriptions of data generalize better.

From Occam's Razor to PAC-Bayes

PAC-Bayes bounds are fundamentally an expression of **Occam's Razor**: simpler descriptions of data generalize better.

- **The PAC-Bayes Shift**: Instead of selecting a single θ, we consider a **distribution of "good" solutions** ν (the posterior).

From Occam's Razor to PAC-Bayes

PAC-Bayes bounds are fundamentally an expression of **Occam's Razor**: simpler descriptions of data generalize better.

- **The PAC-Bayes Shift**: Instead of selecting a single θ, we consider a **distribution of "good" solutions** ν (the posterior).
- **The "Bits Back" Argument**: If we are satisfied with *any* sample from ν, we save bits:

$$\text{Avg. Code Length} = \mathbb{E}_{\theta\sim\nu}\left[\log\frac{1}{\pi(\theta)}\right] \quad \text{(Cross Entropy } \mathbb{H}(\nu,\pi)\text{)}$$

$$\text{Bits Saved} = \mathbb{H}(\nu) \quad \text{(Uncertainty about specific } \theta\text{)}$$

$$\text{Net Complexity} = \mathbb{H}(\nu,\pi) - \mathbb{H}(\nu) = \mathbf{KL}(\nu\|\pi)$$

From Occam's Razor to PAC-Bayes

PAC-Bayes bounds are fundamentally an expression of **Occam's Razor**: simpler descriptions of data generalize better.

- **The PAC-Bayes Shift**: Instead of selecting a single θ, we consider a **distribution of "good" solutions** ν (the posterior).
- **The "Bits Back" Argument**: If we are satisfied with *any* sample from ν, we save bits:

$$\text{Avg. Code Length} = \mathbb{E}_{\theta\sim\nu}\left[\log\frac{1}{\pi(\theta)}\right] \quad \text{(Cross Entropy } \mathbb{H}(\nu,\pi)\text{)}$$

$$\text{Bits Saved} = \mathbb{H}(\nu) \quad \text{(Uncertainty about specific } \theta\text{)}$$

$$\text{Net Complexity} = \mathbb{H}(\nu,\pi) - \mathbb{H}(\nu) = \mathbf{KL}(\nu\|\pi)$$

Result: The complexity term in the bound evolves naturally:

$$\log\frac{1}{\pi(\theta)} \implies \mathrm{KL}(\nu\|\pi)$$

Proof of PAC-Bayesian Theorem

References I

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

Bach, F. (2024).
Learning theory from first principles.
MIT press.

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