# 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: 2026-07-11



# Roadmap

1. Review: Learning Theory
2. Intro to PAC-Bayes Bounds
3. Two Types of PAC Bounds
4. The PAC-Bayes Engine
5. Sub-Gaussian Generalization Gap

## The Basic Problem

We want to minimize the **true risk**

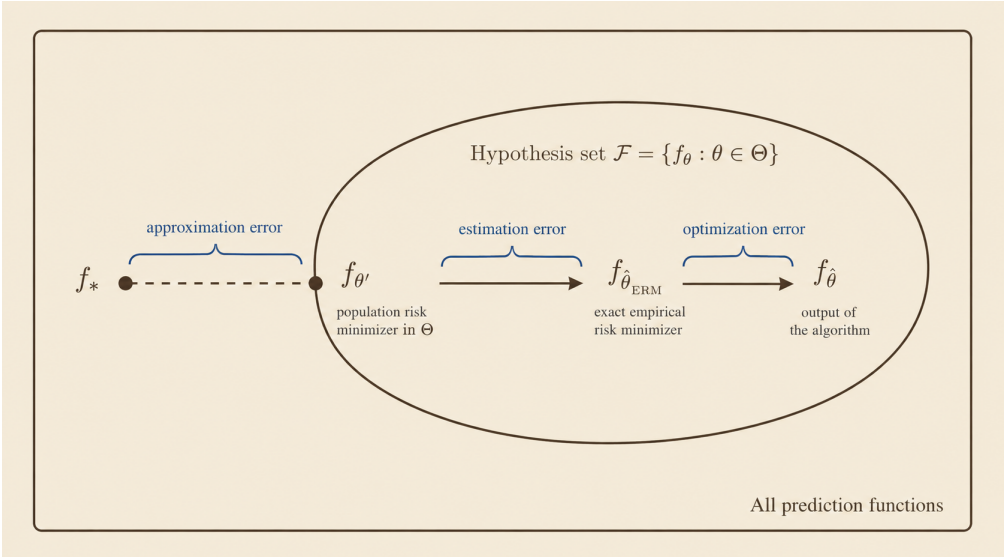
$$R(\theta) = \mathbb{E}_{(X,Y)\sim P}\, \ell(f_\theta(X), Y).$$

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?

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

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

## Occam's Bound

By solving for $\epsilon(\theta)$ using the weighted allocation, we get a $\theta$-dependent bound.
[Fleuret, 2011, van Erven, 2014]

### Theorem (Occam's Bound)

*For any discrete set $\Theta$, any prior distribution $\pi$ on $\Theta$, 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 $\theta$, 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 $\theta$ (in bits) using an optimal code (Shannon/Huffman coding).

## 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 $\Theta$.

## Main Ingredients and Assumptions

1. **IID Samples:** The examples are independent draws from the same distribution.
2. **Sub-Gaussian RVs:** Bounding the MGF using Hoeffding's Lemma or similar inequalities
3. **High probability Tail Bound by Cramér-Chernoff Method:** Exponential + Markov's inequality + optimization
4. **Union bound:** Controlling many bad events costs an extra complexity term.

# Roadmap

1. Review: Learning Theory
2. **Intro to PAC-Bayes Bounds**
3. Two Types of PAC Bounds
4. The PAC-Bayes Engine
5. Sub-Gaussian Generalization Gap

## 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 $\Theta$ equipped with a $\sigma$-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 $\Theta$ equipped with a $\sigma$-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 $\pi$ on $\Theta$.

## Why PAC-Bayes Bounds?

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

- 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 $\pi$ on $\Theta$.

Q: If $\Theta$ 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 $\theta$, we consider a **distribution of "good" solutions** $\nu$ (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 $\theta$, we consider a **distribution of "good" solutions** $\nu$ (the posterior).
- **The "Bits Back" Argument**: If we are satisfied with *any* sample from $\nu$, 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 $\theta$, we consider a **distribution of "good" solutions** $\nu$ (the posterior).
- **The "Bits Back" Argument**: If we are satisfied with *any* sample from $\nu$, 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

## Roadmap

1. Review: Learning Theory
2. Intro to PAC-Bayes Bounds
3. **Two Types of PAC Bounds**
4. The PAC-Bayes Engine
5. Sub-Gaussian Generalization Gap

## Two Types of PAC Bounds

PAC bounds come in two different flavors.

**Empirical PAC bounds**

$$R(f_{\hat{\theta}}) \leq \text{quantity computable from the data.}$$

**Oracle PAC bounds**

$$R(f_{\hat{\theta}}) \leq \inf_{\theta \in \Theta} R(f_\theta) + r_n(\delta).$$

**Main distinction**

Empirical bounds give numerical certificates. Oracle bounds explain statistical optimality.

## Empirical PAC Bounds: Certificates

An empirical PAC bound has the form

$$\mathbb{P}_S\Big(R(f_{\hat{\theta}}) \le B(S, \hat{\theta}, \delta)\Big) \ge 1 - \delta,$$

where $B(S, \hat{\theta}, \delta)$ is computable after observing the data.

Example:

$$\mathbb{P}_S\Big(R(f_{\hat{\theta}_{\mathrm{ERM}}}) \le 0.12\Big) \ge 0.99.$$

### What it tells us

It certifies the performance of the learned predictor.

## Oracle PAC Bounds: Excess Risk

An oracle PAC bound has the form

$$\mathbb{P}_S\left(R(f_{\hat{\theta}_{\mathrm{ERM}}}) \leq \inf_{\theta\in\Theta} R(f_\theta) + r_n(\delta)\right) \geq 1-\delta.$$

Equivalently,

$$R(f_{\hat{\theta}_{\mathrm{ERM}}}) - \inf_{\theta\in\Theta} R(f_\theta) \leq r_n(\delta).$$

### What it tells us

It says how close the learned predictor is to the best predictor in the class.

## Oracle PAC Bounds: Excess Risk

An oracle PAC bound has the form

$$\mathbb{P}_S\left(R(f_{\hat{\theta}_{\mathrm{ERM}}}) \leq \inf_{\theta\in\Theta} R(f_\theta) + r_n(\delta)\right) \geq 1-\delta.$$

Equivalently,

$$R(f_{\hat{\theta}_{\mathrm{ERM}}}) - \inf_{\theta\in\Theta} R(f_\theta) \leq r_n(\delta).$$

### What it tells us

It says how close the learned predictor is to the best predictor in the class.

## Relation to Consistency

Oracle bounds are closely related to consistency.

If

$$r_n(\delta) \to 0 \qquad \text{as} \qquad n \to \infty,$$

then

$$R(f_{\hat{\theta}_{\mathrm{ERM}}}) - \inf_{\theta \in \Theta} R(f_\theta) \to 0.$$

So the algorithm becomes as good as the best predictor in $\mathcal{F}$.

**Important**

This is consistency **relative to the hypothesis class**. To get consistency with respect to the Bayes risk $R_*$, we also need the approximation error to vanish.

# Roadmap

1. Review: Learning Theory
2. Intro to PAC-Bayes Bounds
3. Two Types of PAC Bounds
4. **The PAC-Bayes Engine**
5. Sub-Gaussian Generalization Gap

## From Concentration to PAC-Bayes

For a fixed hypothesis, concentration controls

$$L(h) - \hat{L}(h, \mathbf{x}).$$

PAC-Bayes asks for something stronger:

control the gap after averaging over a data-dependent distribution $\rho$.

### Main ingredients

prior $\pi$ + posterior $\rho$ + complexity $\mathrm{KL}(\rho \| \pi)$

Here $\rho \ll \pi$, and we write

$$\nu(h) = \frac{\mathrm{d}\rho}{\mathrm{d}\pi}(h).$$

## The Main Change-of-Measure Bound

Let

$$F : \mathcal{H} \times \mathcal{X}^n \to \mathbb{R}$$

be a measurable function.

With probability at least $1 - \delta$ over $\mathbf{x} \sim \mu^n$,

$$\mathbb{E}_{h\sim\rho} F(h, \mathbf{x}) \leq \mathrm{KL}(\rho\|\pi) + \log \mathbb{E}_{\mathbf{x}} \mathbb{E}_{g\sim\pi} e^{F(g,\mathbf{x})} + \log \frac{1}{\delta}.$$

### Interpretation

PAC-Bayes turns an exponential-moment bound under the prior $\pi$ into a high-probability bound for a posterior $\rho$.

## A Single-Draw Version

The same argument also gives a pointwise statement.

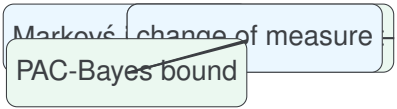
With probability at least $1 - \delta$ over $\mathbf{x} \sim \mu^n$ and $h \sim \rho$,

$$F(h, \mathbf{x}) \leq \log \nu(h) + \log \mathbb{E}_{\mathbf{x}} \mathbb{E}_{g \sim \pi} e^{F(g, \mathbf{x})} + \log \frac{1}{\delta}.$$

### Two viewpoints

- averaged posterior: complexity is $\mathrm{KL}(\rho \| \pi)$,
- one sampled hypothesis: complexity is $\log \nu(h)$.

## Why Does This Work?




For any random variable $Y$,

$$\mathbb{P}\left(Y > \log \mathbb{E}e^{Y} + \log \frac{1}{\delta}\right) \leq \delta.$$

This is the whole probabilistic engine.

## What Remains to Prove?

The theorem is flexible because $F$ is arbitrary.

To get a concrete learning bound, we only need to control

$$\log \mathbb{E}_{\mathbf{x}} \mathbb{E}_{g \sim \pi} e^{F(g, \mathbf{x})}.$$

Typical choices of $F$:

- sub-Gaussian generalization gaps,
- bounded losses and kl bounds,
- U-statistics,
- dependent data, e.g. Markov chains.

# Roadmap

1. Review: Learning Theory
2. Intro to PAC-Bayes Bounds
3. Two Types of PAC Bounds
4. The PAC-Bayes Engine
5. **Sub-Gaussian Generalization Gap**

## Sub-Gaussian Losses

Assume that for every $h \in \mathcal{H}$,

$$f(h, X) - \mathbb{E} f(h, X)$$

is $\sigma$-sub-Gaussian.

Define

$$\hat{L}(h, \mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} f(h, x_i), \qquad L(h) = \mathbb{E}_X f(h, X).$$

Then

$$\hat{L}(h, \mathbf{x}) - L(h)$$

is $\sigma / \sqrt{n}$-sub-Gaussian.

## Choosing the Function $F$

Choose

$$F(h, \mathbf{x}) = \lambda\big(L(h) - \hat{L}(h, \mathbf{x})\big), \qquad \lambda > 0.$$

By the sub-Gaussian assumption,

$$\log \mathbb{E}_{\mathbf{x}} e^{F(h,\mathbf{x})} \leq \frac{\lambda^2 \sigma^2}{2n}.$$

Since $\pi$ is a probability measure, the same bound holds after averaging over $h \sim \pi$.

## A PAC-Bayes Generalization Bound

With probability at least $1 - \delta$,

$$\mathbb{E}_{h\sim\rho} L(h) \leq \mathbb{E}_{h\sim\rho} \hat{L}(h, \mathbf{x}) + \frac{\lambda\sigma^2}{2n} + \frac{\mathrm{KL}(\rho\|\pi) + \log(1/\delta)}{\lambda}.$$

### Message

The posterior risk is controlled by three terms:

$$\text{training loss} \quad + \quad \text{concentration price} \quad + \quad \text{complexity price.}$$

## Single-Draw Generalization Bound

With probability at least $1-\delta$ over $\mathbf{x} \sim \mu^n$ and $h \sim \rho$,

$$L(h) \leq \hat{L}(h, \mathbf{x}) + \frac{\lambda \sigma^2}{2n} + \frac{\log \nu(h) + \log(1/\delta)}{\lambda}.$$

### Interpretation

For a sampled hypothesis, the complexity is the log-density ratio

$$\log \nu(h) = \log \frac{\mathrm{d}\rho}{\mathrm{d}\pi}(h).$$

# Roadmap

1. Review: Learning Theory
2. Intro to PAC-Bayes Bounds
3. Two Types of PAC Bounds
4. The PAC-Bayes Engine
5. Sub-Gaussian Generalization Gap

## Which Posterior Minimizes the Bound?

Ignoring constants independent of $\rho$, the averaged bound asks us to minimize

$$J(\rho) = \mathbb{E}_{h\sim\rho}\hat{L}(h, \mathbf{x}) + \frac{1}{\lambda}\mathrm{KL}(\rho\|\pi).$$

This is an entropy-regularized empirical risk minimization problem.

### Trade-off

- low empirical loss pushes $\rho$ toward good predictors;
- the KL term keeps $\rho$ close to the prior $\pi$.

## Deriving the Gibbs Posterior

Write $\nu = \mathrm{d}\rho / \mathrm{d}\pi$. Then

$$J(\nu) = \int \hat{L}(h, \mathbf{x}) \nu(h) \, \mathrm{d}\pi(h) + \frac{1}{\lambda} \int \nu(h) \log \nu(h) \, \mathrm{d}\pi(h),$$

subject to

$$\int \nu(h) \, \mathrm{d}\pi(h) = 1.$$

The first-order condition gives

$$\nu(h) = C \exp\{-\lambda \hat{L}(h, \mathbf{x})\}.$$

Thus

$$\frac{\mathrm{d}\widehat{\rho}_\lambda}{\mathrm{d}\pi}(h) = \frac{e^{-\lambda \hat{L}(h, \mathbf{x})}}{\mathbb{E}_{g \sim \pi} e^{-\lambda \hat{L}(g, \mathbf{x})}}.$$

## Partition Function

Define the partition function

$$Z_\lambda(\mathbf{x}) = \mathbb{E}_{g\sim\pi} e^{-\lambda \hat{L}(g,\mathbf{x})}.$$

Then the Gibbs posterior is

$$G_\lambda(\,\mathrm{d}h \mid \mathbf{x}) = \frac{e^{-\lambda \hat{L}(h,\mathbf{x})}}{Z_\lambda(\mathbf{x})}\,\pi(\,\mathrm{d}h).$$

Equivalently,

$$\log \nu(h) = -\lambda \hat{L}(h,\mathbf{x}) - \log Z_\lambda(\mathbf{x}).$$

# Roadmap

1. Review: Learning Theory
2. Intro to PAC-Bayes Bounds
3. Two Types of PAC Bounds
4. The PAC-Bayes Engine
5. Sub-Gaussian Generalization Gap

## A Useful Identity for $-\log Z_\beta$

Let

$$A(\beta) = -\log Z_\beta(\mathbf{x}).$$

Then

$$A'(\beta) = \mathbb{E}_{h\sim G_\beta(\mathbf{x})}\hat{L}(h,\mathbf{x}),$$

and

$$A''(\beta) = -\mathrm{Var}_{h\sim G_\beta(\mathbf{x})}\big(\hat{L}(h,\mathbf{x})\big) \leq 0.$$

Therefore,

$$-\log Z_\beta(\mathbf{x}) = \int_0^\beta \mathbb{E}_{h\sim G_\gamma(\mathbf{x})}\hat{L}(h,\mathbf{x})\ \mathrm{d}\gamma.$$

## Discretizing the Integral

For

$$0 = \beta_0 < \beta_1 < \cdots < \beta_K = \beta,$$

concavity gives

$$-\log Z_\beta(\mathbf{x}) \leq \sum_{k=1}^{K} (\beta_k - \beta_{k-1}) \mathbb{E}_{g \sim G_{\beta_{k-1}}(\mathbf{x})} \hat{L}(g, \mathbf{x}).$$

### Why this matters

It replaces the unknown log-partition function by expectations under Gibbs distributions.

## A Data-Dependent Upper Bound on $\log \nu(h)$

Define

$$\Gamma(h, \mathbf{x}, \boldsymbol{\beta}) = -\beta_K \hat{L}(h, \mathbf{x}) + \sum_{k=1}^{K} (\beta_k - \beta_{k-1}) \mathbb{E}_{g \sim G_{\beta_{k-1}}(\mathbf{x})} \hat{L}(g, \mathbf{x}).$$

Then

$$-\beta_K \hat{L}(h, \mathbf{x}) - \log Z_{\beta_K}(\mathbf{x}) \leq \Gamma(h, \mathbf{x}, \boldsymbol{\beta}).$$

So $\Gamma$ gives a computable upper bound on the log-density ratio of the Gibbs posterior.

## Statistical Physics Dictionary

For the Gibbs posterior,

$$G_\beta(\,\mathrm{d}h \mid \mathbf{x}) \propto \exp\{-\beta \hat{L}(h, \mathbf{x})\}\pi(\,\mathrm{d}h).$$

| Learning notation | Physics analogy |
|---|---|
| hypothesis $h$ | state |
| empirical loss $\hat{L}(h, \mathbf{x})$ | energy |
| $\beta$ | inverse temperature |
| $Z_\beta$ | partition function |
| $-\log Z_\beta$ | free energy scale |
| $-A''(\beta)$ | heat-capacity-like fluctuation |

# Roadmap

1. Review: Learning Theory
2. Intro to PAC-Bayes Bounds
3. Two Types of PAC Bounds
4. The PAC-Bayes Engine
5. Sub-Gaussian Generalization Gap

## Bounded Losses and Binary KL

Now assume

$$f : \mathcal{H} \times \mathcal{X} \to [0, 1].$$

Let

$$\hat{p}_h = \frac{1}{n} \sum_{i=1}^{n} f(h, x_i), \qquad p_h = \mathbb{E}_X f(h, X).$$

For $a, b \in [0, 1]$, define the binary KL

$$\kappa(a, b) = a \log \frac{a}{b} + (1 - a) \log \frac{1 - a}{1 - b}.$$

This gives a sharper way to compare empirical and true risks for bounded losses.

## kl-PAC-Bayes Bound

Using the exponential moment bound

$$\mathbb{E}_{\mathbf{x}} \exp\left(n\kappa(\hat{p}_h, p_h)\right) \leq 2\sqrt{n}, \qquad n \geq 8,$$

we obtain, with probability at least $1-\delta$,

$$\kappa\left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{h\sim\rho} f(h, x_i), \mathbb{E}_{h\sim\rho}\mathbb{E}_X f(h, X)\right) \leq \frac{\mathrm{KL}(\rho\|\pi) + \log\frac{2\sqrt{n}}{\delta}}{n}.$$

If you want a free parameter $\lambda$, state explicitly the corresponding exponential-moment version.

## From kl Bound to a Risk Bound

The kl bound can be inverted:

$$\mathbb{E}_{h\sim\rho}\mathbb{E}_X f(h,X) \leq \kappa^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{h\sim\rho}f(h,x_i), \frac{\mathrm{KL}(\rho\|\pi)+\log\frac{2\sqrt{n}}{\delta}}{n}\right).$$

A simpler but looser version is

$$\mathbb{E}_{h\sim\rho}\mathbb{E}_X f(h,X) \leq \frac{1}{n}\sum_{i=1}^{n}{}_{h\sim\rho}f(h,x_i) + \sqrt{\frac{\mathrm{KL}(\rho\|\pi)+\log\frac{2\sqrt{n}}{\delta}}{2n}}.$$

## Takeaway

PAC-Bayes bounds follow a reusable recipe.

1. Choose a function $F(h, \mathbf{x})$ encoding the gap of interest.
2. Prove an exponential-moment bound under the prior $\pi$.
3. Pay $\mathrm{KL}(\rho\|\pi)$ to move from prior to posterior.
4. Optimize the posterior: the Gibbs posterior naturally appears.

### Big picture

Concentration controls fixed hypotheses; PAC-Bayes controls randomized, data-dependent hypotheses.

# References I



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


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


van Erven, T. (2014).
Pac-bayes mini-tutorial: A continuous union bound.
*arXiv preprint arXiv:1405.1580*.