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

Fourth Session: 2026-07-12

Roadmap

1. Review
2. PAC-Bayes Framework
3. Sub-Gaussian PAC-Bayes Bound
4. kl-PAC-Bayes for Bounded Losses
5. The Gibbs Posterior

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!

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. **??** 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:

- 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*.

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.

- **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:

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

Two Types of PAC Bounds

PAC bounds usually appear in two flavors.

Empirical PAC bound:

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

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

Oracle PAC bound:

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

Main distinction

Empirical bounds give certificates. Oracle bounds explain statistical optimality.

Oracle Bounds and Consistency

Oracle bounds control the excess risk inside the class:

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

If

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

then the learned predictor becomes as good as the best predictor in $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$.

Important

This is consistency **relative to the hypothesis class**. For Bayes consistency, we also need the approximation error to vanish:

$$\inf_{\theta \in \Theta} R(f_\theta) - R_* \to 0.$$

Roadmap

1. Review
2. **PAC-Bayes Framework**
3. Sub-Gaussian PAC-Bayes Bound
4. kl-PAC-Bayes for Bounded Losses
5. The Gibbs Posterior

From Occam Bounds to PAC-Bayes

Occam's bound controls one selected hypothesis:

$$R(\theta) \lesssim r(\theta) + \sqrt{\frac{\log \frac{1}{\pi(\theta)}}{n}}.$$

PAC-Bayes changes the object of study:

$$\text{one hypothesis } \theta \quad \longrightarrow \quad \text{a distribution } \rho_{\mathbf{x}} \text{ over hypotheses.}$$

Key shift

Instead of paying for the code length of one hypothesis, we pay for how much the posterior $\rho_{\mathbf{x}}$ moves away from the prior π:

$$\mathrm{KL}(\rho_{\mathbf{x}} \| \pi).$$

Notation

Let Θ be the hypothesis space and let π be a prior distribution on Θ, chosen before observing the sample.

Given data $\mathbf{x} = (x_1, \ldots, x_n) \sim \mu^n$, define

$$r(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, x_i), \qquad R(\theta) = \mathbb{E}_{X \sim \mu} \left[\ell(\theta, X) \right].$$

A learning algorithm may return a posterior distribution

$$\rho_{\mathbf{x}} \in \mathcal{P}(\Theta).$$

We assume $\rho_{\mathbf{x}} \ll \pi$ and write

$$\frac{d\rho_{\mathbf{x}}}{d\pi}(\theta)$$

for the density of the posterior with respect to the prior.

Randomized Predictors

A PAC-Bayes posterior $\rho_{\mathbf{x}}$ can be used in two ways.

Single-draw predictor:

$$\theta \sim \rho_{\mathbf{x}}, \qquad \text{then predict with } \theta.$$

Posterior-mean risk:

$$\mathbb{E}_{\theta\sim\rho_{\mathbf{x}}}[R(\theta)], \qquad \mathbb{E}_{\theta\sim\rho_{\mathbf{x}}}[r(\theta)].$$

What PAC-Bayes bounds

PAC-Bayes gives high-probability control of risks after the posterior has been chosen from the data.

The PAC-Bayes Change-of-Measure Bound

Let

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

be a measurable function.

Theorem
([Shawe-Taylor and Williamson, 1997, McAllester, 1999, Rivasplata et al., 2020])

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

$$\mathbb{E}_{\theta \sim \rho_{\mathbf{x}}}[F(\theta, \mathbf{x})] \leq \mathrm{KL}(\rho_{\mathbf{x}} \| \pi) + \log \mathbb{E}_{\mathbf{x}} \mathbb{E}_{\theta \sim \pi}[\exp\{F(\theta, \mathbf{x})\}] + \log \frac{1}{\delta}.$$

Interpretation

An exponential-moment bound under the prior becomes a high-probability bound for a data-dependent posterior.

Proof: Posterior-Mean version

Single-Draw Version

Theorem

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

$$F(\theta, \mathbf{x}) \leq \log \frac{d\rho_{\mathbf{x}}}{d\pi}(\theta) + \log \mathbb{E}_{\mathbf{x}} \mathbb{E}_{\vartheta \sim \pi}[\exp\{F(\vartheta, \mathbf{x})\}] + \log \frac{1}{\delta}.$$

Viewpoint	Randomness	Complexity term
posterior mean	$\mathbf{x} \sim \mu^n$	$\mathrm{KL}(\rho_{\mathbf{x}} \| \pi)$
single draw	$\mathbf{x} \sim \mu^n, \theta \sim \rho_{\mathbf{x}}$	$\log \frac{d\rho_{\mathbf{x}}}{d\pi}(\theta)$

Proof:

Why the Theorem Works

The proof is essentially Markov's inequality applied to an exponential.

For any random variable Y,

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

For the posterior-mean version, use

$$Y = \mathbb{E}_{\theta \sim \rho_{\mathbf{x}}}\left[F(\theta, \mathbf{x})\right].$$

Then Jensen's inequality and the change of measure $d\rho = (d\rho/d\pi)d\pi$ convert the moment under $\rho_{\mathbf{x}}$ into a moment under π.

The Reusable PAC-Bayes Recipe

PAC-Bayes bounds are obtained by choosing F.

1. Choose $F(\theta, \mathbf{x})$ to encode the generalization quantity of interest.
2. Prove or use a bound on
$$\log \mathbb{E}_{\mathbf{x}} \mathbb{E}_{\theta \sim \pi}[e^{F(\theta, \mathbf{x})}].$$
3. Insert the moment bound into the PAC-Bayes theorem.

Two standard choices

- sub-Gaussian losses: choose F proportional to $R(\theta) - r(\theta)$,
- bounded losses: choose F using the binary kl divergence.

Roadmap

1. Review
2. PAC-Bayes Framework
3. **Sub-Gaussian PAC-Bayes Bound**
4. kl-PAC-Bayes for Bounded Losses
5. The Gibbs Posterior

Sub-Gaussian Generalization Gap

Assume that for every $\theta \in \Theta$, the loss fluctuation

$$\ell(\theta, X) - R(\theta)$$

is σ-sub-Gaussian.

Sub-Gaussian Generalization Gap

Assume that for every $\theta \in \Theta$, the loss fluctuation

$$\ell(\theta, X) - R(\theta)$$

is σ-sub-Gaussian. Then the empirical risk satisfies

$$r(\theta) - R(\theta) \quad \text{is} \quad \frac{\sigma}{\sqrt{n}}\text{-sub-Gaussian.}$$

Sub-Gaussian Generalization Gap

Assume that for every $\theta \in \Theta$, the loss fluctuation

$$\ell(\theta, X) - R(\theta)$$

is σ-sub-Gaussian. Then the empirical risk satisfies

$$r(\theta) - R(\theta) \quad \text{is} \quad \frac{\sigma}{\sqrt{n}}\text{-sub-Gaussian.}$$

Choose

$$F(\theta, \mathbf{x}) = \lambda\big(R(\theta) - r(\theta)\big), \qquad \lambda > 0.$$

Then

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

PAC-Bayes Bound for Sub-Gaussian Losses

For any fixed $\lambda > 0$, with probability at least $1 - \delta$,

$$\mathbb{E}_{\theta \sim \rho_{\mathbf{x}}}[R(\theta)] \leq \mathbb{E}_{\theta \sim \rho_{\mathbf{x}}}[r(\theta)] + \frac{\lambda \sigma^2}{2n} + \frac{\mathrm{KL}(\rho_{\mathbf{x}} \| \pi) + \log(1/\delta)}{\lambda}.$$

Message

true risk $\leq$ empirical risk $+$ concentration price $+$ complexity price.

Single-Draw Sub-Gaussian Bound

For any fixed $\lambda > 0$, with probability at least $1 - \delta$ over $\mathbf{x} \sim \mu^n$ and $\theta \sim \rho_{\mathbf{x}}$,

$$R(\theta) \leq r(\theta) + \frac{\lambda \sigma^2}{2n} + \frac{\log \frac{d\rho_{\mathbf{x}}}{d\pi}(\theta) + \log(1/\delta)}{\lambda}.$$

Posterior mean vs. single draw

Averaging over the posterior replaces the pointwise density ratio by its posterior average:

$$\mathbb{E}_{\theta \sim \rho_{\mathbf{x}}}[\log \frac{d\rho_{\mathbf{x}}}{d\pi}(\theta)] = \mathrm{KL}(\rho_{\mathbf{x}} \| \pi).$$

Roadmap

1. Review
2. PAC-Bayes Framework
3. Sub-Gaussian PAC-Bayes Bound
4. **kl-PAC-Bayes for Bounded Losses**
5. The Gibbs Posterior

Binary kl Divergence

Now assume

$$\ell(\theta, x) \in [0, 1].$$

For $p, q \in [0, 1]$, define

$$\mathrm{kl}(p, q) = p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q}.$$

Binary kl Divergence

Now assume

$$\ell(\theta, x) \in [0, 1].$$

For $p, q \in [0, 1]$, define

$$\mathrm{kl}(p, q) = p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q}.$$

For bounded losses, a sharper choice is

$$F(\theta, \mathbf{x}) = n \, \mathrm{kl}\big(r(\theta), R(\theta)\big).$$

Binary kl Divergence

Now assume

$$\ell(\theta, x) \in [0, 1].$$

For $p, q \in [0, 1]$, define

$$\mathrm{kl}(p, q) = p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q}.$$

For bounded losses, a sharper choice is

$$F(\theta, \mathbf{x}) = n \, \mathrm{kl}\big(r(\theta), R(\theta)\big).$$

The key moment bound is, [Maurer, 2004]

$$\mathbb{E}_{\mathbf{x}}[\exp\Big(n \, \mathrm{kl}(r(\theta), R(\theta))\Big)] \leq 2\sqrt{n}, \qquad n \geq 8.$$

kl-PAC-Bayes Bound: Posterior Mean

Theorem

Using joint convexity of kl*, with probability at least* $1-\delta$*,*

$$\mathrm{kl}\left(\mathbb{E}_{\theta\sim\rho_{\mathbf{x}}}[r(\theta)], \mathbb{E}_{\theta\sim\rho_{\mathbf{x}}}[R(\theta)]\right) \leq \frac{\mathrm{KL}(\rho_{\mathbf{x}}\|\pi) + \log\frac{2\sqrt{n}}{\delta}}{n}.$$

Interpretation

This directly relates the posterior-averaged empirical risk to the posterior-averaged true risk.

kl-PAC-Bayes Bound: Single Draw

Theorem

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

$$\mathrm{kl}\big(r(\theta), R(\theta)\big) \leq \frac{\log \frac{d\rho_{\mathbf{x}}}{d\pi}(\theta) + \log \frac{2\sqrt{n}}{\delta}}{n}.$$

Comparison

- posterior-mean bound: pays $\mathrm{KL}(\rho_{\mathbf{x}} \| \pi)$,
- single-draw bound: pays $\log \frac{d\rho_{\mathbf{x}}}{d\pi}(\theta)$.

Turning the kl Bound into a Risk Bound

Define

$$\mathrm{kl}^{-1}(q, \varepsilon) = \sup\{p \in [q, 1] : \mathrm{kl}(q, p) \leq \varepsilon\}.$$

Turning the kl Bound into a Risk Bound

Define

$$\mathrm{kl}^{-1}(q, \varepsilon) = \sup\{p \in [q, 1] : \mathrm{kl}(q, p) \leq \varepsilon\}.$$

Then the posterior-mean kl bound implies

$$\mathbb{E}_{\theta \sim \rho_{\mathbf{x}}}[R(\theta)] \leq \mathrm{kl}^{-1}\left(\mathbb{E}_{\theta \sim \rho_{\mathbf{x}}}[r(\theta)], \frac{\mathrm{KL}(\rho_{\mathbf{x}} \| \pi) + \log \frac{2\sqrt{n}}{\delta}}{n}\right).$$

Turning the kl Bound into a Risk Bound

Define

$$\mathrm{kl}^{-1}(q, \varepsilon) = \sup\{p \in [q, 1] : \mathrm{kl}(q, p) \leq \varepsilon\}.$$

Then the posterior-mean kl bound implies

$$\mathbb{E}_{\theta \sim \rho_{\mathbf{x}}}[R(\theta)] \leq \mathrm{kl}^{-1}\left(\mathbb{E}_{\theta \sim \rho_{\mathbf{x}}}[r(\theta)], \frac{\mathrm{KL}(\rho_{\mathbf{x}} \| \pi) + \log \frac{2\sqrt{n}}{\delta}}{n}\right).$$

A looser but more readable version is

$$\mathbb{E}_{\theta \sim \rho_{\mathbf{x}}}[R(\theta)] \leq \mathbb{E}_{\theta \sim \rho_{\mathbf{x}}}[r(\theta)] + \sqrt{\frac{\mathrm{KL}(\rho_{\mathbf{x}} \| \pi) + \log \frac{2\sqrt{n}}{\delta}}{2n}}.$$

A tighter but messier version is

$$\mathbb{E}_{\theta \sim \rho_{\mathbf{x}}}[R(\theta)] \leq \mathbb{E}_{\theta \sim \rho_{\mathbf{x}}}[r(\theta)] + \sqrt{\frac{2\mathbb{E}_{\theta \sim \rho_{\mathbf{x}}}[r(\theta)](\mathrm{KL}(\rho_{\mathbf{x}} \| \pi) + \log \frac{2\sqrt{n}}{\delta})}{n}} + \frac{2(\mathrm{KL}(\rho_{\mathbf{x}} \| \pi) + \log \frac{2\sqrt{n}}{\delta})}{n}.$$

Pinsker's Inequality

Define

$$\mathrm{kl}^{-1}(q, \varepsilon) = \sup\{p \in [q, 1] : \mathrm{kl}(q, p) \leq \varepsilon\}.$$

Then the posterior-mean kl bound implies

$$\mathbb{E}_{\theta \sim \rho_{\mathbf{x}}}[R(\theta)] \leq \mathrm{kl}^{-1}\left(\mathbb{E}_{\theta \sim \rho_{\mathbf{x}}}[r(\theta)], \frac{\mathrm{KL}(\rho_{\mathbf{x}} \| \pi) + \log \frac{2\sqrt{n}}{\delta}}{n}\right).$$

- Using Pinsker's inequality [Boucheron et al., 2013, Theorem 4.19] we know $\kappa(p, q) \geq 2(p - q)^2$, and it implies thats $\kappa^{-1}(p, b) \leq p + \frac{\sqrt{b}}{2}$.

Pinsker's Inequality

Define

$$\mathrm{kl}^{-1}(q, \varepsilon) = \sup\{p \in [q, 1] : \mathrm{kl}(q, p) \leq \varepsilon\}.$$

Then the posterior-mean kl bound implies

$$\mathbb{E}_{\theta \sim \rho_{\mathbf{x}}}[R(\theta)] \leq \mathrm{kl}^{-1}\left(\mathbb{E}_{\theta \sim \rho_{\mathbf{x}}}[r(\theta)], \frac{\mathrm{KL}(\rho_{\mathbf{x}} \| \pi) + \log \frac{2\sqrt{n}}{\delta}}{n}\right).$$

- Using Pinsker's inequality [Boucheron et al., 2013, Theorem 4.19] we know $\kappa(p, q) \geq 2(p - q)^2$, and it implies thats $\kappa^{-1}(p, b) \leq p + \frac{\sqrt{b}}{2}$.
- Also, using the relaxed version of Pisnker's inequality [Boucheron et al., 2013, Lemma 8.4] we know $\kappa(p, q) \geq \frac{(p-q)^2}{2q}$ one can show that $\kappa^{-1}(p, b) \leq p + \sqrt{2pb} + 2b$.

Roadmap

1. Review
2. PAC-Bayes Framework
3. Sub-Gaussian PAC-Bayes Bound
4. kl-PAC-Bayes for Bounded Losses
5. **The Gibbs Posterior**

Why Minimize the Bound?

A finite-class PAC bound motivates ERM:

$$R(\hat{\theta}_{\mathrm{ERM}}) \lesssim r(\hat{\theta}_{\mathrm{ERM}}) + \text{complexity}.$$

Why Minimize the Bound?

A finite-class PAC bound motivates ERM:

$$R(\hat{\theta}_{\mathrm{ERM}}) \lesssim r(\hat{\theta}_{\mathrm{ERM}}) + \text{complexity}.$$

So we choose the predictor that minimizes the empirical risk:

$$\hat{\theta}_{\mathrm{ERM}} \in \arg\min_{\theta \in \Theta} r(\theta).$$

Why Minimize the Bound?

A finite-class PAC bound motivates ERM:

$$R(\hat{\theta}_{\mathrm{ERM}}) \lesssim r(\hat{\theta}_{\mathrm{ERM}}) + \text{complexity}.$$

So we choose the predictor that minimizes the empirical risk:

$$\hat{\theta}_{\mathrm{ERM}} \in \arg\min_{\theta \in \Theta} r(\theta).$$

PAC-Bayes gives a bound over probability measures ρ:

$$\mathbb{E}_{\theta\sim\rho}[R(\theta)] \lesssim \mathbb{E}_{\theta\sim\rho}[r(\theta)] + \frac{1}{\lambda}\mathrm{KL}(\rho\|\pi).$$

Why Minimize the Bound?

A finite-class PAC bound motivates ERM:

$$R(\hat{\theta}_{\mathrm{ERM}}) \lesssim r(\hat{\theta}_{\mathrm{ERM}}) + \text{complexity}.$$

So we choose the predictor that minimizes the empirical risk:

$$\hat{\theta}_{\mathrm{ERM}} \in \arg\min_{\theta \in \Theta} r(\theta).$$

PAC-Bayes gives a bound over probability measures ρ:

$$\mathbb{E}_{\theta\sim\rho}[R(\theta)] \lesssim \mathbb{E}_{\theta\sim\rho}[r(\theta)] + \frac{1}{\lambda}\mathrm{KL}(\rho\|\pi).$$

Natural PAC-Bayes learning rule

Choose the posterior that minimizes the right-hand side:

$$\hat{\rho}_\lambda \in \arg\min_{\rho\in\mathcal{P}(\Theta)} \left\{ \mathbb{E}_{\theta\sim\rho}[r(\theta)] + \frac{1}{\lambda}\mathrm{KL}(\rho\|\pi) \right\}.$$

Which Posterior Minimizes the Bound?

Ignoring constants, the sub-Gaussian PAC-Bayes bound suggests minimizing

$$J(\rho) = \mathbb{E}_{\theta\sim\rho}[r(\theta)] + \frac{1}{\beta}\mathrm{KL}(\rho\|\pi), \qquad \beta > 0.$$

Interpretation

- $r(\theta)$: fit the data well;
- $\mathrm{KL}(\rho\|\pi)$: stay close to the prior;
- β: inverse temperature controlling the trade-off.

Donsker–Varadhan Variational Formula

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

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

The supremum is reached at the Gibbs measure π_h:

$$\frac{d\pi_h}{d\pi}(\theta) = \frac{e^{h(\theta)}}{\mathbb{E}_{\vartheta\sim\pi}\left[e^{h(\vartheta)}\right]}.$$

Key message

Link to our bound: min $\Rightarrow$ max, $h(\theta) = -\beta r(\theta)$

Key message

Exponentials turn optimization over distributions into a Gibbs measure.

Proof Idea I: KL Non-negativity

Define

$$\frac{d\pi_h}{d\pi}(\theta) = \frac{e^{h(\theta)}}{\mathbb{E}_{\vartheta\sim\pi}[e^{h(\vartheta)}]}.$$

For any $\rho \in \mathcal{P}(\Theta)$,

$$\mathrm{KL}(\rho\|\pi_h) = \mathbb{E}_\rho\left[\log\frac{d\rho}{d\pi_h}\right].$$

Using the definition of π_h,

$$\mathrm{KL}(\rho\|\pi_h) = \mathrm{KL}(\rho\|\pi) - \mathbb{E}_\rho[h(\theta)] + \log\mathbb{E}_\pi[e^{h(\theta)}].$$

Since $\mathrm{KL}(\rho\|\pi_h) \geq 0$,

$$\mathbb{E}_\rho[h(\theta)] - \mathrm{KL}(\rho\|\pi) \leq \log\mathbb{E}_\pi[e^{h(\theta)}].$$

Equality holds for $\rho = \pi_h$.

Proof Idea II: Change of Measure + Jensen

For any $\rho \ll \pi$,

$$\log \mathbb{E}_\pi[e^{h(\theta)}] = \log \mathbb{E}_\rho \left[e^{h(\theta)} \frac{d\pi}{d\rho}(\theta) \right].$$

By Jensen's inequality,

$$\log \mathbb{E}_\rho \left[e^{h(\theta)} \frac{d\pi}{d\rho}(\theta) \right] \geq \mathbb{E}_\rho \left[h(\theta) + \log \frac{d\pi}{d\rho}(\theta) \right].$$

Therefore,

$$\log \mathbb{E}_\pi[e^{h(\theta)}] \geq \mathbb{E}_\rho[h(\theta)] - \mathrm{KL}(\rho \| \pi).$$

Equality holds when

$$\frac{d\rho}{d\pi}(\theta) \propto e^{h(\theta)}.$$

Proof Idea III: Variational Derivation

Write $u(\theta) = d\rho/d\pi(\theta)$. Then

$$J(u) = \int r(\theta)u(\theta)d\pi(\theta) + \frac{1}{\beta}\int u(\theta)\log u(\theta)d\pi(\theta),$$

subject to

$$\int u(\theta)d\pi(\theta) = 1.$$

The first-order condition gives

$$r(\theta) + \frac{1}{\beta}(\log u(\theta) + 1) + \mu = 0,$$

hence

$$u(\theta) = Ce^{-\beta r(\theta)}.$$

Normalization gives the Gibbs posterior.

Intuition Behind Donsker–Varadhan

Donsker–Varadhan says:

$$\log \mathbb{E}_\pi \left[e^{h(\theta)} \right] = \sup_\rho \left\{ \mathbb{E}_\rho[h(\theta)] - \mathrm{KL}(\rho \| \pi) \right\}.$$

Interpretation

To make the average value of $h(\theta)$ large, we may move from the prior π to another distribution ρ, but we must pay the price $\mathrm{KL}(\rho \| \pi)$.

$$\text{reward} - \text{complexity penalty}$$

The optimal trade-off is obtained by the Gibbs distribution:

$$\frac{d\rho^*}{d\pi}(\theta) \propto e^{h(\theta)}.$$

Applying Donsker–Varadhan

Choose

$$h(\theta) = -\beta r(\theta).$$

Then

$$Z_\beta(\mathbf{x}) = \mathbb{E}_{\vartheta\sim\pi}\left[e^{-\beta r(\vartheta)}\right].$$

Donsker–Varadhan gives

$$\log Z_\beta(\mathbf{x}) = \sup_\rho \left\{-\beta\mathbb{E}_{\theta\sim\rho}[r(\theta)] - \mathrm{KL}(\rho\|\pi)\right\}.$$

Equivalently,

$$-\frac{1}{\beta}\log Z_\beta(\mathbf{x}) = \inf_\rho \left\{\mathbb{E}_{\theta\sim\rho}[r(\theta)] + \frac{1}{\beta}\mathrm{KL}(\rho\|\pi)\right\}.$$

Thus the minimizer of $J(\rho)$ is the Gibbs posterior.

The Gibbs Posterior

The minimizer has density

$$\frac{dG_\beta(\mathbf{x})}{d\pi}(\theta) = \frac{e^{-\beta r(\theta)}}{Z_\beta(\mathbf{x})},$$

where

$$Z_\beta(\mathbf{x}) = \mathbb{E}_{\vartheta\sim\pi}\left[e^{-\beta r(\vartheta)}\right].$$

So

$$G_\beta(d\theta \mid \mathbf{x}) = \frac{e^{-\beta r(\theta)}}{Z_\beta(\mathbf{x})}\,\pi(d\theta).$$

Interpretation

Small empirical risk gets high weight, but the prior still regularizes the posterior.

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.


Maurer, A. (2004).
A note on the PAC Bayesian theorem.
arXiv preprint cs/0411099.


McAllester, D. A. (1999).
PAC-Bayesian model averaging.
In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, pages 164–170.

References II


Rivasplata, O., Kuzborskij, I., Szepesvári, C., and Shawe-Taylor, J. (2020).
PAC-Bayes analysis beyond the usual bounds.
Advances in Neural Information Processing Systems, 33:16833–16845.


Shawe-Taylor, J. and Williamson, R. C. (1997).
A PAC analysis of a Bayesian estimator.
In *Proceedings of the tenth annual conference on Computational learning theory*, pages 2–9.