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

Fifth Session: 2026-07-19

Roadmap

1. Review
2. The Gibbs Posterior
3. Motivation: The generalization puzzle in the interpolation regime
4. Key idea: The temperature path is all you need!
5. Experiments: Main principle and quantitative bounds
6. Takeaways

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-dep quantities the supremum in Eq. **??** needs to be replaced by:

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

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:

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

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:

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

The PAC-Bayes 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* $\mathbf{x} \sim \mu^n$,

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

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

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

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.

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

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

Then

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

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

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

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

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

Binary kl Divergence

Now assume $\ell(\theta, x) \in [0,1]$. For $p, q \in [0,1]$, define

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

The key moment bound is, [Maurer, 2004]

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

Theorem

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

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

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

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

Roadmap

1. Review
2. **The Gibbs Posterior**
3. Motivation: The generalization puzzle in the interpolation regime
4. Key idea: The temperature path is all you need!
5. Experiments: Main principle and quantitative bounds
6. Takeaways

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

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

Proof Idea I: KL Non-negativity

Define

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

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

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

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,

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

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

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

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 $\rho \ll \pi$,

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

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,

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

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,

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

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

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

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.

Log-Density Ratio and the Partition Function

For $\rho_{\mathbf{x}} = G_\beta(\mathbf{x})$,

$$\log \frac{dG_\beta(\mathbf{x})}{d\pi}(\theta) = -\beta r(\theta) - \log Z_\beta(\mathbf{x}).$$

The term $-\log Z_\beta(\mathbf{x})$ is the normalizing-price of concentrating the posterior around low empirical risk hypotheses.

Why this matters

In single-draw PAC-Bayes bounds, the density ratio appears directly. For Gibbs posteriors, this density ratio can be expressed through the empirical risk and the partition function.

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

Define

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

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

Define

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

Then

$$A'(\beta) = \mathbb{E}_{\theta \sim G_\beta(\mathbf{x})}[r(\theta)],$$

and

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

Define

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

Then

$$A'(\beta) = \mathbb{E}_{\theta \sim G_\beta(\mathbf{x})}[r(\theta)],$$

and

$$A''(\beta) = -\mathrm{Var}_{\theta \sim G_\beta(\mathbf{x})}(r(\theta)) \leq 0.$$

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

Define

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

Then

$$A'(\beta) = \mathbb{E}_{\theta \sim G_\beta(\mathbf{x})}[r(\theta)],$$

and

$$A''(\beta) = -\mathrm{Var}_{\theta \sim G_\beta(\mathbf{x})}(r(\theta)) \leq 0.$$

Therefore,

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

A Computable Upper Bound

Let $0 = \beta_0 < \beta_1 < \cdots < \beta_K = \beta$. Since A is concave,

$$-\log Z_\beta(\mathbf{x}) \leq \sum_{k=1}^{K} (\beta_k - \beta_{k-1}) \mathbb{E}_{\theta \sim G_{\beta_{k-1}}(\mathbf{x})}[r(\theta)].$$

A Computable Upper Bound

Let $0 = \beta_0 < \beta_1 < \cdots < \beta_K = \beta$. Since A is concave,

$$-\log Z_\beta(\mathbf{x}) \leq \sum_{k=1}^{K} (\beta_k - \beta_{k-1}) \mathbb{E}_{\theta \sim G_{\beta_{k-1}}(\mathbf{x})}[r(\theta)].$$

Thus, for a sampled θ,

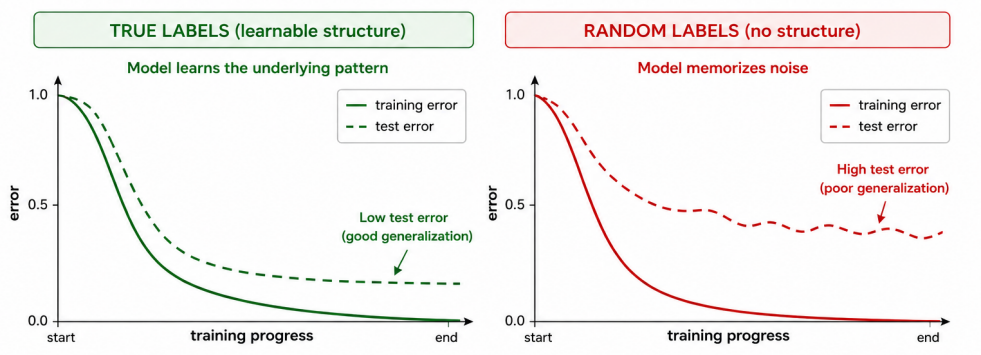
$$\log \frac{dG_\beta(\mathbf{x})}{d\pi}(\theta) \leq -\beta r(\theta) + \sum_{k=1}^{K} (\beta_k - \beta_{k-1}) \mathbb{E}_{\vartheta \sim G_{\beta_{k-1}}(\mathbf{x})}[r(\vartheta)].$$

Roadmap

1. Review
2. The Gibbs Posterior
3. **Motivation: The generalization puzzle in the interpolation regime**
4. Key idea: The temperature path is all you need!
5. Experiments: Main principle and quantitative bounds
6. Takeaways

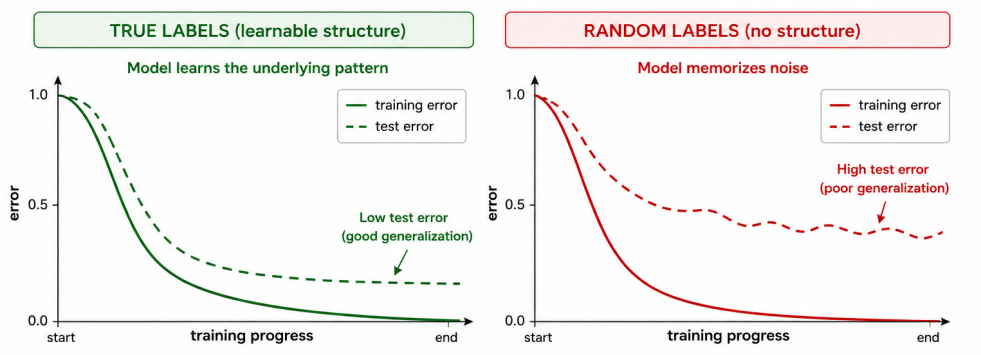
The Puzzle

Both models fit the training data. Only one generalizes.



The Puzzle

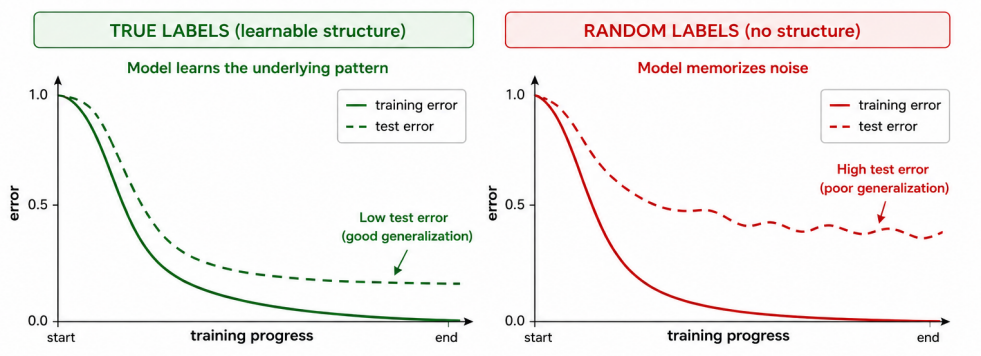
Both models fit the training data. Only one generalizes.



When train error is zero for any data, what predicts test error?

The Puzzle

Both models fit the training data. Only one generalizes.

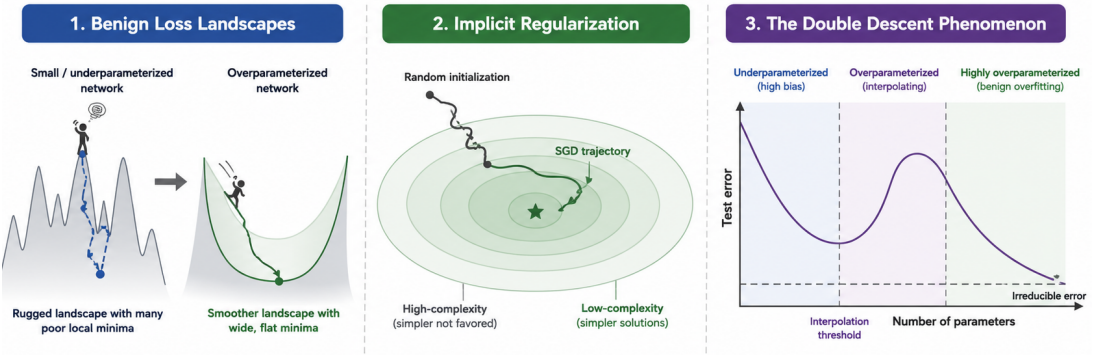


When train error is zero for any data, what predicts test error?

$\Rightarrow$ the key to generalization must be buried in the **structure of the data**

Why It Matters?

overparameterized models change the behavior of learning



Roadmap

1. Review
2. The Gibbs Posterior
3. Motivation: The generalization puzzle in the interpolation regime
4. **Key idea: The temperature path is all you need!**
5. Experiments: Main principle and quantitative bounds
6. Takeaways

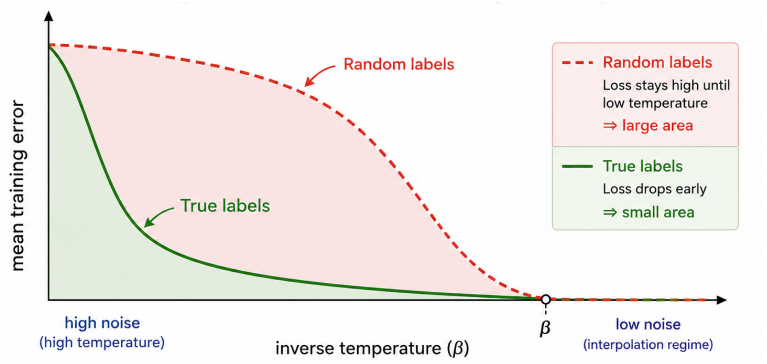
The Temperature Path Is All You Need?

Do not look only at the final model. Look at the whole **temperature path**.

The Temperature Path Is All You Need?

Do not look only at the final model. Look at the whole **temperature path**.

How early does the loss drop as inverse temperature increases?



Notation

- **Hypothesis Space:** $\mathcal{H}$ (e.g., neural network weights, $w \in \mathbb{R}^d$)

Notation

- **Hypothesis Space:** $\mathcal{H}$ (e.g., neural network weights, $w \in \mathbb{R}^d$)
- **Measures:** $\mathcal{P}(\mathcal{H})$ space of all probability measures on $\mathcal{H}$
  - **Prior:** fixed $\pi \in \mathcal{P}(\mathcal{H})$, independent of data

Notation

- **Hypothesis Space:** $\mathcal{H}$ (e.g., neural network weights, $w \in \mathbb{R}^d$)
- **Measures:** $\mathcal{P}(\mathcal{H})$ space of all probability measures on $\mathcal{H}$
  - **Prior:** fixed $\pi \in \mathcal{P}(\mathcal{H})$, independent of data
- **Information Divergences:** For $\nu, \pi \in \mathcal{P}(\mathcal{H})$:
  - **Relative Entropy:** $KL(\nu, \pi) = \mathbb{E}_{h\sim\nu}\left[\ln(\frac{d\nu}{d\pi}(h))\right]$
  - **Rényi-infinity:** $R_\infty(\nu, \pi) = \sup_{h\in\mathcal{H}} \ln(\frac{d\nu}{d\pi}(h))$

Notation

- **Hypothesis Space:** $\mathcal{H}$ (e.g., neural network weights, $w \in \mathbb{R}^d$)
- **Measures:** $\mathcal{P}(\mathcal{H})$ space of all probability measures on $\mathcal{H}$
  - **Prior:** fixed $\pi \in \mathcal{P}(\mathcal{H})$, independent of data
- **Information Divergences:** For $\nu, \pi \in \mathcal{P}(\mathcal{H})$:
  - **Relative Entropy:** $KL(\nu, \pi) = \mathbb{E}_{h\sim\nu}\left[\ln(\frac{d\nu}{d\pi}(h))\right]$
  - **Rényi-infinity:** $R_\infty(\nu, \pi) = \sup_{h\in\mathcal{H}} \ln(\frac{d\nu}{d\pi}(h))$
- **The Learning Process:**
  - **Data**: $\mathbf{x} \in \mathcal{X}^n$ (typically an i.i.d. vector $\mathbf{x} \sim \mu^n$ for $\mu \in \mathcal{P}(\mathcal{X})$)
  - **Empirical Error**: $\hat{L}(h, \mathbf{x})$, non-negative(typically $\frac{1}{n}\sum \ell(h, x_i)$)
  - **True Error**: $L(h) = \mathbb{E}_{\mathbf{x}}[\hat{L}(h, \mathbf{x})]$
  - **Stochastic Learning Algorithm:** $\nu : \mathcal{X}^n \to \mathcal{P}(\mathcal{H})$, aka data-dependent prob measure

PAC-Bayes Bounds & Gibbs Posterior

Typical PAC-Bayes bounds give upper bounds on **true error**, which include:

1. An **empirical risk** term: $\hat{L}(h, \mathbf{x})$
2. A **complexity penalty** term: $KL(\nu(\mathbf{x}), \pi)$

$$\mathbb{E}_{h\sim\nu(\mathbf{x})}[L(h)] \leq \Psi\left(\mathbb{E}_{h\sim\nu(\mathbf{x})}[\hat{L}(h, \mathbf{x})], KL(\nu(\mathbf{x}), \pi)\right)$$

The Gibbs posterior, $G_\beta(\mathbf{x})$, is the unique minimizer of the RHS of these bounds.

$$G_\beta(\mathbf{x})(dh) = \frac{\exp(-\beta\hat{L}(h, \mathbf{x}))}{Z_\beta(\mathbf{x})}\pi(dh), \quad \text{with} \quad Z_\beta(\mathbf{x}) = \int_{\mathcal{H}} e^{-\beta\,\hat{L}(h,\mathbf{x})} d\pi(h).$$

Integral Representation of Gibbs Complexity

Lemma (Integral Representation)

For every $\beta \geq 0$, $\mathbf{x} \in \mathcal{X}^n$, and $h \in \mathcal{H}$,

$$
\begin{aligned}
-\ln Z_\beta(\mathbf{x}) &= \int_0^\beta \mathbb{E}_{h \sim G_\gamma(\mathbf{x})}\left[\hat{L}(h, \mathbf{x})\right] d\gamma \\
KL(G_\beta(\mathbf{x}), \pi) &= -\beta\, \mathbb{E}_{g \sim G_\beta(\mathbf{x})}\left[\hat{L}(g, \mathbf{x})\right] + \int_0^\beta \mathbb{E}_{g \sim G_\gamma(\mathbf{x})}\left[\hat{L}(g, \mathbf{x})\right] d\gamma
\end{aligned}
$$

Integral Representation of Gibbs Complexity

Lemma (Integral Representation)

For every $\beta \geq 0$, $\mathbf{x} \in \mathcal{X}^n$, *and* $h \in \mathcal{H}$,

$$
\begin{aligned}
-\ln Z_\beta(\mathbf{x}) &= \int_0^\beta \mathbb{E}_{h \sim G_\gamma(\mathbf{x})}\left[\hat{L}(h, \mathbf{x})\right] d\gamma \\
KL(G_\beta(\mathbf{x}), \pi) &= -\beta\, \mathbb{E}_{g \sim G_\beta(\mathbf{x})}\left[\hat{L}(g, \mathbf{x})\right] + \int_0^\beta \mathbb{E}_{g \sim G_\gamma(\mathbf{x})}\left[\hat{L}(g, \mathbf{x})\right] d\gamma
\end{aligned}
$$

How can this lemma explain generalization in the interpolation regime?

Integral Representation of Gibbs Complexity

Lemma (Integral Representation)

For every $\beta \geq 0$, $\mathbf{x} \in \mathcal{X}^n$, *and* $h \in \mathcal{H}$,

$$
\begin{aligned}
-\ln Z_\beta(\mathbf{x}) &= \int_0^\beta \mathbb{E}_{h \sim G_\gamma(\mathbf{x})}\left[\hat{L}(h, \mathbf{x})\right] d\gamma \\
KL(G_\beta(\mathbf{x}), \pi) &= -\beta\, \mathbb{E}_{g \sim G_\beta(\mathbf{x})}\left[\hat{L}(g, \mathbf{x})\right] + \int_0^\beta \mathbb{E}_{g \sim G_\gamma(\mathbf{x})}\left[\hat{L}(g, \mathbf{x})\right] d\gamma
\end{aligned}
$$

How can this lemma explain generalization in the interpolation regime?

- In interpolation: empirical risk disappears

Integral Representation of Gibbs Complexity

Lemma (Integral Representation)

For every $\beta \geq 0$, $\mathbf{x} \in \mathcal{X}^n$, *and* $h \in \mathcal{H}$,

$$
\begin{aligned}
-\ln Z_\beta(\mathbf{x}) &= \int_0^\beta \mathbb{E}_{h \sim G_\gamma(\mathbf{x})}\left[\hat{L}(h, \mathbf{x})\right] d\gamma \\
KL(G_\beta(\mathbf{x}), \pi) &= -\beta\, \mathbb{E}_{g \sim G_\beta(\mathbf{x})}\left[\hat{L}(g, \mathbf{x})\right] + \int_0^\beta \mathbb{E}_{g \sim G_\gamma(\mathbf{x})}\left[\hat{L}(g, \mathbf{x})\right] d\gamma
\end{aligned}
$$

How can this lemma explain generalization in the interpolation regime?

- In interpolation: empirical risk disappears

$$\text{test error} \lesssim \text{complexity}$$

Integral Representation of Gibbs Complexity

Lemma (Integral Representation)

For every $\beta \geq 0$, $\mathbf{x} \in \mathcal{X}^n$, and $h \in \mathcal{H}$,

$$
\begin{aligned}
-\ln Z_\beta(\mathbf{x}) &= \int_0^\beta \mathbb{E}_{h \sim G_\gamma(\mathbf{x})}\left[\hat{L}(h, \mathbf{x})\right] d\gamma \\
KL(G_\beta(\mathbf{x}), \pi) &= -\beta\, \mathbb{E}_{g \sim G_\beta(\mathbf{x})}\left[\hat{L}(g, \mathbf{x})\right] + \int_0^\beta \mathbb{E}_{g \sim G_\gamma(\mathbf{x})}\left[\hat{L}(g, \mathbf{x})\right] d\gamma
\end{aligned}
$$

How can this lemma explain generalization in the interpolation regime?

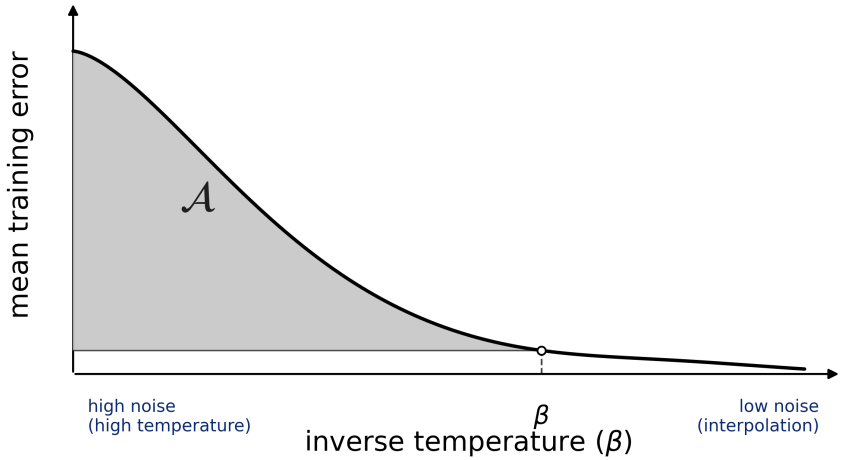
- In interpolation: empirical risk disappears

$$\text{test error} \lesssim \text{complexity}$$

- Previous bounds on the complexity term are vacuous for this regime

Geometric Intuition

For Gibbs posterior: complexity = area under the mean training-loss curve

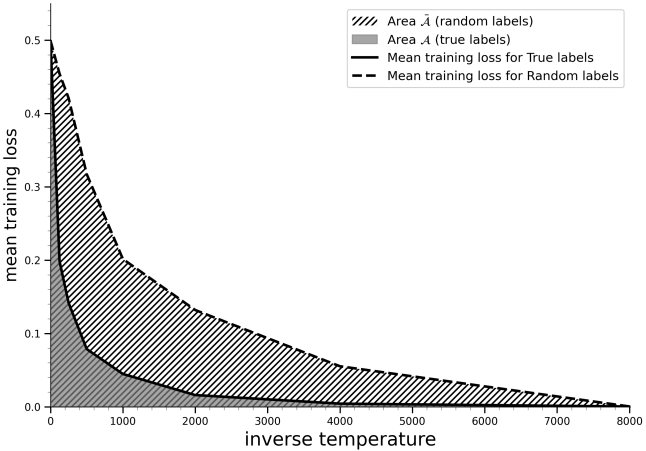


Roadmap

1. Review
2. The Gibbs Posterior
3. Motivation: The generalization puzzle in the interpolation regime
4. Key idea: The temperature path is all you need!
5. **Experiments: Main principle and quantitative bounds**
6. Takeaways

The Main Principle

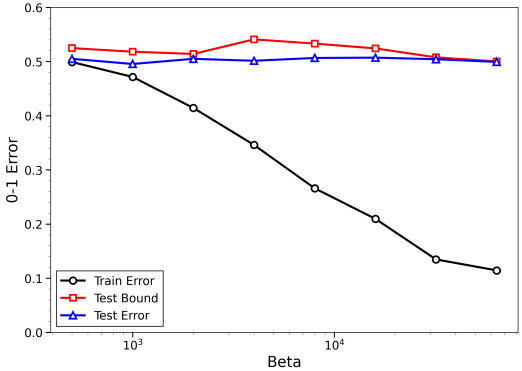
The Main Principle



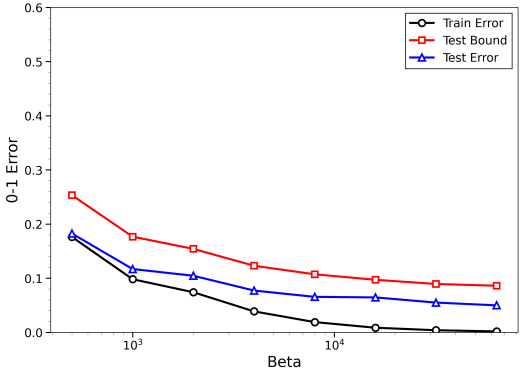
Good generalization at low temperature is announced by low training loss at high temperature

The Bound Tracks Test Error

The Bound Tracks Test Error



(a) Random labels



(b) True labels

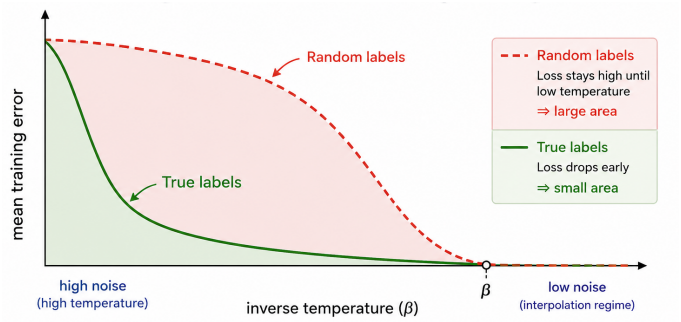
The method predicts the generalization gap from training data

Roadmap

1. Review
2. The Gibbs Posterior
3. Motivation: The generalization puzzle in the interpolation regime
4. Key idea: The temperature path is all you need!
5. Experiments: Main principle and quantitative bounds
6. Takeaways

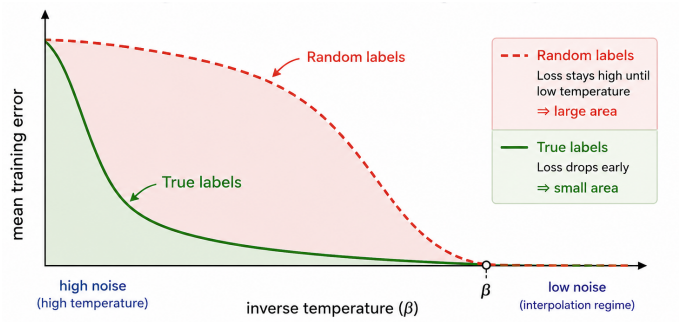
Takeaways

Good generalization at low temperature is announced by low training loss at high temperature



Takeaways

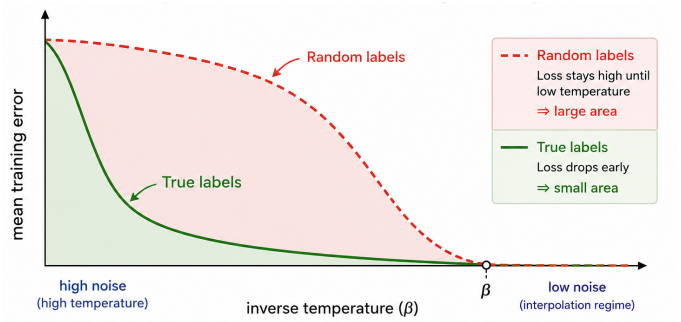
Good generalization at low temperature is announced by low training loss at high temperature



Gibbs complexity = area under the mean training-loss curve

Takeaways

Good generalization at low temperature is announced by low training loss at high temperature

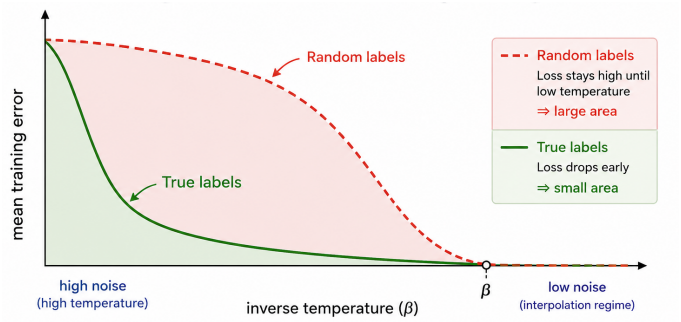


Gibbs complexity = area under the mean training-loss curve

Small area $\Rightarrow$ Better generalization

Takeaways

Good generalization at low temperature is announced by low training loss at high temperature



Gibbs complexity = area under the mean training-loss curve

Small area ⇒ Better generalization

LMC lets us estimate this path in neural networks

References I


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


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.


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.

References II


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.