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

Sixth Session: 2026-07-26

Roadmap

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

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

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

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:

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

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.

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

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

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,

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

Thus, for a sampled θ,

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

Radon–Nikodym Theorem

Let ρ and π be probability measures on Θ.

If ρ is absolutely continuous with respect to π,

$$\rho \ll \pi,$$

then there exists a measurable function

$$\frac{d\rho}{d\pi} : \Theta \to [0, \infty)$$

such that for every measurable set $A \subseteq \Theta$,

$$\rho(A) = \int_A \frac{d\rho}{d\pi}(\theta)\, \pi(d\theta).$$

Meaning

The posterior ρ can be written as a reweighting of the prior π.

Change of Measure Formula

A direct consequence is that for any integrable function $F : \Theta \to \mathbb{R}$,

$$\mathbb{E}_{\theta\sim\rho}[F(\theta)] = \int F(\theta)\,\rho(d\theta) = \int F(\theta)\frac{d\rho}{d\pi}(\theta)\,\pi(d\theta).$$

Equivalently,

$$\mathbb{E}_{\theta\sim\rho}[F(\theta)] = \mathbb{E}_{\theta\sim\pi}\left[F(\theta)\frac{d\rho}{d\pi}(\theta)\right].$$

Why we need this

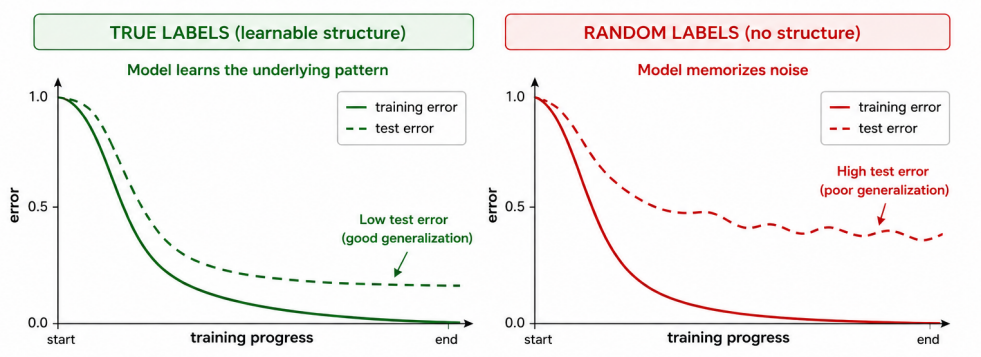
This is what allows PAC-Bayes proofs to move expectations from the posterior ρ back to the prior π.

Roadmap

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

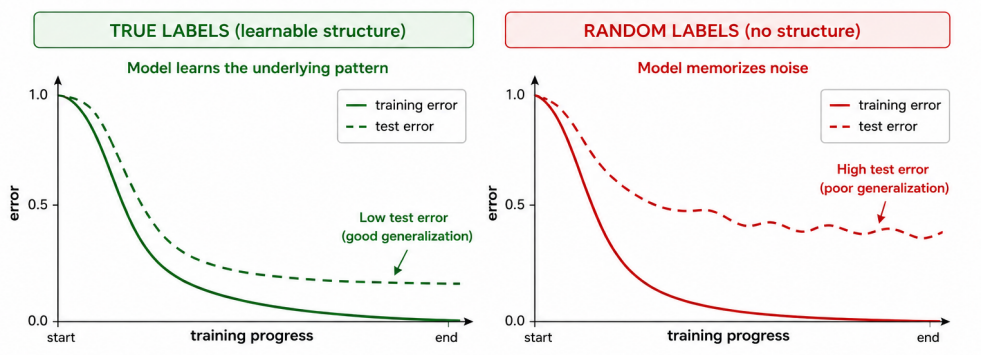
The Puzzle

Both models fit the training data. Only one generalizes.



The Puzzle

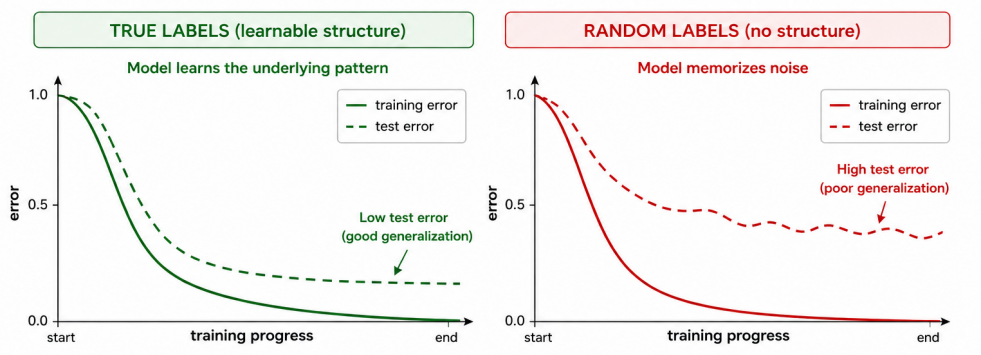
Both models fit the training data. Only one generalizes.



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

The Puzzle

Both models fit the training data. Only one generalizes.

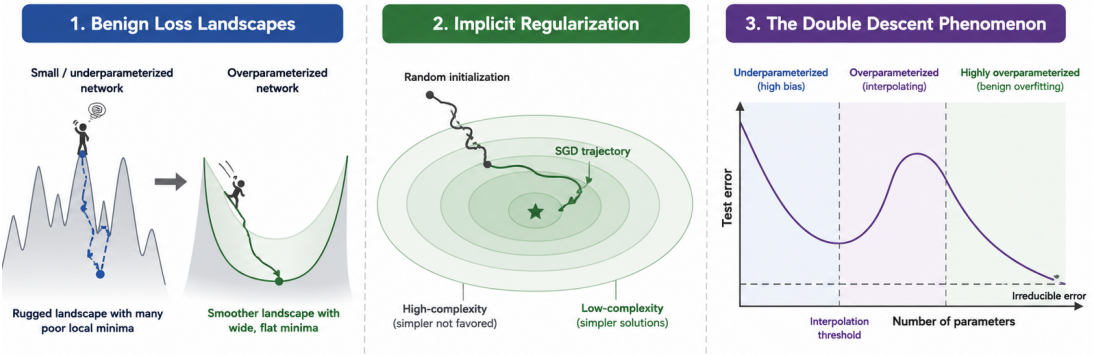


When train error is zero for any data, what predicts test error?

$\Rightarrow$ 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. Motivation: The generalization puzzle in the interpolation regime
3. **Key idea: The temperature path is all you need!**
4. Experiments: Main principle and quantitative bounds
5. Takeaways
6. Gibbs Approximation via Langevin Monte Carlo Algorithms

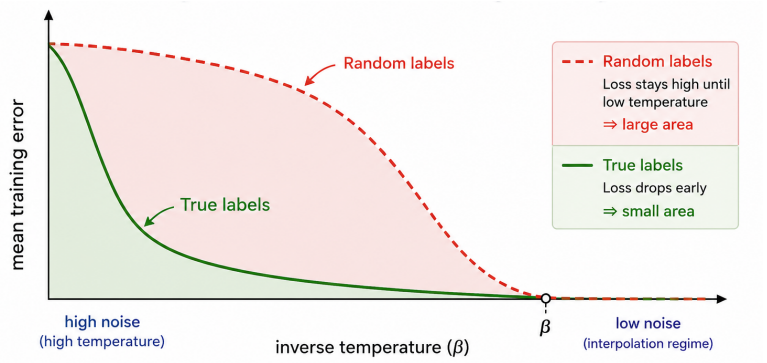
The Temperature Path Is All You Need?

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?

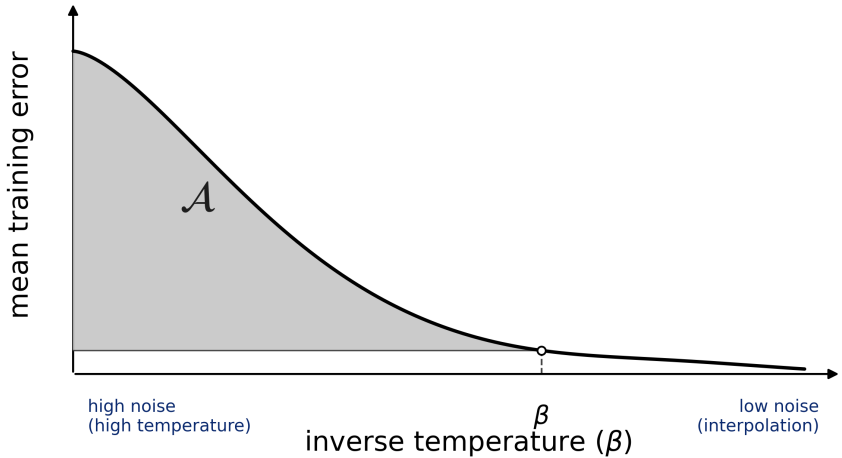
- In interpolation: empirical risk disappears

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

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

Geometric Intuition

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

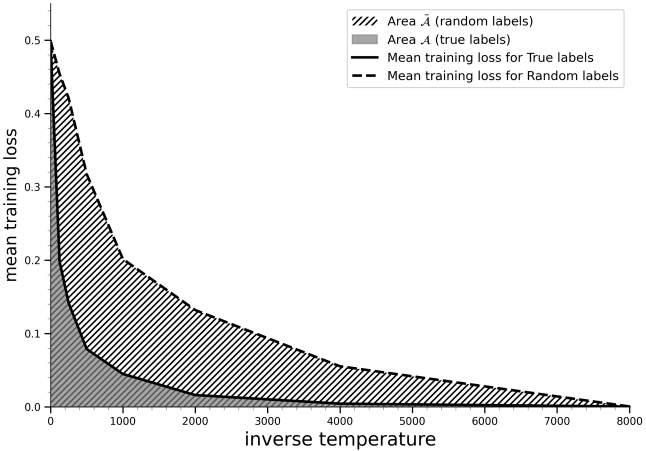


Roadmap

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

The Main Principle

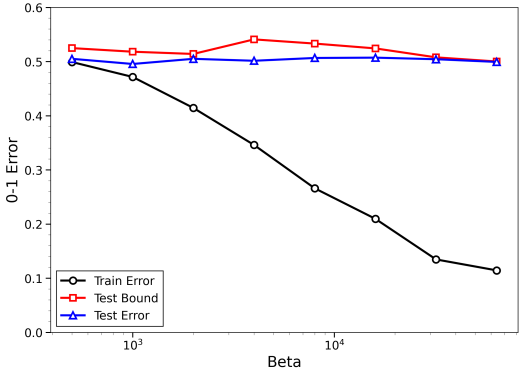
The Main Principle



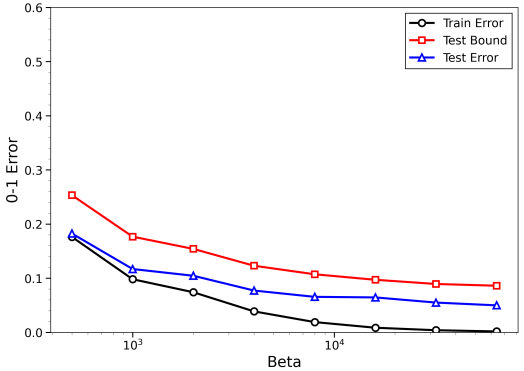
Good generalization at low temperature is announced by low training loss at high temperature

The Bound Tracks Test Error

The Bound Tracks Test Error



(a) Random labels



(b) True labels

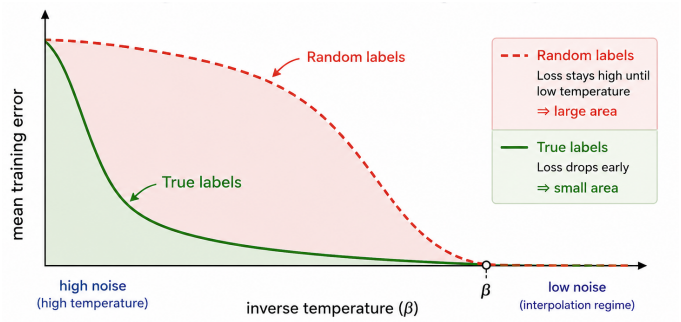
The method predicts the generalization gap from training data

Roadmap

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

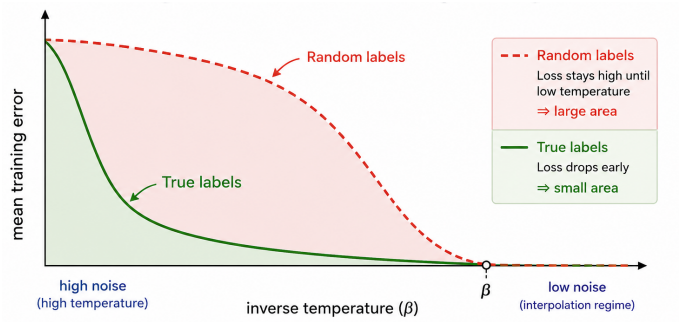
Takeaways

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



Takeaways

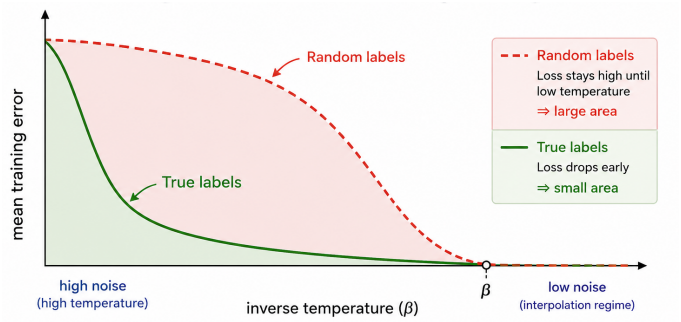
Good generalization at low temperature is announced by low training loss at high temperature



Gibbs complexity = area under the mean training-loss curve

Takeaways

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

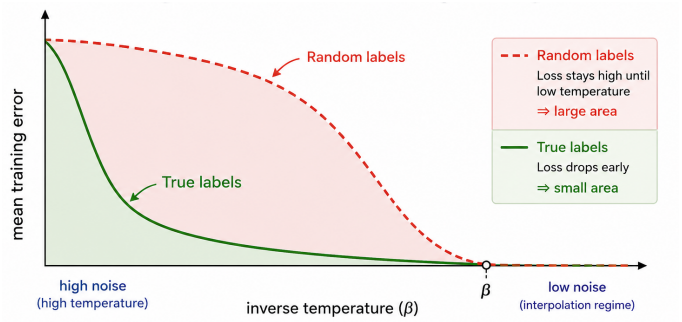


Gibbs complexity = area under the mean training-loss curve

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

Roadmap

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

Langevin Monte Carlo Algorithms

Time-homogeneous Markov processes converging to the Gibbs posterior (**G**) or to some nearby distributions (**N**). They are all some form of gradient descent with noise (β^{-1}) added.

- **CLD** (Continuous Langevin Dynamics, **G**), the classic, in continuous time
- **ULA** (Unadjusted Langevin Algorithm, **N**), discretization of CLD
- **SGLD** (Stochastic Gradient Langevin Dynamics, **N**), acceleration of ULA
- **MALA** (Metropolis Adjusted Langevin Algorithm, **G**) like ULA with Metropolis-type accept/reject step
- **HMCMC** (Hamiltonian Monte Carlo Markov Chain, **G**) a refinement of MALA

We expect our principle to apply to all these algorithms near their limiting distributions.

Langevin dynamics: noisy gradient descent

$$G_\beta(dh) \;\propto\; \exp\!\left(-\beta \hat{L}(h, \mathbf{x}) - \frac{\|h\|^2}{2\sigma^2}\right) dh$$

$$h_{t+1} = h_t \underbrace{- \eta\, \widehat{g}_t}_{\text{fit data}} \underbrace{- \frac{\eta}{\beta\sigma^2} h_t}_{\text{prior pull}} \underbrace{+ \sqrt{\frac{2\eta}{\beta}}\, \xi_t}_{\text{temperature noise}}$$

Loss g moves train

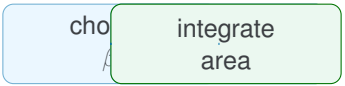
Noise explores the posterior

$$\widehat{g}_t = \begin{cases} \nabla_h \hat{L}(h_t, \mathbf{x}) & \text{ULA / full gradient,} \\ \text{unbiased minibatch estimate} & \text{SGLD.} \end{cases}$$

From SGLD samples to the temperature curve

For each temperature, estimate one number:

$$\tilde{m}_{\mathbf{x}}(\beta) = \frac{1}{M}\sum_{j=1}^{M} \hat{L}(h_j, \mathbf{x}), \qquad h_j \sim \nu_{\beta,\eta,t}$$



$$\text{Gibbs complexity} \approx \int_0^{\beta} \Big(\tilde{m}_{\mathbf{x}}(\gamma) - \tilde{m}_{\mathbf{x}}(\beta)\Big)\, d\gamma$$

Small area $\Rightarrow$ low complexity $\Rightarrow$ better generalization

Backup: why the approximation is controlled

$$\nu_{\beta,\eta,t} \approx G_\beta \qquad \text{in relative entropy}$$

$$\mathrm{KL}(\nu_{\beta,\eta,t} \,\|\, G_\beta) \le \underbrace{e^{-\frac{\alpha\eta t}{\beta}} D(\beta)}_{\text{finite-time error}} + \underbrace{\frac{8\eta d}{\beta\alpha}\left(\beta R + \frac{1}{\sigma^2}\right)^2}_{\text{finite-step bias}}$$

More steps

$t \uparrow$

reduces initialization error

Smaller step size

$\eta \downarrow$

reduces discretization bias

Exact-gradient ULA is controlled by this bound; SGLD is the minibatch version used for scalable experiments.

Markov Processes and the 2nd Law of Thermodynamics

We now study generalization along the trajectory of iterative algorithms.

Markov Processes and the 2nd Law of Thermodynamics

We now study generalization along the trajectory of iterative algorithms.

$\{h_t\}_{t\in I}$ = time homogeneous Markov process with values in $\mathcal{H}$ and $I = [0, \infty)$ or $I = \mathbb{N}$.
The distribution of h_t is denoted ν_t.

Markov Processes and the 2nd Law of Thermodynamics

We now study generalization along the trajectory of iterative algorithms.

$\{h_t\}_{t\in I}$ = time homogeneous Markov process with values in $\mathcal{H}$ and $I = [0, \infty)$ or $I = \mathbb{N}$. The distribution of h_t is denoted ν_t.

Lemma (Second Law of Thermodynamics)

If ν is a stationary distribution of $\{h_t\}_{t\in I}$ and $s < t$, then $KL(\nu_t, \nu) \leq KL(\nu_s, \nu)$ and $R_\infty(\nu_t, \nu) \leq R_\infty(\nu_s, \nu)$, with equality in either case if and only if $\nu_s = \nu$.

Markov Processes and the 2nd Law of Thermodynamics

We now study generalization along the trajectory of iterative algorithms.

$\{h_t\}_{t\in I}$ = time homogeneous Markov process with values in $\mathcal{H}$ and $I = [0, \infty)$ or $I = \mathbb{N}$. The distribution of h_t is denoted ν_t.

Lemma (Second Law of Thermodynamics)

If ν is a stationary distribution of $\{h_t\}_{t\in I}$ and $s < t$, then $KL(\nu_t, \nu) \leq KL(\nu_s, \nu)$ and $R_\infty(\nu_t, \nu) \leq R_\infty(\nu_s, \nu)$, with equality in either case if and only if $\nu_s = \nu$.

[Harel et al., 2025] use this, when $\nu_0 = \pi$ and G_β is invariant, to conclude $KL(\nu_t, \pi) \leq \beta\, \mathbb{E}_\pi\left[\hat{L}(h)\right]$ for all $t \in I$.

Markov Processes and the 2nd Law of Thermodynamics

We now study generalization along the trajectory of iterative algorithms.

$\{h_t\}_{t\in I}$ = time homogeneous Markov process with values in $\mathcal{H}$ and $I = [0, \infty)$ or $I = \mathbb{N}$. The distribution of h_t is denoted ν_t.

Lemma (Second Law of Thermodynamics)

If ν is a stationary distribution of $\{h_t\}_{t\in I}$ and $s < t$, then $KL(\nu_t, \nu) \leq KL(\nu_s, \nu)$ and $R_\infty(\nu_t, \nu) \leq R_\infty(\nu_s, \nu)$, with equality in either case if and only if $\nu_s = \nu$.

[Harel et al., 2025] use this, when $\nu_0 = \pi$ and G_β is invariant, to conclude $KL(\nu_t, \pi) \leq \beta\, \mathbb{E}_\pi\left[\hat{L}(h)\right]$ for all $t \in I$.

The bound remains valid along the entire training trajectory. However, it is largely independent of distribution and data, and vacuous for $\beta > n$.

Markov Processes; Initialized at the Prior

Theorem

Suppose $\nu_0 = \pi$.
(i) Let $\lambda_t = R_\infty(\nu_t, G_\beta) / R_\infty(\pi, G_\beta)$. *Then for all* $t \in I$

$$R_\infty(\nu_t, \pi) \leq \lambda_t \beta \|\hat{L}\|_\infty + (1 - \lambda_t) \int_0^\beta \mathbb{E}_{g \sim G_\gamma} \left[\hat{L}(g) \right] d\gamma.$$

(ii) If instead $\lambda_t = KL(\nu_t, G_\beta) / KL(\pi, G_\beta)$, *then*

$$KL(\nu_t, \pi) \leq \lambda_t \beta \, \mathbb{E}_{g \sim \pi} \left[\hat{L}(g) \right] + (1 - \lambda_t) \int_0^\beta \mathbb{E}_{g \sim G_\gamma} \left[\hat{L}(g) \right] d\gamma.$$

Markov Processes; Initialized at the Prior

Theorem

Suppose $\nu_0 = \pi$.
(i) Let $\lambda_t = R_\infty(\nu_t, G_\beta)/R_\infty(\pi, G_\beta)$. *Then for all* $t \in I$

$$R_\infty(\nu_t, \pi) \leq \lambda_t \beta \|\hat{L}\|_\infty + (1-\lambda_t)\int_0^\beta \mathbb{E}_{g\sim G_\gamma}\left[\hat{L}(g)\right] d\gamma.$$

(ii) If instead $\lambda_t = KL(\nu_t, G_\beta)/KL(\pi, G_\beta)$, *then*

$$KL(\nu_t, \pi) \leq \lambda_t \beta\, \mathbb{E}_{g\sim\pi}\left[\hat{L}(g)\right] + (1-\lambda_t)\int_0^\beta \mathbb{E}_{g\sim G_\gamma}\left[\hat{L}(g)\right] d\gamma.$$

Remark: If G_β is the invariant distribution, then by the 2nd law $\lambda_t \leq 1$,and λ_t is strictly decreasing, except when $\nu_t = G_\beta$, so $\lambda_t = 0$.

Markov Processes; Initialized at the Prior

Theorem

Suppose $\nu_0 = \pi$.
(i) Let $\lambda_t = R_\infty(\nu_t, G_\beta) / R_\infty(\pi, G_\beta)$. *Then for all* $t \in I$

$$R_\infty(\nu_t, \pi) \leq \lambda_t \beta \|\hat{L}\|_\infty + (1 - \lambda_t) \int_0^\beta \mathbb{E}_{g \sim G_\gamma} \left[\hat{L}(g) \right] d\gamma.$$

(ii) If instead $\lambda_t = KL(\nu_t, G_\beta) / KL(\pi, G_\beta)$, *then*

$$KL(\nu_t, \pi) \leq \lambda_t \beta \, \mathbb{E}_{g \sim \pi} \left[\hat{L}(g) \right] + (1 - \lambda_t) \int_0^\beta \mathbb{E}_{g \sim G_\gamma} \left[\hat{L}(g) \right] d\gamma.$$

Remark: If G_β is the invariant distribution, then by the 2nd law $\lambda_t \leq 1$,and λ_t is strictly decreasing, except when $\nu_t = G_\beta$, so $\lambda_t = 0$.
If the process converges to G_β in relative entropy, then $\lambda_t \to 0$, and the bound converges to the temperature integral (CLD, MALA, HMCMC).

Proof

$$
\begin{aligned}
KL\left(\nu_t, \pi\right) &= E_{\nu_t}\left[\ln \frac{d\nu_t}{dG_\beta}\right] + E_{\nu_t}\left[\frac{dG_\beta}{d\pi}\right] \\
&= KL\left(\nu_t, dG_\beta\right) - \beta\, E_{\nu_t}\left[\hat{L}\right] - \ln Z_\beta \\
&\leq \lambda_t\, KL\left(\pi, G_\beta\right) - \ln Z_\beta \\
&= \lambda_t \left(\beta\, E_\pi\left[\hat{L}\right] + \ln Z_\beta\right) - \ln Z_\beta \\
&= \lambda_t\, \beta\, E_\pi\left[\hat{L}\right] + (1-\lambda_t) \int_0^\beta \mathbb{E}_{g\sim G_\gamma}\left[\hat{L}\left(g\right)\right] d\gamma.
\end{aligned}
$$

This is (ii). The proof of (i) is similar.

A Bound for any Iterative Algorithm

For any $\beta \geq 0$ and any t

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

A Bound for any Iterative Algorithm

For any $\beta \geq 0$ and any t

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

For versions of ULA or SGLD, which are not time-homogeneous (decreasing stepsizes), $KL(\nu_t, G_\beta) \to 0$ as $t \to \infty$.

So we also converge to the integral bound, but we don't have a full trajectory guarantee.

References I

Harel, I., Wolanowsky, Y., Vardi, G., Srebro, N., and Soudry, D. (2025).
Temperature is all you need for generalization in Langevin dynamics and other markov processes.
arXiv preprint arXiv:2505.19087.

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.