PAC-Bayes Generalization Bound:
A Derivation Sketch

Generated by Alethfeld v5.1

January 5, 2026

Abstract

We derive a PAC-Bayes bound on the expected generalization error for hypotheses that achieve zero training error. The bound relates the expected population loss to the KL divergence between posterior and prior, which equals the negative log of the Bayesian evidence.

1 Setup

Definition 1 (Weight Space and Distributions)**.** Let $\mathcal{W}$ be the weight space of a learning algorithm. We define:

- P: prior distribution on $\mathcal{W}$
- $\mathcal{S} = \{w \in \mathcal{W} : \text{train\_error}(w) = 0\}$: the set of solutions
- Q: posterior distribution, defined as P restricted to $\mathcal{S}$
- $h(w)$: generalization error (population loss) of weight vector w
- $\beta = m - 1$ where m is the number of training points

2 Main Result

Theorem 2 (PAC-Bayes Bound for Zero Training Error)**.** *Under the setup above:*

$$\mathbb{E}_{w\sim Q}[h(w)] \leq \frac{\mathrm{KL}(Q\|P) + \ln \mathbb{E}_{w\sim P}[e^{\beta h(w)}]}{\beta}$$

Proof. We derive this bound by applying the general PAC-Bayes theorem with specific choices for the prior and posterior.

⟨1⟩**1. Assumption.** Let P be a prior distribution on the weight space $\mathcal{W}$. [assumption]

⟨1⟩**2. Assumption.** Let Q be the posterior distribution, defined as P restricted to the set of solutions $\mathcal{S} = \{w : \text{train\_error}(w) = 0\}$. [assumption]

⟨1⟩**3. Assumption.** Let $h(w)$ denote the generalization error (population loss) of weight vector w. [assumption]

⟨1⟩**4. Assumption.** Let $\beta = m - 1$ where m is the number of training points. [assumption]

⟨1⟩**5. PAC-Bayes Theorem** (McAllester, 1999): For any prior P, posterior Q, and $\beta > 0$:

$$\mathbb{E}_{w \sim Q}[h(w)] \leq \frac{\mathrm{KL}(Q \| P) + \ln \mathbb{E}_{w \sim P}[e^{\beta h(w)}]}{\beta}$$

[⋆ ADMITTED]

This is a standard result in PAC-Bayes theory. See McAllester (1999) or Catoni (2007) for proofs.

⟨1⟩**6.** Since Q is P restricted to solutions $\mathcal{S}$, we have:

$$Q(w) = \frac{P(w)}{P(\mathcal{S})} \quad \text{for } w \in \mathcal{S}, \qquad Q(w) = 0 \text{ otherwise}$$

[definition; from 2, 2]

⟨1⟩**7.** Computing the KL divergence:

$$\begin{aligned}\mathrm{KL}(Q \| P) &= \mathbb{E}_{w \sim Q}\left[\ln \frac{Q(w)}{P(w)}\right] \\ &= \mathbb{E}_{w \sim Q}\left[\ln \frac{P(w)/P(\mathcal{S})}{P(w)}\right] \\ &= \mathbb{E}_{w \sim Q}\left[\ln \frac{1}{P(\mathcal{S})}\right] \\ &= \ln \frac{1}{P(\mathcal{S})}\end{aligned}$$

[algebraic; from 2]

⟨1⟩**8.** $P(\mathcal{S})$ is the prior probability of all solutions, i.e., the **evidence**. Thus:

$$\mathrm{KL}(Q \| P) = \ln \frac{1}{\text{evidence}}$$

[definition; from 2]

⟨1⟩**9.** For $w \in \mathrm{supp}(Q) = \mathcal{S}$, the training error is 0, so $h(w)$ measures the population (generalization) error directly. [modus ponens; from 2, 2]

⟨1⟩**10.** Applying the PAC-Bayes theorem with our setup:

$$\mathbb{E}_{w \sim Q}[h(w)] \leq \frac{\ln(1/\text{evidence}) + \ln \mathbb{E}_{w \sim P}[e^{(m-1)h(w)}]}{m-1}$$

[substitution; from 2, 2, 2]

⟨1⟩**11. QED.** The expected generalization error over solutions is bounded by the sum of log-inverse-evidence and the log-moment-generating-function of the loss under the prior, normalized by $\beta = m - 1$. [from 2, 2]

□

3 Interpretation

The bound decomposes into two terms:

1. **Complexity term**: $\ln(1/\text{evidence}) = -\ln P(\mathcal{S})$

   This measures how "surprising" it is that a random weight vector from the prior achieves zero training error. Smaller solution sets (more complex hypotheses) yield larger complexity penalties.

2. **Prior MGF term**: $\ln \mathbb{E}_{w\sim P}[e^{\beta h(w)}]$

   This is the log-moment-generating-function of the generalization error under the prior. It captures how well the prior concentrates on low-error hypotheses.

4 Proof Metadata

Graph ID:	`graph-18d478-4253a9`
Version:	12
Proof Mode:	strict-mathematics
Total Steps:	11
Assumptions:	4
Claims:	6
Admitted Steps:	1 (PAC-Bayes Theorem)
Taint Status:	Tainted (depends on admitted step)

5 References

- McAllester, D. A. (1999). PAC-Bayesian model averaging. *Proceedings of the 12th Annual Conference on Computational Learning Theory.*
- Catoni, O. (2007). *PAC-Bayesian Supervised Classification.* Institute of Mathematical Statistics Lecture Notes.

Generated by Alethfeld Proof Orchestrator v5.1