UTN

HELMHOLTZ AI

Occam's Razor in Bayesian Inference: Evidence, Compression, and Generalization

Vincent Fortuin

Workshop on "Model Choice in Bayesian Inference", March 2026

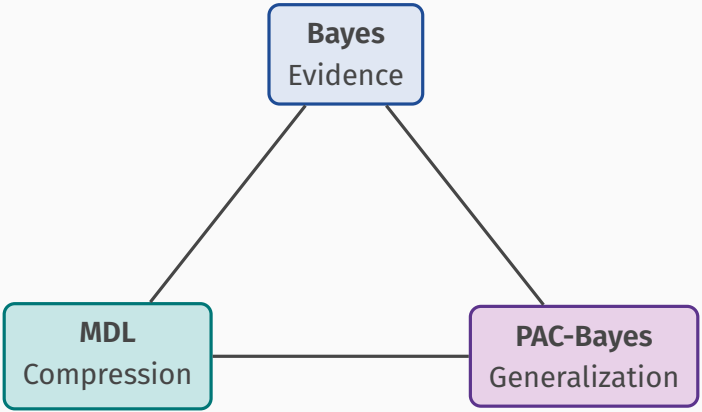
What does it mean for a model to be simple?

- We often say Bayesian inference “implements Occam’s razor”
- But how exactly is this simplicity/complexity measured?
- Simplicity of *what*: parameter count, code length, prior mass, generalization gap?

Claim: **There is no practical way to universally define simplicity; it is always defined relative to a prior belief or representational system.**

C. E. Rasmussen and Z. Ghahramani, “Occam’s Razor,” in *Advances in Neural Information Processing Systems 13*, 2001.

The Three Perspectives



Bayesian Model Selection

Evidence as an Occam score

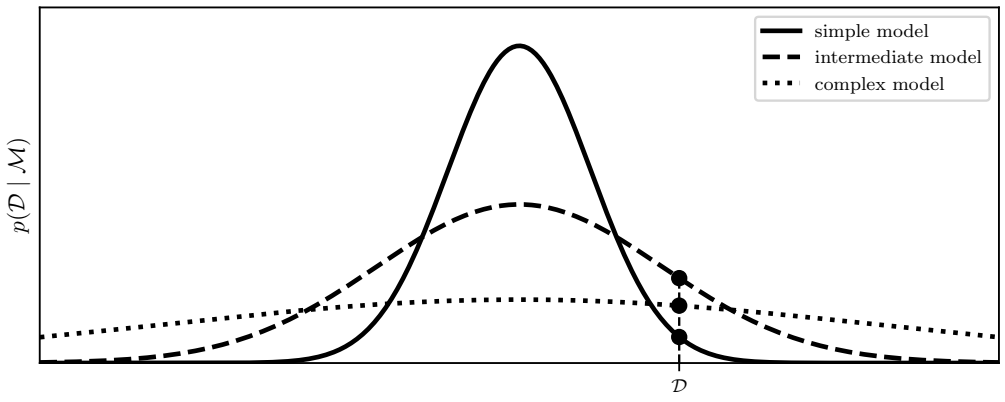
Bayesian Model Selection

$$p(M \mid D) \propto p(D \mid M)\, p(M)$$

$$p(D \mid M) = \int p(D \mid \theta, M)\, p(\theta \mid M)\, d\theta$$

- Model comparison is driven by the **marginal likelihood** (evidence)
- The key object is **average fit over the prior**, not best fit

Why does the Marginal Likelihood implement Occam's Razor?



D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, 2003, Ch. 28.

Laplace Approximation and the Occam Factor

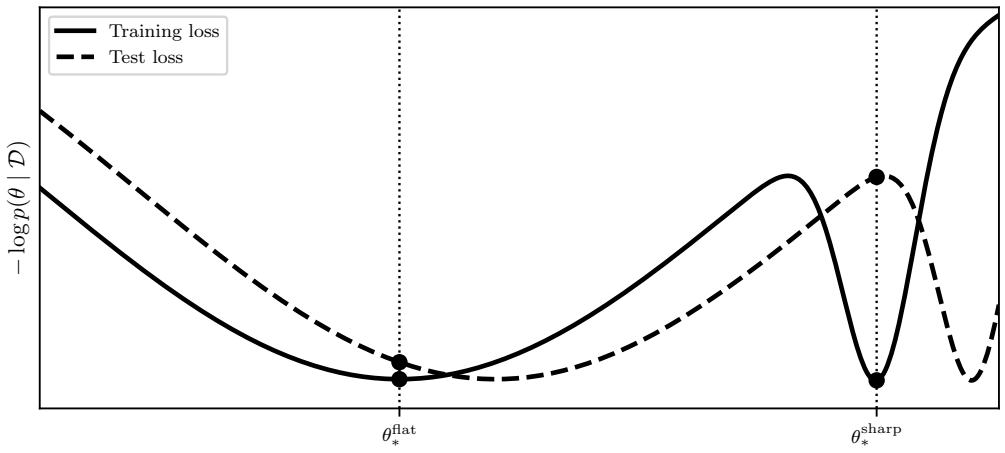
$$p(D \mid M) \approx p(D \mid \hat{\theta}, M)\, p(\hat{\theta} \mid M)\, (2\pi)^{k/2} \left| H(\hat{\theta}) \right|^{-1/2}$$

Peak likelihood term: how well the best local fit explains data

Volume term (Occam factor): posterior concentration relative to prior spread

Simplicity in the Bayesian sense is surviving prior volume after updating

Flat Minima and Marginal Likelihood



Minimum Description Length (MDL)

Model selection through compression

MDL: Model Selection as Coding

$$L(D, M) = L(M) + L(D \mid M)$$

- *Idea:* If Alice wants to send the data to Bob, she can send him her model first and then use it to compress the data
- Best model = shortest total code length for the message
- Complexity = bits needed to specify a data-explaining model
- Goodness of fit = bits needed to encode data under model

Simplicity becomes measurable through compression performance

P. D. Grünwald, *The Minimum Description Length Principle*, MIT Press, 2007.

Idealized code length: Kolmogorov and Solomonoff

- Algorithmic probability favors models generated by short programs
- Kolmogorov complexity $L(M) = K(M)$: shortest program that outputs M
- This is the purest Occam principle: shorter explanations get higher weight

But:

- depends on a universal machine (up to additive constants)
- is not computable in general

R. J. Solomonoff, "A Formal Theory of Inductive Inference. Part I," *Information and Control*, 7(1):1–22, 1964.

Traditional Two-Part MDL

1. Choose a discretized parameter or model index
2. Encode that choice
3. Encode data given that choice

Interpretation:

- often behaves like an implicit (uniform) prior over encoded parameters
- but only after fixing parametrization and coding precision

Even “objective” code lengths hide representational decisions

Bayesian MDL: Prior-Based Mixture Coding

$$L(D) = -\log p(D) \quad \text{for a mixture code with prior } p(\theta \mid M)$$

- Shared-code assumption: sender and receiver must fix the same prior/codebook before observing data
- Code length: with explicit priors, this is exactly the negative log marginal likelihood

Prequential Coding and Bayesian Prediction

$$L(D) = \sum_{t=1}^{n} -\log p(y_t \mid y_{<t}).$$

- Prequential code length: cumulative sequential predictive performance
- Bayesian updating: this telescopes to $-\log p(D)$
- This relates to the classic result that the marginal likelihood can be decomposed prequentially, and to modern predictive Bayes approaches, such as martingale posteriors

E. Fong and C. Holmes, "On the Marginal Likelihood and Cross-Validation," *Biometrika*, 107(2):489–496, 2020.

PAC-Bayes

Generalization bounds

Classic PAC-Bayes Bound

$$\mathcal{R}(Q) \lesssim \widehat{\mathcal{R}}(Q) + \sqrt{\frac{\mathrm{KL}(Q\|P) + \log(1/\delta)}{n}}$$

- *Assumption:* fixed reference distribution P and target distribution Q over predictors
- Generalization bound: we want predictors to perform well on unseen data
- Fit term: empirical risk controls in-sample performance
- Complexity term: information cost relative to reference via $\mathrm{KL}(Q\|P)$

Simple solutions (compared to a reference) are guaranteed to generalize

D. A. McAllester, "Some PAC-Bayesian Theorems," *Machine Learning*, 37:355–363, 1999.

Marginal Likelihood and PAC-Bayes

$$\mathcal{R}(Q^*) \leq \frac{s^2}{2(1-c)} - \frac{1}{n} \log p(D)$$

- *Assumptions:* optimal Bayes posterior Q^*, sub-Gamma loss with variance factor s^2 and scale $c < 1$
- In this case, optimizing the PAC-Bayes bound is **exactly equivalent** to optimizing the Bayesian marginal likelihood

P. Germain, F. Bach, A. Lacoste, and S. Lacoste-Julien, "PAC-Bayesian Theory Meets Bayesian Inference," in *Advances in Neural Information Processing Systems 29*, 2016.

Modern Compression Bounds and Deep Learning

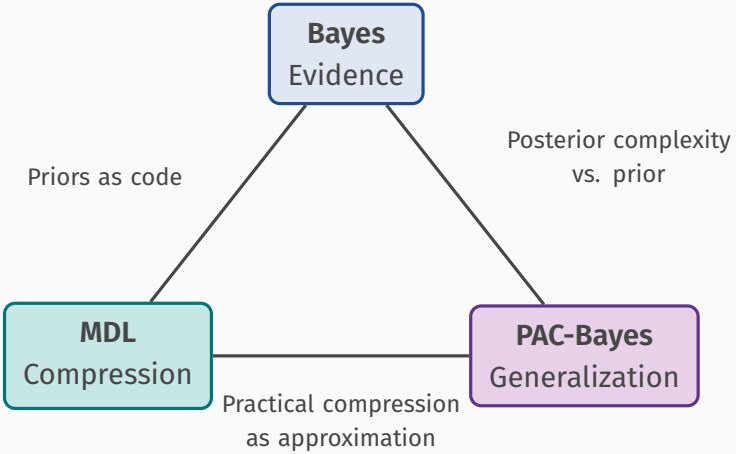
$$\mathcal{R}(M) \lesssim \widehat{\mathcal{R}}(M) + \sqrt{\frac{\log|M| + \log(1/\delta)}{2n}}$$

- Recent bounds explain generalization through compressibility of trained networks
- This is often closer in spirit to MDL than to Bayesian marginal likelihoods
- However: compression on disk still requires a choice of representational algorithm (e.g., Lempel-Ziv-Welch)

Common lesson: complexity is always representation-relative

S. Lotfi, M. Finzi, S. Kapoor, A. Potapczynski, M. Goldblum, and A. G. Wilson, "PAC-Bayes Compression Bounds So Tight That They Can Explain Generalization," in *Advances in Neural Information Processing Systems 35*, 2022.

Recap: The Three Perspectives



Caveats: Computability and Approximation

- Exact marginal likelihood is usually intractable
- Practical Bayesian inference relies on Laplace, variational methods, Markov Chain Monte Carlo, and other approximations
- Kolmogorov complexity is uncomputable
- Practical MDL depends on chosen codes and surrogates
- PAC-Bayes bounds may be loose or difficult to optimize tightly

Caveats: Simplicity depends on the Model Class

Often, we might want to compare models from different classes:

- Neural nets vs. Gaussian processes
- Symbolic programs vs. parametric probabilistic models
- Different architectures of neural networks

However, there is no common metric to compare them!

- Prior distributions are only meaningful within their parametric support
- Compression codes depend on representation of the model
- KL-divergences rely on shared support in parameters or predictive

There is no model-class-independent complexity metric: Bayes needs a prior, MDL needs a code, PAC-Bayes needs a divergence

Final Thoughts

- Occam's razor survives in modern inference only in operational forms
- Bayes: simplicity = high predictive concentration under a prior
- MDL: simplicity = short code length under a representation
- PAC-Bayes: simplicity = small divergence relative to a reference
- These are deeply related, but none is free of priors or representations
- Comparing different model classes requires philosophical as well as mathematical choices

Simplicity is not universal; it is measured relative to assumptions.