Solomonoff Induction

David Quarel
Australian National University

Leon Lang
University of Amsterdam

August 22, 2026

What you'll learn

- Define the Bayesian mixture over a class of environments and prove it is a proper predictor whose posterior updates multiplicatively.
- Prove the cumulative prediction-error bound and specialize it to the Solomonoff prior to recover the Occam bound.
- Show the mixture makes boundedly many mistakes and is Pareto-optimal under KL and squared loss.
- Bound the misspecified case when the truth lies outside the model class.

Much of this material is drawn from [HQC24, Chapter 3] and the earlier [Hut05]; both books provide fuller explanations, additional context, and proofs that we gloss over here.

Difficulty ratings. Each problem is tagged with a rating in square brackets following Knuth's exercise rating scheme [Knu73]: roughly, **[00]** trivial, **[10]** 15–minute pencil-and-paper, **[20]** 1–2 hours, **[30]** several hours to a day, **[40]** a significant research result. Intermediate values are possible.

Overview

Solomonoff induction is the *prediction* half of **Universal AI**: what would an *optimally intelligent* agent do if it only had to predict, given unlimited compute and the weakest possible assumptions? The answer is a single **Bayesian mixture** ξ over a countable class $\mathcal{M}$ of candidate environments, weighted by a prior w_ν. It (eventually) predicts as well as the true environment μ, and under the **universal** choice — $\mathcal{M}$ = all computable

environments, prior $w_\nu = 2^{-K(\nu)}$ with K the Kolmogorov complexity — the assumption "$\mu \in \mathcal{M}$" becomes "the universe is computable," and Occam's razor drops out of the mathematics.

This module is a worksheet: you build the theory yourself, one problem at a time. The sequel module, AIXI, lifts the same mixture to sequential decision-making (learning to *act*).

Prerequisites

- Comfort with discrete probability: conditional distributions, the chain rule, expectations.
- Kullback–Leibler divergence and basic information theory (helpful, developed as needed).

Goal and Roadmap

A recipe for prediction: Solomonoff induction is an attempt to mathematically formalize the *problem of induction* from philosophy: How to make predictions about the future based on past observations?

We formalize this as sequence prediction: There is some true environment μ that generates a sequence $x_1, x_2, \dots$ of (binary) symbols. Each next symbol $x_t \sim \mu(\cdot \mid x_{<t})$ is sampled from μ conditioned on the past $x_{<t}$. A predictor P takes the history $x_{<t}$ and gives a distribution $P(\cdot \mid x_{<t})$ over the next symbol x_t.

We measure the quality of a predictor P by the μ-expected squared prediction error, summed over every timestep:

$$S_\infty^\mu := \sum_{t=1}^{\infty} \sum_{x_{<t} \in \mathbb{B}^*} \mu(x_{<t}) \sum_{x_t \in \mathbb{B}} \big(P(x_t \mid x_{<t}) - \mu(x_t \mid x_{<t})\big)^2. \tag{1}$$

How to construct a predictor P such that S_∞^μ is small (or at least finite)? Bayesian inference to the rescue: we choose as our predictor a *Bayesian mixture* ξ over a countable class $\mathcal{M} = \{\nu_1, \nu_2, \dots\}$ of candidate environments, weighted by prior beliefs w_ν (formalized in Definition 1.2). The main results we build up to are:

- **Cumulative bound** (Section 4): Assuming $\mu \in \mathcal{M}$, $S_\infty^\mu \leq -\ln w_\mu$. The higher the prior on μ, the lower the prediction error.

- **Explicit bound** (Exercise 4.2): specialize to the Solomonoff prior $w_\nu = 2^{-K(\nu)}$ to get $S_\infty^\mu \leq K(\mu) \ln 2$, where K is the *Kolmogorov complexity*.
- **Pareto optimality** (Exercises 6.2 and 8.2): no other predictor weakly dominates ξ on every $\nu \in \mathcal{M}$, for either KL or squared loss.
- **Misspecified version** (Section 7): If $\mu \notin \mathcal{M}$, the cumulative bound becomes $-\ln w_{\hat{\mu}} + D_n(\mu \parallel \hat{\mu})$: the constant complexity term plus an approximation term $D_n(\mu \parallel \hat{\mu}) = \sum_{t=1}^n d_t(\mu \parallel \hat{\mu})$ that in general grows linearly in n, so S_∞^μ diverges. Here $\hat{\mu} \in \mathcal{M}$ is the "closest" environment to μ.

Notation

For more background, see the corresponding post on Solomonoff induction.

Note

Symbols used throughout.

- $\mathbb{B} = \{0, 1\}$: the binary alphabet; $\mathbb{B}^*$ all finite binary strings; $\mathbb{B}^n$ length-n strings.
- ϵ: the empty string. $x_{1:n} := x_1 x_2 \cdots x_n$; $x_{<t} := x_1 x_2 \cdots x_{t-1}$.
- $X_{1:n}, X_{<t}$: the corresponding random variables.
- ν, ρ: generic environments / predictors. μ: the true (unknown) environment generating the data. ξ: a Bayesian mixture of environments.
- $\mathcal{M} = \{\nu_1, \nu_2, \ldots\}$: a countable class of candidate environments.
- $w_\nu > 0$ with $\sum_{\nu \in \mathcal{M}} w_\nu = 1$: the prior over $\mathcal{M}$. $w(\nu \mid x_{<t})$: the posterior after observing $x_{<t}$.
- $\Delta\mathcal{X}$: the set of probability distributions over a finite set $\mathcal{X}$.

1. The mixture is a predictor

Definition 1.1 (Environment)**.** An **environment** ν assigns to each history $x_{<t} \in \mathbb{B}^*$ a predictive distribution $\nu(\cdot \mid x_{<t}) \in \Delta\mathbb{B}$ over the next symbol. The joint probability of a string is defined by the chain rule

$$\nu(x_{1:n}) \;:=\; \prod_{t=1}^{n} \nu(x_t \mid x_{<t}), \qquad \nu(\epsilon) := 1,$$

so that $\nu(x_{1:t}) = \nu(x_{<t}) \cdot \nu(x_t \mid x_{<t})$ and $\nu(x) = \nu(x0) + \nu(x1)$ for all $x \in \mathbb{B}^*$. The true (unknown) environment generating the data is denoted μ.

Definition 1.2 (Bayesian mixture ξ)**.** Let $\mathcal{M} = \{\nu_1, \nu_2, \dots\}$ be a countable class of environments with prior weights $w_\nu > 0$ satisfying $\sum_{\nu \in \mathcal{M}} w_\nu = 1$. The **Bayesian mixture** is defined as a prior-weighted mixture over all environments ν in $\mathcal{M}$:

$$\xi(x_{1:t}) \;:=\; \sum_{\nu \in \mathcal{M}} w_\nu \, \nu(x_{1:t}). \tag{2}$$

Its one-step predictive distribution is

$$\xi(x_t \mid x_{<t}) \;=\; \sum_{\nu \in \mathcal{M}} w(\nu \mid x_{<t}) \, \nu(x_t \mid x_{<t}),$$

with posterior weights

$$w(\nu \mid x_{<t}) \;:=\; w_\nu \, \frac{\nu(x_{<t})}{\xi(x_{<t})} \qquad w(\nu \mid \epsilon) := w_\nu.$$

Here $w(\nu \mid x_{<t})$ is the *posterior* belief in ν after observing $x_{<t}$.

The mixture ξ is defined as a sum of *joint* probabilities. To use it as a predictor we need its one-step conditional $\xi(x_t \mid x_{<t})$, and we need to know it is a genuine probability distribution (so that, later, Pinsker's inequality applies to it).

Exercise 1.1 (Generalized chain rule). [05]
The two-term chain rule baked into Definition 1.1 extends to arbitrary contiguous blocks. Show that for any $1 \le p \le q \le r$ and any history $x_{<p}$ with $\nu(x_{<p}) > 0$,

$$\nu(x_{p:r} \mid x_{<p}) \;=\; \nu(x_{p:q} \mid x_{<p}) \cdot \nu(x_{q+1:r} \mid x_{<p}\, x_{p:q}).$$

Solution to Exercise 1.1

Expand the conditional as a ratio of joints; the chain rule (Definition 1.1) telescopes the ratio to a product over indices $k = p, \dots, r$:

$$\nu(x_{p:r} \mid x_{<p}) \;=\; \frac{\nu(x_{1:r})}{\nu(x_{<p})} \;=\; \frac{\prod_{k=1}^{r} \nu(x_k \mid x_{<k})}{\prod_{k=1}^{p-1} \nu(x_k \mid x_{<k})} \;=\; \prod_{k=p}^{r} \nu(x_k \mid x_{<k}).$$

For any $k \ge p$, the past $x_{<k}$ is just $x_{<p}$ extended by $x_{p:k-1}$, so the same expansion identifies every contiguous sub-block as a conditional with the right history.

Splitting the product at index q:

$$\nu(x_{p:r} \mid x_{<p}) \;=\; \underbrace{\prod_{k=p}^{q} \nu(x_k \mid x_{<k})}_{=\,\nu(x_{p:q} \mid x_{<p})} \cdot \underbrace{\prod_{k=q+1}^{r} \nu(x_k \mid x_{<k})}_{=\,\nu(x_{q+1:r} \mid x_{<p}\, x_{p:q})}.$$

Exercise 1.2 (Mixture is a predictor). [10]
Starting from $\xi(x_{1:t}) = \sum_{\nu \in \mathcal{M}} w_\nu\, \nu(x_{1:t})$, show that

$$\xi(x_t \mid x_{<t}) \;=\; \sum_{\nu \in \mathcal{M}} w(\nu \mid x_{<t})\, \nu(x_t \mid x_{<t}), \qquad w(\nu \mid x_{<t}) := w_\nu\, \frac{\nu(x_{<t})}{\xi(x_{<t})},$$

and conclude that $\sum_{x_t \in \mathbb{B}} \xi(x_t \mid x_{<t}) = 1$, i.e. $\xi(\cdot \mid x_{<t})$ is a probability distribution over $\mathbb{B}$.

Hint: Write $\xi(x_t \mid x_{<t}) = \xi(x_{1:t})/\xi(x_{<t})$, expand the numerator, and use the chain rule $\nu(x_{1:t}) = \nu(x_{<t})\, \nu(x_t \mid x_{<t})$ (Definition 1.1). For the last part, note that the posterior weights sum to 1.

Solution to Exercise 1.2

The one-step conditional is the ratio of joint to marginal, $\xi(x_t \mid x_{<t}) := \xi(x_{1:t})/\xi(x_{<t})$. Expanding the numerator and applying the chain rule $\nu(x_{1:t}) = \nu(x_{<t})\, \nu(x_t \mid x_{<t})$ to each term:

$$\begin{aligned} \xi(x_t \mid x_{<t}) \;&=\; \frac{\sum_{\nu \in \mathcal{M}} w_\nu\, \nu(x_{1:t})}{\xi(x_{<t})} \;=\; \frac{\sum_{\nu \in \mathcal{M}} w_\nu\, \nu(x_{<t})\, \nu(x_t \mid x_{<t})}{\xi(x_{<t})} \\ &=\; \sum_{\nu \in \mathcal{M}} \underbrace{\frac{w_\nu\, \nu(x_{<t})}{\xi(x_{<t})}}_{=\,w(\nu \mid x_{<t})}\, \nu(x_t \mid x_{<t}) \;=\; \sum_{\nu \in \mathcal{M}} w(\nu \mid x_{<t})\, \nu(x_t \mid x_{<t}). \end{aligned}$$

The posterior weights are non-negative and sum to 1:

$$\sum_{\nu \in \mathcal{M}} w(\nu \mid x_{<t}) \;=\; \frac{\sum_{\nu \in \mathcal{M}} w_\nu\, \nu(x_{<t})}{\xi(x_{<t})} \;=\; \frac{\xi(x_{<t})}{\xi(x_{<t})} \;=\; 1.$$

Therefore $\xi(\cdot \mid x_{<t})$ is a convex combination of the probability distributions $\nu(\cdot \mid x_{<t})$, so it is itself a probability distribution over $\mathbb{B}$:

$$\sum_{x_t \in \mathbb{B}} \xi(x_t \mid x_{<t}) \;=\; \sum_{\nu \in \mathcal{M}} w(\nu \mid x_{<t}) \underbrace{\sum_{x_t \in \mathbb{B}} \nu(x_t \mid x_{<t})}_{=\,1} \;=\; \sum_{\nu \in \mathcal{M}} w(\nu \mid x_{<t}) \;=\; 1.$$

Exercise 1.3 (Posterior update). [10]
Show that the posterior weight $w(\cdot \mid \cdot)$ updates *multiplicatively* as new symbols arrive. For every $x_{1:t} \in \mathbb{B}^*$ with $\xi(x_{1:t}) > 0$,

$$w(\nu \mid x_{1:t}) \;=\; w(\nu \mid x_{<t}) \cdot \frac{\nu(x_t \mid x_{<t})}{\xi(x_t \mid x_{<t})}.$$

That is: the new posterior equals the old posterior, scaled by the **likelihood ratio** of how well ν predicted the just-observed symbol relative to the mixture's own prediction.

Solution to Exercise 1.3

By Definition 1.2,

$$w(\nu \mid x_{1:t}) \;=\; w_\nu \frac{\nu(x_{1:t})}{\xi(x_{1:t})}, \qquad w(\nu \mid x_{<t}) \;=\; w_\nu \frac{\nu(x_{<t})}{\xi(x_{<t})}.$$

Apply the chain rule (Definition 1.1) to both ν and ξ: $\nu(x_{1:t}) = \nu(x_{<t})\,\nu(x_t \mid x_{<t})$ and $\xi(x_{1:t}) = \xi(x_{<t})\,\xi(x_t \mid x_{<t})$. Dividing,

$$\frac{w(\nu \mid x_{1:t})}{w(\nu \mid x_{<t})} \;=\; \frac{\nu(x_{1:t})}{\nu(x_{<t})} \cdot \frac{\xi(x_{<t})}{\xi(x_{1:t})} \;=\; \frac{\nu(x_t \mid x_{<t})}{\xi(x_t \mid x_{<t})}.$$

Rearranging gives the claim.

Exercise 1.4 (Multi-step posterior linearity). [10]
Exercise 1.2 showed that ξ's one-step prediction $\xi(x_t \mid x_{<t})$ is the posterior-weighted average of the per-model one-step predictions. The same identity extends to predictions over a *whole future segment* $x_{t:m}$: for any $t \le m$,

$$\xi(x_{t:m} \mid x_{<t}) \;=\; \sum_{\nu \in \mathcal{M}} w(\nu \mid x_{<t})\,\nu(x_{t:m} \mid x_{<t}).$$

That is: the *same* posterior $w(\nu \mid x_{<t})$ governs ξ's predictions for arbitrarily many steps into the future, not just the next one. Show this.

Hint: Write $\xi(x_{t:m} \mid x_{<t}) = \xi(x_{1:m})/\xi(x_{<t})$, expand the numerator using the mixture definition, and apply the generalized chain rule (Exercise 1.1 with $p = 1$, $q = t-1$, $r = m$) to factor each $\nu(x_{1:m}) = \nu(x_{<t})\,\nu(x_{t:m} \mid x_{<t})$.

Solution to Exercise 1.4

By definition of the conditional and the mixture form $\xi(x_{1:m}) = \sum_\nu w_\nu \nu(x_{1:m})$ (Definition 1.2),

$$\xi(x_{t:m} \mid x_{<t}) \;=\; \frac{\xi(x_{1:m})}{\xi(x_{<t})} \;=\; \frac{\sum_\nu w_\nu\, \nu(x_{1:m})}{\xi(x_{<t})}.$$

The generalized chain rule (Exercise 1.1 with $p = 1$, $q = t-1$, $r = m$) gives $\nu(x_{1:m}) = \nu(x_{<t})\,\nu(x_{t:m} \mid x_{<t})$. Substituting,

$$\begin{aligned}\xi(x_{t:m} \mid x_{<t}) \;&=\; \sum_\nu \underbrace{\frac{w_\nu\, \nu(x_{<t})}{\xi(x_{<t})}}_{= w(\nu \mid x_{<t})}\, \nu(x_{t:m} \mid x_{<t}) \\ &=\; \sum_{\nu \in \mathcal{M}} w(\nu \mid x_{<t})\, \nu(x_{t:m} \mid x_{<t}).\end{aligned}$$

2. KL divergence

The KL divergence between two joint distributions over a length-n history splits, by an inductive application of the chain rule for probabilities, into a sum of per-step conditional KLs. This telescoping identity lets us trade a single joint KL over an entire history for a sum of per-step KLs (and vice versa): exactly the bridge we will need in Section 4.

We introduce two pieces of notation that will be used throughout the rest of the worksheet:

Definition 2.1 (Cumulative and per-step KL)**.** For environments / predictors ν, ρ, horizon $n \geq 1$, and $1 \leq t \leq n$, the **cumulative joint KL** and the **expected per-step KL at step** t are

$$D_n(\nu \,\|\, \rho) \;:=\; \sum_{x_{1:n} \in \mathbb{B}^n} \nu(x_{1:n}) \ln \frac{\nu(x_{1:n})}{\rho(x_{1:n})},$$

$$d_t(\nu \,\|\, \rho) \;:=\; \sum_{x_{<t} \in \mathbb{B}^{t-1}} \nu(x_{<t}) \sum_{x_t \in \mathbb{B}} \nu(x_t \mid x_{<t}) \ln \frac{\nu(x_t \mid x_{<t})}{\rho(x_t \mid x_{<t})},$$

with the conventions $0 \ln \frac{0}{q} := 0$ for $q \geq 0$ and $p \ln \frac{p}{0} := \infty$ for $p > 0$. We write $D_\infty(\nu \,\|\, \rho) := \lim_{t \to \infty} D_t(\nu \,\|\, \rho)$ and $D_\infty^\mu := D_\infty(\mu \,\|\, \xi)$.

Theorem 2.2 (KL divergence is non-negative)**.** *For any environments ν, ρ and any horizon $n \geq 1$ and step $1 \leq t \leq n$:*

(a) $D_n(\nu \parallel \rho) \geq 0$, *with equality iff* $\nu(x_{1:n}) = \rho(x_{1:n})$ *for every* $x_{1:n} \in \mathbb{B}^n$ *with* $\nu(x_{1:n}) > 0$.

(b) $d_t(\nu \parallel \rho) \geq 0$, *with equality iff* $\nu(x_t \mid x_{<t}) = \rho(x_t \mid x_{<t})$ *for every* $x_{<t}$ *with* $\nu(x_{<t}) > 0$ *and every* $x_t \in \mathbb{B}$.

Proof in Appendix B.

Exercise 2.1 (Telescoping KL). [10]
Show by induction on n that for all environments ν, ρ,

$$D_n(\nu \parallel \rho) \;=\; \sum_{t=1}^{n} d_t(\nu \parallel \rho),$$

where on the right-hand side, the $t = 1$ summand uses the conventions $\nu(x_1 \mid x_{<1}) := \nu(x_1)$, $\rho(x_1 \mid x_{<1}) := \rho(x_1)$.

Hint: For the induction step, factor $\nu(x_{1:n}) = \nu(x_{<n})\,\nu(x_n \mid x_{<n})$ (and likewise for ρ) inside the log; the log splits additively into two pieces matching $D_{n-1}(\nu \parallel \rho)$ and $d_n(\nu \parallel \rho)$.

Solution to Exercise 2.1

By induction on n.
Base case ($n = 1$):

$$D_1(\nu \parallel \rho) \;=\; \sum_{x_1} \nu(x_1) \ln \frac{\nu(x_1)}{\rho(x_1)} \;=\; d_1(\nu \parallel \rho),$$

using $\nu(x_1 \mid x_{<1}) := \nu(x_1)$ and likewise for ρ.
Induction step. Apply the chain rule for probabilities (Definition 1.1) inside the

log: $\nu(x_{1:n}) = \nu(x_{<n})\,\nu(x_n \mid x_{<n})$, and similarly for ρ. The log splits additively:

$$
\begin{aligned}
& D_n(\nu \,\|\, \rho) \\
&= \sum_{x_{1:n} \in \mathbb{B}^n} \nu(x_{1:n}) \,\ln \frac{\nu(x_{1:n})}{\rho(x_{1:n})} \\
&= \sum_{x_{1:n}} \nu(x_n \mid x_{<n}) \nu(x_{<n}) \,\ln \frac{\nu(x_n \mid x_{<n})\,\nu(x_{<n})}{\rho(x_n \mid x_{<n})\,\rho(x_{<n})} \\
&= \sum_{x_{1:n}} \nu(x_n \mid x_{<n}) \nu(x_{<n}) \left[\ln \frac{\nu(x_n \mid x_{<n})}{\rho(x_n \mid x_{<n})} \;+\; \ln \frac{\nu(x_{<n})}{\rho(x_{<n})} \right]. \\
&= \underbrace{\sum_{x_{<n}} \nu(x_{<n}) \sum_{x_n} \nu(x_n \mid x_{<n}) \ln \frac{\nu(x_n \mid x_{<n})}{\rho(x_n \mid x_{<n})}}_{d_n(\nu \| \rho)} \\
&\qquad + \sum_{x_{1:n}} \nu(x_n \mid x_{<n}) \nu(x_{<n}) \ln \frac{\nu(x_{<n})}{\rho(x_{<n})} \\
&= d_n(\nu \,\|\, \rho) \;+\; \underbrace{\sum_{x_{<n}} \nu(x_{<n}) \ln \frac{\nu(x_{<n})}{\rho(x_{<n})}}_{D_{n-1}(\nu \| \rho)} \cancel{\sum_{x_n} \nu(x_n \mid x_{<n})} \\
&= d_n(\nu \,\|\, \rho) + D_{n-1}(\nu \,\|\, \rho) \;=\; d_n(\nu \,\|\, \rho) + \sum_{t=1}^{n-1} d_t(\nu \,\|\, \rho) = \sum_{t=1}^{n} d_t(\nu \,\|\, \rho)
\end{aligned}
$$

3. Mixture dominance

The mixture ξ never assigns much less probability than any single environment weighted by its prior. This single inequality is the engine behind the cumulative bound of Section 4 below: dividing through gives $\mu(x)/\xi(x) \le 1/w_\mu$, which is then fed into Pinsker and the chain rule in the proof of the main cumulative bound.

Exercise 3.1 (Mixture dominance). [05]
Show that for every $\nu \in \mathcal{M}$ and every $x \in \mathbb{B}^*$,

$$\xi(x) \;\ge\; w_\nu \cdot \nu(x).$$

Solution to Exercise 3.1

By Definition 1.2, $\xi(x) = \sum_{\nu' \in \mathcal{M}} w_{\nu'} \nu'(x)$. Every term in this sum is non-negative, so dropping all terms except the one with $\nu' = \nu$ can only decrease the sum:

$$\xi(x) \;=\; \sum_{\nu' \in \mathcal{M}} w_{\nu'} \nu'(x) \;\geq\; w_\nu \nu(x).$$

Remark (Solomonoff specialization)

Under the Solomonoff prior, $w_\nu = 2^{-K(\nu)}$, so the bound becomes

$$\xi_U(x) \;\geq\; 2^{-K(\nu)} \cdot \nu(x) \qquad \text{for every computable } \nu.$$

This is what makes ξ_U *universal*: a single predictor dominates the entire computable model class up to a factor that depends only on the model's description length.

4. Cumulative prediction error bound (main result)

The next ingredient relates squared prediction error to KL divergence [CT06, Lemma 11.6.1]. We require the following inequality:

Theorem 4.1 (Pinsker's inequality)**.** *Let P and Q be probability distributions over $\mathbb{B} = \{0, 1\}$, i.e., $q_0, q_1 \geq 0$ with $q_0 + q_1 = 1$. Write $p_x = P(x)$ and $q_x = Q(x)$. Then*

$$\sum_{x \in \mathbb{B}} \big(q_x - p_x\big)^2 \;\leq\; \sum_{x \in \mathbb{B}} p_x \ln \frac{p_x}{q_x}. \tag{$\star$}$$

The proof of this inequality is mostly tedious algebra but is included in Appendix A.

Definition 4.2 (Cumulative expected and per step squared prediction error)**.** For environments / predictors ν, ρ and horizon $n \geq 1$, the **cumulative expected squared error** and the **per-step squared error at step** t are

$$\begin{aligned} S_n(\nu \,\|\, \rho) \;&:=\; \sum_{t=1}^{n} s_t(\nu \,\|\, \rho) \\ s_t(\nu \,\|\, \rho) \;&:=\; \sum_{x_{<t} \in \mathbb{B}^{t-1}} \nu(x_{<t}) \sum_{x_t \in \mathbb{B}} \big(\nu(x_t \mid x_{<t}) - \rho(x_t \mid x_{<t})\big)^2 \end{aligned}$$

Write $S_\infty(\nu \,\|\, \rho) := \lim_{n \to \infty} S_n(\nu \,\|\, \rho)$ and $S_\infty^\mu := S_\infty(\mu \,\|\, \xi)$.

Exercise 4.1 (Cumulative prediction bound). [15]
Show that

$$S_\infty^\mu \leq -\ln w_\mu.$$

Hint: You need Theorem 4.1, Exercise 2.1 and Exercise 3.1.

Solution to Exercise 4.1

The proof is a three-step chain:

$$S_\infty^\mu \overset{(1)}{\leq} D_\infty^\mu \overset{(2)}{\leq} \lim_{n\to\infty} \sum_{x_{1:n}} \mu(x_{1:n}) \ln \tfrac{1}{w_\mu} \overset{(3)}{=} -\ln w_\mu.$$

Step (1): Pinsker, pointwise. For each t and $x_{<t}$, both $\mu(\cdot \mid x_{<t})$ and $\xi(\cdot \mid x_{<t})$ are probability distributions on $\mathbb{B}$ (ξ by Section 1), so Theorem 4.1 gives

$$\sum_{x_t} \big(\xi(x_t \mid x_{<t}) - \mu(x_t \mid x_{<t})\big)^2 \leq \sum_{x_t} \mu(x_t \mid x_{<t}) \ln \tfrac{\mu(x_t|x_{<t})}{\xi(x_t|x_{<t})}.$$

Multiplying by $\mu(x_{<t}) \geq 0$ and summing over $x_{<t}$ gives $s_t(\mu \parallel \xi) \leq d_t(\mu \parallel \xi)$, so $S_t(\mu \parallel \xi) \leq \sum_{i=1}^t d_i(\mu \parallel \xi) = D_t(\mu \parallel \xi)$ (Exercise 2.1), from which we take $t \to \infty$.

Step (2): Mixture dominance. By Exercise 3.1, $\xi(x_{1:n}) \geq w_\mu \mu(x_{1:n})$, so $\mu(x_{1:n})/\xi(x_{1:n}) \leq 1/w_\mu$, hence $\ln(\mu/\xi) \leq \ln(1/w_\mu)$ pointwise. Substituting into the definition $D_n(\mu \parallel \xi) = \sum_{x_{1:n}} \mu(x_{1:n}) \ln(\mu/\xi)$ gives the bound.

Step (3): Total mass. Pull the constant out and use $\sum_{x_{1:n}} \mu(x_{1:n}) = 1$ (μ is an environment).

Exercise 4.2 (Explicit complexity bound). [05]
Specialize Section 4 to the **Solomonoff prior** $w_\nu = 2^{-K(\nu)}$ (where $K(\nu)$ is the length of the shortest program computing ν) to show

$$S_\infty^\mu \leq K(\mu) \ln 2.$$

Solution to Exercise 4.2

By Section 4, for *any* prior, $S_\infty^\mu \leq -\ln w_\mu$. The Solomonoff prior assigns μ the weight $w_\mu = 2^{-K(\mu)}$, so

$$S_\infty^\mu \;\leq\; -\ln w_\mu \;=\; -\ln 2^{-K(\mu)} \;=\; K(\mu) \ln 2.$$

Exercise 4.3 (Per-step error). [05]
Hence prove that the per-step expected prediction error

$$s_t := \sum_{x_{<t} \in \mathbb{B}^{t-1}} \mu(x_{<t}) \sum_{x_t \in \mathbb{B}} \big(\xi(x_t \mid x_{<t}) - \mu(x_t \mid x_{<t})\big)^2$$

converges to zero as $t \to \infty$.

Solution to Exercise 4.3

Follows immediately as $\sum_{t=1}^{\infty} s_t \leq -\ln w_\mu < \infty$ implies $s_t \to 0$.

5. Bounded number of prediction mistakes

So far we have bounded *squared error*. For a deterministic environment (i.e. a single infinite binary sequence $x_1^*, x_2^*, \ldots$ in $\mathcal{M}$), one can ask the sharper 0/1-loss question: how often does the mixture predict the *wrong bit*? The cumulative bound is strong enough to give an answer.

Let $\mu \in \mathcal{M}$ be a deterministic environment, so $\mu(x_t \mid x_{<t}) \in \{0, 1\}$ on every history; write x_t^* for the unique bit on which $\mu(\cdot \mid x_{<t}^*)$ puts all its mass. Define the **threshold predictor** associated with ξ: at each step, predict the more-likely bit,

$$\hat{x}_t \;:=\; \arg\max_{x \in \mathbb{B}} \xi(x \mid x_{<t}) \qquad \text{(break ties arbitrarily)}.$$

We say the mixture makes a **mistake** at time t if $\hat{x}_t \neq x_t^*$.

Exercise 5.1 (Bounded prediction mistakes). [15]
Show that the number of mistakes made by the threshold predictor on the μ-trajectory is at most

$$\#\{\, t \geq 1 \;:\; \hat{x}_t \neq x_t^* \,\} \;\leq\; -2 \ln w_\mu.$$

Hint: At a mistake step, $\xi(x_t^* \mid x_{<t}) \leq \frac{1}{2}$ (why?). The per-step squared error along the deterministic μ-trajectory simplifies dramatically; bound it from below, then sum.

Solution to Exercise 5.1

Step 1: per-step squared error along the μ-trajectory. Because μ is deterministic, the outer sum over $x_{<t}$ in d_t collapses onto the unique trajectory

$x^*_{<t} := x^*_1 \cdots x^*_{t-1}$, with weight $\mu(x^*_{<t}) = 1$. Write $\xi_t := \xi(x^*_t \mid x^*_{<t}) \in [0,1]$ for the mass the mixture places on the *correct* bit. Since μ assigns mass 1 to x^*_t and 0 to the other bit,

$$\begin{aligned} s_t &= \sum_{x_t \in \mathbb{B}} \big(\xi(x_t \mid x^*_{<t}) - \mu(x_t \mid x^*_{<t})\big)^2 \\ &= (\xi_t - 1)^2 + (1 - \xi_t - 0)^2 \;=\; 2(1-\xi_t)^2. \end{aligned}$$

Step 2: a mistake step contributes at least $\frac{1}{2}$. A mistake at time t means $\hat{x}_t \neq x^*_t$, i.e. the threshold predictor chose the *other* bit. This requires $\xi(\hat{x}_t \mid x^*_{<t}) \geq \xi(x^*_t \mid x^*_{<t})$, i.e. $1 - \xi_t \geq \xi_t$, i.e. $\xi_t \leq \frac{1}{2}$. So at every mistake step,

$$s_t \;=\; 2(1-\xi_t)^2 \;\geq\; 2 \cdot \tfrac{1}{4} \;=\; \tfrac{1}{2}.$$

Step 3: sum and apply the cumulative bound. Let N be the total number of mistakes. Summing d_t over mistake steps (non-negative contributions from non-mistake steps only help):

$$\tfrac{1}{2} \cdot N \;\leq\; \sum_{t\,:\,\hat{x}_t \neq x^*_t} s_t \;\leq\; \sum_{t=1}^{\infty} s_t \;=\; S^{\mu}_{\infty} \;\leq\; -\ln w_\mu,$$

where the last step is Section 4. Rearranging, $N \leq -2 \ln w_\mu$.

Remark (Solomonoff specialization)

Under the Solomonoff prior $w_\mu = 2^{-K(\mu)}$, so $-\ln w_\mu = K(\mu) \ln 2$ and the bound becomes

$$\#\{\, t \geq 1 \;:\; \hat{x}_t \neq x^*_t \,\} \;\leq\; 2K(\mu) \ln 2.$$

With $\ln 2 \approx 0.693$ the constant is just under 1.4, so if the true environment μ is computable, and can be described by a program at most k bits long, then the predictor ξ will make at worst $\approx 1.4k$ mistakes over predicting the entire sequence $x^*_{1:\infty}$. Note that the bound is purely existence-style: it does not say *when* the mistakes happen. They could all occur at the start, or be arbitrarily far into the future.

6. Pareto optimality under KL loss

The final two sections establish that the Bayesian mixture ξ is *uniquely Pareto-optimal* under both KL and squared prediction loss: no other predictor can do at least as well as ξ on every environment $\nu \in \mathcal{M}$ without coinciding with ξ entirely.

Definition 6.1 (Pareto domination)**.** For some measure of **loss** $\mathcal{F}(\nu, \rho)$ that takes an environment ν and a predictor ρ, a predictor ρ **weakly Pareto-dominates** ξ with respect to loss $\mathcal{F}$ and class $\mathcal{M}$ if

$$\mathcal{F}(\nu, \rho) \;\leq\; \mathcal{F}(\nu, \xi) \qquad \text{for every } \nu \in \mathcal{M}.$$

ξ is **Pareto-optimal** (w.r.t. $\mathcal{F}$) if the only predictor that weakly dominates it is ξ itself.
We specialize $\mathcal{F}$ to $D_n(\nu \parallel \rho)$ in Section 6 and to $S_n(\nu \parallel \rho)$ in Section 8.

Exercise 6.1 (KL Pythagorean identity). [10]
For *any* predictor ρ with $\rho(x_{1:n}) > 0$ on every $x_{1:n} \in \mathbb{B}^n$, show that

$$D_n(\xi \parallel \rho) \;+\; \sum_{\nu \in \mathcal{M}} w_\nu \, D_n(\nu \parallel \xi) \;=\; \sum_{\nu \in \mathcal{M}} w_\nu \, D_n(\nu \parallel \rho).$$

Hint: Take the difference between the two summation terms.

Solution to Exercise 6.1

Take the difference $\sum_\nu w_\nu D_n(\nu \parallel \rho) - \sum_\nu w_\nu D_n(\nu \parallel \xi)$. Inside each D_n the $\nu(x_{1:n}) \ln \nu(x_{1:n})$ pieces cancel, leaving

$$\begin{aligned}
&\sum_\nu w_\nu D_n(\nu \parallel \rho) - \sum_\nu w_\nu D_n(\nu \parallel \xi) \\
&\quad = \sum_\nu w_\nu \sum_{x_{1:n}} \nu(x_{1:n}) \left[\ln \tfrac{\nu(x_{1:n})}{\rho(x_{1:n})} - \ln \tfrac{\nu(x_{1:n})}{\xi(x_{1:n})} \right] \\
&\quad = \sum_\nu w_\nu \sum_{x_{1:n}} \nu(x_{1:n}) \, \ln \tfrac{\xi(x_{1:n})}{\rho(x_{1:n})}.
\end{aligned}$$

The factor $\ln(\xi/\rho)$ does not depend on ν, so we swap the order of summation and pull it out, then use $\sum_\nu w_\nu \nu(x_{1:n}) = \xi(x_{1:n})$:

$$\begin{aligned}
\sum_\nu w_\nu \sum_{x_{1:n}} \nu(x_{1:n}) \, \ln \tfrac{\xi(x_{1:n})}{\rho(x_{1:n})} &= \sum_{x_{1:n}} \ln \tfrac{\xi(x_{1:n})}{\rho(x_{1:n})} \underbrace{\sum_\nu w_\nu \nu(x_{1:n})}_{= \xi(x_{1:n})} \\
&= \sum_{x_{1:n}} \xi(x_{1:n}) \, \ln \tfrac{\xi(x_{1:n})}{\rho(x_{1:n})} \;=\; D_n(\xi \parallel \rho).
\end{aligned}$$

Rearranging gives the claimed identity.

Exercise 6.2 (KL Pareto-optimality). [05]
Show that ξ is Pareto-optimal under $D_n(\nu \parallel \rho)$ in the sense of Definition 6.1: if ρ is any predictor with

$$D_n(\nu \parallel \rho) \;\leq\; D_n(\nu \parallel \xi) \qquad \text{for every } \nu \in \mathcal{M},$$

then $\rho = \xi$.

Hint: Multiply the assumed inequality by $w_\nu \geq 0$ and sum over $\nu \in \mathcal{M}$; then apply the KL Pythagorean identity to recognize the extra non-negative term, which must vanish.

Solution to Exercise 6.2

Multiply the assumed inequality by $w_\nu \geq 0$ and sum over $\nu \in \mathcal{M}$:

$$\sum_\nu w_\nu \, D_n(\nu \parallel \rho) \;\leq\; \sum_\nu w_\nu \, D_n(\nu \parallel \xi). \tag{3}$$

The KL Pythagorean identity rewrites the LHS as

$$D_n(\xi \parallel \rho) \;+\; \sum_\nu w_\nu \, D_n(\nu \parallel \xi),$$

so Equation (3) forces $D_n(\xi \parallel \rho) \leq 0$. KL is non-negative, hence $D_n(\xi \parallel \rho) = 0$, which forces $\rho(x_{1:n}) = \xi(x_{1:n})$ for every $x_{1:n} \in \mathbb{B}^n$. By chain rule, the conditionals agree at every history $x_{<t}$ with $\xi(x_{<t}) > 0$.

7. Misspecified models: when $\mu \notin \mathcal{M}$

The cumulative bound Section 4 assumed the true environment μ lives in the model class $\mathcal{M}$. What happens when it does not? We now show the bound *degrades gracefully*: it splits cleanly into a complexity term ("cost of not knowing which model is best"), exactly as before, plus a linear-in-time approximation term ("cost of no model being right"). See [Hut05, §3.2.8] for the original treatment.

Throughout this section, we no longer assume $\mu \in \mathcal{M}$. We will state the bound in terms of an arbitrary $\hat{\mu} \in \mathcal{M}$, so taking the infimum over $\hat{\mu}$ on the right-hand side gives the tightest version by using the "closest" approximation of $\hat{\mu} \in \mathcal{M}$ to μ.

Exercise 7.1 (Misspecified KL bound). [10]

Show that for all $\hat{\mu} \in \mathcal{M}$,

$$D_n(\mu \,\|\, \xi) \;\leq\; -\ln w_{\hat{\mu}} \;+\; D_n(\mu \,\|\, \hat{\mu}).$$

Hint: Use Exercise 3.1.

Solution to Exercise 7.1

Mixture dominance applied to $\hat{\mu} \in \mathcal{M}$ (Section 3) gives $\xi(x_{1:n}) \geq w_{\hat{\mu}}\, \hat{\mu}(x_{1:n})$ for every $x_{1:n}$. Hence $\mu(x_{1:n})/\xi(x_{1:n}) \leq \mu(x_{1:n})/(w_{\hat{\mu}}\, \hat{\mu}(x_{1:n}))$, and taking logs,

$$\ln \frac{\mu(x_{1:n})}{\xi(x_{1:n})} \;\leq\; -\ln w_{\hat{\mu}} \;+\; \ln \frac{\mu(x_{1:n})}{\hat{\mu}(x_{1:n})}.$$

Now take the μ-expectation. The LHS becomes $D_n(\mu \,\|\, \xi)$; the second RHS term becomes $D_n(\mu \,\|\, \hat{\mu})$; the constant $-\ln w_{\hat{\mu}}$ survives because $\mu(X_{1:n})$ has total mass 1.

Remark (Reading the two terms)

The right-hand side splits into two qualitatively different terms:

- $-\ln w_{\hat{\mu}}$ is *constant in* n: under the Solomonoff prior, this is $K(\hat{\mu}) \ln 2$. The familiar "complexity" cost of search.
- The approximation term $D_n(\mu \,\|\, \hat{\mu}) = \sum_{t=1}^{n} d_t(\mu \,\|\, \hat{\mu})$ (by Exercise 2.1) is a sum of n per-step KLs; in general it grows with n. It vanishes identically iff $\hat{\mu}$ matches μ on every μ-reachable history, in particular when $\mu \in \mathcal{M}$ and we pick $\hat{\mu} = \mu$, recovering Section 4.

Combining with Pinsker (Theorem 4.1) gives the corresponding bound on S_∞^μ, which is infinite in general but inherits the rate of $\hat{\mu}$.

Details on how to define the best choice of $\hat{\mu}$ are in Appendix D.

8. Pareto optimality under squared loss

The squared specialization $\mathcal{F}(\nu, \rho) = S_n(\nu \,\|\, \rho)$ (Definition 4.2) lives one level down from KL: at the conditional distributions $\nu(\cdot \mid x_{<t})$. The natural Pythagorean decomposition at a fixed history uses *posterior* weights $w(\nu \mid x_{<t})$: those are the weights that make $\xi(x_t \mid x_{<t})$ the mean of the $\nu(x_t \mid x_{<t})$'s (via the posterior-predictive form, Definition 1.2). The Pareto-optimality aggregation, however, uses prior weights w_ν. Bridging the two takes one extra Bayes-rule step.

Exercise 8.1 (Squared Pythagorean identity). [10]
Fix any $t \in \{1, \dots, n\}$ and any history $x_{<t} \in \mathbb{B}^{t-1}$: *the history is arbitrary, not sampled from any environment.* For any $x_t \in \mathbb{B}$ and any predictor ρ, show

$$\big(\xi(x_t \mid x_{<t}) - \rho(x_t \mid x_{<t})\big)^2 \;+\; \sum_{\nu \in \mathcal{M}} w(\nu \mid x_{<t}) \big(\nu(x_t \mid x_{<t}) - \xi(x_t \mid x_{<t})\big)^2$$
$$= \sum_{\nu \in \mathcal{M}} w(\nu \mid x_{<t}) \big(\nu(x_t \mid x_{<t}) - \rho(x_t \mid x_{<t})\big)^2.$$

Hint: Add and subtract $\xi(x_t \mid x_{<t})$ inside the squared term, then expand. Use $\sum_\nu w(\nu \mid x_{<t}) = 1$ and the posterior-predictive form of ξ (Definition 1.2), $\sum_\nu w(\nu \mid x_{<t})\, \nu(x_t \mid x_{<t}) = \xi(x_t \mid x_{<t})$, to kill the cross-term. The weighting must be *posterior* weights $w(\nu \mid x_{<t})$ (not prior weights w_ν), because $\xi(x_t \mid x_{<t})$ is the posterior-weighted mean of the $\nu(x_t \mid x_{<t})$'s, not the prior one.

Solution to Exercise 8.1

Throughout, the history $x_{<t}$ and symbol x_t are fixed but arbitrary. Abbreviate $\nu := \nu(x_t \mid x_{<t})$, $\xi := \xi(x_t \mid x_{<t})$, $\rho := \rho(x_t \mid x_{<t})$, and $\tilde{w}_\nu := w(\nu \mid x_{<t})$ for the duration of the calculation.

Step 1: ξ is the posterior-weighted mean of the ν's. By the posterior-predictive form of the mixture (Definition 1.2) and the posterior normalization $\sum_\nu \tilde{w}_\nu = 1$,

$$\sum_\nu \tilde{w}_\nu \nu \;=\; \xi.$$

Step 2: the cross-term vanishes. A direct consequence of Step 1, again using $\sum_\nu \tilde{w}_\nu = 1$:

$$\sum_\nu \tilde{w}_\nu (\nu - \xi) \;=\; \sum_\nu \tilde{w}_\nu \nu \;-\; \xi \sum_\nu \tilde{w}_\nu \;=\; \xi - \xi \;=\; 0.$$

Step 3: expand the square. Write $\nu - \rho = (\nu - \xi) + (\xi - \rho)$ and expand:

$$\begin{aligned}
&\sum_\nu \tilde{w}_\nu (\nu - \rho)^2 \\
&\quad = \sum_\nu \tilde{w}_\nu \big[(\nu - \xi)^2 + 2(\nu - \xi)(\xi - \rho) + (\xi - \rho)^2\big] \\
&\quad = \sum_\nu \tilde{w}_\nu (\nu - \xi)^2 \;+\; 2(\xi - \rho) \sum_\nu \tilde{w}_\nu (\nu - \xi) \;+\; (\xi - \rho)^2 \sum_\nu \tilde{w}_\nu.
\end{aligned}$$

The middle sum is zero by Step 2; the trailing $\sum_\nu \tilde{w}_\nu = 1$. So

$$(\xi - \rho)^2 \;+\; \sum_\nu \tilde{w}_\nu (\nu - \xi)^2 \;=\; \sum_\nu \tilde{w}_\nu (\nu - \rho)^2.$$

Exercise 8.2 (Squared Pareto-optimality). [05]
Show that ξ is Pareto-optimal under $S_n(\nu \parallel \rho)$ in the sense of Definition 6.1: if ρ is any predictor with

$$S_n(\nu \parallel \rho) \;\leq\; S_n(\nu \parallel \xi) \qquad \text{for every } \nu \in \mathcal{M},$$

then $\rho(x_t \mid x_{<t}) = \xi(x_t \mid x_{<t})$ at every history $x_{<t} \in \mathbb{B}^{t-1}$ *with* $\xi(x_{<t}) > 0$ (equivalently, ξ-almost surely) and every $x_t \in \mathbb{B}$. Histories that the mixture never reaches ($\xi(x_{<t}) = 0$, i.e., reached by no $\nu \in \mathcal{M}$ either) are invisible to the loss and are not pinned down.

Hint: Multiply the assumed inequality by $w_\nu \geq 0$ and sum over ν; then for each fixed history $x_{<t}$, multiply the per-history Pythagorean identity from above by $\xi(x_{<t})$ and use the Bayes identity $w(\nu \mid x_{<t})\,\xi(x_{<t}) = w_\nu\,\nu(x_{<t})$ to bridge from posterior weights (inside the identity) to prior weights (in the aggregated Pareto inequality).

Solution to Exercise 8.2

Multiply the assumed inequality by $w_\nu \geq 0$ and sum over $\nu \in \mathcal{M}$:

$$\sum_\nu w_\nu\, S_n(\nu \parallel \rho) \;\leq\; \sum_\nu w_\nu\, S_n(\nu \parallel \xi). \tag{4}$$

Now bridge to the per-history Pythagorean. At each *fixed* history $x_{<t} \in \mathbb{B}^{t-1}$ and symbol $x_t \in \mathbb{B}$, the identity says

$$(\xi - \rho)^2 \;+\; \sum_\nu w(\nu \mid x_{<t})\,(\nu - \xi)^2 \;=\; \sum_\nu w(\nu \mid x_{<t})\,(\nu - \rho)^2,$$

with the abbreviations $\nu := \nu(x_t \mid x_{<t})$, $\xi := \xi(x_t \mid x_{<t})$, $\rho := \rho(x_t \mid x_{<t})$. Multiply by $\xi(x_{<t})$ and use Bayes, $w(\nu \mid x_{<t})\,\xi(x_{<t}) = w_\nu\,\nu(x_{<t})$:

$$\xi(x_{<t})(\xi - \rho)^2 \;+\; \sum_\nu w_\nu\,\nu(x_{<t})\,(\nu - \xi)^2 \;=\; \sum_\nu w_\nu\,\nu(x_{<t})\,(\nu - \rho)^2.$$

Sum over $t = 1, \ldots, n$ and all histories $x_{<t} \in \mathbb{B}^{t-1}$ and symbols $x_t \in \mathbb{B}$. The two $\sum_\nu w_\nu\,\nu(x_{<t})(\cdots)$ terms become exactly $\sum_\nu w_\nu\, S_n(\nu \parallel \xi)$ and $\sum_\nu w_\nu\, S_n(\nu \parallel \rho)$ respectively, so

$$\|\xi - \rho\|_\xi^2 \;+\; \sum_\nu w_\nu\, S_n(\nu \parallel \xi) \;=\; \sum_\nu w_\nu\, S_n(\nu \parallel \rho),$$

where

$$\|\xi - \rho\|_\xi^2 \;:=\; \sum_{t=1}^{n} \sum_{x_{<t}} \xi(x_{<t}) \sum_{x_t} \big(\xi(x_t \mid x_{<t}) - \rho(x_t \mid x_{<t})\big)^2 \;\geq\; 0.$$

Combining with Equation (4),

$$\|\xi - \rho\|_\xi^2 \;+\; \sum_\nu w_\nu \, S_n(\nu \,\|\, \xi) \;\le\; \sum_\nu w_\nu \, S_n(\nu \,\|\, \xi) \;\;\Longrightarrow\;\; \|\xi - \rho\|_\xi^2 \;\le\; 0.$$

Since $\|\xi - \rho\|_\xi^2$ is a sum of non-negative terms and is itself ≤ 0, every term vanishes: $\xi(x_{<t})\big(\xi(x_t \mid x_{<t}) - \rho(x_t \mid x_{<t})\big)^2 = 0$ at every $(t, x_{<t}, x_t)$. So $\rho(x_t \mid x_{<t}) = \xi(x_t \mid x_{<t})$ at every history $x_{<t}$ with $\xi(x_{<t}) > 0$.

A. Proof of Pinsker's inequality

We first prove the following Lemma:

Lemma A.1 (Pinsker's binary inequality)**.** *For $0 \le p \le 1$ and $0 < q < 1$ we have*

$$2(p-q)^2 \;\le\; p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q}.$$

Define $f(x) := p \ln x + (1-p) \ln(1-x)$, so that

$$p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q} = f(p) - f(q) = \int_q^p f'(x)\, dx = \int_q^p \frac{p-x}{x(1-x)}\, dx,$$

using $f_p'(x) = \frac{p}{x} - \frac{1-p}{1-x} = \frac{p-x}{x(1-x)}$. Recall also that $x(1-x) \le \frac{1}{4}$ on $[0,1]$, so $\frac{1}{x(1-x)} \ge 4$.

Case $q \le p$. For every $x \in [q,p]$ we have $x \le p$, hence $p - x \ge 0$.

$$\frac{p-x}{x(1-x)} \;\ge\; 4(p-x) \qquad \text{for } x \in [q,p].$$

Integrating over $[q,p]$,

$$\int_q^p \frac{p-x}{x(1-x)}\, dx \;\ge\; 4 \int_q^p (p-x)\, dx \;=\; 2(p-q)^2.$$

Case $q \ge p$. For every $x \in [p,q]$ we have $x \ge p$, hence $x - p \ge 0$.

$$\frac{x-p}{x(1-x)} \;\ge\; 4(x-p) \qquad \text{for } x \in [p,q].$$

Integrating over $[q,p]$,

$$\begin{aligned} \int_q^p \frac{p-x}{x(1-x)}\, dx &= \int_p^q \frac{x-p}{x(1-x)}\, dx \\ &\ge 4 \int_p^q (x-p)\, dx = 4 \int_q^p (p-x)\, dx = 2(p-q)^2 \end{aligned}$$

Proof of Theorem 4.1. We now reduce Theorem 4.1 to the binary inequality Lemma A.1. Recall $p_0 + p_1 = 1$ and $q_0 + q_1 = 1$. We handle the boundary case $q_x = 0$ first, then reduce the interior case directly to Lemma A.1.

Case 1: $q_x = 0$ for some $x \in \mathbb{B}$. If $p_x > 0$, then the right-hand side of $(\star)$ contains $p_x \ln \frac{p_x}{0} = +\infty$, so $(\star)$ holds trivially. If instead $p_x = 0$, that atom contributes 0 to both sides (using the convention $0 \ln \frac{0}{0} = 0$). The other atom x' then satisfies $p_{x'} = 1$ and $q_{x'} = 1 - q_x = 1$, so both sides vanish and $(\star)$ holds.

Case 2: $q_0, q_1 > 0$**.** Since $q_1 = 1 - q_0$ and $p_1 = 1 - p_0$, the two sides of $(\star)$ become

$$\sum_{x \in \mathbb{B}} (q_x - p_x)^2 = (q_0 - p_0)^2 + \big((1 - q_0) - (1 - p_0)\big)^2 = 2(p_0 - q_0)^2,$$

$$\sum_{x \in \mathbb{B}} p_x \ln \tfrac{p_x}{q_x} = p_0 \ln \tfrac{p_0}{q_0} + (1 - p_0) \ln \tfrac{1-p_0}{1-q_0}.$$

So $(\star)$ is exactly $2(p_0 - q_0)^2 \le p_0 \ln \frac{p_0}{q_0} + (1 - p_0) \ln \frac{1-p_0}{1-q_0}$, which is Lemma A.1 with $p = p_0$ and $q = q_0$ (here $0 < q_0 < 1$, since $q_0, q_1 > 0$). □

B. Proof of KL non-negativity

We prove Theorem 2.2. The core is the elementary inequality

$$\ln x \;\le\; x - 1 \qquad \text{for all } x > 0, \quad \text{equality iff } x = 1.$$

(Let $f(x) := (x - 1) - \ln x$; then $f'(x) = 1 - 1/x$, vanishing only at $x = 1$, with $f''(x) = 1/x^2 > 0$, so f is strictly convex and attains its unique minimum $f(1) = 0$ at $x = 1$.)

Lemma B.1 (Gibbs' inequality)**.** *Let P, Q be probability distributions over a finite set $\mathcal{X}$, with the conventions $0 \ln \frac{0}{q} := 0$ and $p \ln \frac{p}{0} := \infty$. Then*

$$\sum_{x \in \mathcal{X}} P(x) \ln \frac{P(x)}{Q(x)} \;\ge\; 0,$$

with equality iff $P(x) = Q(x)$ for every x with $P(x) > 0$.

Proof. If $Q(x) = 0$ for some x with $P(x) > 0$, the LHS is $+\infty$ and the inequality is strict. Otherwise restrict the sum to $\{x : P(x) > 0\}$ (terms with $P(x) = 0$ contribute 0 by convention), where every ratio $Q(x)/P(x)$ is in $(0, \infty)$. Apply $\ln y \le y - 1$ to $y = Q(x)/P(x)$:

$$\begin{aligned} -\sum_x P(x) \ln \frac{P(x)}{Q(x)} \;&=\; \sum_x P(x) \ln \frac{Q(x)}{P(x)} \;\le\; \sum_x P(x) \left(\frac{Q(x)}{P(x)} - 1 \right) \\ &=\; \sum_x Q(x) - \sum_x P(x) \;\le\; 0, \end{aligned}$$

where the last step uses $\sum_x Q(x) \le 1 = \sum_x P(x)$ (the sum over $\{x : P(x) > 0\}$ may miss some Q-mass). Negating gives the claim. Equality in $\ln y \le y - 1$ requires $y = 1$ for every x with $P(x) > 0$, i.e. $Q(x) = P(x)$ there; equality in $\sum Q \le \sum P$ then forces Q to have no mass outside $\operatorname{supp} P$. Both together give $P = Q$ on all of $\mathcal{X}$. □

Proof of Theorem 2.2. Part (a). Apply Lemma B.1 with $\mathcal{X} = \mathbb{B}^n$, $P(x_{1:n}) = \nu(x_{1:n})$, $Q(x_{1:n}) = \rho(x_{1:n})$. Both are probability distributions on $\mathbb{B}^n$ since ν and ρ are environments. The lemma gives $D_n(\nu \parallel \rho) \geq 0$ with the stated equality condition.

Part (b). For each fixed history $x_{<t}$ with $\nu(x_{<t}) > 0$, apply Lemma B.1 with $\mathcal{X} = \mathbb{B}$, $P = \nu(\cdot \mid x_{<t})$, $Q = \rho(\cdot \mid x_{<t})$ (both are probability distributions on $\mathbb{B}$):

$$\sum_{x_t \in \mathbb{B}} \nu(x_t \mid x_{<t}) \ln \frac{\nu(x_t \mid x_{<t})}{\rho(x_t \mid x_{<t})} \;\geq\; 0,$$

with equality iff $\nu(\cdot \mid x_{<t}) = \rho(\cdot \mid x_{<t})$. Multiply by $\nu(x_{<t}) \geq 0$ and sum over $x_{<t}$: this gives $d_t(\nu \parallel \rho) \geq 0$, as a sum of non-negative terms. Equality forces every $x_{<t}$ with $\nu(x_{<t}) > 0$ to contribute a vanishing inner sum, which is exactly the stated condition. □

C. Model Class $\mathcal{M}$

All results in this sheet hold for any countable class $\mathcal{M}$ and prior w_ν. The canonical **Solomonoff** choice takes $\mathcal{M}$ to be the class of all computable environments and the **Solomonoff prior**

$$w_\nu \;:=\; 2^{-K(\nu)},$$

where $K(\nu)$ is the *Kolmogorov complexity* of ν: the length of the shortest program that computes ν. The Bayesian mixture ξ under this particular prior is the *universal mixture*, written ξ_U. With this class the assumption $\mu \in \mathcal{M}$ reduces to "the universe is computable": as weak an assumption as one can make. For technical details on Kolmogorov complexity, see [HQC24, §2.7].

The class of all computable measures is countable but not enumerable (by the halting problem, you can't decide which Turing machines define total functions), so one can't even formally write $\sum_{\nu \in \mathcal{M}} \ldots$ as the decision problem of determining if $\nu \overset{?}{\in} \mathcal{M}$ is undecidable. A faithful construction of ξ_U requires expanding $\mathcal{M}$ to all *lower-semicomputable semimeasures* (Levin), at which point ξ_U becomes a semimeasure rather than a measure: $\sum_x \xi_U(x) \leq 1$ instead of $= 1$, conditionals may not sum to 1, and every other proof in this sheet gains a side-case. This also gives the nice benefit that $\xi \in \mathcal{M}$.

We dodge all of this: $\mathcal{M}$ is just *some* countable set of proper measures with $\mu \in \mathcal{M}$, and we assume $K(\nu)$ is defined for each $\nu \in \mathcal{M}$ when we need it. We never need ξ itself to be in $\mathcal{M}$, nor do we need its computability, so we don't ask. See [Hut05, §2.4.3] for the full Levin construction.

D. Best choice of $\hat{\mu} \in \mathcal{M}$.

The bound is valid for every $\hat{\mu} \in \mathcal{M}$. A tempting question: which $\hat{\mu}$ gives the tightest bound? The answer is genuinely setting-dependent.

Finite horizon n. The minimizer of the RHS is

$$\hat{\mu}_n \;:=\; \arg\min_{\nu \in \mathcal{M}} \Big\{ -\ln w_\nu \;+\; D_n(\mu \,\|\, \nu) \Big\}.$$

This is well-defined (finitely many terms, attained when $\mathcal{M}$ is finite), but the answer depends on n: for small n the prior term dominates and a high-prior coarse model wins; for large n the fit term dominates and the best per-step approximator wins. Both regimes are correct.

Asymptotic rate. A horizon-free analogue minimizes the per-step KL rate

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t \le n} d_t(\mu \,\|\, \nu).$$

This is the right notion when one cares about the linear-growth slope of the bound. But it has two drawbacks: the limit need not exist for non-stationary μ (one can substitute lim inf but at the cost of clarity), and it discards the prior w_ν entirely, so the minimizer is not the same as $\hat{\mu}_n$ at any finite n, but only of its slope.

Stationary μ. The per-step KLs $d_t(\mu \,\|\, \nu)$ are constant in t, so all three notions essentially agree: $\hat{\mu} = \arg\min_{\nu \in \mathcal{M}} D_{\mathrm{KL}}(\mu \,\|\, \nu)$, where the KL is taken under the (common) one-step conditional law. This is the case where "$\hat{\mu}$ = projection of μ onto $\mathcal{M}$ in KL" has unambiguous meaning.

Upshot. The bound holds for all $\hat{\mu} \in \mathcal{M}$, and one is free to take the infimum on the RHS over $\hat{\mu}$. Which $\hat{\mu}$ actually attains that infimum is a separate question whose answer depends on horizon, prior, and any structural assumptions on μ.

References

[CT06] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley-Intersience, 2nd edition, 2006.

[HQC24] Marcus Hutter, David Quarel, and Elliot Catt. *An Introduction to Universal Artificial Intelligence.* Chapman & Hall, 2024. URL: http://www.hutter1.net/ai/uaibook2.htm.

[Hut05] Marcus Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability.* Springer, Berlin, 2005. URL: http://www.hutter1.net/ai/uaibook.htm, doi:10.1007/b138233.

[Knu73] D. E. Knuth. *The Art of Computer Programming, Volume I: Fundamental Algorithms*. Addison-Wesley, Reading, MA, 1973.