Iliad Intensive **London Initiative for Safe AI**
April 2026 Leon Lang

Solomonoff Induction Exercises April 7, 2026

Setup and Definitions

For more setup and definitions, see the corresponding post on Solomonoff induction.

A **semiprobability distribution** over a finite set $\mathcal{X}$ is a function $Q : \mathcal{X} \to [0, 1]$ with $Q(x) \geq 0$ and $\sum_{x\in\mathcal{X}} Q(x) \leq 1$.

A **semimeasure** over infinite binary sequences is a function $\nu : \{0, 1\}^* \to [0, 1]$ with $\nu(\epsilon) \leq 1$ and $\nu(x) \geq \nu(x0) + \nu(x1)$ for all $x \in \{0, 1\}^*$. A semimeasure is a **measure** if equality holds for both equations.

The **universal a priori distribution** M is a lower semicomputable semimeasure, which can be written in the form

$$M(x_{\leq n}) = \sum_{\nu\in\mathcal{M}_{sol}} P_{ap}(\nu)\nu(x_{\leq n}) \tag{1}$$

for all finite strings $x_{\leq n}$, where $\mathcal{M}_{sol}$ is the set of all lower semicomputable semimeasures and $P_{ap}(\nu)$ is the **a priori probability** of ν.

With $\nu_i, i = 1, 2, \ldots$ an effective enumeration of all lower semicomputable discrete semimeasures and the **Solomonoff prior** $P_{sol}(i) = 2^{-K(i)}$, there also exist two constants $0 < c < C$ such that

$$c \cdot \sum_{i\in\mathbb{N}} P_{sol}(i)\nu_i(x_{\leq n}) < M(x_{\leq n}) < C \cdot \sum_{i\in\mathbb{N}} P_{sol}(i)\nu_i(x_{\leq n}). \tag{2}$$

Problem 1

Let P be a probability distribution over $\mathcal{X}$ and Q be a semiprobability distribution over the same set. The *Kullback–Leibler (KL) divergence* from P to Q is defined as

$$D_{\mathrm{KL}}\big(P \parallel Q\big) := \sum_{x\in\mathcal{X}} P(x) \ln \frac{P(x)}{Q(x)},$$

where we use the conventions $0 \ln \frac{0}{q} = 0$ for any $q \geq 0$, and $p \ln \frac{p}{0} = \infty$ for $p > 0$.

Now let (X, Y) be a pair of random variables taking values in $\mathcal{X} \times \mathcal{Y}$, and let P and Q be two joint distributions over (X, Y) (with Q only being a *semi*probability distribution). Define the *conditional KL divergence* as

$$D_{\mathrm{KL}}\big(P(Y \mid X) \parallel Q(Y \mid X)\big) := \sum_{x \in \mathcal{X}} P(x) \sum_{y \in \mathcal{Y}} P(y \mid x) \ln \frac{P(y \mid x)}{Q(y \mid x)}.$$

Show that the KL divergence satisfies the *chain rule*:

$$D_{\mathrm{KL}}\big(P(X, Y) \parallel Q(X, Y)\big) = D_{\mathrm{KL}}\big(P(X) \parallel Q(X)\big) + D_{\mathrm{KL}}\big(P(Y \mid X) \parallel Q(Y \mid X)\big).$$

Solution

We expand the joint KL divergence using $P(x, y) = P(x)P(y \mid x)$ and $Q(x, y) = Q(x)Q(y \mid x)$:

$$\begin{aligned}
D_{\mathrm{KL}}\big(P(X, Y) \parallel Q(X, Y)\big) &= \sum_{x,y} P(x, y) \ln \frac{P(x, y)}{Q(x, y)} \\
&= \sum_{x,y} P(x)P(y \mid x) \ln \frac{P(x)P(y \mid x)}{Q(x)Q(y \mid x)} \\
&= \sum_{x,y} P(x)P(y \mid x) \left[\ln \frac{P(x)}{Q(x)} + \ln \frac{P(y \mid x)}{Q(y \mid x)} \right] \\
&= \sum_{x} P(x) \ln \frac{P(x)}{Q(x)} \underbrace{\sum_{y} P(y \mid x)}_{=1} + \sum_{x} P(x) \sum_{y} P(y \mid x) \ln \frac{P(y \mid x)}{Q(y \mid x)} \\
&= D_{\mathrm{KL}}\big(P(X) \parallel Q(X)\big) + D_{\mathrm{KL}}\big(P(Y \mid X) \parallel Q(Y \mid X)\big).
\end{aligned}$$

Problem 2

Let P be a joint probability distribution over random variables $X_1, \ldots, X_n$ taking values in $\mathcal{X}_1 \times \cdots \times \mathcal{X}_n$, and let Q be a *semi*probability distribution over the same set of variables. Using the notation $X_{<k} := (X_1, \ldots, X_{k-1})$, show by induction on n (using the result of Problem 1) that

$$D_{\mathrm{KL}}\big(P(X_1, \ldots, X_n) \parallel Q(X_1, \ldots, X_n)\big) = \sum_{t=1}^{n} D_{\mathrm{KL}}\big(P(X_t \mid X_{<t}) \parallel Q(X_t \mid X_{<t})\big),$$

where $P(X_1 \mid X_{<1}) := P(X_1)$ and similarly for Q.

Solution

We prove this by induction on n.
Base case ($n = 1$): The sum reduces to $D_{\mathrm{KL}}\big(P(X_1) \parallel Q(X_1)\big)$, which is just the left-hand side.
Induction step ($n{-}1 \to n$): Apply the chain rule from Problem 1 with $X = X_{<n}$ and $Y = X_n$:

$$\begin{aligned} &D_{\mathrm{KL}}\big(P(X_1, \ldots, X_n) \parallel Q(X_1, \ldots, X_n)\big) \\ &= D_{\mathrm{KL}}\big(P(X_{<n}) \parallel Q(X_{<n})\big) + D_{\mathrm{KL}}\big(P(X_n \mid X_{<n}) \parallel Q(X_n \mid X_{<n})\big). \end{aligned}$$

By the induction hypothesis, the first term satisfies

$$D_{\mathrm{KL}}\big(P(X_{<n}) \parallel Q(X_{<n})\big) = \sum_{t=1}^{n-1} D_{\mathrm{KL}}\big(P(X_t \mid X_{<t}) \parallel Q(X_t \mid X_{<t})\big).$$

Combining both gives

$$D_{\mathrm{KL}}\big(P(X_1, \ldots, X_n) \parallel Q(X_1, \ldots, X_n)\big) = \sum_{t=1}^{n} D_{\mathrm{KL}}\big(P(X_t \mid X_{<t}) \parallel Q(X_t \mid X_{<t})\big).$$

Problem 3

Show that for all $z > 0$,

$$\ln(z) \geq 1 - \frac{1}{z}.$$

Hint: Consider the function $f(z) = \ln(z) - 1 + \frac{1}{z}$ and show that it attains its global minimum at $z = 1$.

Solution

Define $f(z) = \ln(z) - 1 + \frac{1}{z}$. We want to show $f(z) \geq 0$ for all $z > 0$. We have

$$f'(z) = \frac{1}{z} - \frac{1}{z^2} = \frac{z-1}{z^2}.$$

Thus $f'(z) < 0$ for $z \in (0, 1)$ and $f'(z) > 0$ for $z > 1$, so f has a global minimum at $z = 1$. Since

$$f(1) = \ln(1) - 1 + 1 = 0,$$

we conclude $f(z) \geq 0$ for all $z > 0$, i.e., $\ln(z) \geq 1 - \frac{1}{z}$.

Problem 4

Let P be a probability distribution over $\mathbb{B} = \{0, 1\}$ and let Q be a semiprobability distribution over $\mathbb{B}$. Write $p_0 = P(0)$, $p_1 = P(1) = 1 - p_0$, $q_0 = Q(0)$, $q_1 = Q(1)$. We want to show that

$$\sum_{x \in \mathbb{B}} \big(Q(x) - P(x)\big)^2 \leq \sum_{x \in \mathbb{B}} P(x) \ln \frac{P(x)}{Q(x)}.$$

i) Show that the inequality holds whenever $Q(x) = 0$ for some $x \in \mathbb{B}$.

Hint: Distinguish the cases $P(x) > 0$ and $P(x) = 0$, and use Problem 3 for the latter.

ii) Assuming $q_0, q_1 > 0$, show that it suffices to prove the inequality for the case $q_0 + q_1 = 1$, i.e., when Q is a full probability distribution.

Hint: Define $F(q_0, q_1) := p_0 \ln \frac{p_0}{q_0} + p_1 \ln \frac{p_1}{q_1} - (q_0 - p_0)^2 - (q_1 - p_1)^2$ and show that F is decreasing in q_1.

iii) Now assume $q_0 + q_1 = 1$. Show that

$$2(p_0 - q_0)^2 \leq p_0 \ln \frac{p_0}{q_0} + p_1 \ln \frac{p_1}{1 - q_0}$$

and observe this finishes the proof.

Hint: Define $g(q_0)$ as the right-hand side minus the left-hand side, compute $g'(q_0)$, and use that $q_0(1 - q_0) \leq \frac{1}{4}$.

Solution

Part (i): If $P(x) > 0$ and $Q(x) = 0$ for some $x \in \mathbb{B}$, then the right-hand side contains the term $P(x) \ln \frac{P(x)}{0} = +\infty$, so the inequality holds trivially.
If $P(x) = 0$ and $Q(x) = 0$ for some x, then that x contributes 0 to both sides (using the convention $0 \ln \frac{0}{0} = 0$). The other value $x' \neq x$ has $P(x') = 1$ and $Q(x') \leq 1$, so the inequality reduces to

$$(Q(x') - 1)^2 \leq \ln \frac{1}{Q(x')}.$$

By Problem 3 applied to $z = \frac{1}{Q(x')}$, we get $\ln \frac{1}{Q(x')} \geq 1 - Q(x')$. Since $1 - Q(x') \in [0, 1]$, we have $(1 - Q(x'))^2 \leq 1 - Q(x') \leq \ln \frac{1}{Q(x')}$.

Part (ii): Define

$$F(q_0, q_1) := p_0 \ln \frac{p_0}{q_0} + p_1 \ln \frac{p_1}{q_1} - (q_0 - p_0)^2 - (q_1 - p_1)^2.$$

We want to show $F(q_0, q_1) \geq 0$. We compute

$$\frac{\partial F}{\partial q_1} = -\frac{p_1}{q_1} - 2(q_1 - p_1).$$

This has no real roots as a function of q_1: setting it to zero gives $2q_1^2 - 2p_1q_1 + p_1 = 0$, whose discriminant is $4p_1(p_1 - 2) < 0$. Since $\frac{\partial F}{\partial q_1}\big|_{q_1=p_1} = -1 < 0$, the derivative is strictly negative for all $q_1 > 0$.
Thus F is decreasing in q_1. For any semiprobability distribution with $q_0 + q_1 \leq 1$, increasing q_1 to $1 - q_0$ only decreases F:

$$F(q_0, q_1) \geq F(q_0, 1 - q_0).$$

So it suffices to show $F(q_0, 1 - q_0) \geq 0$.

Part (iii): Setting $q_1 = 1 - q_0$, we have $(q_0 - p_0)^2 + (q_1 - p_1)^2 = 2(q_0 - p_0)^2$, so we need to show

$$g(q_0) := p_0 \ln \frac{p_0}{q_0} + p_1 \ln \frac{p_1}{1 - q_0} - 2(p_0 - q_0)^2 \geq 0$$

for all $q_0 \in (0, 1)$. Differentiating:

$$g'(q_0) = -\frac{p_0}{q_0} + \frac{p_1}{1 - q_0} + 4(p_0 - q_0) = (q_0 - p_0)\left[\frac{1}{q_0(1 - q_0)} - 4\right].$$

Since $q_0(1 - q_0) \leq \frac{1}{4}$, the bracket is non-negative. Thus $g'(q_0) \leq 0$ for $q_0 < p_0$ and $g'(q_0) \geq 0$ for $q_0 > p_0$, so g has a global minimum at $q_0 = p_0$. Since $g(p_0) = 0$, we conclude $g(q_0) \geq 0$ for all $q_0 \in (0, 1)$.

Problem 5

Let μ be any lower semicomputable *measure* (!) on binary sequences, and assume it generates our actual universe. Let M be the Solomonoff mixture distribution.

Define the cumulative expected prediction error of predicting sequences sampled by μ via M as:

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

Show that

$$S_\infty^\mu \leq -\ln P_{ap}(\mu).$$

Hint: Apply Problem 4 to each summand to pass from squared error to KL divergence, then make the infinite sum explicit as a limit of finite sums up to n, then use Problem 2 to telescope the KL divergences, and finally use the mixture representation of M.

Solution

For each t and $x_{<t}$, the conditional $\mu(\cdot \mid x_{<t})$ is a probability distribution over $\mathbb{B}$ (since μ is a measure), while $M(\cdot \mid x_{<t})$ is a semiprobability distribution over $\mathbb{B}$. Thus, by Problem 4:

$$\sum_{x_t \in \mathbb{B}} \big(M(x_t \mid x_{<t}) - \mu(x_t \mid x_{<t})\big)^2 \leq \sum_{x_t \in \mathbb{B}} \mu(x_t \mid x_{<t}) \ln \frac{\mu(x_t \mid x_{<t})}{M(x_t \mid x_{<t})}.$$

Thus:

$$\begin{aligned} S_\infty^\mu &\leq \sum_{t=1}^{\infty} \sum_{x_{<t} \in \mathbb{B}^*} \mu(x_{<t}) \sum_{x_t \in \mathbb{B}} \mu(x_t \mid x_{<t}) \ln \frac{\mu(x_t \mid x_{<t})}{M(x_t \mid x_{<t})} \\ &= \lim_{n \to \infty} \sum_{t=1}^{n} D_{\mathrm{KL}}\big(\mu(X_t \mid X_{<t}) \parallel M(X_t \mid X_{<t})\big), \end{aligned}$$

where the inner sum over $x_{<t}$ weighted by $\mu(x_{<t})$ gives exactly the conditional KL divergence. By Problem 2 (the telescoped chain rule), this equals

$$S_\infty^\mu \leq \lim_{n \to \infty} D_{\mathrm{KL}}\big(\mu(X_1, \ldots, X_n) \parallel M(X_1, \ldots, X_n)\big) = \lim_{n \to \infty} \sum_{x_{\leq n} \in \mathbb{B}^n} \mu(x_{\leq n}) \ln \frac{\mu(x_{\leq n})}{M(x_{\leq n})}.$$

Now, since $\mu \in \mathcal{M}_{sol}$, the mixture representation gives

$$M(x_{\leq n}) = \sum_{\nu \in \mathcal{M}_{sol}} P_{ap}(\nu) \nu(x_{\leq n}) \geq P_{ap}(\mu) \mu(x_{\leq n}).$$

Therefore $\frac{\mu(x_{\leq n})}{M(x_{\leq n})} \leq \frac{1}{P_{ap}(\mu)}$, and so

$$S_\infty^\mu \leq \lim_{n \to \infty} \sum_{x_{\leq n}} \mu(x_{\leq n}) \ln \frac{1}{P_{ap}(\mu)} = -\ln P_{ap}(\mu) \cdot \lim_{n \to \infty} \underbrace{\sum_{x_{\leq n}} \mu(x_{\leq n})}_{=1} = -\ln P_{ap}(\mu),$$

where $\sum_{x_{\leq n}} \mu(x_{\leq n}) = 1$ since μ is a measure.

Problem 5'

Under the same setup as Problem 5, show that there exists a constant $\kappa \in \mathbb{R}$ (independent of μ) such that

$$S_\infty^\mu \leq K(\mu) \ln 2 + \kappa,$$

where $K(\mu) := \min\{K(i) : \nu_i = \mu\}$ is the Kolmogorov complexity of μ (minimized over all indices i in the effective enumeration from the setup that give rise to μ).

Hint: Follow the same strategy as Problem 5, but instead of the mixture representation Equation (1), use the lower bound $M(x_{\leq n}) > c \cdot \sum_i P_{sol}(i)\, \nu_i(x_{\leq n})$ from Equation (2), together with $P_{sol}(i) = 2^{-K(i)}$.

Solution

The proof follows the same steps as Problem 5 up to the KL divergence expression:

$$S_\infty^\mu \leq \lim_{n\to\infty} \sum_{x_{\leq n}} \mu(x_{\leq n}) \ln \frac{\mu(x_{\leq n})}{M(x_{\leq n})}.$$

Now, by Equation (2), we have

$$M(x_{\leq n}) > c \cdot \sum_{i\in\mathbb{N}} P_{sol}(i)\, \nu_i(x_{\leq n}) \geq c \cdot P_{sol}(i^*)\, \mu(x_{\leq n}),$$

where i^* is any index with $\nu_{i^*} = \mu$. Therefore $\frac{\mu(x_{\leq n})}{M(x_{\leq n})} < \frac{1}{c \cdot P_{sol}(i^*)}$, and so

$$S_\infty^\mu \leq \lim_{n\to\infty} \sum_{x_{\leq n}} \mu(x_{\leq n}) \ln \frac{1}{c \cdot P_{sol}(i^*)} = -\ln c - \ln P_{sol}(i^*) = -\ln c + K(i^*) \ln 2,$$

using $P_{sol}(i^*) = 2^{-K(i^*)}$ and $\sum_{x_{\leq n}} \mu(x_{\leq n}) = 1$. Since this holds for every i^* with $\nu_{i^*} = \mu$, we may take the minimum over all such indices to obtain

$$S_\infty^\mu \leq K(\mu) \ln 2 + \kappa,$$

where $\kappa = -\ln c$.

Interpretation

Problems 5 and 5' show that the *cumulative* expected prediction error S_∞^μ is finite. Since S_∞^μ is a sum of non-negative terms, this immediately implies that the per-step expected prediction error

$$d_t := \sum_{x_{<t}\in\mathbb{B}^*} \mu(x_{<t}) \sum_{x_t\in\mathbb{B}} \big(M(x_t \mid x_{<t}) - \mu(x_t \mid x_{<t})\big)^2$$

converges to zero as $t \to \infty$. In other words, M's predictions become indistinguishable from μ's predictions on average over histories.