

AIXI

David Quarel

ARENA

October 15, 2026

Prerequisites

- [Solomonoff Induction](#) — **start there first.** The Bayesian mixture, posterior updates, and dominance carry over directly; AIXI reuses them in the action setting.
- Comfort with discrete probability (conditional distributions, chain rule, expectations).
- Familiarity with sequential decision-making / reinforcement learning (agents, rewards, discounting, value functions) is useful but not assumed.

What you'll learn

- Formalize the agent-environment interaction, value functions, and the Bayes-optimal policy that defines AIXI.
- Derive the explicit expectimax form of AIXI.
- Prove on-policy value convergence and that AIXI cannot be fooled in deterministic environments.
- Prove the self-optimizing property via likelihood-ratio martingales and change of measure.

Much of this material is drawn from [\[HQC24\]](#). The book provides fuller explanations, additional context, and proofs omitted here.

Difficulty ratings. Each subproblem is tagged with a rating in square brackets using Knuth's scale [\[Knu73\]](#): 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. See [Appendix B](#) for the full scale.

Problems marked (*) are less interesting/insightful, and while the result may be used later, I would recommend skipping them on a first pass.

Overview

AIXI is the *action* half of **Universal AI**: it lifts the Bayesian mixture from [Solomonoff induction](#) — learning to predict — to learning to act. Now the agent takes actions that shape what it observes, and AIXI is the **Bayes-optimal policy** over a universal mixture ξ of environments. It learns to act as well as if it knew the true environment μ . Under the universal choice — $\mathcal{M}$ = all lower-semicomputable environments, prior $w_\nu = 2^{-K(\nu)}$ — the assumption “ $\mu \in \mathcal{M}$ ” again becomes “the universe is computable.”

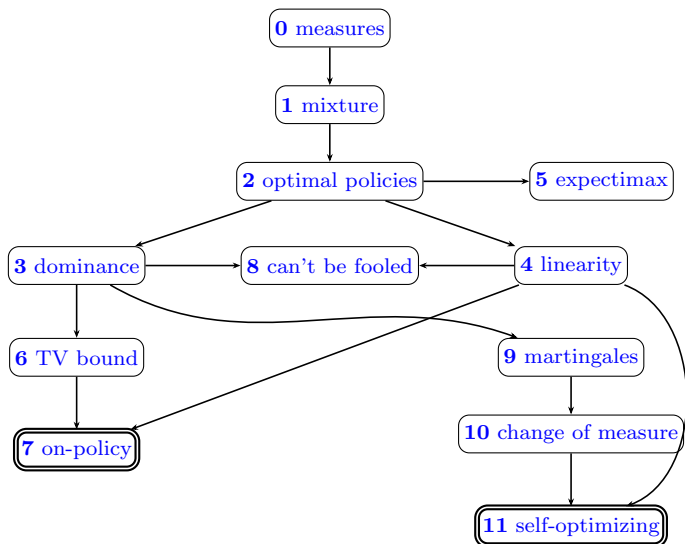
Goal and Roadmap

This exercise sheet builds towards three main results:

- **On-policy value convergence** (Section 7): the Bayesian mixture ξ learns to predict the value of any fixed policy as well as the true environment μ .
- **AIXI can’t be fooled** (Section 8): in deterministic environments, the Bayes-optimal agent is guaranteed non-zero value whenever optimal value is non-zero.
- **Self-optimizing property** (Section 11, advanced stretch goal): the Bayes-optimal policy π_ξ^* learns to *act* as well as if it knew μ , provided that any learnable policy can achieve this. This is the central theoretical justification for the AIXI agent.

A good target is to complete Sections 7 and 8. The self-optimizing property (Sections 9–11) requires substantial additional machinery (supermartingales, change of measure) and is an advanced stretch goal.

Critical path. The three results share a common foundation (Sections 0–4) and then diverge:



In summary:

- Sections 0–4 establish the Bayesian RL framework: measure algebra, mixture properties, existence of optimal policies, dominance, and linearity.
- Section 5 derives the explicit expectimax form of AIXI.
- Sections 6–7 prove *on-policy value convergence*: V_ξ^π and V_μ^π become indistinguishable for any fixed π .
- Section 8 shows *AIXI can't be fooled*: the Bayes-optimal agent achieves non-zero value whenever optimal value is non-zero (in deterministic environments).
- Sections 9–11 (advanced stretch goal) prove the *self-optimizing property*: if any policy can learn to act optimally, π_ξ^* inherits this.

Setup

An **agent** interacts with an **environment** in discrete time steps $t = 1, 2, \dots$

Definition 0.1 (Spaces and notation).

- $\mathcal{A}$: finite set of **actions**
- $\mathcal{O}$: finite set of **observations**
- $\mathcal{R} \subset [0, 1]$: finite set of **rewards**
- $\mathcal{E} := \mathcal{O} \times \mathcal{R}$: set of **percepts**; $e_t = (o_t, r_t) \equiv o_t r_t$

- $\mathcal{H}^t := (\mathcal{A} \times \mathcal{E})^t$: set of all histories of length t
- $\mathcal{H}^* := \cup_{t=0}^{\infty} \mathcal{H}^t$: set of all finite **histories**
- $\mathcal{H} \equiv \mathcal{H}^*$: set of all finite histories
- $\mathcal{H}^\infty := (\mathcal{A} \times \mathcal{E})^\infty$: set of all infinite histories
- ΔS : set of all probability distributions over set S
- $\llbracket \cdot \rrbracket$: **Iverson bracket**: $\llbracket P \rrbracket = 1$ if P is true, 0 if false
- t : current time step; m : finite horizon; $1 \leq t \leq m$
- i, j, k : arbitrary integer indices
- xy : concatenation
- ϵ : the empty string/history
- $\mathfrak{x}_{i:j} := a_i e_i a_{i+1} e_{i+1} \cdots a_j e_j$: history segment from time i to j
- $\mathfrak{x}_{<t} := a_1 e_1 a_2 e_2 \cdots a_{t-1} e_{t-1}$: history up to (but not including) time t

Definition 0.2 (Policy π). A **policy** $\pi : \mathcal{H} \rightarrow \Delta \mathcal{A}$ maps each history to a probability distribution over actions. Given history $\mathfrak{x}_{<t}$:

- $\pi(\cdot \mid \mathfrak{x}_{<t})$ is a distribution over $\mathcal{A}$,
- $\pi(a_t \mid \mathfrak{x}_{<t}) \in [0, 1]$ is the probability of choosing action a_t ,
- the agent samples $a_t \sim \pi(\cdot \mid \mathfrak{x}_{<t})$.

A policy is **deterministic** if $\pi(a \mid \mathfrak{x}_{<t}) \in \{0, 1\}$ for all $a, \mathfrak{x}_{<t}$. We write $\pi(\mathfrak{x}_{i:j}) := \prod_{k=i}^j \pi(a_k \mid \mathfrak{x}_{<k})$.

Definition 0.3 (Environment ν). An **environment** $\nu : \mathcal{H} \times \mathcal{A} \rightarrow \Delta \mathcal{E}$ maps each history–action pair to a distribution over percepts. Given history $\mathfrak{x}_{<t}$ and action a_t :

- $\nu(\cdot \mid \mathfrak{x}_{<t} a_t)$ is a distribution over $\mathcal{E}$,
- $\nu(e_t \mid \mathfrak{x}_{<t} a_t) \in [0, 1]$ is the probability of percept e_t ,
- the environment samples $e_t \sim \nu(\cdot \mid \mathfrak{x}_{<t} a_t)$.

We write $\nu(\mathfrak{x}_{i:j}) := \prod_{k=i}^j \nu(e_k \mid \mathfrak{x}_{<k} a_k)$. This satisfies the chain rule: $\nu(\mathfrak{x}_{i:j}) = \nu(\mathfrak{x}_{i:j-1}) \cdot \nu(e_j \mid \mathfrak{x}_{i:j-1} a_j)$ or in the form we will usually use, $\nu(\mathfrak{x}_{1:t}) = \nu(\mathfrak{x}_{<t}) \cdot \nu(e_t \mid \mathfrak{x}_{<t} a_t)$. An environment is **deterministic** if $\nu(e_t \mid \mathfrak{x}_{<t} a_t) \in \{0, 1\}$ for

all $e_t, \mathfrak{a}_{<t}, a_t$ (each percept is produced with certainty). We denote the true (unknown) environment by μ .

Definition 0.4 (Interaction measure ν^π). When policy π interacts with environment ν , the joint probability of a history segment $\mathfrak{a}_{i:j}$ given past $\mathfrak{a}_{<i}$ is

$$\nu^\pi(\mathfrak{a}_{i:j} \mid \mathfrak{a}_{<i}) := \prod_{k=i}^j \pi(a_k \mid \mathfrak{a}_{<k}) \nu(e_k \mid \mathfrak{a}_{<k} a_k).$$

0. Properties of Measures

Exercise 0.1 (Factorization). [05] Show that $\nu^\pi(\mathfrak{a}_{1:t}) = \pi(\mathfrak{a}_{1:t}) \cdot \nu(\mathfrak{a}_{1:t})$.

Exercise 0.2 (Chain rule). [05] Show that $\nu(\mathfrak{a}_{1:t}) = \nu(\mathfrak{a}_{<t}) \cdot \nu(e_t \mid \mathfrak{a}_{<t} a_t)$.

Exercise 0.3 (Marginalizing percepts). [03] (*) Show that $\sum_{e_t} \nu^\pi(a_t e_t \mid \mathfrak{a}_{<t}) = \pi(a_t \mid \mathfrak{a}_{<t})$.

Exercise 0.4 (Deterministic interaction measure). [05] (*) Let π be a deterministic policy, i.e. at each history $\mathfrak{a}_{<k}$ there is a unique action a_k^* with $\pi(a_k^* \mid \mathfrak{a}_{<k}) = 1$. Show that, provided $\mathfrak{a}_{i:j}$ is consistent with π (i.e. $a_k = a_k^*$ for every $k \in \{i, \dots, j\}$):

$$\nu^\pi(\mathfrak{a}_{i:j} \mid \mathfrak{a}_{<i}) = \nu(\mathfrak{a}_{i:j} \mid \mathfrak{a}_{<i}),$$

i.e. on π -consistent futures the interaction measure reduces to the environment measure: every policy factor is 1.

Exercise 0.5. [05] (General chain rule for ν^π) Show that for any $p \leq q \leq r$:

$$\nu^\pi(\mathfrak{a}_{p:r} \mid \mathfrak{a}_{<p}) = \nu^\pi(\mathfrak{a}_{p:q} \mid \mathfrak{a}_{<p}) \cdot \nu^\pi(\mathfrak{a}_{q+1:r} \mid \mathfrak{a}_{<p} \mathfrak{a}_{p:q}).$$

The Bayesian Mixture and Value Function

Definition 0.1 (Model class, prior, and Bayesian mixture ξ). Let $\mathcal{M} = \{\nu_1, \nu_2, \dots\}$ be a countable class of environments with prior weights $w_\nu > 0$ satisfying $\sum_{\nu \in \mathcal{M}} w_\nu \leq 1$. We assume $\mu \in \mathcal{M}$.

The **Bayesian mixture** ξ is defined as

$$\xi(\mathbf{a}_{1:t}) := \sum_{\nu \in \mathcal{M}} w_\nu \nu(\mathbf{a}_{1:t}) \quad \text{and} \quad \xi(e_t \mid \mathbf{a}_{<t} a_t) := \xi(\mathbf{a}_{1:t}) / \xi(\mathbf{a}_{<t}).$$

One can show that the one-step predictive distribution can be written as

$$\xi(e_t \mid \mathbf{a}_{<t} a_t) = \sum_{\nu \in \mathcal{M}} w(\nu \mid \mathbf{a}_{<t}) \nu(e_t \mid \mathbf{a}_{<t} a_t),$$

where the **posterior weight** is

$$w(\nu \mid \mathbf{a}_{<t}) := w_\nu \frac{\nu(\mathbf{a}_{<t})}{\xi(\mathbf{a}_{<t})}, \quad w(\nu \mid \epsilon) := w_\nu.$$

See Appendix A for a worked example.

Definition 0.2 (Expectation $\mathbb{E}_\nu^\pi$). For a function $f : \mathcal{H} \rightarrow \mathbb{R}$ of a finite future segment $\mathbf{a}'_{t:m}$:

$$\mathbb{E}_\nu^\pi[f \mid \mathbf{a}_{<t}] := \mathbb{E}_{\mathbf{a}'_{t:m} \sim \nu^\pi(\cdot \mid \mathbf{a}_{<t})} [f] = \sum_{\mathbf{a}'_{t:m}} \nu^\pi(\mathbf{a}'_{t:m} \mid \mathbf{a}_{<t}) f(\mathbf{a}'_{t:m}).$$

For functions f of the infinite future (like the value function), with a sequence $f_1, f_2, \dots \rightarrow f$, where f_m depends only on $\mathbf{a}'_{t:m}$, we define $\mathbb{E}_\nu^\pi[f \mid \mathbf{a}_{<t}] := \lim_{m \rightarrow \infty} \mathbb{E}_\nu^\pi[f_m \mid \mathbf{a}_{<t}]$.

Definition 0.3 (Discounted return $G_{t:m}$). Fix a discount factor $\gamma \in (0, 1)$, held constant throughout. The **discounted return** at time step t , up to horizon m , is

$$G_{t:m}^\gamma := \sum_{k=t+1}^m \gamma^{k-(t+1)} r_k = r_{t+1} + \gamma r_{t+2} + \dots + \gamma^{m-t} r_m.$$

Since γ is fixed we drop it and write $G_{t:m} \equiv G_{t:m}^\gamma$. It satisfies the recursion $G_{t:m} = r_{t+1} + \gamma G_{t+1:m}$, with $G_{m-1:m} = r_m$ and $G_{n:m} = 0$ for $n \geq m$. The

infinite-horizon return is the pointwise limit

$$G_{\geq t} \equiv G_{t:\infty} := \lim_{m \rightarrow \infty} G_{t:m} = \sum_{k=t+1}^{\infty} \gamma^{k-(t+1)} r_k,$$

with the clean recursion $G_{\geq t} = r_{t+1} + \gamma G_{\geq t+1}$ and no boundary cases.

Definition 0.4 (Value function^a). The **value** of policy π in environment ν with horizon $m \geq t$ given history $\mathfrak{x}_{<t}$ is

$$\begin{aligned} V_{\nu}^{\pi,m}(\mathfrak{x}_{<t}) &:= (1 - \gamma) \mathbb{E}_{\nu}^{\pi}[G_{t-1:m} \mid \mathfrak{x}_{<t}] \\ &= (1 - \gamma) \sum_{\mathfrak{x}_{t:m}} \nu^{\pi}(\mathfrak{x}_{t:m} \mid \mathfrak{x}_{<t}) G_{t-1:m} \\ &= (1 - \gamma) \sum_{\mathfrak{x}_{t:m}} \nu^{\pi}(\mathfrak{x}_{t:m} \mid \mathfrak{x}_{<t}) \left[\sum_{k=t}^m \gamma^{k-t} r_k \right]. \end{aligned}$$

The **infinite-horizon value** is the pointwise limit $V_{\nu}^{\pi,\infty}(\mathfrak{x}_{<t}) := \lim_{m \rightarrow \infty} V_{\nu}^{\pi,m}(\mathfrak{x}_{<t}) = (1 - \gamma) \mathbb{E}_{\nu}^{\pi}[G_{\geq t-1} \mid \mathfrak{x}_{<t}]$. We write $V_{\nu}^{\pi} \equiv V_{\nu}^{\pi,\infty}$ for short. The **optimal value** is $V_{\nu}^{*,m}(\mathfrak{x}_{<t}) := \sup_{\pi} V_{\nu}^{\pi,m}(\mathfrak{x}_{<t})$ for $m \in \mathbb{N} \cup \{\infty\}$, with $V_{\nu}^* \equiv V_{\nu}^{*,\infty}$.^b An **optimal policy** π_{ν}^* satisfies $V_{\nu}^{\pi_{\nu}^*} = V_{\nu}^*$. The **Bayes-optimal policy** is $\pi_{\xi}^* \in \arg \max_{\pi} V_{\xi}^{\pi}$.

^aSee Appendix A for a worked example. For simplicity, we consider only geometric discounting.

^bThe optimal value is defined as a sup over policies. In general, a supremum need not be attained (e.g. $\sup_{x \in (0,1)} x = 1$ but no $x \in (0,1)$ achieves it). Section 2 shows that the sup is attained in our setting.

Remark (AIXI)

All results in this sheet hold for any countable $\mathcal{M}$ and prior weights w_{ν} . **AIXI** is a special case of the Bayes-optimal policy π_{ξ}^* for the particular choice $\mathcal{M} :=$ the class of all lower-semicomputable chronological semimeasures, and 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 resulting agent is written $\pi_{\xi_U}^*$, or simply AIXI .

Why this $\mathcal{M}$? By including every computable environment, the assumption $\mu \in \mathcal{M}$ reduces to “the universe is computable”: as weak an assumption as one can make.

Why this prior? The prior $2^{-K(\nu)}$ is *dominant*: for any other computable prior w'_{ν} , there exists a constant $c > 0$ such that $2^{-K(\nu)} \geq c \cdot w'_{\nu}$ for all ν .

Caveat: The constant c depends on the choice of universal Turing machine U , and adversarial choices of U can make AIXI behave arbitrarily badly [LH15].

See [HQC24]: Chapter 2.7 for Kolmogorov complexity, Chapters 3.7–3.8 for the model class and universal prior, and Chapter 7.4 for AIXI itself.

1. Properties of the Bayesian Mixture

Exercise 1.1 (Posterior update). [10] (*) Show that the posterior updates multiplicatively:

$$w(\nu \mid \mathfrak{a}_{1:t}) = w(\nu \mid \mathfrak{a}_{<t}) \frac{\nu(e_t \mid \mathfrak{a}_{<t} a_t)}{\xi(e_t \mid \mathfrak{a}_{<t} a_t)}.$$

Hint: Use the definitions of $w(\nu \mid \mathfrak{a}_{1:t})$ and $w(\nu \mid \mathfrak{a}_{<t})$, and apply Exercise 0.2.

Exercise 1.2. [15] (One-step predictive distribution of ξ) Starting from $\xi(\mathfrak{a}_{1:t}) = \sum_{\nu \in \mathcal{M}} w_\nu \nu(\mathfrak{a}_{1:t})$, derive the one-step predictive form:

$$\xi(e_t \mid \mathfrak{a}_{<t} a_t) = \sum_{\nu \in \mathcal{M}} w(\nu \mid \mathfrak{a}_{<t}) \nu(e_t \mid \mathfrak{a}_{<t} a_t).$$

Hint: Write $\xi(e_t \mid \mathfrak{a}_{<t} a_t) = \xi(\mathfrak{a}_{1:t}) / \xi(\mathfrak{a}_{<t})$, expand the numerator, and use Exercise 0.2.

Exercise 1.3 (Bounded value). [10] Show that $V_\nu^{\pi, m}(\mathfrak{a}_{<t}) \in [0, 1]$ for any $\nu, \pi, m \in \mathbb{N} \cup \{\infty\}, \mathfrak{a}_{<t}$.

Hint: Recall the formula for geometric series: $(1 - \gamma) \sum_{k=0}^n \gamma^k = 1 - \gamma^{n+1}$.

Exercise 1.4. [15] (*) (Multi-step posterior linearity of ξ^π) Extend Exercise 1.2 to multi-step histories: show that for finite $m \geq t$,

$$\xi^\pi(\mathfrak{a}_{t:m} \mid \mathfrak{a}_{<t}) = \sum_{\nu \in \mathcal{M}} w(\nu \mid \mathfrak{a}_{<t}) \nu^\pi(\mathfrak{a}_{t:m} \mid \mathfrak{a}_{<t}).$$

Hint: Apply the general chain rule (Exercise 0.5) to $\xi(\mathfrak{a}_{1:m}) = \sum_\nu w_\nu \nu(\mathfrak{a}_{1:m})$, then use factorization (Exercise 0.1).

2. Existence of Optimal Policies

Recall that $V_\nu^*(\mathfrak{x}_{<t}) := \sup_\pi V_\nu^\pi(\mathfrak{x}_{<t})$ is defined as a supremum over all policies. In general, a supremum need not be achieved: for example, $\sup_{x \in (0,1)} x = 1$, but no $x \in (0,1)$ attains this value. This problem shows that in our setting, the supremum *is* attained, so an optimal policy π_ν^* with $V_\nu^{\pi_\nu^*} = V_\nu^*$ exists.

Exercise 2.1. [05] Consider a single time step. Given history $\mathfrak{x}_{<t}$ and a function $Q : \mathcal{A} \rightarrow \mathbb{R}$, show that $\sup_\pi \sum_{a \in \mathcal{A}} \pi(a \mid \mathfrak{x}_{<t}) Q(a) = \max_{a \in \mathcal{A}} Q(a)$, and that the supremum is attained by the deterministic policy that places all probability on an action achieving the maximum.

Exercise 2.2 (Bellman equation). [15] Show that for finite m and $t \leq m$:

$$V_\nu^{\pi,m}(\mathfrak{x}_{<t}) = \sum_{\mathfrak{x}'_t} \nu^\pi(\mathfrak{x}'_t \mid \mathfrak{x}_{<t}) \left[(1 - \gamma) r'_t + \gamma V_\nu^{\pi,m}(\mathfrak{x}_{<t} \mathfrak{x}'_t) \right],$$

Hint: Use the general chain rule (Exercise 0.5) to peel off the *first* step,

$$\nu^\pi(\mathfrak{x}_{t:m} \mid \mathfrak{x}_{<t}) = \nu^\pi(\mathfrak{x}_t \mid \mathfrak{x}_{<t}) \cdot \nu^\pi(\mathfrak{x}_{t+1:m} \mid \mathfrak{x}_{1:t}).$$

Break up the sum $\sum_{\mathfrak{x}_{t:m}} = \sum_{\mathfrak{x}_t} \sum_{\mathfrak{x}_{t+1:m}}$, and factor out r'_t using $\sum_{\mathfrak{x}'_{t+1:m}} \nu^\pi(\mathfrak{x}'_{t+1:m} \mid \mathfrak{x}) = 1$ for any history $\mathfrak{x}$.

Exercise 2.3 (Backward induction). [20] (*) Using the Bellman equation from Exercise 2.2 and Exercise 2.1, show by backward induction on $t = m, m-1, \dots, 1$ that for each finite m , a deterministic optimal policy exists.

Exercise 2.4 (Bellman optimality equation). [10] Using Exercise 2.3, show that for finite m :

$$V_\nu^{*,m}(\mathfrak{x}_{<t}) = \max_{a'_t} \sum_{e'_t} \nu(e'_t \mid \mathfrak{x}_{<t} a'_t) \left[(1 - \gamma) r'_t + \gamma V_\nu^{*,m}(\mathfrak{x}_{<t} \mathfrak{x}'_t) \right].$$

where $e'_t = o'_t r'_t$ and $\mathfrak{x}'_t = a'_t e'_t$.

Exercise 2.5. [15] (*) (Existence of V_ν^*) Show that the pointwise limit $V_\nu^*(\mathfrak{x}_{<t}) := \lim_{m \rightarrow \infty} V_\nu^{*,m}(\mathfrak{x}_{<t})$ exists for all histories $\mathfrak{x}_{<t}$.

Hint: Show $V_\nu^{*,m+1}(\mathbf{x}_{<t}) \geq V_\nu^{*,m}(\mathbf{x}_{<t})$ by expanding $V_\nu^{*,m+1}$. Use this and Exercise 1.3 and Monotone Convergence Theorem to obtain the result.

Exercise 2.6 (Infinite-horizon Bellman optimality equation). [10] Show that the Bellman optimality equation (Exercise 2.4) extends to $m = \infty$:

$$V_\nu^*(\mathbf{x}_{<t}) = \max_{a'_t} \sum_{e'_t} \nu(e'_t | \mathbf{x}_{<t} a'_t) \left[(1 - \gamma) r'_t + \gamma V_\nu^*(\mathbf{x}_{<t} a'_t e'_t) \right].$$

Exercise 2.7 (Optimal policy for infinite horizon). [25] (*) Show that a deterministic policy π_ν^* exists achieving $V_\nu^{\pi_\nu^*}(\mathbf{x}_{<t}) = V_\nu^*(\mathbf{x}_{<t})$ for all $\mathbf{x}_{<t}$.
Approach:

- (a) Define π_ν^* as the greedy policy (the arg max action from Exercise 2.6).
- (b) Write two Bellman equations: one for V_ν^* , one for $V_\nu^{\pi_\nu^*,m}$.
- (c) Define $\Delta V_\nu^m := V_\nu^* - V_\nu^{\pi_\nu^*,m}$ and obtain a recursive equation for it.
- (d) Iterate out to the finite horizon, noting $V_\nu^{\pi_\nu^*,m}(\mathbf{x}_{1:m}) = 0$ (why?)
- (e) Bound the tail, take $m \rightarrow \infty$.

3. Dominance of the Bayesian Mixture

Exercise 3.1. [03] Show that $\xi(\mathbf{x}_{<t}) \geq w_\nu \cdot \nu(\mathbf{x}_{<t})$ for every $\nu \in \mathcal{M}$.

Exercise 3.2. [10] Conclude that $\xi^\pi(\mathbf{x}_{1:m}) \geq w_\nu \cdot \nu^\pi(\mathbf{x}_{1:m})$ for any $\nu \in \mathcal{M}$ and any policy π .

4. Properties of V_ξ

Exercise 4.1. [15] Show that for finite m :

$$V_\xi^{\pi,m}(\mathbf{x}_{<t}) = \sum_{\nu \in \mathcal{M}} w(\nu | \mathbf{x}_{<t}) V_\nu^{\pi,m}(\mathbf{x}_{<t}).$$

Hint: Use the multi-step posterior linearity of ξ^π (Exercise 1.4).

Fact 4.1 (Dominated convergence for sums). *If $a_{\nu,m} \rightarrow a_\nu$ as $m \rightarrow \infty$ for each ν , and $|a_{\nu,m}| \leq b_\nu$ for all m with $\sum_\nu b_\nu < \infty$, then $\sum_\nu a_{\nu,m} \rightarrow \sum_\nu a_\nu$. (For finite sums this is trivial.)*

Exercise 4.2 (Infinite-horizon linearity). [17] (*) Show that the result extends to $m = \infty$:

$$V_\xi^\pi(\mathfrak{a}_{<t}) = \sum_{\nu \in \mathcal{M}} w(\nu \mid \mathfrak{a}_{<t}) V_\nu^\pi(\mathfrak{a}_{<t}).$$

Hint: Take $m \rightarrow \infty$ in the finite-horizon result. You will need Fact 4.1 to exchange limit and sum for countably infinite $\mathcal{M}$.

Exercise 4.1 shows V_ξ^π is linear in ν . The optimal value V_ξ^* is only convex: a single policy must perform well across all $\nu \in \mathcal{M}$ simultaneously, rather than being tailored to each ν individually.

Exercise 4.3. [10] (*) (Convexity of V_ξ^*) Using Exercise 4.1, show that for finite m :

$$V_\xi^{*,m}(\mathfrak{a}_{<t}) \leq \sum_{\nu \in \mathcal{M}} w(\nu \mid \mathfrak{a}_{<t}) V_\nu^{*,m}(\mathfrak{a}_{<t}).$$

Hint: Apply Exercise 4.1 with $\pi = \pi_\xi^*$, the Bayes-optimal policy for ξ .

Exercise 4.4. [15] (*) (Non-linearity of V_ξ^*) Show by example that the inequality in Exercise 4.3 can be strict, i.e. $V_\xi^{*,m}$ is *not* linear in ν .

Hint: Consider $\mathcal{M} = \{\nu_0, \nu_1\}$: predicting a two-headed coin vs. a two-tailed coin.

5. The Expectimax Form of AIXI

We have specified AIXI only implicitly, as the Bayes-optimal policy π_ξ^* (Definition 0.4). We now unroll this definition into the explicit **expectimax** expression: an alternating sequence of maximizations over actions and ξ -expectations over percepts.

Exercise 5.1 (Expectimax form of AIXI). [20] By iterating the finite-horizon Bellman optimality equation (Exercise 2.4), show that

$$\begin{aligned} V_{\xi}^{*,m}(\mathfrak{a}_{<t}) &= (1 - \gamma) \max_{a_t} \sum_{e_t} \xi(e_t \mid \mathfrak{a}_{<t} a_t) \\ &\quad \max_{a_{t+1}} \sum_{e_{t+1}} \xi(e_{t+1} \mid \mathfrak{a}_{<t+1} a_{t+1}) \cdots \\ &\quad \cdots \max_{a_m} \sum_{e_m} \xi(e_m \mid \mathfrak{a}_{<m} a_m) G_{t-1:m} \end{aligned}$$

and hence, collecting the percept factors with the chain rule (Exercise 0.2) and taking $m \rightarrow \infty$ (Exercises 2.5 and 2.7), that AIXI selects the action

$$\begin{aligned} a_t &= \pi_{\xi}^*(\mathfrak{a}_{<t}) \\ &\in \arg \max_{a_t} \lim_{m \rightarrow \infty} \sum_{e_t} \max_{a_{t+1}} \sum_{e_{t+1}} \cdots \max_{a_m} \sum_{e_m} \xi(e_{t:m} \mid \mathfrak{a}_{<t} a_{t:m}) G_{t-1:m} \end{aligned}$$

Hint: Prove the first display by backward induction on t from m down to 1, using the finite-horizon Bellman optimality equation (Exercise 2.4) and the return recursion $G_{t-1:m} = r_t + \gamma G_{t:m}$ (Definition 0.3). At each step the leftover reward term $(1 - \gamma) r_t$ is constant in the deeper variables; carry it inward using $\sum_{e_k} \xi(e_k \mid \mathfrak{a}_{<k} a_k) = 1$ and $c + \max(\cdot) = \max(c + \cdot)$.

On-Policy Value Convergence

The following problems prove the first two main results: on-policy value convergence (Section 7), and that AIXI can't be fooled in deterministic environments (Section 8). The path to the self-optimizing property (advanced stretch goal) resumes at Section 9.

6. Bounding Expectation Differences by Total Variation

Definition 6.1 (Total variation distance). The **total variation distance** between probability measures P and Q on a countable set Ω is $\text{TV}_{\Omega}(P, Q) := \sup_{S \subseteq \Omega} |P(S) - Q(S)|$, where $P(S) := \sum_{\omega \in S} P(\omega)$.

Definition 6.2 (Expectation under P). For a probability measure P on a countable set Ω and a function $f : \Omega \rightarrow \mathbb{R}$, the **expectation** of f under P is $\mathbb{E}_P[f] := \sum_{\omega \in \Omega} f(\omega) P(\omega)$.

The following technical lemma is needed for the on-policy value convergence proof.

Exercise 6.1. [15] (*) Let $f : \Omega \rightarrow [0, c]$. Show that $|\mathbb{E}_P[f] - \mathbb{E}_Q[f]| \leq c \cdot \text{TV}_\Omega(P, Q)$.

Hint: Define $A^+ = \{\omega \in \Omega : P(\omega) \geq Q(\omega)\}$ and decompose the expectation difference as a sum over A^+ and its complement $A^- = \Omega \setminus A^+$.

7. On-Policy Value Convergence of Bayes

The following definitions are needed for the on-policy value convergence theorem.

Definition 7.1 (Probability of events). A **finite-length event** is a set $A \subseteq (\mathcal{A} \times \mathcal{E})^t$ of histories of fixed length t . Its probability under ν^π is

$$\nu^\pi(A) := \sum_{\mathbf{a}_{1:t} \in A} \nu^\pi(\mathbf{a}_{1:t}).$$

An **event on infinite histories** is any set built from finite-length events by countable unions, intersections, and complements.^a Probabilities of such events are uniquely determined by two properties:

- **Normalization:** $\nu^\pi((\mathcal{A} \times \mathcal{E})^\infty) = 1$.
- **Countable additivity:** if $A_1, A_2, \dots$ are pairwise disjoint events, then $\nu^\pi(\bigcup_{n=1}^\infty A_n) = \sum_{n=1}^\infty \nu^\pi(A_n)$.

From these, all standard rules of probability can be derived (e.g. $\nu^\pi(A \cup B) = \nu^\pi(A) + \nu^\pi(B) - \nu^\pi(A \cap B)$, etc.), though we will not prove them here. Two derived properties we use explicitly:

- **Complement:** $\nu^\pi(A^c) = 1 - \nu^\pi(A)$.
- **Monotone limits:** if $A_1 \supseteq A_2 \supseteq \dots$, then $\nu^\pi(\bigcap_{n=1}^\infty A_n) = \lim_{n \rightarrow \infty} \nu^\pi(A_n)$.

Further useful notions:

- **Conditional probability:** The probability of event A given observed history $\mathfrak{x}_{<t}$ is

$$\nu^\pi(A \mid \mathfrak{x}_{<t}) := \frac{\nu^\pi(\{\mathfrak{x}_{<t}h : h \in A\})}{\nu^\pi(\mathfrak{x}_{<t})}.$$

^aWe are deliberately avoiding a formal treatment of measure theory here. Not all subsets of $(\mathcal{A} \times \mathcal{E})^\infty$ are measurable; we restrict to “nice” (measurable) sets built from finite-prefix conditions via countable set operations, which suffice for everything in this sheet. For a rigorous treatment using σ -algebras and probability measures, see [HQC24, Chapter 2.2].

Example 7.2 (A fair coin shows 1 infinitely often). Let $\mathcal{A} = \mathcal{O} = \{0, 1\}$, $\mathcal{R} = \{0\}$, and let ν be a fair coin that ignores the action: $\nu(o_t = 1 \mid \mathfrak{x}_{<t} a_t) = \frac{1}{2}$ for all $\mathfrak{x}_{<t}, a_t$. Consider the event $F := “o_t = 1 \text{ for infinitely many } t”$.

Step 1: Build F^c from finite-prefix events. For each t , the set $\{o_t = 0\} := \{\mathfrak{x}_{1:t} : o_t = 0\}$ is a finite-length event. For each N and $T \geq N$, the set $D_{N,T} := \bigcap_{t=N}^T \{o_t = 0\}$ is also a finite-length event (determined by time T). Define $C_N := \bigcap_{t=N}^\infty \{o_t = 0\}$ (“all zeros from time N onwards”): this is an infinite-history event, built as a countable intersection of finite-length events. Then $F^c = \bigcup_{N=1}^\infty C_N$ (“eventually all zeros”).

Step 2: Compute $\nu^\pi(C_N) = 0$. Since $D_{N,T} \supseteq D_{N,T+1} \supseteq \dots$ (adding more constraints), the monotone limit property (for decreasing sets) gives:

$$\nu^\pi(C_N) = \lim_{T \rightarrow \infty} \nu^\pi(D_{N,T}) = \lim_{T \rightarrow \infty} \left(\frac{1}{2}\right)^{T-N+1} = 0.$$

(The probability $\nu^\pi(D_{N,T}) = (1/2)^{T-N+1}$ is independent of π , since ν ignores actions.)

Step 3: By countable additivity: $\nu^\pi(F^c) \leq \sum_{N=1}^\infty \nu^\pi(C_N) = 0$, so $\nu^\pi(F) = 1$. A fair coin shows 1 infinitely often, almost surely.

As such, F^c (the set of all infinite histories containing only finitely many 1s) is a ν^π -measure-zero set. Such histories “exist” as infinite sequences, but occur with probability zero.

Definition 7.3 (Covering¹). We say Q **covers** P if Q is positive everywhere P is: $P(A) > 0 \implies Q(A) > 0$ for all events A . Equivalently: any event that Q rules out, P also rules out. By Section 3, ξ^π covers μ^π .

¹The standard name is *absolute continuity* of P with respect to Q , written $P \ll Q$.

Definition 7.4 (Convergence ν^π -almost surely). Let $f_t : \mathcal{H}^{t-1} \rightarrow \mathbb{R}$ be a sequence of functions, where f_t depends on the history $\mathfrak{x}_{<t}$ of length $t - 1$. We write $f_t(\mathfrak{x}_{<t}) \rightarrow 0$ ν^π -almost surely (ν^π -a.s.) if

$$\nu^\pi\left(\left\{\mathfrak{x}_{1:\infty} : f_t(\mathfrak{x}_{<t}) \not\rightarrow 0\right\}\right) = 0.$$

This set is built from finite-prefix conditions: $\{f_t \not\rightarrow 0\} = \bigcup_{n=1}^{\infty} \bigcap_{N=1}^{\infty} \bigcup_{t=N}^{\infty} \{|f_t(\mathfrak{x}_{<t})| > \frac{1}{n}\}$, so its probability is well-defined (Definition 7.1).

Example 7.5 (Unpacking “ $f_t \not\rightarrow 0$ ”). The set $\{f_t \not\rightarrow 0\}$ looks intimidating, but it reads naturally from the inside out:

- $\{|f_t(\mathfrak{x}_{<t})| > \frac{1}{n}\}$ is a finite-length event: “at time t , the function is at least $\frac{1}{n}$ away from zero.”
- $\bigcup_{t=N}^{\infty} \{|f_t| > \frac{1}{n}\}$: “at *some* time $t \geq N$, f_t is at least $\frac{1}{n}$ away from zero.”
- $\bigcap_{N=1}^{\infty} \bigcup_{t=N}^{\infty} \{|f_t| > \frac{1}{n}\}$: “for *every* N , there is some $t \geq N$ where f_t is at least $\frac{1}{n}$ from zero”, i.e., f_t exceeds $\frac{1}{n}$ *infinitely often*.
- $\bigcup_{n=1}^{\infty} \bigcap_{N=1}^{\infty} \bigcup_{t=N}^{\infty} \{|f_t| > \frac{1}{n}\}$: “for *some* $\frac{1}{n} > 0$, f_t exceeds $\frac{1}{n}$ infinitely often.”

This last condition is exactly $\{f_t \not\rightarrow 0\}$: convergence $f_t \rightarrow 0$ means that for *every* $\varepsilon > 0$, $|f_t| \leq \varepsilon$ for all sufficiently large t . Its negation is that *some* $\varepsilon > 0$ is exceeded infinitely often.

Each layer is a countable union or intersection of the previous, so the whole set is a well-defined event on infinite histories.

The Bayesian agent uses the mixture ξ because the true environment μ is unknown. A natural question: does planning with ξ eventually become as good as planning with μ ? The following theorem says *yes*: the value of any fixed policy π , as evaluated by ξ , converges to the value under the true environment μ , along histories that μ^π actually generates.² This tells us that the Bayesian mixture “learns” to predict the true environment’s value, on policy.

Theorem 7.6 (On-policy value convergence; [HQC24, Theorem 7.3.1]). *For any $\mu \in \mathcal{M}$ and any policy π : $V_\xi^\pi(\mathfrak{x}_{<t}) - V_\mu^\pi(\mathfrak{x}_{<t}) \rightarrow 0$ as $t \rightarrow \infty$, μ^π -almost surely. That is, the set of infinite histories along which the value difference does not*

²If the true environment is μ then we don’t care about the behaviour of the Bayesian agent on histories that have μ^π -probability zero: such histories will never be observed anyway.

vanish has μ^π -probability zero:

$$\mu^\pi\left(\left\{\mathfrak{x}_{1:\infty} : V_\xi^\pi(\mathfrak{x}_{<t}) - V_\mu^\pi(\mathfrak{x}_{<t}) \not\rightarrow 0\right\}\right) = 0.$$

The proof reduces to a finite-horizon TV bound (Exercise 7.1), a covering argument (Exercise 7.2), and the Blackwell–Dubins theorem (Exercise 7.3). The only ingredient we do not prove is Blackwell–Dubins itself, which we state as a given fact.

Exercise 7.1 (Finite-horizon TV bound on value difference). [10] Using Section 6, show that for finite m :

$$|V_\xi^{\pi,m}(\mathfrak{x}_{<t}) - V_\mu^{\pi,m}(\mathfrak{x}_{<t})| \leq \text{TV}_{\mathcal{H}^{m-t+1}}(\xi^\pi(\cdot | \mathfrak{x}_{<t}), \mu^\pi(\cdot | \mathfrak{x}_{<t})),$$

where $\mathcal{H}^{m-t+1} := (\mathcal{A} \times \mathcal{E})^{m-t+1}$ is the set of all future history segments $\mathfrak{x}_{t:m}$.

Exercise 7.2 (Covering). [05] Show that ξ^π covers μ^π (Definition 7.3).

Hint: Use Section 3.

To complete the proof, we need the TV distance on finite segments to vanish μ^π -a.s. as $t \rightarrow \infty$. This follows from the following deep result, which we state without proof:

Fact 7.7 (Blackwell–Dubins merging of opinions [BD62]). *If Q covers P , then $\sup_S |P(S | \mathfrak{x}_{<t}) - Q(S | \mathfrak{x}_{<t})| \rightarrow 0$ P -almost surely, where the supremum ranges over all measurable events S (including events on infinite histories).*

The proof of Blackwell–Dubins requires the Radon–Nikodym derivative and Lévy’s martingale convergence theorem: tools from measure theory that are beyond the scope of this sheet. See [HQC24, Chapter 3.9] for discussion.

Exercise 7.3 (On-policy convergence). [15] Using Fact 7.7, Exercise 7.1 and Exercise 7.2, prove Theorem 7.6.

Exercise 7.4. [40] (*) Prove Fact 7.7.

Hint: See [BD62]. The proof uses the Radon–Nikodym derivative dP/dQ , Lévy’s martingale convergence theorem, and the Lebesgue decomposition. No elementary proof is known that avoids graduate-level measure theory.

8. AIXI Cannot Be Fooled in Deterministic Environments

Theorem 8.1 (AIXI cannot act poorly in good environments). *If $\mu \in \mathcal{M}$ is deterministic and $V_\mu^*(\mathfrak{a}_{<t}) > \varepsilon > 0$ for all t along the history generated by μ and π_ξ^* , then $V_\xi^*(\mathfrak{a}_{<t}) \geq w_\mu \varepsilon > 0$ for all t . [HQC24, Theorem 7.4.10]*

Exercise 8.1. [10] Show that $V_\xi^*(\mathfrak{a}_{<t}) \geq w(\mu \mid \mathfrak{a}_{<t}) \cdot V_\mu^*(\mathfrak{a}_{<t})$.

Hint: Use linearity of V^π (Exercise 4.2).

Exercise 8.2. [15] Show $w(\mu \mid \mathfrak{a}_{<t}) \geq w_\mu$ when μ is deterministic, and combine with Exercise 8.1 to prove Theorem 8.1.

Remark

With the Solomonoff prior, $w_\mu = 2^{-K(\mu)}$ may be astronomically small. The result guarantees non-zero value, not near-optimal value.

Sections 9–11: Self-Optimizing Policy (Finite Model Class)

We now prove: if *any* policy can learn to act optimally, π_ξ^* also learns. We restrict to finite $\mathcal{M} = \{\nu_1, \dots, \nu_N\}$ where each ν_i is a proper probability measure.

9. Likelihood Ratios Are Martingales

The following definition and theorem state the main goal of Sections 9–11.

Definition 9.1 (Self-optimizing). Fix a historic policy π . A policy $\tilde{\pi}$ is **self-optimizing** for $\mathcal{M}$ with respect to π if for every $\nu \in \mathcal{M}$: $V_\nu^*(\mathfrak{a}_{<t}) - V_\nu^{\tilde{\pi}}(\mathfrak{a}_{<t}) \rightarrow 0$ as $t \rightarrow \infty$, ν^π -almost surely.

The percepts $e_{1:\infty}$ are sampled from ν ; the historic actions $a_{<t}$ come from π , which may differ from $\tilde{\pi}$.

Theorem 9.2 (Self-optimizing; [HQC24, Theorem 7.5.2]). *Let $\mathcal{M}$ be finite and fix a historic policy π . If there exists a $\tilde{\pi}$ self-optimizing for $\mathcal{M}$ with respect to π , then π_ξ^* is also self-optimizing for $\mathcal{M}$ with respect to π : for any $\mu \in \mathcal{M}$,*

$$V_{\mu}^*(\mathfrak{a}_{<t}) - V_{\mu}^{\pi^*}(\mathfrak{a}_{<t}) \rightarrow 0 \text{ } \mu^{\pi}\text{-almost surely.}$$

Definition 9.3 (Supermartingale). A sequence of functions $X_t : \mathcal{H}^t \rightarrow \mathbb{R}_{\geq 0}$ (each depending on the history $\mathfrak{a}_{1:t}$) is a μ^{π} -**supermartingale** if $\mathbb{E}_{\mu^{\pi}}[X_t \mid \mathfrak{a}_{<t}] \leq X_{t-1}$ for all t ; that is,

$$\sum_{\mathfrak{a}_t} \pi(a_t \mid \mathfrak{a}_{<t}) \mu(e_t \mid \mathfrak{a}_{<t} a_t) X_t(\mathfrak{a}_{1:t}) \leq X_{t-1}(\mathfrak{a}_{<t}).$$

If equality holds, it is a μ^{π} -**martingale**.

Intuition. A supermartingale is a quantity whose expected future value, given the present, is no larger than its current value: it cannot “drift upward on average.” A martingale is the equality case: expected future value equals current value. The key example is the **likelihood ratio** $X_{\nu,t}(\mathfrak{a}_{1:t}) := \nu(\mathfrak{a}_{1:t})/\mu(\mathfrak{a}_{1:t})$: how much more (or less) likely the observed history is under ν than under the true environment μ . Under μ , this ratio is a martingale: the true environment does not, on average, favour any alternative ν over itself. Non-negativity plus the (super)martingale structure forces $X_{\nu,t}$ to converge (by Doob’s theorem below), which is the engine behind the self-optimizing proof: it lets us separate environments that remain plausible ($X_{\nu,\infty} > 0$) from those that are eventually ruled out ($X_{\nu,\infty} = 0$), and handle each case differently in Section 11.

Fact 9.4 (Doob’s supermartingale convergence). *If $(X_t)_{t \geq 0}$ is a non-negative μ^{π} -supermartingale, then there exists a function $X_{\infty} : \mathcal{H}^{\infty} \rightarrow \mathbb{R}_{\geq 0}$ of the infinite history such that $X_t(\mathfrak{a}_{1:t}) \rightarrow X_{\infty}(\mathfrak{a}_{1:\infty})$ as $t \rightarrow \infty$, with $X_{\infty}(\mathfrak{a}_{1:\infty}) < \infty$, for μ^{π} -almost every infinite history $\mathfrak{a}_{1:\infty}$. That is:*

$$\mu^{\pi}\left(\left\{\mathfrak{a}_{1:\infty} \in (\mathcal{A} \times \mathcal{E})^{\infty} : X_t(\mathfrak{a}_{1:t}) \not\rightarrow X_{\infty}(\mathfrak{a}_{1:\infty})\right\}\right) = 0.$$

For each $\nu \in \mathcal{M}$, define the likelihood ratio

$$X_{\nu,t} : \mathcal{H}^t \rightarrow \mathbb{R}_{\geq 0}, \quad X_{\nu,t}(\mathfrak{a}_{1:t}) := \frac{\nu(\mathfrak{a}_{1:t})}{\mu(\mathfrak{a}_{1:t})}, \quad X_{\nu,0} := 1.$$

$X_{\nu,t}$ is a *function* of the history $\mathfrak{a}_{1:t}$, not a number: different histories give different values. Throughout this section and the next, any unqualified statement of the form “ $X_{\nu,t} \geq c$ ” or “ $X_{\nu,t} \rightarrow X_{\nu,\infty}$ ” is shorthand for “ $X_{\nu,t}(\mathfrak{a}_{1:t}) \geq c$ ” or “ $X_{\nu,t}(\mathfrak{a}_{1:t}) \rightarrow X_{\nu,\infty}(\mathfrak{a}_{1:\infty})$ ” holding μ^{π} -almost surely, i.e. on every infinite history outside a set of μ^{π} -measure zero. We will not track these null sets explicitly; their countable union over all claims is still a μ^{π} -null set.

Exercise 9.1. [15] Show that $X_{\nu,t}$ is a μ^π -martingale: $\mathbb{E}_{\mu^\pi}[X_{\nu,t} \mid \mathfrak{a}_{<t}] = X_{\nu,t-1}$.

Hint: Write $X_{\nu,t}(\mathfrak{a}_{1:t}) = X_{\nu,t-1}(\mathfrak{a}_{<t}) \cdot \nu(e_t \mid \mathfrak{a}_{<t}a_t) / \mu(e_t \mid \mathfrak{a}_{<t}a_t)$. Sum over a_t, e_t ; μ cancels, leaving $\sum_{e_t} \nu(e_t \mid \mathfrak{a}_{<t}a_t) = 1$.

Exercise 9.2. [05] Apply Fact 9.4 to conclude $X_{\nu,t} \rightarrow X_{\nu,\infty} < \infty$ μ^π -a.s.

Exercise 9.3. [10] Define $X_{\xi,t} := \xi(\mathfrak{a}_{1:t}) / \mu(\mathfrak{a}_{1:t})$. Show $X_{\xi,t} = \sum_\nu w_\nu X_{\nu,t}$ and $X_{\xi,t} \geq w_\mu > 0$.

10. Change of Measure

Exercise 10.1. [15] Let A be a set of finite histories of length m . Show:

$$\nu^\pi[\mathfrak{a}_{1:m} \in A] = \mathbb{E}_{\mu^\pi}[X_{\nu,m} \cdot \mathbb{I}[\mathfrak{a}_{1:m} \in A]].$$

Hint: Use Exercise 0.1 to show $\nu^\pi / \mu^\pi = \nu / \mu = X_{\nu,m}$.

Exercise 10.2. (Given.) The identity extends to infinite histories: for any event $A \subseteq (\mathcal{A} \times \mathcal{E})^\infty$,

$$\nu^\pi[A] = \mathbb{E}_{\mu^\pi}[X_{\nu,\infty} \cdot \mathbb{I}[\mathfrak{a}_{1:\infty} \in A]].$$

This says that $X_{\nu,\infty}$ plays the role of the Radon–Nikodym derivative $d\nu^\pi / d\mu^\pi$ on the space of infinite histories. The proof rests on two ingredients beyond the scope of this worksheet:

- **L^1 martingale convergence** (closed martingale / Lévy’s upward theorem): since $\mathbb{E}_{\mu^\pi}[X_{\nu,t}] = 1$ for every t , the non-negative μ^π -martingale $(X_{\nu,t})$ is uniformly integrable, so $X_{\nu,t} \rightarrow X_{\nu,\infty}$ in $L^1(\mu^\pi)$, not merely μ^π -a.s. This lets us pass the $t \rightarrow \infty$ limit through the expectation.
- **Uniqueness of measure extension** (Carathéodory / π - λ theorem): both sides of the identity are finite measures on $(\mathcal{A} \times \mathcal{E})^\infty$; the finite-history identity (Exercise 10.1) shows they agree on all cylinder events $A = A_0 \times (\mathcal{A} \times \mathcal{E})^\infty$, and cylinders generate the full event σ -algebra, so agreement extends uniquely to every event.

We omit the proof and use this identity freely below.

Exercise 10.3 (Markov inequality for change of measure). [15] Let E be an event of infinite histories. For any $\varepsilon > 0$, show that

$$\mu^\pi[E \cap \{X_{\nu,\infty} \geq \varepsilon\}] \leq \frac{\nu^\pi[E]}{\varepsilon}.$$

Hint: On $E \cap \{X_{\nu,\infty} \geq \varepsilon\}$, $X_{\nu,\infty} \geq \varepsilon$ pointwise. Take μ^π -expectations and apply Exercise 10.2.

11. Proving the Self-Optimizing Property

Fix a policy $\tilde{\pi}$ that is self-optimizing for $\mathcal{M}$ with respect to π (Definition 9.1): its existence is the hypothesis of Theorem 9.2. For each $\nu \in \mathcal{M}$, define the **suboptimality gap**

$$\delta_{\nu,t} : \mathcal{H}^{t-1} \rightarrow [0, 1], \quad \delta_{\nu,t}(\mathfrak{x}_{<t}) := V_\nu^*(\mathfrak{x}_{<t}) - V_{\tilde{\pi}}^{\tilde{\pi}}(\mathfrak{x}_{<t}).$$

As with $X_{\nu,t}$, we write $\delta_{\nu,t}$ without its argument when convenient: “ $\delta_{\nu,t} \rightarrow 0$ μ^π -a.s.” means $\delta_{\nu,t}(\mathfrak{x}_{<t}) \rightarrow 0$ on every infinite history outside a μ^π -null set, and similarly for “ $\delta_{\nu,t} \leq 1$ ” etc.

Exercise 11.1 (Chain of inequalities). [20] Show:

$$0 \leq w(\mu \mid \mathfrak{x}_{<t})[V_\mu^* - V_\mu^{\pi_\xi^*}] \leq \sum_{\nu \in \mathcal{M}} w(\nu \mid \mathfrak{x}_{<t}) \delta_{\nu,t}.$$

Hint: Use $V_\xi^{\pi_\xi^*} \geq V_\xi^{\tilde{\pi}}$ and Exercise 4.2.

Exercise 11.2 (Vanishing posterior). [10] Show: if $X_{\nu,\infty} = 0$ then $w(\nu \mid \mathfrak{x}_{<t}) \rightarrow 0$ μ^π -a.s.

Hint: Show $w(\nu \mid \mathfrak{x}_{<t}) = w_\nu X_{\nu,t-1} / X_{\xi,t-1}$ and use Exercise 9.3.

Exercise 11.3 (Transferring convergence). [20] Let $F := \{\mathfrak{x}_{1:\infty} : \delta_{\nu,t}(\mathfrak{x}_{<t}) \not\rightarrow 0\}$, the set of infinite histories on which the suboptimality gap fails to vanish. Show that $\mu^\pi[F \cap \{X_{\nu,\infty} > 0\}] = 0$.

Hint: Self-optimizing (Definition 9.1) gives $\nu^\pi[F] = 0$. Apply Exercise 10.3 with $E = F$ and $\varepsilon = 1/n$ to conclude $\mu^\pi[F \cap \{X_{\nu,\infty} \geq 1/n\}] = 0$ for each $n \geq 1$. Then note $\{X_{\nu,\infty} > 0\} = \bigcup_{n \geq 1} \{X_{\nu,\infty} \geq 1/n\}$ and take the countable union (Definition 7.1).

Exercise 11.4 (Single-term convergence). [10] Combine Exercise 11.2 and Exercise 11.3 to show $w(\nu \mid \mathfrak{a}_{<t})\delta_{\nu,t} \rightarrow 0$ μ^π -a.s.

Exercise 11.5 (Self-Optimizing Theorem). [15] Combine Exercise 11.4, Exercise 11.1, Exercise 9.3 to finally prove the main result of Theorem 9.2.

Hint: $w(\mu \mid \mathfrak{a}_{<t}) = w_\mu / X_{\xi,t-1}$.

Remark

We do not need to know which policy $\tilde{\pi}$ is self-optimizing, or whether it is computable. Mere existence suffices. For countable $\mathcal{M}$, this final step is harder: a countable sum of μ^π -a.s.-convergent sequences need not converge μ^π -a.s. The general proof uses a “convergence of mixture tails” argument [Hut05, Lem. 5.28].

Further reading

- Hutter, *An Introduction to Universal Artificial Intelligence* (2024): Chapter 2.7 (Kolmogorov complexity), Chapters 3.7–3.8 (the model class and universal prior), and Chapter 7.4 (AIXI).
- Hutter, *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability* (Springer, 2005) — the original book-length treatment; Lem. 5.28 handles the countable- $\mathcal{M}$ self-optimizing case.
- Blackwell & Dubins, *Merging of Opinions with Increasing Information* (Ann. Math. Statist., 1962) — the merging-of-opinions theorem behind on-policy value convergence.
- Leike & Hutter, *Bad Universal Priors and Notions of Optimality* (COLT 2015) — adversarial choices of the universal Turing machine can make AIXI behave arbitrarily badly.

A. Worked Example: Bayesian Mixture and Value Function

Note

Setup.

- **Actions:** $\mathcal{A} = \{H, T\}$ (predict the next coin flip)
- **Observations:** $\mathcal{O} = \{H, T\}$ (actual coin flip)
- **Rewards:** $\mathcal{R} = \{0, 1\}$, with $r_t = \llbracket a_t = o_t \rrbracket$
- **Model class:** $\mathcal{M} = \{\nu_{HH}, \nu_{HT}\}$ (two-headed coin, fair coin)
- **Prior:** $w_{\nu_{HH}} = w_{\nu_{HT}} = \frac{1}{2}$

Before any interaction ($t = 1$, $\mathfrak{x}_{<1} = \epsilon$):

$$\xi(o_1 = H \mid a_1) = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot \frac{1}{2} = \frac{3}{4}.$$

After observing a head ($t = 2$), the posterior updates:

$$w(\nu_{HH} \mid \mathfrak{x}_1) = \frac{1}{2} \cdot \frac{1}{3/4} = \frac{2}{3}, \quad w(\nu_{HT} \mid \mathfrak{x}_1) = \frac{1}{2} \cdot \frac{1/2}{3/4} = \frac{1}{3}.$$

Updated prediction: $\xi(o_2 = H \mid \mathfrak{x}_1 a_2) = \frac{2}{3} \cdot 1 + \frac{1}{3} \cdot \frac{1}{2} = \frac{5}{6}$.

Value function. Continuing the same setup, suppose the agent always predicts H (policy π_H).

Under ν_{HH} : always correct, $r_t = 1$ every step: $V_{\nu_{HH}}^{\pi_H}(\epsilon) = (1 - \gamma) \sum_{k=0}^{\infty} \gamma^k \cdot 1 = 1$.

Under ν_{HT} : correct half the time: $V_{\nu_{HT}}^{\pi_H}(\epsilon) = (1 - \gamma) \sum_{k=0}^{\infty} \gamma^k \cdot \frac{1}{2} = \frac{1}{2}$.

The mixture value (by Exercise 4.2) is: $V_{\xi}^{\pi_H}(\epsilon) = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot \frac{1}{2} = \frac{3}{4}$.

As the agent observes more heads ($\mu = \nu_{HH}$), the posterior on ν_{HH} increases towards 1, and $V_{\xi}^{\pi_H} \rightarrow V_{\nu_{HH}}^{\pi_H} = 1$. This is on-policy value convergence (Theorem 7.6) in action.

B. Knuth's Difficulty Scale

Each subproblem carries a difficulty rating in square brackets, following Knuth's rating scheme for exercises [Knu73] in slightly adapted form. The rating assumes that the material in the preceding problems (on which the subproblem depends) has been understood. In-between values are possible.

- [00] *Very easy.* Solvable from the top of your head.
- [10] *Easy.* Needs 15 minutes to think, possibly pencil and paper.
- [20] *Average.* May take 1–2 hours to answer completely.
- [30] *Moderately difficult or lengthy.* May take several hours to a day.
- [40] *Quite difficult or lengthy.* Often a significant research result.
- [50] *Open research problem.* An obtained solution should be published.

References

- [BD62] D. Blackwell and L. Dubins. Merging of opinions with increasing information. *Annals of Mathematical Statistics*, 33:882–887, 1962. URL: <http://www.dklevine.com/archive/refs4565.pdf>.
- [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.
- [LH15] Jan Leike and Marcus Hutter. Bad universal priors and notions of optimality. *CoRR*, abs/1510.04931, 2015. URL: <https://arxiv.org/abs/1510.04931>.