

AIXI

David Quarel

Australian National University

September 8, 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.”

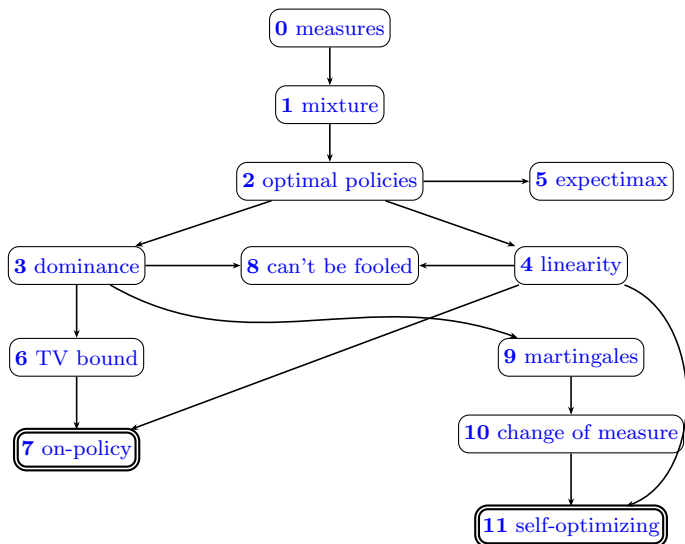
Goal and Roadmap

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**
- $\mathcal{H}^{\infty} := (\mathcal{A} \times \mathcal{E})^{\infty}$: 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 $e_t, \mathfrak{x}_{<t}, a_t$ (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 $\nu^\pi(\mathfrak{a}_{1:t}) = \pi(\mathfrak{a}_{1:t}) \cdot \nu(\mathfrak{a}_{1:t})$.

Solution to Exercise 0.1

Expanding the definition: $\nu^\pi(\mathfrak{a}_{1:t}) = \prod_{k=1}^t \pi(a_k \mid \mathfrak{a}_{<k}) \nu(e_k \mid \mathfrak{a}_{<k} a_k) = \underbrace{\prod_{k=1}^t \pi(a_k \mid \mathfrak{a}_{<k})}_{\pi(\mathfrak{a}_{1:t})} \cdot \underbrace{\prod_{k=1}^t \nu(e_k \mid \mathfrak{a}_{<k} a_k)}_{\nu(\mathfrak{a}_{1:t})}$.

Key observation: $\pi(\mathfrak{a}_{1:t})$ is the same regardless of the environment. When comparing ν^π and μ^π , the policy factors cancel.

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

Solution to Exercise 0.2

From the definition $\nu(\mathfrak{a}_{1:t}) = \prod_{k=1}^t \nu(e_k \mid \mathfrak{a}_{<k} a_k)$, split off the last factor:

$$\nu(\mathfrak{a}_{1:t}) = \prod_{k=1}^{t-1} \nu(e_k \mid \mathfrak{a}_{<k} a_k) \cdot \nu(e_t \mid \mathfrak{a}_{<t} a_t) = \nu(\mathfrak{a}_{<t}) \cdot \nu(e_t \mid \mathfrak{a}_{<t} a_t).$$

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

Solution to Exercise 0.3

From Definition 0.4, the one-step interaction is $\nu^\pi(a_t e_t \mid \mathfrak{x}_{<t}) = \pi(a_t \mid \mathfrak{x}_{<t}) \nu(e_t \mid \mathfrak{x}_{<t} a_t)$. Summing over e_t :

$$\sum_{e_t} \nu^\pi(a_t e_t \mid \mathfrak{x}_{<t}) = \pi(a_t \mid \mathfrak{x}_{<t}) \underbrace{\sum_{e_t} \nu(e_t \mid \mathfrak{x}_{<t} a_t)}_{=1} = \pi(a_t \mid \mathfrak{x}_{<t}).$$

Interpretation: after summing out the environment's response, only the agent's action probability remains.

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

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

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

Solution to Exercise 0.4

Since π is deterministic, $\pi(a_k \mid \mathfrak{x}_{<k}) = \llbracket a_k = a_k^* \rrbracket$. By hypothesis $\mathfrak{x}_{i:j}$ is π -consistent, so every such bracket equals 1. From Definition 0.4:

$$\begin{aligned} \nu^\pi(\mathfrak{x}_{i:j} \mid \mathfrak{x}_{<i}) &= \prod_{k=i}^j \underbrace{\pi(a_k \mid \mathfrak{x}_{<k})}_{=1} \nu(e_k \mid \mathfrak{x}_{<k} a_k) \\ &= \prod_{k=i}^j \nu(e_k \mid \mathfrak{x}_{<k} a_k) = \nu(\mathfrak{x}_{i:j} \mid \mathfrak{x}_{<i}), \end{aligned}$$

where the last equality is the natural segment extension of $\nu(\mathfrak{x}_{1:t}) := \prod_{k=1}^t \nu(e_k \mid \mathfrak{x}_{<k} a_k)$.

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

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

Solution to Exercise 0.5

Split the defining product $\nu^\pi(\mathfrak{a}_{p:r} \mid \mathfrak{a}_{< p}) = \prod_{k=p}^r \pi(a_k \mid \mathfrak{a}_{< k}) \nu(e_k \mid \mathfrak{a}_{< k} a_k)$ at index q :

$$\begin{aligned} \nu^\pi(\mathfrak{a}_{p:r} \mid \mathfrak{a}_{< p}) &= \prod_{k=p}^r \pi(a_k \mid \mathfrak{a}_{< k}) \nu(e_k \mid \mathfrak{a}_{< k} a_k) \\ &= \underbrace{\prod_{k=p}^q \pi(a_k \mid \mathfrak{a}_{< k}) \nu(e_k \mid \mathfrak{a}_{< k} a_k)}_{\nu^\pi(\mathfrak{a}_{p:q} \mid \mathfrak{a}_{< p})} \cdot \underbrace{\prod_{k=q+1}^r \pi(a_k \mid \mathfrak{a}_{< k}) \nu(e_k \mid \mathfrak{a}_{< k} a_k)}_{\nu^\pi(\mathfrak{a}_{q+1:r} \mid \mathfrak{a}_{< p} \mathfrak{a}_{p:q})}, \end{aligned}$$

identifying each block as a conditional via the same expansion: the past for step $k \geq q+1$ is the given history $\mathfrak{a}_{< p}$ extended by the first chunk $\mathfrak{a}_{p:q}$.

Note: for a deterministic policy, Exercise 0.4 simplifies this further: each factor $\pi(a_k \mid \mathfrak{a}_{< k}) \nu(e_k \mid \mathfrak{a}_{< k} a_k)$ becomes just $\nu(e_k \mid \mathfrak{a}_{< k} a_k^*)$.

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

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

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

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

where the **posterior weight** is

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

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 $\mathfrak{a}'_{t:m}$:

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

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 $\mathfrak{a}'_{t:m}$, we define $\mathbb{E}_\nu^\pi[f \mid \mathfrak{a}_{<t}] := \lim_{m \rightarrow \infty} \mathbb{E}_\nu^\pi[f_m \mid \mathfrak{a}_{<t}]$.

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 $\mathfrak{a}_{<t}$ is

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

The **infinite-horizon value** is the pointwise limit $V_\nu^{\pi,\infty}(\mathfrak{a}_{<t}) := \lim_{m \rightarrow \infty} V_\nu^{\pi,m}(\mathfrak{a}_{<t}) = (1 - \gamma) \mathbb{E}_\nu^\pi[G_{\geq t-1} \mid \mathfrak{a}_{<t}]$. We write $V_\nu^\pi \equiv V_\nu^{\pi,\infty}$ for short. The **optimal value** is $V_\nu^{*,m}(\mathfrak{a}_{<t}) := \sup_\pi V_\nu^{\pi,m}(\mathfrak{a}_{<t})$ 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 **AI ξ** .

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:

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

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

Solution to Exercise 1.1

From the definition: $w(\nu \mid \mathfrak{x}_{1:t}) = w_\nu \cdot \nu(\mathfrak{x}_{1:t}) / \xi(\mathfrak{x}_{1:t})$ and $w(\nu \mid \mathfrak{x}_{<t}) = w_\nu \cdot \nu(\mathfrak{x}_{<t}) / \xi(\mathfrak{x}_{<t})$.

Dividing:

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

where we used the chain rule (Exercise 0.2) for both ν and ξ .

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

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

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

Solution to Exercise 1.2

The one-step conditional is the ratio of joint to marginal: $\xi(e_t \mid \mathbf{a}_{<t} a_t) := \xi(\mathbf{a}_{1:t}) / \xi(\mathbf{a}_{<t})$.

Expanding the numerator using $\xi(\mathbf{a}_{1:t}) = \sum_{\nu} w_\nu \nu(\mathbf{a}_{1:t})$:

$$\begin{aligned} \xi(e_t \mid \mathbf{a}_{<t} a_t) &= \frac{\sum_{\nu \in \mathcal{M}} w_\nu \nu(\mathbf{a}_{1:t})}{\xi(\mathbf{a}_{<t})} \\ &= \frac{\sum_{\nu} w_\nu \nu(\mathbf{a}_{<t}) \cdot \nu(e_t \mid \mathbf{a}_{<t} a_t)}{\xi(\mathbf{a}_{<t})} \quad (\text{chain rule, Exercise 0.2}) \\ &= \sum_{\nu} \underbrace{\frac{w_\nu \nu(\mathbf{a}_{<t})}{\xi(\mathbf{a}_{<t})}}_{= w(\nu \mid \mathbf{a}_{<t})} \cdot \nu(e_t \mid \mathbf{a}_{<t} a_t) \\ &= \sum_{\nu \in \mathcal{M}} w(\nu \mid \mathbf{a}_{<t}) \nu(e_t \mid \mathbf{a}_{<t} a_t). \end{aligned}$$

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

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

Solution to Exercise 1.3

Since $r_k \in [0, 1]$, each discounted reward sum is bounded:

$$0 \leq (1 - \gamma) G_{t-1:m} \leq (1 - \gamma) \sum_{k=t}^m \gamma^{k-t} = 1 - \gamma^{m-t+1} \leq 1.$$

The interaction measure satisfies $\nu^\pi(\mathbf{a}'_{t:m} \mid \mathbf{a}_{<t}) \geq 0$ and $\sum_{\mathbf{a}'_{t:m}} \nu^\pi(\mathbf{a}'_{t:m} \mid \mathbf{a}_{<t}) = 1$, so the value function is a weighted average of terms in $[0, 1]$:

$$0 \leq V_\nu^{\pi,m}(\mathbf{a}_{<t}) = \sum_{\mathbf{a}'_{t:m}} \nu^\pi(\mathbf{a}'_{t:m} \mid \mathbf{a}_{<t}) (1 - \gamma) G_{t-1:m} \leq 1.$$

For $m = \infty$: since $0 \leq V_\nu^{\pi,m}(\mathbf{a}_{<t}) \leq 1$ for all finite m , the limit $V_\nu^\pi(\mathbf{a}_{<t}) = \lim_{m \rightarrow \infty} V_\nu^{\pi,m}(\mathbf{a}_{<t})$ also lies in $[0, 1]$.

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

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

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

Solution to Exercise 1.4

Start from the definition $\xi(\mathbf{a}_{1:m}) = \sum_{\nu \in \mathcal{M}} w_\nu \nu(\mathbf{a}_{1:m})$. Apply the general chain rule (Exercise 0.5) to both sides: $\xi(\mathbf{a}_{1:m}) = \xi(\mathbf{a}_{<t}) \cdot \xi(\mathbf{a}_{t:m} \mid \mathbf{a}_{<t})$ and $\nu(\mathbf{a}_{1:m}) = \nu(\mathbf{a}_{<t}) \cdot \nu(\mathbf{a}_{t:m} \mid \mathbf{a}_{<t})$. Dividing by $\xi(\mathbf{a}_{<t})$:

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

Now multiply both sides by $\pi(\mathbf{a}_{t:m} \mid \mathbf{a}_{<t})$. By factorization (Exercise 0.1), $\pi(\mathbf{a}_{t:m} \mid \mathbf{a}_{<t}) \cdot \xi(\mathbf{a}_{t:m} \mid \mathbf{a}_{<t}) = \xi^\pi(\mathbf{a}_{t:m} \mid \mathbf{a}_{<t})$ and likewise for each ν :

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

2. Existence of Optimal Policies

Recall that $V_\nu^*(\mathbf{a}_{<t}) := \sup_\pi V_\nu^\pi(\mathbf{a}_{<t})$ 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 $\mathbf{a}_{<t}$ and a function $Q : \mathcal{A} \rightarrow \mathbb{R}$, show that $\sup_\pi \sum_{a \in \mathcal{A}} \pi(a \mid \mathbf{a}_{<t}) Q(a) = \max_{a \in \mathcal{A}} Q(a)$, and that the supremum is attained by the deterministic policy that places all probability on an action achieving the maximum.

Solution to Exercise 2.1

Let $a^* \in \arg \max_{a' \in \mathcal{A}} Q(a')$, which exists because $\mathcal{A}$ is finite. So $Q(a^*) = \max_{a'} Q(a')$.

Upper bound. For any π : $\sum_a \pi(a \mid \mathbf{a}_{<t}) Q(a) \leq \sum_a \pi(a \mid \mathbf{a}_{<t}) Q(a^*) = Q(a^*)$.

Lower bound. The deterministic π^* with $\pi^*(a^* \mid \mathbf{a}_{<t}) = 1$ achieves $Q(a^*)$.

Combining: $\sup_\pi \sum_a \pi(a) Q(a) = Q(a^*) = \max_{a'} Q(a')$, attained by π^* .

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

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

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

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

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

Solution to Exercise 2.2

Step 1: Factor the future. Write $\mathbf{a}'_{t:m} = \mathbf{a}'_t \mathbf{a}'_{t+1:m}$: $\nu^\pi(\mathbf{a}'_{t:m} \mid \mathbf{a}_{<t}) = \nu^\pi(\mathbf{a}'_t \mid \mathbf{a}_{<t}) \cdot \nu^\pi(\mathbf{a}'_{t+1:m} \mid \mathbf{a}_{<t} \mathbf{a}'_t)$.

Step 2: Split the reward sum. This is just the return recursion (Definition 0.3):

$G_{t-1:m} = r'_t + \gamma G_{t:m}$, where $G_{t-1:m} = \sum_{k=t}^m \gamma^{k-t} r'_k$ is the return given $\mathbf{a}_{<t}$.

Step 3: Substitute into the definition of $V_\nu^{\pi, m}(\mathbf{a}_{<t})$.

$$\begin{aligned} V_\nu^{\pi, m}(\mathbf{a}_{<t}) &= (1 - \gamma) \sum_{\mathbf{a}'_t} \nu^\pi(\mathbf{a}'_t \mid \mathbf{a}_{<t}) \sum_{\mathbf{a}'_{t+1:m}} \nu^\pi(\mathbf{a}'_{t+1:m} \mid \mathbf{a}_{<t} \mathbf{a}'_t) \\ &\quad \times [r'_t + \gamma G_{t:m}]. \end{aligned}$$

Consider the inner term $\sum_{\mathbf{a}'_{t+1:m}} \nu^\pi(\mathbf{a}'_{t+1:m} \mid \mathbf{a}_{<t} \mathbf{a}'_t) [r'_t + \gamma G_{t:m}]$. We can expand as:

$$\begin{aligned}
&= \sum_{\mathbf{a}'_{t+1:m}} \nu^\pi(\mathbf{a}'_{t+1:m} \mid \mathbf{a}_{<t} \mathbf{a}'_t) r'_t + \gamma \sum_{\mathbf{a}'_{t+1:m}} \nu^\pi(\mathbf{a}'_{t+1:m} \mid \mathbf{a}_{<t} \mathbf{a}'_t) G_{t:m} \\
&= r'_t \sum_{\mathbf{a}'_{t+1:m}} \cancel{\nu^\pi(\mathbf{a}'_{t+1:m} \mid \mathbf{a}_{<t} \mathbf{a}'_t)} + \gamma \sum_{\mathbf{a}'_{t+1:m}} \nu^\pi(\mathbf{a}'_{t+1:m} \mid \mathbf{a}_{<t} \mathbf{a}'_t) G_{t:m} \\
&= r'_t + \gamma \sum_{\mathbf{a}'_{t+1:m}} \nu^\pi(\mathbf{a}'_{t+1:m} \mid \mathbf{a}_{<t} \mathbf{a}'_t) G_{t:m}
\end{aligned}$$

as r'_t doesn't depend on $\mathbf{a}'_{t+1:m}$, and $\sum_{\mathbf{a}'_{t+1:m}} \nu^\pi(\mathbf{a}'_{t+1:m} \mid \mathbf{a}_{<t} \mathbf{a}'_t) = 1$. Inserting this above, we obtain:

$$\begin{aligned}
V_\nu^{\pi,m}(\mathbf{a}_{<t}) &= \sum_{\mathbf{a}'_t} \nu^\pi(\mathbf{a}'_t \mid \mathbf{a}_{<t}) \left[(1 - \gamma) r'_t \right. \\
&\quad \left. + \gamma (1 - \gamma) \underbrace{\sum_{\mathbf{a}'_{t+1:m}} \nu^\pi(\mathbf{a}'_{t+1:m} \mid \mathbf{a}_{<t} \mathbf{a}'_t) G_{t:m}}_{= V_\nu^{\pi,m}(\mathbf{a}_{<t} \mathbf{a}'_t)} \right].
\end{aligned}$$

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.

Solution to Exercise 2.3

We induct on $t = m, m - 1, \dots, 1$, showing that for every history $\mathbf{a}_{<t}$, a deterministic policy achieves $V_\nu^{*,m}(\mathbf{a}_{<t})$.

Base case ($t = m$): At $t = m$ the continuation term vanishes (the value from time $m + 1$ is an empty sum), so the Bellman equation gives $V_\nu^{\pi,m}(\mathbf{a}_{<m}) = (1 - \gamma) \sum_{\mathbf{a}'_m} \nu^\pi(\mathbf{a}'_m \mid \mathbf{a}_{<m}) r'_m$. Splitting the step $\mathbf{a}'_m = a'_m e'_m$ via $\nu^\pi(\mathbf{a}'_m \mid \mathbf{a}_{<m}) = \pi(a'_m \mid \mathbf{a}_{<m}) \nu(e'_m \mid \mathbf{a}_{<m} a'_m)$ gives $V_\nu^{\pi,m}(\mathbf{a}_{<m}) = (1 - \gamma) \sum_{a'_m} \pi(a'_m \mid \mathbf{a}_{<m}) \sum_{e'_m} \nu(e'_m \mid \mathbf{a}_{<m} a'_m) r'_m$. This has the form $\sum_a \pi(a) Q(a)$ with $Q(a) = (1 - \gamma) \sum_{e'_m} \nu(e'_m \mid \mathbf{a}_{<m} a) r'_m$, which is independent of π . By Exercise 2.1, the supremum over π is $\max_a Q(a)$, attained by the deterministic policy playing $a^* \in \arg \max_a Q(a)$.

Inductive step: Suppose that for every history of length t , there exists a deterministic policy π_{t+1}^* achieving $V_\nu^{\pi_{t+1}^*,m} = V_\nu^{*,m}$ from time $t + 1$ onwards. Substituting

this optimal continuation into the Bellman equation from Exercise 2.2 gives

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

Splitting the step $\mathfrak{x}'_t = a'_t e'_t$ via $\nu^{\pi}(\mathfrak{x}'_t \mid \mathfrak{x}_{<t}) = \pi(a'_t \mid \mathfrak{x}_{<t}) \nu(e'_t \mid \mathfrak{x}_{<t} a'_t)$ and grouping the percept sum into the action weight:

$$V_{\nu}^{\pi,m}(\mathfrak{x}_{<t}) = \sum_{a'_t} \pi(a'_t \mid \mathfrak{x}_{<t}) Q(a'_t),$$

where $Q(a'_t) := \sum_{e'_t} \nu(e'_t \mid \mathfrak{x}_{<t} a'_t) [(1 - \gamma) r'_t + \gamma V_{\nu}^{*,m}(\mathfrak{x}_{<t} a'_t e'_t)]$ is independent of π (the continuation value $V_{\nu}^{*,m}$ is fixed by the inductive hypothesis). By Exercise 2.1, $\sup_{\pi} \sum_{a'_t} \pi(a'_t) Q(a'_t) = \max_{a'_t} Q(a'_t)$, attained by the deterministic policy playing $a_t^* \in \arg \max_{a'_t} Q(a'_t)$ at time t , then following π_{t+1}^* .

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

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

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

Solution to Exercise 2.4

By Exercise 2.3, a deterministic optimal policy π_{ν}^* exists with $V_{\nu}^{\pi_{\nu}^*,m} = V_{\nu}^{*,m}$. Substituting $\pi = \pi_{\nu}^*$ into the Bellman equation from Exercise 2.2 (using $V_{\nu}^{\pi_{\nu}^*,m} = V_{\nu}^{*,m}$ on both sides):

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

Splitting the step $\mathfrak{x}'_t = a'_t e'_t$ via $\nu^{\pi_{\nu}^*}(\mathfrak{x}'_t \mid \mathfrak{x}_{<t}) = \pi_{\nu}^*(a'_t \mid \mathfrak{x}_{<t}) \nu(e'_t \mid \mathfrak{x}_{<t} a'_t)$:

$$\begin{aligned} V_{\nu}^{*,m}(\mathfrak{x}_{<t}) &= \sum_{a'_t} \pi_{\nu}^*(a'_t \mid \mathfrak{x}_{<t}) \\ &\quad \times \sum_{e'_t} \nu(e'_t \mid \mathfrak{x}_{<t} a'_t) [(1 - \gamma) r'_t + \gamma V_{\nu}^{*,m}(\mathfrak{x}_{<t} a'_t e'_t)]. \end{aligned}$$

Since π_ν^* is deterministic, let a_t^* denote the unique action with $\pi_\nu^*(a_t^* \mid \mathfrak{a}_{<t}) = 1$. Then:

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

Define $Q(a) := \sum_{e'_t} \nu(e'_t \mid \mathfrak{a}_{<t} a) [(1-\gamma) r'_t + \gamma V_\nu^{*,m}(\mathfrak{a}_{<t} a e'_t)]$. Then $V_\nu^{*,m}(\mathfrak{a}_{<t}) = Q(a_t^*)$. We must have $Q(a_t^*) = \max_a Q(a)$: if some $\tilde{a}$ had $Q(\tilde{a}) > Q(a_t^*)$, then the policy that plays $\tilde{a}$ at $\mathfrak{a}_{<t}$ and follows π_ν^* elsewhere would achieve value $Q(\tilde{a}) > V_\nu^{*,m}(\mathfrak{a}_{<t})$, contradicting $V_\nu^{*,m} = \sup_\pi V_\nu^{\pi,m}$. Hence:

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

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

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

Solution to Exercise 2.5

We show $V_\nu^{*,m+1}(\mathfrak{a}_{<t}) \geq V_\nu^{*,m}(\mathfrak{a}_{<t})$ for all $\mathfrak{a}_{<t}$. Since $V_\nu^{*,m+1} = \sup_\pi V_\nu^{\pi,m+1}$, it suffices to show $V_\nu^{\pi,m+1}(\mathfrak{a}_{<t}) \geq V_\nu^{*,m}(\mathfrak{a}_{<t})$ for every π . Expanding the definition:

$$\begin{aligned} V_\nu^{\pi,m+1}(\mathfrak{a}_{<t}) &= (1-\gamma) \sum_{\mathfrak{a}_{t:m+1}} \nu^\pi(\mathfrak{a}_{t:m+1} \mid \mathfrak{a}_{<t}) G_{t-1:m+1} \\ &= (1-\gamma) \sum_{\mathfrak{a}_{t:m+1}} \nu^\pi(\mathfrak{a}_{t:m+1} \mid \mathfrak{a}_{<t}) [G_{t-1:m} + \gamma^{m+1-t} r_{m+1}] \\ &= (1-\gamma) \sum_{\mathfrak{a}_{t:m+1}} \nu^\pi(\mathfrak{a}_{t:m+1} \mid \mathfrak{a}_{<t}) G_{t-1:m} \\ &\quad + \underbrace{(1-\gamma) \gamma^{m+1-t} \sum_{\mathfrak{a}_{t:m+1}} \nu^\pi(\mathfrak{a}_{t:m+1} \mid \mathfrak{a}_{<t}) r_{m+1}}_{\geq 0}. \end{aligned}$$

For the first term, marginalize over a_{m+1}, e_{m+1} (which $G_{t-1:m}$ does not depend on):

$$\sum_{\mathfrak{a}_{t:m+1}} \nu^\pi(\mathfrak{a}_{t:m+1} \mid \mathfrak{a}_{<t}) G_{t-1:m} = \sum_{\mathfrak{a}_{t:m}} \nu^\pi(\mathfrak{a}_{t:m} \mid \mathfrak{a}_{<t}) G_{t-1:m}.$$

So $V_\nu^{\pi, m+1}(\mathbf{x}_{<t}) \geq (1 - \gamma) \sum_{\mathbf{x}_{t:m}} \nu^\pi(\mathbf{x}_{t:m} \mid \mathbf{x}_{<t}) G_{t-1:m} = V_\nu^{\pi, m}(\mathbf{x}_{<t})$. Since $V_\nu^{*, m}(\mathbf{x}_{<t})$ is non-decreasing in m and bounded in $[0, 1]$ by Exercise 1.3, the monotone convergence theorem gives that $V_\nu^*(\mathbf{x}_{<t}) := \lim_{m \rightarrow \infty} V_\nu^{*, m}(\mathbf{x}_{<t})$ exists for each $\mathbf{x}_{<t}$.

Exercise 2.6 (Infinite-horizon Bellman optimality equation). [10]

Show that the Bellman optimality equation (Exercise 2.4) extends to $m = \infty$:

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

Solution to Exercise 2.6

By Exercise 2.4, for every finite m :

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

By Exercise 2.5, each $V_\nu^{*, m}(\cdot)$ converges pointwise to $V_\nu^*(\cdot)$ as $m \rightarrow \infty$. Since $\mathcal{A}$ and $\mathcal{E}$ are finite, the max and $\sum$ are over finitely many convergent terms, so:

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

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:

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

Solution to Exercise 2.7

Step 1: Define the greedy policy. For each history $\mathbf{a}_{<t}$, let

$$a_t^* \in \arg \max_{a_t'} \sum_{e_t'} \nu(e_t' \mid \mathbf{a}_{<t} a_t') [(1 - \gamma) r_t' + \gamma V_\nu^*(\mathbf{a}_{<t} a_t' e_t')],$$

which exists because $\mathcal{A}$ is finite. Define the deterministic policy $\pi_\nu^*(a \mid \mathbf{a}_{<t}) := \mathbb{I}[a = a_t^*]$. By Exercise 2.6, the infinite-horizon Bellman optimality equation becomes:

$$V_\nu^*(\mathbf{a}_{<t}) = \sum_{e_t'} \nu(e_t' \mid \mathbf{a}_{<t} a_t^*) \left[(1 - \gamma) r_t' + \gamma V_\nu^*(\mathbf{a}_{<t} a_t^* e_t') \right].$$

Step 2: Error recursion. The Bellman equation (Exercise 2.2) holds for any policy, so for π_ν^* at finite horizon m :

$$V_\nu^{\pi_\nu^*, m}(\mathbf{a}_{<t}) = \sum_{e_t'} \nu(e_t' \mid \mathbf{a}_{<t} a_t^*) \left[(1 - \gamma) r_t' + \gamma V_\nu^{\pi_\nu^*, m}(\mathbf{a}_{<t} a_t^* e_t') \right].$$

Subtracting this from the equation in Step 1, the $(1 - \gamma) r_t'$ terms cancel:

$$\begin{aligned} V_\nu^*(\mathbf{a}_{<t}) - V_\nu^{\pi_\nu^*, m}(\mathbf{a}_{<t}) &= \gamma \sum_{e_t'} \nu(e_t' \mid \mathbf{a}_{<t} a_t^*) \\ &\quad \times \left[V_\nu^*(\mathbf{a}_{<t} a_t^* e_t') - V_\nu^{\pi_\nu^*, m}(\mathbf{a}_{<t} a_t^* e_t') \right]. \end{aligned}$$

The error at time t is γ times a weighted average of errors at time $t + 1$.

Step 3: Iterate the contraction. Write $\Delta V_\nu^m(\mathbf{a}_{<t}) := V_\nu^*(\mathbf{a}_{<t}) - V_\nu^{\pi_\nu^*, m}(\mathbf{a}_{<t}) \geq 0$ for the error. Step 2 gives:

$$\Delta V_\nu^m(\mathbf{a}_{<t}) = \gamma \sum_{e_t'} \nu(e_t' \mid \mathbf{a}_{<t} a_t^*) \Delta V_\nu^m(\mathbf{a}_{<t} a_t^* e_t').$$

Applying the same recursion at time $t + 1$ (with greedy action a_{t+1}^* at history $\mathbf{a}_{<t} a_t^* e_t'$):

$$\begin{aligned} \Delta V_\nu^m(\mathbf{a}_{<t}) &= \gamma^2 \sum_{e_t'} \nu(e_t' \mid \mathbf{a}_{<t} a_t^*) \sum_{e_{t+1}'} \nu(e_{t+1}' \mid \mathbf{a}_{<t} a_t^* e_t' a_{t+1}^*) \\ &\quad \times \Delta V_\nu^m(\mathbf{a}_{<t} a_t^* e_t' a_{t+1}^* e_{t+1}'). \end{aligned}$$

After $m - t + 1$ iterations, summing over all percept sequences $e_{t:m}'$:

$$\Delta V_\nu^m(\mathbf{a}_{<t}) = \gamma^{m-t+1} \sum_{e_{t:m}'} \left(\prod_{k=t}^m \nu(e_k' \mid \mathbf{a}_{<k} a_k^*) \right) \Delta V_\nu^m(\mathbf{a}_{<t} \mathbf{a}_{t:m}'),$$

where $\mathbf{a}'_{t:m} = a_t^* e'_t \cdots a_m^* e'_m$ is the history segment generated by π_ν^* . By Exercise 0.4, the product of ν -conditionals is $\nu^{\pi_\nu^*}(\mathbf{a}'_{t:m} \mid \mathbf{a}_{<t})$. At the boundary, $V_\nu^{\pi_\nu^*, m}(\mathbf{a}_{<t} \mathbf{a}'_{t:m}) = 0$ (summing over no rewards past horizon m), so $\Delta V_\nu^m(\mathbf{a}_{<t} \mathbf{a}'_{t:m}) = V_\nu^*(\mathbf{a}_{<t} \mathbf{a}'_{t:m}) \in [0, 1]$. Therefore:

$$\Delta V_\nu^m(\mathbf{a}_{<t}) = \gamma^{m-t+1} \sum_{\mathbf{a}'_{t:m}} \nu^{\pi_\nu^*}(\mathbf{a}'_{t:m} \mid \mathbf{a}_{<t}) V_\nu^*(\mathbf{a}_{<t} \mathbf{a}'_{t:m}) \leq \gamma^{m-t+1}.$$

Step 4: Take $m \rightarrow \infty$. Since $0 \leq V_\nu^*(\mathbf{a}_{<t}) - V_\nu^{\pi_\nu^*, m}(\mathbf{a}_{<t}) \leq \gamma^{m-t+1} \rightarrow 0$, we have $V_\nu^{\pi_\nu^*, m}(\mathbf{a}_{<t}) \rightarrow V_\nu^*(\mathbf{a}_{<t})$, i.e. $V_\nu^{\pi_\nu^*} = V_\nu^*$.

3. Dominance of the Bayesian Mixture

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

Solution to Exercise 3.1

$$\xi(\mathbf{a}_{<t}) = \sum_{\nu' \in \mathcal{M}} w_{\nu'} \nu'(\mathbf{a}_{<t}) \geq w_\nu \nu(\mathbf{a}_{<t}).$$

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

Solution to Exercise 3.2

By Exercise 0.1: $\xi^\pi(\mathbf{a}_{1:m}) = \pi(\mathbf{a}_{1:m}) \cdot \xi(\mathbf{a}_{1:m}) \geq \pi(\mathbf{a}_{1:m}) \cdot w_\nu \nu(\mathbf{a}_{1:m}) = w_\nu \cdot \nu^\pi(\mathbf{a}_{1:m})$.

4. Properties of V_ξ

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

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

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

Solution to Exercise 4.1

By Exercise 1.4: $\xi^\pi(\mathfrak{a}_{t:m} \mid \mathfrak{a}_{<t}) = \sum_\nu w(\nu \mid \mathfrak{a}_{<t}) \nu^\pi(\mathfrak{a}_{t:m} \mid \mathfrak{a}_{<t})$. Multiply both sides by $(1 - \gamma)G_{t-1:m}$ and sum over all $\mathfrak{a}_{t:m}$:

$$\begin{aligned} V_\xi^{\pi,m}(\mathfrak{a}_{<t}) &= (1 - \gamma) \sum_{\mathfrak{a}_{t:m}} \xi^\pi(\mathfrak{a}_{t:m} \mid \mathfrak{a}_{<t}) G_{t-1:m} \\ &= (1 - \gamma) \sum_{\mathfrak{a}_{t:m}} \left[\sum_\nu w(\nu \mid \mathfrak{a}_{<t}) \nu^\pi(\mathfrak{a}_{t:m} \mid \mathfrak{a}_{<t}) \right] G_{t-1:m} \\ &= \sum_\nu w(\nu \mid \mathfrak{a}_{<t}) (1 - \gamma) \underbrace{\sum_{\mathfrak{a}_{t:m}} \nu^\pi(\mathfrak{a}_{t:m} \mid \mathfrak{a}_{<t}) G_{t-1:m}}_{= V_\nu^{\pi,m}(\mathfrak{a}_{<t})} . \end{aligned}$$

In the last step we exchanged $\sum_{\mathfrak{a}_{t:m}}$ and $\sum_\nu$; this is valid because both are sums of non-negative terms (or, for finite m , the sum over $\mathfrak{a}_{t:m}$ is finite). So:

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

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 $\mathcal{M}$.

Solution to Exercise 4.2

The left side converges: $V_\xi^{\pi,m}(\mathfrak{a}_{<t}) \rightarrow V_\xi^\pi(\mathfrak{a}_{<t})$ by definition.

For the right side, we need to push $\lim_{m \rightarrow \infty}$ through $\sum_\nu$. If $\mathcal{M}$ is finite this is immediate. For countably infinite $\mathcal{M}$, we apply Fact 4.1 (dominated convergence for sums):

- For each ν : $a_{\nu,m} := w(\nu \mid \mathfrak{a}_{<t}) V_\nu^{\pi,m}(\mathfrak{a}_{<t}) \rightarrow w(\nu \mid \mathfrak{a}_{<t}) V_\nu^\pi(\mathfrak{a}_{<t})$ as $m \rightarrow \infty$ by definition.

- Domination: $|a_{\nu,m}| \leq w(\nu \mid \mathfrak{a}_{<t}) \cdot 1 =: b_\nu$ for all m , since $V_\nu^{\pi,m} \in [0, 1]$ (Exercise 1.3).
- Summability: $\sum_\nu b_\nu = \sum_\nu w(\nu \mid \mathfrak{a}_{<t}) = 1 < \infty$.

So Fact 4.1 gives:

$$V_\xi^\pi(\mathfrak{a}_{<t}) = \lim_{m \rightarrow \infty} \sum_\nu w(\nu \mid \mathfrak{a}_{<t}) V_\nu^{\pi,m}(\mathfrak{a}_{<t}) = \sum_\nu w(\nu \mid \mathfrak{a}_{<t}) V_\nu^\pi(\mathfrak{a}_{<t}).$$

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 .

Solution to Exercise 4.3

By Exercise 2.3, a deterministic Bayes-optimal policy π_ξ^* exists with $V_\xi^{\pi_\xi^*,m} = V_\xi^{*,m}$. Applying Exercise 4.1 with $\pi = \pi_\xi^*$:

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

Since π_ξ^* is optimal for ξ but not necessarily for each individual ν , we have $V_\nu^{\pi_\xi^*,m}(\mathfrak{a}_{<t}) \leq V_\nu^{*,m}(\mathfrak{a}_{<t})$ for each ν . Since $w(\nu \mid \mathfrak{a}_{<t}) \geq 0$:

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

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.

Solution to Exercise 4.4

Coin-flip prediction. Let $\mathcal{A} = \mathcal{O} = \mathcal{R} = \{0, 1\}$, $\mathcal{M} = \{\nu_0, \nu_1\}$ with $w_{\nu_0} = w_{\nu_1} = \frac{1}{2}$, horizon $m = 1$. Environment ν_i always shows outcome i , and the agent is rewarded for a correct prediction:

$$\nu_i(e_t \mid \mathfrak{x}_{<t} a_t) : o_t = i, \quad r_t = \llbracket a_t = i \rrbracket.$$

In ν_i the optimal policy plays $a_t = i$, achieving $V_{\nu_i}^{*,m} = 1$. In $\xi = \frac{1}{2}\nu_0 + \frac{1}{2}\nu_1$ the outcome is a fair coin flip, so no policy predicts better than chance: $V_{\xi}^{*,m} = \frac{1}{2}$. Therefore:

$$\sum_{\nu} w_{\nu} V_{\nu}^{*,m}(\mathfrak{x}_{<t}) = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 1 = 1 > \frac{1}{2} = V_{\xi}^{*,m}(\mathfrak{x}_{<t}).$$

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

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

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

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

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 $\sum_{e_k} \xi(e_k \mid \mathfrak{x}_{<k} a_k) = 1$ and $c + \max(\cdot) = \max(c + \cdot)$.

Solution to Exercise 5.1

Abbreviate the one-step predictor $\xi_k := \xi(e_k \mid \mathfrak{a}_{<k} a_k)$, and write the **expecti-max operator** over steps $i, \dots, j$ as

$$\sum_{i:j}^{\max} := \max_{a_i} \sum_{e_i} \xi_i \max_{a_{i+1}} \sum_{e_{i+1}} \xi_{i+1} \cdots \max_{a_j} \sum_{e_j} \xi_j,$$

the alternating “maximize over the action, average over the percept” chain, with the ξ_k weights built in; the single step is $\sum_k^{\max} := \max_{a_k} \sum_{e_k} \xi_k$. Two facts:

- **(E1) Composition:** $\sum_{i:j}^{\max} = \sum_i^{\max} \sum_{i+1:j}^{\max}$, by nesting the definition.
- **(E2) Affine pass-through:** for any $c \in \mathbb{R}$ and $\lambda \geq 0$ not depending on $(a_i, e_i, \dots, a_j, e_j)$,

$$c + \lambda \sum_{i:j}^{\max} X = \sum_{i:j}^{\max} (c + \lambda X),$$

since each layer has $\sum_{e_k} \xi_k = 1$ (so $\sum_{e_k} \xi_k (c + \lambda Y) = c + \lambda \sum_{e_k} \xi_k Y$) and $\max_a (c + \lambda g(a)) = c + \lambda \max_a g(a)$ for $\lambda \geq 0$.

We prove by backward induction on $t = m, m-1, \dots, 1$ that

$$V_\xi^{*,m}(\mathfrak{a}_{<t}) = (1 - \gamma) \sum_{t:m}^{\max} G_{t-1:m}.$$

Base case $t = m$. Here $G_{m-1:m} = r_m$ and $V_\xi^{*,m}(\mathfrak{a}_{1:m}) = 0$ (the return past the horizon is empty, $G_{m:m} = 0$). The Bellman optimality equation (Exercise 2.4) for $\nu = \xi$ gives

$$V_\xi^{*,m}(\mathfrak{a}_{<m}) = \max_{a_m} \sum_{e_m} \xi_m [(1 - \gamma) r_m + \gamma \cdot 0] = (1 - \gamma) \sum_{m:m}^{\max} G_{m-1:m}.$$

Inductive step. Assume the claim at time $t+1$, i.e. $V_\xi^{*,m}(\mathfrak{a}_{1:t}) = (1 - \gamma) \sum_{t+1:m}^{\max} G_{t:m}$ for every length- t history (recall $\mathfrak{a}_{<t+1} = \mathfrak{a}_{1:t}$). The Bellman optimality equation (Exercise 2.4) at time t reads

$$V_\xi^{*,m}(\mathfrak{a}_{<t}) = \sum_t^{\max} [(1 - \gamma) r_t + \gamma V_\xi^{*,m}(\mathfrak{a}_{1:t})].$$

Substituting the inductive hypothesis and pulling out $(1 - \gamma)$,

$$V_\xi^{*,m}(\mathfrak{a}_{<t}) = (1 - \gamma) \sum_t^{\max} [r_t + \gamma \sum_{t+1:m}^{\max} G_{t:m}].$$

As r_t and γ do not depend on $(a_{t+1}, e_{t+1}, \dots, a_m, e_m)$, property (E2) carries them inside the inner operator, and the return recursion $G_{t-1:m} = r_t + \gamma G_{t:m}$ (Definition 0.3) gives

$$r_t + \gamma \max_{t+1:m} G_{t:m} \stackrel{(E2)}{=} \max_{t+1:m} (r_t + \gamma G_{t:m}) = \max_{t+1:m} G_{t-1:m}.$$

Recombining the two operators by (E1),

$$V_{\xi}^{*,m}(\mathfrak{a}_{<t}) = (1 - \gamma) \max_t \sum_{t+1:m} G_{t-1:m} = (1 - \gamma) \sum_{t:m} G_{t-1:m},$$

which expands back into the $\max / \sum$ layers of the first display.

Infinite horizon and the action. Taking $m \rightarrow \infty$ (Exercises 2.5 and 2.7), $G_{t-1:m} \rightarrow G_{\geq t-1}$ and the chain extends indefinitely, so $V_{\xi}^*(\mathfrak{a}_{<t}) = (1 - \gamma) \sum_{t:\infty} G_{\geq t-1}$. Collecting the per-step factors by the chain rule (Exercise 0.2, iterated), $\prod_{k=t}^m \xi_k = \xi(e_{t:m} \mid \mathfrak{a}_{<t} a_{t:m})$. A Bayes-optimal action maximizes this expression; peeling the outer $\max_{a_t}$ off as an $\arg \max$ and dropping the positive constant $(1 - \gamma)$ (which does not move the maximizer) gives the stated action.

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 $A^- = \Omega \setminus A^+$.

Solution to Exercise 6.1

Write out the expectation difference using Definition 6.2:

$$\mathbb{E}_P[f] - \mathbb{E}_Q[f] = \sum_{\omega \in \Omega} f(\omega)(P(\omega) - Q(\omega)).$$

Define $A^+ := \{\omega \in \Omega : P(\omega) \geq Q(\omega)\}$ and $A^- := \Omega \setminus A^+$, and split the sum:

$$\mathbb{E}_P[f] - \mathbb{E}_Q[f] = \sum_{\omega \in A^+} f(\omega)(P(\omega) - Q(\omega)) + \sum_{\omega \in A^-} f(\omega)(P(\omega) - Q(\omega)).$$

Bounding the sum over A^+ . On A^+ , $P(\omega) - Q(\omega) \geq 0$ and $f(\omega) \leq c$, so:

$$\begin{aligned} \sum_{\omega \in A^+} f(\omega)(P(\omega) - Q(\omega)) &\leq c \sum_{\omega \in A^+} (P(\omega) - Q(\omega)) \\ &= c \cdot (P(A^+) - Q(A^+)) \\ &\leq c \cdot \sup_S |P(S) - Q(S)| \\ &= c \cdot \text{TV}_\Omega(P, Q). \end{aligned}$$

Bounding the sum over A^- . On A^- , $P(\omega) - Q(\omega) < 0$ and $f(\omega) \geq 0$, so every term $f(\omega)(P(\omega) - Q(\omega)) \leq 0$:

$$\sum_{\omega \in A^-} f(\omega)(P(\omega) - Q(\omega)) \leq 0.$$

Combining. $\mathbb{E}_P[f] - \mathbb{E}_Q[f] \leq c \cdot \text{TV}_\Omega(P, Q) + 0 = c \cdot \text{TV}_\Omega(P, Q)$.

The other direction. Swapping P and Q : define $\tilde{A}^+ = \{\omega : Q(\omega) \geq P(\omega)\} = A^-$.

The same argument gives $\mathbb{E}_Q[f] - \mathbb{E}_P[f] \leq c \cdot \text{TV}_\Omega(Q, P) = c \cdot \text{TV}_\Omega(P, Q)$.

Therefore $|\mathbb{E}_P[f] - \mathbb{E}_Q[f]| \leq c \cdot \text{TV}_\Omega(P, Q)$.

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

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

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 $A_1 \subseteq A_2 \subseteq \dots$, then $\nu^\pi(\bigcup_{n=1}^\infty A_n) = \lim_{n \rightarrow \infty} \nu^\pi(A_n)$.

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 .

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 $\mathfrak{a}_{<t}$ of length $t-1$. We write $f_t(\mathfrak{a}_{<t}) \rightarrow 0$ ν^π -**almost surely** (ν^π -a.s.) if

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

This set is built from finite-prefix conditions: $\{f_t \not\rightarrow 0\} = \bigcup_{n=1}^{\infty} \bigcap_{N=1}^{\infty} \bigcup_{t=N}^{\infty} \{|f_t(\mathfrak{a}_{<t})| > \frac{1}{n}\}$, 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:

- $\{|f_t(\mathfrak{a}_{<t})| > \frac{1}{n}\}$ 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*.

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

- $\bigcup_{n=1}^{\infty} \bigcap_{N=1}^{\infty} \bigcup_{t=N}^{\infty} \{|f_t| > \frac{1}{n}\}$: “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 vanish has μ^π -probability zero:*

$$\mu^\pi\left(\left\{\mathfrak{x}_{1:\infty} : \lim_{t \rightarrow \infty} (V_\xi^\pi(\mathfrak{x}_{<t}) - V_\mu^\pi(\mathfrak{x}_{<t})) \neq 0\right\}\right) = 0.$$

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 :

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

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

Solution to Exercise 7.1

We want to apply Section 6. Identify:

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

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.

- $P := \mu^\pi(\cdot \mid \mathfrak{a}_{<t})$ and $Q := \xi^\pi(\cdot \mid \mathfrak{a}_{<t})$, the conditional measures over Ω .
- $f(\mathfrak{a}_{t:m}) := (1 - \gamma)G_{t-1:m}$, which satisfies $f \in [0, 1]$ by Exercise 1.3 (so $c = 1$).

Then $V_\nu^{\pi,m}(\mathfrak{a}_{<t}) = \sum_{\mathfrak{a}_{t:m}} \nu^\pi(\mathfrak{a}_{t:m} \mid \mathfrak{a}_{<t}) f(\mathfrak{a}_{t:m})$ is exactly $\mathbb{E}_P[f]$ (for $\nu = \mu$) or $\mathbb{E}_Q[f]$ (for $\nu = \xi$). Applying Section 6:

$$|V_\xi^{\pi,m}(\mathfrak{a}_{<t}) - V_\mu^{\pi,m}(\mathfrak{a}_{<t})| = |\mathbb{E}_Q[f] - \mathbb{E}_P[f]| \leq 1 \cdot \text{TV}_{\mathcal{H}^{m-t+1}}(P, Q).$$

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

Hint: Use Section 3.

Solution to Exercise 7.2

By Section 3, $\xi^\pi(\mathfrak{a}_{1:m}) \geq w_\mu \cdot \mu^\pi(\mathfrak{a}_{1:m})$ for all histories and all m . So for any event A : $\mu^\pi(A) > 0 \implies \xi^\pi(A) \geq w_\mu \cdot \mu^\pi(A) > 0$. Hence ξ^π covers μ^π .

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 $\sup_S |P(S \mid \mathfrak{a}_{<t}) - Q(S \mid \mathfrak{a}_{<t})| \rightarrow 0$ 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.

Solution to Exercise 7.3

Step 1: Finite-horizon bound. By Exercise 7.1, for each finite m :

$$\begin{aligned} |V_\xi^{\pi,m}(\mathfrak{a}_{<t}) - V_\mu^{\pi,m}(\mathfrak{a}_{<t})| &\leq \text{TV}_{\mathcal{H}^{m-t+1}}(\xi^\pi(\cdot \mid \mathfrak{a}_{<t}), \mu^\pi(\cdot \mid \mathfrak{a}_{<t})) \\ &\leq \sup_S |\xi^\pi(S \mid \mathfrak{a}_{<t}) - \mu^\pi(S \mid \mathfrak{a}_{<t})|, \end{aligned}$$

where the second inequality uses that the sup over subsets of $\mathcal{H}^{m-t+1}$ is at most the sup over all measurable events.

Step 2: Take $m \rightarrow \infty$. The LHS converges to $|V_\xi^\pi(\mathfrak{x}_{<t}) - V_\mu^\pi(\mathfrak{x}_{<t})|$ (by definition of the infinite-horizon value as the pointwise limit). The RHS does not depend on m . Hence

$$|V_\xi^\pi(\mathfrak{x}_{<t}) - V_\mu^\pi(\mathfrak{x}_{<t})| \leq \sup_S |\xi^\pi(S \mid \mathfrak{x}_{<t}) - \mu^\pi(S \mid \mathfrak{x}_{<t})|.$$

Step 3: Take $t \rightarrow \infty$. By Exercise 7.2, ξ^π covers μ^π . By Fact 7.7 (Blackwell–Dubins) with $P = \mu^\pi$ and $Q = \xi^\pi$, the RHS tends to 0 as $t \rightarrow \infty$, μ^π -a.s. Therefore

$$|V_\xi^\pi(\mathfrak{x}_{<t}) - V_\mu^\pi(\mathfrak{x}_{<t})| \rightarrow 0 \quad \mu^\pi\text{-a.s.}$$

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 $V_\mu^*(\mathfrak{x}_{<t}) > \varepsilon > 0$ for all t along the history generated by μ and π_ξ^* , then $V_\xi^*(\mathfrak{x}_{<t}) \geq w_\mu \varepsilon > 0$ for all t . [HQC24, Theorem 7.4.10]*

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

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

Solution to Exercise 8.1

By definition, $V_\xi^*(\mathfrak{x}_{<t}) = \sup_\pi V_\xi^\pi(\mathfrak{x}_{<t}) \geq V_\xi^{\pi'}(\mathfrak{x}_{<t})$ for any policy π' . By Exercise 4.2 (linearity of V^π in the environment):

$$V_\xi^{\pi'}(\mathfrak{x}_{<t}) = \sum_{\nu \in \mathcal{M}} w(\nu \mid \mathfrak{x}_{<t}) V_\nu^{\pi'}(\mathfrak{x}_{<t}).$$

Choose $\pi' = \pi_\mu^*$, the deterministic optimal policy for μ whose existence is guaranteed by Exercise 2.7:

$$V_\xi^*(\mathfrak{x}_{<t}) \geq \sum_{\nu \in \mathcal{M}} w(\nu \mid \mathfrak{x}_{<t}) V_\nu^{\pi_\mu^*}(\mathfrak{x}_{<t}).$$

Every term is non-negative ($w(\nu \mid \mathfrak{x}_{<t}) \geq 0$ and $V_\nu^{\pi_\mu^*} \geq 0$). Dropping all terms except $\nu = \mu$:

$$V_\xi^*(\mathfrak{x}_{<t}) \geq w(\mu \mid \mathfrak{x}_{<t}) V_\mu^{\pi_\mu^*}(\mathfrak{x}_{<t}) = w(\mu \mid \mathfrak{x}_{<t}) V_\mu^*(\mathfrak{x}_{<t}).$$

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

Solution to Exercise 8.2

From Definition 0.1, the posterior weight is:

$$w(\mu \mid \mathfrak{x}_{<t}) = w_\mu \cdot \frac{\mu(\mathfrak{x}_{<t})}{\xi(\mathfrak{x}_{<t})} = w_\mu \cdot \frac{\prod_{k=1}^{t-1} \mu(e_k \mid \mathfrak{x}_{<k} a_k)}{\prod_{k=1}^{t-1} \xi(e_k \mid \mathfrak{x}_{<k} a_k)}.$$

Since μ is deterministic, for the history that μ actually generates, $\mu(e_k \mid \mathfrak{x}_{<k} a_k) = 1$ for each k (the environment produces each percept with certainty). So:

$$\mu(\mathfrak{x}_{<t}) = \prod_{k=1}^{t-1} 1 = 1.$$

Meanwhile, $\xi(e_k \mid \mathfrak{x}_{<k} a_k) \leq 1$ for each k (it is a probability), so $\xi(\mathfrak{x}_{<t}) = \prod_{k=1}^{t-1} \xi(e_k \mid \mathfrak{x}_{<k} a_k) \leq 1$.

Therefore $w(\mu \mid \mathfrak{x}_{<t}) = w_\mu \cdot 1 / \xi(\mathfrak{x}_{<t}) \geq w_\mu$.

Combining with Exercise 8.1:

$$V_\xi^*(\mathfrak{x}_{<t}) \geq w(\mu \mid \mathfrak{x}_{<t}) \cdot V_\mu^*(\mathfrak{x}_{<t}) \geq w_\mu \cdot V_\mu^*(\mathfrak{x}_{<t}) > w_\mu \cdot \varepsilon > 0.$$

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 $\mu \in \mathcal{M}$, $V_{\mu}^*(\mathfrak{a}_{<t}) - V_{\mu}^{\pi_{\xi}^*}(\mathfrak{a}_{<t}) \rightarrow 0$ μ^{π} -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 $\mathfrak{x}_{1:t}$, 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 “ $X_{\nu,t}(\mathfrak{x}_{1:t}) \geq c$ ” or “ $X_{\nu,t}(\mathfrak{x}_{1:t}) \rightarrow X_{\nu,\infty}(\mathfrak{x}_{1:\infty})$ ” 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: $\mathbb{E}_{\mu^\pi}[X_{\nu,t} \mid \mathfrak{x}_{<t}] = X_{\nu,t-1}$.

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

Solution to Exercise 9.1

We need to compute $\mathbb{E}_{\mu^\pi}[X_{\nu,t} \mid \mathfrak{x}_{<t}]$. Conditioning on $\mathfrak{x}_{<t}$ means a_t and e_t are the random quantities (sampled from π and μ respectively). First, write $X_{\nu,t}$ pointwise as a function of history in terms of $X_{\nu,t-1}$:

$$\begin{aligned} X_{\nu,t}(\mathfrak{x}_{1:t}) &= \frac{\nu(\mathfrak{x}_{1:t})}{\mu(\mathfrak{x}_{1:t})} = \frac{\nu(\mathfrak{x}_{<t})}{\mu(\mathfrak{x}_{<t})} \cdot \frac{\nu(e_t \mid \mathfrak{x}_{<t}a_t)}{\mu(e_t \mid \mathfrak{x}_{<t}a_t)} \\ &= X_{\nu,t-1}(\mathfrak{x}_{<t}) \cdot \frac{\nu(e_t \mid \mathfrak{x}_{<t}a_t)}{\mu(e_t \mid \mathfrak{x}_{<t}a_t)}, \end{aligned}$$

using the chain rule (Exercise 0.2) for both ν and μ . Now take the conditional expectation. Since $X_{\nu,t-1}$ depends only on $\mathfrak{x}_{<t}$, it comes out of the expectation:

$$\begin{aligned} \mathbb{E}_{\mu^\pi}[X_{\nu,t} \mid \mathfrak{x}_{<t}] &= X_{\nu,t-1} \cdot \mathbb{E}_{\mu^\pi} \left[\frac{\nu(e_t \mid \mathfrak{x}_{<t}a_t)}{\mu(e_t \mid \mathfrak{x}_{<t}a_t)} \mid \mathfrak{x}_{<t} \right] \\ &= X_{\nu,t-1} \sum_{a_t \in \mathcal{A}} \pi(a_t \mid \mathfrak{x}_{<t}) \sum_{e_t \in \mathcal{E}} \mu(e_t \mid \mathfrak{x}_{<t}a_t) \cdot \frac{\nu(e_t \mid \mathfrak{x}_{<t}a_t)}{\mu(e_t \mid \mathfrak{x}_{<t}a_t)} \\ &= X_{\nu,t-1} \sum_{a_t} \pi(a_t \mid \mathfrak{x}_{<t}) \sum_{e_t} \nu(e_t \mid \mathfrak{x}_{<t}a_t). \end{aligned}$$

In the last step, $\mu(e_t \mid \mathfrak{x}_{<t}a_t)$ cancels. Since ν is a probability measure, $\sum_{e_t} \nu(e_t \mid \mathfrak{x}_{<t}a_t) = 1$ for every $\mathfrak{x}_{<t}, a_t$; and $\sum_{a_t} \pi(a_t \mid \mathfrak{x}_{<t}) = 1$. Therefore:

$$\mathbb{E}_{\mu^\pi}[X_{\nu,t} \mid \mathfrak{x}_{<t}] = X_{\nu,t-1} \cdot 1 \cdot 1 = X_{\nu,t-1}.$$

Since $X_{\nu,t} \geq 0$ (a ratio of non-negative quantities), $(X_{\nu,t})_{t \geq 0}$ is a non-negative μ^π -martingale, in particular, a non-negative μ^π -supermartingale.

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

Solution to Exercise 9.2

$X_{\nu,t}$ is a non-negative μ^π -supermartingale by Exercise 9.1. By Fact 9.4 (Doob's convergence theorem), $X_{\nu,t}$ converges μ^π -a.s. to a finite limit: $X_{\nu,t} \rightarrow X_{\nu,\infty} < \infty$.

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 $X_{\xi,t} \geq w_\mu > 0$.

Solution to Exercise 9.3

From the definition:

$$X_{\xi,t} := \frac{\xi(\mathfrak{a}_{1:t})}{\mu(\mathfrak{a}_{1:t})} = \frac{\sum_\nu w_\nu \nu(\mathfrak{a}_{1:t})}{\mu(\mathfrak{a}_{1:t})} = \sum_\nu w_\nu \cdot \frac{\nu(\mathfrak{a}_{1:t})}{\mu(\mathfrak{a}_{1:t})} = \sum_\nu w_\nu X_{\nu,t}.$$

By Section 3: $\xi(\mathfrak{a}_{1:t}) \geq w_\mu \cdot \mu(\mathfrak{a}_{1:t})$, so $X_{\xi,t} = \xi(\mathfrak{a}_{1:t})/\mu(\mathfrak{a}_{1:t}) \geq w_\mu > 0$.

Since $\mathcal{M}$ is finite, $X_{\xi,t} = \sum_\nu w_\nu X_{\nu,t}$ is a finite sum of convergent sequences, so $X_{\xi,t} \rightarrow X_{\xi,\infty} := \sum_\nu w_\nu X_{\nu,\infty} < \infty$ μ^π -a.s., with $X_{\xi,\infty} \geq w_\mu > 0$.

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 $\nu^\pi/\mu^\pi = \nu/\mu = X_{\nu,m}$.

Solution to Exercise 10.1

By Exercise 0.1, $\nu^\pi(\mathfrak{a}'_{1:m}) = \pi(\mathfrak{a}'_{1:m}) \cdot \nu(\mathfrak{a}'_{1:m})$ and $\mu^\pi(\mathfrak{a}'_{1:m}) = \pi(\mathfrak{a}'_{1:m}) \cdot \mu(\mathfrak{a}'_{1:m})$. For any history $\mathfrak{a}'_{1:m}$ with $\mu^\pi(\mathfrak{a}'_{1:m}) > 0$, the policy factors cancel:

$$\begin{aligned} \nu^\pi(\mathfrak{a}'_{1:m}) &= \frac{\nu^\pi(\mathfrak{a}'_{1:m})}{\mu^\pi(\mathfrak{a}'_{1:m})} \cdot \mu^\pi(\mathfrak{a}'_{1:m}) = \frac{\nu(\mathfrak{a}'_{1:m})}{\mu(\mathfrak{a}'_{1:m})} \cdot \mu^\pi(\mathfrak{a}'_{1:m}) \\ &= X_{\nu,m}(\mathfrak{a}'_{1:m}) \cdot \mu^\pi(\mathfrak{a}'_{1:m}). \end{aligned}$$

For histories with $\mu^\pi(\mathfrak{a}'_{1:m}) = 0$: since $\mu^\pi = \pi \cdot \mu$, some $\pi(a_k \mid \mathfrak{a}'_{<k}) = 0$, which forces $\nu^\pi(\mathfrak{a}'_{1:m}) = 0$ too (the same π -factor appears in $\nu^\pi = \pi \cdot \nu$). So both sides

are zero.

Summing over $\mathfrak{a}'_{1:m} \in A$:

$$\begin{aligned}\nu^\pi[\mathfrak{a}_{1:m} \in A] &= \sum_{\mathfrak{a}'_{1:m} \in A} \nu^\pi(\mathfrak{a}'_{1:m}) = \sum_{\mathfrak{a}'_{1:m} \in A} X_{\nu,m}(\mathfrak{a}'_{1:m}) \cdot \mu^\pi(\mathfrak{a}'_{1:m}) \\ &= \mathbb{E}_{\mu^\pi}[X_{\nu,m} \cdot \mathbb{I}[\mathfrak{a}_{1:m} \in A]].\end{aligned}$$

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.

Solution to Exercise 10.3

Let $E_\varepsilon := E \cap \{X_{\nu,\infty} \geq \varepsilon\}$. On E_ε , $X_{\nu,\infty} \geq \varepsilon$, so pointwise

$$X_{\nu,\infty} \cdot \mathbb{I}[\mathfrak{a}_{1:\infty} \in E_\varepsilon] \geq \varepsilon \cdot \mathbb{I}[\mathfrak{a}_{1:\infty} \in E_\varepsilon].$$

Take μ^π -expectations and apply Exercise 10.2:

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

Since $E_\varepsilon \subseteq E$, $\nu^\pi[E_\varepsilon] \leq \nu^\pi[E]$. Dividing by ε gives $\mu^\pi[E_\varepsilon] \leq \nu^\pi[E]/\varepsilon$.

Interpretation. A likelihood ratio of at least ε forces ν^π and μ^π measures of an event to be comparable: μ^π of the event can exceed ν^π of it only by the factor $1/\varepsilon$. In particular, if ν^π vanishes on E , so does μ^π on the part where the likelihood ratio stays bounded away from zero.

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**

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

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 $\delta_{\nu,t}(\mathfrak{a}_{<t}) \rightarrow 0$ 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:

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

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

Solution to Exercise 11.1

We establish two inequalities and chain them.

First inequality: isolate the μ -term. For each $\nu \in \mathcal{M}$, the optimal value is at least the value of any policy: $V_\nu^*(\mathfrak{a}_{<t}) \geq V_\nu^{\pi_\xi^*}(\mathfrak{a}_{<t})$, so $V_\nu^* - V_\nu^{\pi_\xi^*} \geq 0$. Since the posterior weights $w(\nu \mid \mathfrak{a}_{<t}) \geq 0$, every term in $\sum_\nu w(\nu \mid \mathfrak{a}_{<t})[V_\nu^* - V_\nu^{\pi_\xi^*}]$ is non-negative. The μ -term is one such term:

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

Second inequality: replace π_ξ^ with $\tilde{\pi}$.* By definition of the optimal value, $V_\xi^*(\mathfrak{a}_{<t}) \geq V_\xi^{\tilde{\pi}}(\mathfrak{a}_{<t})$. Expanding both sides using infinite-horizon linearity (Ex-

ercise 4.2, applied to the fixed policies π_ξ^* and $\tilde{\pi}$:

$$\begin{aligned} \sum_{\nu} w(\nu \mid \mathfrak{a}_{<t}) V_{\nu}^{\pi_\xi^*}(\mathfrak{a}_{<t}) &= V_{\xi}^*(\mathfrak{a}_{<t}) \geq V_{\xi}^{\tilde{\pi}}(\mathfrak{a}_{<t}) \\ &= \sum_{\nu} w(\nu \mid \mathfrak{a}_{<t}) V_{\nu}^{\tilde{\pi}}(\mathfrak{a}_{<t}). \end{aligned}$$

Rearranging: $\sum_{\nu} w(\nu)[V_{\nu}^* - V_{\nu}^{\pi_\xi^*}] \leq \sum_{\nu} w(\nu)[V_{\nu}^* - V_{\nu}^{\tilde{\pi}}] = \sum_{\nu} w(\nu) \delta_{\nu,t}$.

Chaining:

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

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

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

Solution to Exercise 11.2

From the posterior weight formula, dividing numerator and denominator by $\mu(\mathfrak{a}_{<t})$:

$$w(\nu \mid \mathfrak{a}_{<t}) = w_{\nu} \cdot \frac{\nu(\mathfrak{a}_{<t})}{\xi(\mathfrak{a}_{<t})} = w_{\nu} \cdot \frac{\nu(\mathfrak{a}_{<t})/\mu(\mathfrak{a}_{<t})}{\xi(\mathfrak{a}_{<t})/\mu(\mathfrak{a}_{<t})} = \frac{w_{\nu} \cdot X_{\nu,t-1}}{X_{\xi,t-1}},$$

where $X_{\nu,t-1}$ and $X_{\xi,t-1}$ are evaluated at the length- $(t-1)$ history $\mathfrak{a}_{<t}$.

If $X_{\nu,\infty} = 0$, then $X_{\nu,t-1} \rightarrow 0$ μ^π -a.s. (Exercise 9.2). Since $X_{\xi,t-1} \geq w_{\mu} > 0$ always (Exercise 9.3), the ratio $w(\nu \mid \mathfrak{a}_{<t}) = w_{\nu} X_{\nu,t-1} / X_{\xi,t-1} \rightarrow 0$ μ^π -a.s.

Exercise 11.3 (Transferring convergence). [20] Let $F := \{\mathfrak{a}_{1:\infty} : \delta_{\nu,t}(\mathfrak{a}_{<t}) \not\rightarrow 0\}$, 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).

Solution to Exercise 11.3

Step 1: F is ν^π -null. By the self-optimizing assumption (Definition 9.1), under ν^π the suboptimality gap $\delta_{\nu,t}$ vanishes, so $\nu^\pi[F] = 0$.

Step 2: Apply Exercise 10.3. For each $n \geq 1$, take $E = F$ and $\varepsilon = 1/n$:

$$\mu^\pi[F \cap \{X_{\nu,\infty} \geq 1/n\}] \leq \frac{\nu^\pi[F]}{1/n} = 0.$$

Step 3: Countable union. The level sets of $X_{\nu,\infty}$ form a nested increasing family: $\{X_{\nu,\infty} > 0\} = \bigcup_{n \geq 1} \{X_{\nu,\infty} \geq 1/n\}$. Intersecting with F gives $F \cap \{X_{\nu,\infty} > 0\} = \bigcup_{n \geq 1} (F \cap \{X_{\nu,\infty} \geq 1/n\})$, a countable union. By countable subadditivity (Definition 7.1):

$$\begin{aligned} \mu^\pi[F \cap \{X_{\nu,\infty} > 0\}] &= \mu^\pi\left[\bigcup_n F \cap \{X_{\nu,\infty} \geq 1/n\}\right] \\ &\leq \sum_{n \geq 1} \mu^\pi[F \cap \{X_{\nu,\infty} \geq 1/n\}] = 0. \end{aligned}$$

What this says. “ $\delta_{\nu,t} \rightarrow 0$ μ^π -a.s. on $\{X_{\nu,\infty} > 0\}$ ” is shorthand for exactly this event statement: among infinite histories on which the likelihood ratio stays bounded away from zero, all but a μ^π -null subset are ones where $\delta_{\nu,t}(\mathbf{x}_{<t}) \rightarrow 0$. We do *not* claim convergence on histories where $X_{\nu,\infty} = 0$: that case is handled separately in Exercise 11.2 via the posterior weight.

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

Solution to Exercise 11.4

Fix $\nu \in \mathcal{M}$. We show $w(\nu \mid \mathbf{x}_{<t})\delta_{\nu,t} \rightarrow 0$ μ^π -a.s. by considering two exhaustive cases:

Case 1: $X_{\nu,\infty} = 0$. By Exercise 11.2, $w(\nu \mid \mathbf{x}_{<t}) \rightarrow 0$. Since $\delta_{\nu,t} \in [0, 1]$ (Exercise 1.3), the product $w(\nu \mid \mathbf{x}_{<t}) \cdot \delta_{\nu,t} \leq 1 \cdot w(\nu \mid \mathbf{x}_{<t}) \rightarrow 0$.

Case 2: $X_{\nu,\infty} > 0$. By Exercise 11.3, $\delta_{\nu,t} \rightarrow 0$. Since $w(\nu \mid \mathbf{x}_{<t}) \leq 1$ (posterior weights are at most 1), the product $w(\nu \mid \mathbf{x}_{<t}) \cdot \delta_{\nu,t} \leq 1 \cdot \delta_{\nu,t} \rightarrow 0$.

These two cases cover all infinite histories (μ^π -a.s.), so $w(\nu \mid \mathbf{x}_{<t})\delta_{\nu,t} \rightarrow 0$ 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: $w(\mu \mid \mathbf{x}_{<t}) = w_\mu / X_{\xi,t-1}$.

Solution to Exercise 11.5

Since $\mathcal{M} = \{\nu_1, \dots, \nu_N\}$ is finite, we can sum Exercise 11.4 over all $\nu \in \mathcal{M}$:

$$\sum_{\nu \in \mathcal{M}} w(\nu \mid \mathfrak{x}_{<t}) \delta_{\nu,t} \longrightarrow 0 \quad \mu^\pi\text{-a.s.}$$

From Exercise 11.1:

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

So $w(\mu \mid \mathfrak{x}_{<t}) [V_\mu^* - V_\mu^{\pi_\xi^*}] \rightarrow 0$ μ^π -a.s.

It remains to divide by $w(\mu \mid \mathfrak{x}_{<t})$. From Exercise 9.3:

$$w(\mu \mid \mathfrak{x}_{<t}) = \frac{w_\mu}{X_{\xi,t-1}}.$$

Fix an infinite history outside the (single) μ^π -null set on which $X_{\xi,t} \rightarrow X_{\xi,\infty} < \infty$ fails. Along this history:

- $X_{\xi,t-1} \rightarrow X_{\xi,\infty}$, a finite positive number (with $X_{\xi,\infty} \geq w_\mu > 0$ by Exercise 9.3).
- Hence $w(\mu \mid \mathfrak{x}_{<t}) = w_\mu / X_{\xi,t-1} \rightarrow w_\mu / X_{\xi,\infty} > 0$, so eventually $w(\mu \mid \mathfrak{x}_{<t}) \geq w_\mu / (2X_{\xi,\infty}) > 0$.

Dividing the convergence $w(\mu \mid \mathfrak{x}_{<t}) (V_\mu^* - V_\mu^{\pi_\xi^*}) \rightarrow 0$ by this positive (per-history) lower bound, and using $V_\mu^* - V_\mu^{\pi_\xi^*} \geq 0$:

$$V_\mu^*(\mathfrak{x}_{<t}) - V_\mu^{\pi_\xi^*}(\mathfrak{x}_{<t}) \longrightarrow 0 \quad \mu^\pi\text{-a.s.}$$

This completes the proof of Theorem 9.2.

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.

Problems marked (*) are enrichment: they are off the critical path to the three main results (Theorem 7.6, Theorem 8.1, Theorem 9.2) and can be skipped on a first pass without loss of continuity.

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>.