A Model-Free Universal AI

Yegon Kim[1] **Juho Lee**[1]

[1] Graduate School of AI, KAIST, Seoul, South Korea

Abstract

In general reinforcement learning, all established optimal agents, including AIXI, are model-based, explicitly maintaining and using environment models. This paper introduces Universal AI with Q-Induction (AIQI), the first model-free agent proven to be asymptotically ε-optimal in general RL. AIQI performs universal induction over distributional action-value functions, instead of policies or environments like previous works. Under a grain of truth condition, we prove that AIQI is strong asymptotically ε-optimal and asymptotically ε-Bayes-optimal. We also apply our novel proof techniques to show asymptotic ε-optimality of Self-AIXI without any ad-hoc assumptions. Our results significantly expand the diversity of known universal agents.

1 INTRODUCTION

The theory of general reinforcement learning [GRL; Lattimore, 2014] provides a framework for studying agents in the most general class of environments, those that satisfy only minimal structural assumptions. In this setting, AIXI [Hutter, 2005] is a foundational theoretical model: it combines universal induction [Solomonoff, 1964a,b] with sequential decision theory [Von Neumann and Morgenstern, 1944, Bellman, 1957] to define a Bayes-optimal agent. Although AIXI is uncomputable, it serves as a useful theoretical model of powerful goal-driven AIs [Orseau and Ring, 2012a, Everitt et al., 2016], and has also been used to define a universal measure of intelligence, called the Legg-Hutter intelligence [Legg and Hutter, 2007].

Interestingly, AIXI and all other established optimal agents in GRL [Lattimore and Hutter, 2014, Leike et al., 2016a, Cohen et al., 2019, Catt et al., 2023], otherwise known as universal agents or universal AI, are *model-based*: they explicitly infer and make use of a model of the environment (a world model). In contrast, the dominant paradigm in practical RL is *model-free*, learning value functions or policies directly from experience [Watkins, 1989, Williams, 1992, Rummery and Niranjan, 1994, Konda and Tsitsiklis, 1999]. Showing that a model-free algorithm is optimal in GRL has therefore remained elusive and has been repeatedly highlighted as an open problem [Everitt and Hutter, 2018b, Catt, 2022, Hutter et al., 2024].

In this paper, we present Universal **AI** with **Q**-**I**nduction (AIQI), a model-free agent that performs universal induction over *return-predictors*, objects similar to distributional Q-values [Bellemare et al., 2017]. AIQI is essentially an ε-greedy, on-policy distributional Monte Carlo control algorithm. Under a grain of truth condition [Kalai and Lehrer, 1993], we prove that AIQI achieves strong asymptotic ε-optimality and asymptotic ε-Bayes-optimality in GRL. Our proof techniques can also be used to show asymptotic ε-optimality of Self-AIXI without incurring ad-hoc assumptions such as in the proof by Catt et al. [2023]. Our results significantly expand the class of known universal agents, and provide a blueprint for the analysis of other policy iteration algorithms in general environments.

2 BACKGROUND

2.1 GENERAL REINFORCEMENT LEARNING

We establish the basic notation for general reinforcement learning. We also provide a summary of all the important symbols in § A. Let $\mathcal{A}$, $\mathcal{O}$, and $\mathcal{R} \subseteq [0,1]$ denote the finite sets of actions, observations, and rewards, respectively. The percept space is $\mathcal{E} := \mathcal{O} \times \mathcal{R}$, and a history $h_{1:t} \in \mathcal{H} := (\mathcal{A} \times \mathcal{E})^*$ is a sequence of actions and percepts. We write $h_{1:t} = a_1 e_1 \dots a_t e_t$ for the history up to time t. $\mathcal{H}_t := (\mathcal{A} \times \mathcal{E})^t$ denotes the set of all histories up to time t. A policy $\pi : \mathcal{H} \to \Delta\mathcal{A}$ maps histories to action distributions, while an environment $\nu : \mathcal{H} \times \mathcal{A} \to \Delta\mathcal{E}$ maps history-action pairs to percept distributions. When policy π interacts with

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:3012–3035.

environment ν, they induce a distribution ν^π over $\mathcal{H}_t$:

$$\begin{aligned}\nu^\pi(h_{1:t}) &:= \pi(a_1)\,\nu(e_1|a_1)\,\cdots\,\pi(a_t|h_{<t})\,\nu(e_t|h_{<t}a_t)\\ &= \textstyle\prod_{1\le i\le t}\pi(a_i \mid h_{<i})\prod_{1\le i\le t}\nu(e_i \mid h_{<i}a_i)\\ &= \pi(a_{1:t} \parallel e_{<t})\,\nu(e_{1:t} \parallel a_{1:t})\end{aligned}$$

This is a valid distribution, that is, the probabilities sum to 1 across all histories $h_{1:t}$. A direct consequence is that for a deterministic policy π',

$$\sum_{e_{1:t}\in\mathcal{E}^t} \nu\left(e_{1:t} \parallel \pi'\left(e_{<t}\right)\right) = 1, \tag{1}$$

where $\pi'(e_{<t})$ is the sequence of actions $a_{1:t}$ taken by π' under fixed percepts $e_{<t}$. We also define the sample space $\Omega := (\mathcal{A}\times\mathcal{E})^\infty$, which consists of all *infinite* action-percept sequences h. Every history $h_{1:t}$ is a prefix of some outcome h.

Given a discount factor $\gamma \in (0,1)$, the (full) *return* at time t is defined as

$$R_t := (1-\gamma)\sum_{k=0}^{\infty}\gamma^k r_{t+k},$$

where $1-\gamma = \left(\sum_{k=0}^{\infty}\gamma^k\right)^{-1}$ ensures that returns lie in $[0,1]$, like the rewards. This definition assumes a geometric discount, and we discuss the generalization to general discount sequences in § E.

The *value function* for policy π in environment ν is

$$V_\nu^\pi(h_{<t}) := \mathbb{E}_\nu^\pi\left[R_t \mid h_{<t}\right],$$

where $\mathbb{E}_\nu^\pi$ is the expectation with respect to ν^π. The optimal value is defined as $V_\nu^* = \sup_\pi V_\nu^\pi$, and an optimal policy π_ν^* is any policy that achieves this supremum. The action-value, or Q-value function, is $Q_\nu^\pi(h_{<t},a_t) := \mathbb{E}_\nu^\pi\left[R_t \mid h_{<t}a_t\right]$.

Value functions satisfy the Bellman equations, which allow us to express one with the other:

$$\begin{aligned}V_\nu^\pi(h_{<t}) &= \mathbb{E}_{a_t\sim\pi(\cdot|h_{<t})}\left[Q_\nu^\pi(h_{<t},a_t)\right]\\ Q_\nu^\pi(h_{<t},a_t) &= \mathbb{E}_{e_t'\sim\nu(\cdot|h_{<t},a_t)}\left[(1-\gamma)r_t + \gamma V_\nu^\pi(h'_{<t+1})\right]\end{aligned}$$

Obtaining the full return requires simulating or experiencing an infinite length trajectory. Given some horizon length $H\in\mathbb{Z}_+$, the H-step return is defined as

$$R_{t,H} := (1-\gamma)\sum_{k=0}^{H-1}\gamma^k r_{t+k},$$

which can be computed with H steps of agent-environment interactions. The H-step value function is defined as $V_{\nu,H}^\pi(h_{<t}) := \mathbb{E}_\nu^\pi\left[R_{t,H} \mid h_{<t}\right]$. Finally, we define the effective horizon as follows.

Definition 2.1 (Effective horizon)**.** *For $\eta > 0$, the η-effective horizon is*

$$H(\eta) := \min\left\{H\in\mathbb{Z}_+ \;\middle|\; (1-\gamma)\sum_{k=H}^{\infty}\gamma^k \le \eta\right\}.$$

One can show that the $H(\eta)$-step return differs from the full return by at most η, and likewise for the value function: $R_t - R_{t,H(\eta)} = (1-\gamma)\sum_{k=H(\eta)}^{\infty}\gamma^k r_{t+k} \le \eta$.

2.2 AIXI

To define AIXI, we first need to introduce the mixture environment ξ.

Definition 2.2 (Mixture environment)**.** *For a countable environment class $\mathcal{M}$ and a prior distribution $w(\nu) > 0$ over the environments ν in $\mathcal{M}$, the mixture environment ξ is defined as the unique environment that satisfies*

$$\xi(e_{1:t} \parallel a_{1:t}) = \sum_{\nu\in\mathcal{M}} w(\nu)\,\nu(e_{1:t} \parallel a_{1:t}).$$

Equivalently, $\xi^\pi(\cdot) = \sum_\nu w(\nu)\nu^\pi(\cdot)$ for any policy π.

AIXI is then defined as the optimal policy in the mixture environment ξ. By its definition, AIXI is the Bayes-optimal agent with respect to the prior $w(\nu)$.

Definition 2.3 (AIXI)**.** *AIXI is the optimal policy π_ξ^* in the mixture environment ξ:*

$$\pi_\xi^*(h_{<t}) := \arg\max_a Q_\xi^*(h_{<t},a).$$

3 UNIVERSAL AI WITH Q-INDUCTION

We present Universal AI with Q-Induction (AIQI), a model-free approach to general reinforcement learning.

For a discretization level $M\in\mathbb{Z}_+$, the M-discretized H-step return at time t is defined as

$$z_t := \frac{\lfloor MR_{t,H}\rfloor}{M}, \quad z_t \in \mathcal{Z} := \left\{0, \frac{1}{M}, \ldots, \frac{M-1}{M}\right\}.$$

Note that z_t is an approximation of $R_{t,H}$ whose error gets smaller as $M\to\infty$. More precisely, it always holds that $R_{t,H} - z_t < 1/M$. We choose z_t as the target of prediction instead of $R_{t,H}$. Although this choice isn't necessary for our case of geometric discounting, there exist general discount sequences (§ E) where $R_{t,H}$ can take an infinite number of possible values, which would complicate our analysis, and also the implementation.

We would like to be able to predict the discretized return z_t conditioned on any history $h_{<t}$ and action a_t. One can naturally consider augmenting the history with returns,

$$\ldots a_{t-3}z_{t-3}e_{t-3}a_{t-2}z_{t-2}e_{t-2}a_{t-1}z_{t-1}e_{t-1}a_t,$$

which would be passed in to a Bayesian sequence predictor to predict z_t. However, a complication arises: computing the ground truth return z_{t-1} requires the yet to be observed reward r_t in e_t. Similarly, the returns $z_{t-H+1}, \dots, z_{t-1}$ are all unavailable at time t. Therefore, instead of augmenting the sequence with returns at every step, we augment with a period $N \geq H$, at only the positions $i \equiv t \pmod N$:

$$\dots a_{t-2N} z_{t-2N} e_{t-2N} \dots a_{t-N} z_{t-N} e_{t-N} \dots a_{t-1} e_{t-1} a_t.$$

Then, the augmented returns require rewards at times only up to $t - N + H - 1 \leq t - 1$, which are all available. Why we don't simply consider $N = H$ will be revealed in the proof of Lemma 4.2. The periodically augmented sequence can be passed in to a Bayesian sequence predictor, to predict the next return z_t. We formalize this below.

Given an augmentation *period* $N \geq H$, the augmented sample space associated with *phase* $n \in \{0, \dots, N-1\}$, is

$$\tilde{\Omega}^{(n)} := \prod_{i=1}^{\infty} \mathcal{B}_i^{(n)}, \quad \mathcal{B}_i^{(n)} := \begin{cases} \mathcal{A} \times \mathcal{Z} \times \mathcal{E}, & i \equiv n \pmod N, \\ \mathcal{A} \times \mathcal{E}, & \text{otherwise.} \end{cases}$$

In other words, an augmented outcome $\tilde{h}^{(n)} \in \tilde{\Omega}^{(n)}$ contains some $\tilde{z}_{i_k} \in \mathcal{Z}$ at the periodic positions $i_k = n + kN$, in between a_{i_k} and e_{i_k}. Note that these augmented returns do not necessarily have to match the ground truth returns z_{i_k} computed with $r_{i_k:i_k+H-1}$, in order for $\tilde{h}^{(n)}$ to qualify as an element of $\tilde{\Omega}^{(n)}$. We call an augmented outcome $\tilde{h}^{(n)}$ "valid" if its augmented returns match the ground truth returns. The prefix $\tilde{h}^{(n)}_{1:t}$ is called an augmented history. The set of augmented histories is denoted $\tilde{\mathcal{H}}^{(n)}$, and the set of augmented histories up to time t is denoted $\tilde{\mathcal{H}}^{(n)}_t$.

A *return-predictor* with phase n is any mapping $\phi : \bigcup_{i \equiv n} (\tilde{\mathcal{H}}^{(n)}_{i-1} \times \mathcal{A}) \to \Delta\mathcal{Z}$. The domain consists exactly of all phase n augmented histories that end exactly at positions where the augmented return $\tilde{z}_i$ should come next. We define the mixture return-predictor, like the mixture environment in Definition 2.2.

Definition 3.1 (Mixture return-predictor)**.** *Given a class $\mathcal{P}_n$ of phase n return-predictors ϕ, and a prior distribution $\omega_n(\phi)$, a mixture return-predictor ψ_n is the unique phase n return-predictor that satisfies*

$$\prod_{k=0}^{K} \psi_n(\tilde{z}_{i_k} \mid \tilde{h}^{(n)}_{<i_k} a_{i_k}) = \sum_{\phi \in \mathcal{P}_n} \omega_n(\phi) \prod_{k=0}^{K} \phi(\tilde{z}_{i_k} \mid \tilde{h}^{(n)}_{<i_k} a_{i_k}),$$

for all $K \geq 0$, where $i_k = n + kN$.

Equivalently, $\psi_n(\tilde{z}_{i_k} \mid \tilde{h}^{(n)}_{<i_k} a_{i_k})$ is given by the posterior predictive distribution

$$\sum_{\phi \in \mathcal{P}_n} \omega_n(\phi \mid \tilde{h}^{(n)}_{<i_{k-1}} a_{i_{k-1}} \tilde{z}_{i_{k-1}})\, \phi(\tilde{z}_{i_k} \mid \tilde{h}^{(n)}_{<i_k} a_{i_k}),$$

with the posterior weights $\omega_n(\phi \mid \tilde{h}^{(n)}_{<i_k} a_{i_k} \tilde{z}_{i_k})$ updated to

$$\omega_n(\phi \mid \tilde{h}^{(n)}_{<i_{k-1}} a_{i_{k-1}} \tilde{z}_{i_{k-1}}) \frac{\phi(\tilde{z}_{i_k} \mid \tilde{h}^{(n)}_{<i_k} a_{i_k})}{\psi_n(\tilde{z}_{i_k} \mid \tilde{h}^{(n)}_{<i_k} a_{i_k})},$$

starting with $\omega_n(\phi \mid \epsilon) = \omega_n(\phi)$ where ϵ denotes the empty string. In other words, by conditioning the mixture ψ_n on a history, we are implicitly performing Bayesian inference over the return-predictors ϕ.

Let $\mathrm{aug}_n : \Omega \hookrightarrow \tilde{\Omega}^{(n)}$ be the injective mapping that augments an outcome h to its corresponding valid augmented outcome $\mathrm{aug}_n(h)$, by inserting the true returns z_{i_k} computed with the rewards $r_{i_k:i_k+H-1}$ in h. Under a certain condition, a history $h_{<t}$ uniquely determines $\mathrm{aug}_n(h)_{<t}$ regardless of $h_{t:\infty}$, and in that case we abuse the notation aug_n to also denote the mapping $\mathrm{aug}_n(h_{<t}) = \mathrm{aug}_n(h)_{<t}$. Fortunately, we can state this condition explicitly as:

$$(t - 1 - n) \bmod N \geq H - 1. \tag{2}$$

This is because the latest index $i \equiv n \pmod N$ smaller than t is exactly $t - 1 - (t - 1 - n) \bmod N$, and the return at this index requires rewards at times up to $t - 1 - (t - 1 - n) \bmod N + H - 1$, which is smaller than t if and only if Eq. 2 holds. In other words, all the rewards required for augmenting with the true returns z_{i_k} are available in $h_{<t}$. Note that Eq. 2 is trivially satisfied if $n = t \bmod N$. We can hence define a *unified predictor*, that abstracts away the periodic augmentation.

Definition 3.2 (Unified return-predictor)**.** *Given a collection of N mixture return-predictors $\{\psi_n\}_{0 \leq n < N}$, the unified predictor $\psi : \mathcal{H} \times \mathcal{A} \to \Delta\mathcal{Z}$ is given by*

$$\psi(\tilde{z}_t \mid h_{<t} a_t) := \psi_n(\tilde{z}_t \mid \mathrm{aug}_n(h_{<t}) a_t)$$

where $n = t \bmod N$.

Next, we define AIQI as follows. We also provide a pseudocode in Algorithm 1.

Definition 3.3 (AIQI)**.** *Fix a return horizon H, a discretization level M, an augmentation period $N \geq H$, an exploration rate $\tau > 0$, and a unified predictor ψ. Let the Q-value estimate given history $h_{<t}$ be*

$$\hat{Q}(h_{<t}, a_t) = \sum_{\tilde{z}_t \in \mathcal{Z}} \tilde{z}_t \cdot \psi(\tilde{z}_t \mid h_{<t} a_t).$$

AIQI, denoted $\hat{\pi}_\psi^{H,M,N,\tau}$ or $\hat{\pi}$, chooses the action that maximizes this estimate, with random exploration rate τ:

$$\hat{\pi}(a \mid h_{<t}) := (1 - \tau) \mathbb{1}[a = a^*] + \tau / |\mathcal{A}|,$$

where $a^* = \arg\max_{a_t} \hat{Q}(h_{<t}, a_t)$. We allow for stochastic tie-breaking when multiple actions maximize $\hat{Q}$.

Algorithm 1 Universal AI with Q-Induction (AIQI)

Require: Return horizon H, discretization level M, augmentation period $N \geq H$, exploration probability τ, set of unified predictor ψ

History $h_{<1} \leftarrow \epsilon$, time $t \leftarrow 1$
loop
  for each action $a \in \mathcal{A}$ **do**
    $\hat{Q}(h_{<t}, a) \leftarrow \sum_{\tilde{z} \in \mathcal{Z}} \tilde{z} \cdot \psi(\tilde{z} \mid h_{<t}, a)$
  end for
  $a^* \leftarrow \operatorname{argmax}_{a \in \mathcal{A}} \hat{Q}(h_{<t}, a)$
  Sample $p \sim \mathcal{U}[0, 1]$
  if $p < 1 - \tau$ **then**
    $a_t \leftarrow a^*$ ▷ Exploit
  else
    $a_t \sim \mathcal{U}(\mathcal{A})$ ▷ Explore
  end if
  Execute a_t, observe e_t
  $h_{<t+1} \leftarrow h_{<t} a_t e_t$
  $t \leftarrow t + 1$
end loop

Note that AIQI doesn't simulate the environment or plan for the future. Instead, AIQI directly predicts its own action-value, and at every step chooses the action that maximizes it. As evidence accumulates, the value predictions become more and more accurate. In the limit, AIQI chooses the action that maximizes its true action-value, and will keep doing so in all future steps, thereby turning into a globally optimal policy. Another way to look at AIQI is that it is performing policy evaluation at every step, treating the whole trajectory as having been generated by a single policy, and performing policy improvement in choosing the next action.

As is common in the universal AI literature, AIQI faces the grain of truth (self-referential) problem [Kalai and Lehrer, 1993, Leike et al., 2016b, Meulemans et al., 2025, Wyeth et al., 2025]. Roughly speaking, every mixture return-predictor ψ_n in ψ has to contain in its class the true conditional return distribution under the AIQI policy $\hat{\pi}_\psi$, which itself depends on ψ. The formal statement is given below.

Definition 3.4 (Grain of truth)**.** *Fix the AIQI parameters H, M, N, τ. A unified predictor ψ has a grain of truth w.r.t. environment ν if: every ψ_n is a mixture of a class $\mathcal{P}_n$ that contains the return-predictor ϕ^* given by*

$$\phi^*(\tilde{z}_i \mid \tilde{h}^{(n)}_{<i} a_i) = \nu^{\hat{\pi}}(z_i = \tilde{z}_i \mid h_{<i} a_i), \;\; i \equiv n \pmod{N},$$

where $\hat{\pi} = \hat{\pi}_\psi^{H,M,N,\tau}$ and the right hand side is the conditional return distribution under $\nu^{\hat{\pi}}$.

A nontrivial solution can be obtained by fixing a reflective oracle O [Fallenstein et al., 2015, Leike et al., 2016b, Wyeth et al., 2025], setting $\mathcal{P}_n$ to be the class of all O-estimable return-predictors, and assuming that ν is O-estimable. Under this setup, the AIQI policy $\hat{\pi}$ is O-estimable, and $\phi^* = \nu^{\hat{\pi}}$ is also O-estimable, hence inside $\mathcal{P}_n$. Allowing for stochastic tie-breaking in Definition 3.3 is essential for the use of reflective oracles.

We now introduce a notion of optimality that AIQI will be shown to satisfy. It is a more lenient form of strong asymptotic optimality, which is one of the strongest forms of optimality considered in the universal AI literature.

Definition 3.5 (Strong Asymptotic ε-Optimality)**.** *Given $\varepsilon > 0$ and a class of environments $\mathcal{M}$, a policy π is strong asymptotically ε-optimal if, for all environments $\nu \in \mathcal{M}$,*

$$\limsup_{t \to \infty} V^*_\nu(h_{<t}) - V^\pi_\nu(h_{<t}) \leq \varepsilon, \quad \nu^\pi\text{-a.s.}$$

It has been shown by Lattimore and Hutter [2011, Theorem 8] that any deterministic policy, including AIXI, is not strong asymptotically ε-optimal for all $\varepsilon < 1/4$.

4 RESULTS

The aim of this section is to prove that:

- For arbitrarily small $\varepsilon > 0$, one can choose parameters H, M, N, τ, ψ with which AIQI satisfies *strong asymptotic ε-optimality*. (Theorem 4.6)
- With the same choice of parameters, AIQI is asymptotically ε-optimal in the mixture environment ξ. (Theorem 4.8)
- In contrast to AIXI, AIQI is not necessarily *self-optimizing*—a property that roughly translates to "off-policy asymptotic optimality". (Theorem 4.10)

Notably, we do not invoke any novel, ad-hoc assumptions in proving our results. The list of notations can be found in § A, and omitted proofs are in § B. We have also performed experiments with a computable approximation of AIQI, which we report in § F. We also show in § 5 that our novel proof techniques can be applied to the analysis of Self-AIXI [Catt et al., 2023], to get rid of strong assumptions such as off-policy sensibility.

4.1 CONVERGENCE OF RETURN-PREDICTOR

First, we show that the unified return-predictor ψ converges to the true return-predictor $\phi^* = \nu^\pi$, when conditioned on increasingly longer histories generated by ν^π. A useful analytic device is the total variation (TV) distance. Given two probability measures P and Q defined on a measurable space $(\Omega, \mathcal{F})$, the TV distance is defined as

$$D(P, Q) := \sup_{E \in \mathcal{F}} |P(E) - Q(E)|.$$

An elementary property of TV distance is that, for a countable set $\{E_i\}$ of pairwise disjoint events,

$$\textstyle\sum_i |P(E_i) - Q(E_i)| \leq 2 \cdot D(P, Q). \tag{3}$$

Given an environment ν and a policy π, define the phase n augmented distribution $\tilde{\nu}_n^\pi := \nu^\pi \circ \mathrm{aug}_n^{-1}$ as the pushforward of ν^π by aug_n. We can express $\tilde{\nu}_n^\pi$ as a product of conditional probabilities whose dependencies respect the written order in $\tilde{h}^{(n)} = a_1 e_1 \ldots a_n \tilde{z}_n e_n \ldots$, as follows:

$$\tilde{\nu}_n^\pi(\tilde{h}^{(n)}) = \prod_i \tilde{\nu}_n^\pi(a_i \mid \tilde{h}_{<i}^{(n)}) \prod_{i \not\equiv n} \tilde{\nu}_n^\pi(e_i \mid \tilde{h}_{<i}^{(n)} a_i) \cdot \prod_{i \equiv n} \tilde{\nu}_n^\pi(e_i \mid \tilde{h}_{<i}^{(n)} a_i \tilde{z}_i) \prod_{i \equiv n} \tilde{\nu}_n^\pi(\tilde{z}_i \mid \tilde{h}_{<i}^{(n)} a_i). \quad (4)$$

Note that the above characterization of $\tilde{\nu}_n^\pi$ is informal, as $\tilde{\nu}_n^\pi(\tilde{h}^{(n)})$ is just 0 for most $\tilde{h}^{(n)}$. A proper characterization would involve $\tilde{\nu}_n^\pi(\tilde{h}_{<t}^{(n)})$ and all products indexed up to $i = t-1$. We use the informal notation for brevity.

The last factor in Eq. 4 is simply $\nu^\pi(z_i = \tilde{z}_i \mid h_{<i} a_i)$, which we abbreviate as $\nu^\pi(\tilde{z}_i \mid h_{<i} a_i)$. Let us also abbreviate the first three factors in Eq. 4 as $\tilde{\nu}_n^\pi(h \parallel \tilde{z}_{i \equiv n})$, so that

$$\tilde{\nu}_n^\pi(\tilde{h}^{(n)}) = \tilde{\nu}_n^\pi(h \parallel \tilde{z}_{i \equiv n}) \prod_{i \equiv n} \nu^\pi(\tilde{z}_i \mid h_{<i} a_i). \quad (5)$$

Let π be the AIQI policy $\hat{\pi}$ with some parameters H,M,N,τ, ψ. Suppose its mixture return-predictor ψ_n in ψ is a mixture of return-predictors including ϕ^* given by

$$\phi^*(\tilde{z}_i \mid \tilde{h}_{<i}^{(n)} a_i) = \nu^\pi(\tilde{z}_i \mid h_{<i} a_i), \quad i \equiv n \pmod N,$$

which is exactly the grain of truth condition. Then, define

$$q_n(\tilde{h}^{(n)}) := \tilde{\nu}_n^\pi(h \parallel \tilde{z}_{i \equiv n}) \prod_{i \equiv n} \psi_n(\tilde{z}_i \mid \tilde{h}_{<i}^{(n)} a_i),$$

which is Eq. 5 with $\nu^\pi = \phi^*$ replaced by ψ_n. Intuitively, we want to show that our mixture ψ_n converges to ϕ^*, and we do so by first showing that q_n converges to $\tilde{\nu}_n^\pi$.

By the definition of a mixture in Definition 3.1,

$$q_n(\tilde{h}^{(n)}) \geq \tilde{\nu}_n^\pi(h \parallel \tilde{z}_{i \equiv n}) \left(\omega_n(\phi^*) \prod_{i \equiv n} \phi^*(\tilde{z}_i \mid \tilde{h}_{<i}^{(n)} a_i) \right) = \omega_n(\phi^*) \tilde{\nu}_n^\pi(\tilde{h}^{(n)}).$$

By the inequality, $\tilde{\nu}_n^\pi$ is absolutely continuous with respect to q_n, so that by Blackwell and Dubins [1962],

$$\lim_{t \to \infty} D(q_n, \tilde{\nu}_n^\pi \mid \tilde{h}_{<t}^{(n)}) = 0, \quad \tilde{\nu}_n^\pi\text{-a.s.}$$

We can then prove the following lemma. Intuitively, $\tilde{\nu}_n^\pi$ is the pushforward of ν^π by aug_n, so we can replace $\tilde{h}_{<t}^{(n)}$ with $\mathrm{aug}_n(h)_{<t}$, and $\tilde{\nu}_n^\pi$- with ν^π-almost sure convergence.

Lemma 4.1 (Convergence in TV distance)**.** *Let π be the AIQI policy $\hat{\pi}_\psi^{H,M,N,\tau}$ where ψ has a grain of truth w.r.t. an environment ν. There exists a ν^π-probability-one set $S \subseteq \Omega$ such that for all $h \in S$ and $n \in [0, N-1]$,*

$$\lim_{t \to \infty} D(q_n, \tilde{\nu}_n^\pi \mid \mathrm{aug}_n(h)_{<t}) = 0.$$

We are now ready to show that each mixture return-predictor ψ_n, or more generally the unified predictor ψ, converges to the true return-predictor $\phi^* = \nu^\pi$. By "convergence", one would usually mean something along the line of

$$\lim_{t \to \infty} \sum_{\tilde{z}_t \in \mathcal{Z}} |\psi(\tilde{z}_t | h_{<t}\, a_t) - \nu^\pi(\tilde{z}_t | h_{<t}\, a_t)| = 0. \quad (6)$$

Instead, we prove something strictly stronger, involving a sum over hypothetical future trajectories $h'_{t:m-1}$ that extend $h_{<t}$. This stronger form of convergence is essential for proving Theorem 4.6 without invoking any ad-hoc assumptions, e.g., off-policy sensibility [Catt et al., 2023, Theorem 16].

Let us denote the return-predictor error by

$$\delta_\psi(h_{<t} a_t) := \sum_{\tilde{z}_t \in \mathcal{Z}} |\psi(\tilde{z}_t \mid h_{<t}\, a_t) - \nu^\pi(\tilde{z}_t \mid h_{<t}\, a_t)|,$$

and let $\mathcal{T}$ be the set of hypothetical trajectories $h'_{<m}$,

$$\mathcal{T} := \{h'_{<m} \mid h'_{<t} = h_{<t},\, e'_{t:m-1} \in \mathcal{E}^{m-t}, a'_{t:m-1} = \pi'(h_{<t}, e'_{t:m-2})\}, \quad (7)$$

formed by $h_{<t}$ appended with all possible percept sequences of length $m-t$ and reactions from some *deterministic policy* π'. We assume for now that π' is an arbitrary deterministic policy, and specify it later in the proof of Theorem 4.6. We can show the following result.

Lemma 4.2 (Convergence of return-predictor)**.** *Let π be the AIQI policy $\hat{\pi}_\psi^{H,M,N,\tau}$ where ψ has a grain of truth w.r.t. an environment ν. For every $\beta > 0$, there exists a ν^π-probability-one set $S \subseteq \Omega$ where, for every outcome $h \in S$, there exists t_0 such that: for $t \geq t_0$ and $m \in [t, t+N-H]$,*

$$\sum_{h'_{<m} \in \mathcal{T}, a'_m \in \mathcal{A}} \nu(e'_{t:m-1} \mid h_{<t} \parallel a'_{t:m-1}) \cdot \delta_\psi(h'_{<m} a'_m) < 2\beta\, (\tau / |\mathcal{A}|)^{-(m-t+1)}.$$

A corollary of our result is Eq. 6, obtained by setting $m = t$. Remarkably, our result is stronger, bounding the sum over hypothetical trajectories for all $m \in [t, t+N-H]$. This is possible due to the subtle fact that $\mathrm{aug}_n(h_{<t})$ is well-defined for $n = m \bmod N$, according to Eq. 2. This subtle fact also reveals why we can't simply set $N = H$ and have to add a buffer of length $N - H$. In fact, the length of this buffer is exactly the maximum length of hypothetical future trajectories that can extend $h_{<t}$ in our bound.

4.2 ONE-STEP OPTIMALITY

We continue to denote the AIQI policy $\hat{\pi}$ as π. Let the AIQI parameter H be the η-effective horizon $H(\eta)$ for some $\eta > 0$. We can show that the estimated action-value $\hat{Q}$ in Definition 3.3 converges to the true action-value Q_ν^π. Denote the Q-value estimation error by

$$\delta_Q(h_{<t} a_t) := |\hat{Q}(h_{<t} a_t) - Q_\nu^\pi(h_{<t} a_t)|.$$

Lemma 4.3 (Convergence of Q-value prediction)**.** *Under the conditions established in Lemma 4.2,*

$$\sum_{h'_{<m}\in\mathcal{T},a'_m\in\mathcal{A}} \nu(e'_{t:m-1} \mid h_{<t} \parallel a'_{t:m-1}) \cdot \delta_Q(h'_{<m}a'_m) < 2\beta\,(\tau/|\mathcal{A}|)^{-(m-t+1)} + M^{-1} + \eta.$$

Next, we show that the difference between the on-policy value and maximum Q-value becomes small. We call this difference the one-step optimality gap,

$$\delta_1(h_{<t}) := \max_a Q^\pi_\nu(h_{<t}, a) - V^\pi_\nu(h_{<t}),$$

which is non-negative due to $V(\cdot) = \mathbb{E}_a[Q(\cdot, a)]$.

Lemma 4.4 (Convergence of one-step optimality gap)**.** *Under the conditions established in Lemma 4.2,*

$$\sum_{h'_{<m}\in\mathcal{T}} \nu(e'_{t:m-1} \mid h_{<t} \parallel a'_{t:m-1}) \cdot \delta_1(h'_{<m}) < 2 \cdot (2\beta\,(\tau/|\mathcal{A}|)^{-(m-t+1)} + M^{-1} + \eta) + 2\tau.$$

Intuitively, since AIQI chooses the action a^* that maximizes the estimate $\hat{Q}$, the value V of AIQI is approximately equal to its Q-value under the action a^*, with only a slight difference due to the exploration probability τ. As the estimate $\hat{Q}$ converges to the true Q-value due to Lemma 4.3, the Q-value $Q(\cdot, a^*)$, and hence the value $V(\cdot)$, converges to the true maximum $\max_a Q(\cdot, a)$.

4.3 ASYMPTOTIC OPTIMALITY

Let the global optimality gap be

$$\delta_\infty(h_{<t}) := V^*_\nu(h_{<t}) - V^\pi_\nu(h_{<t}).$$

We now want to show the convergence of the global optimality gap, which is equivalent to asymptotic optimality. First, we present a lemma that relates the global optimality gap to the one-step optimality gap.

Lemma 4.5 (Bound on global optimality gap)**.** *For any environment ν, policy π, and history $h_{<t}$,*

$$\delta_\infty(h_{<t}) \le \gamma \sum_{e'_t} \nu(e'_t \mid h_{<t}\, a'_t)\, \delta_\infty(h_{<t}\, a'_t\, e'_t) + \delta_1(h_{<t}),$$

where $a'_t = \arg\max_a (Q^*_\nu(h_{<t}, a) - Q^\pi_\nu(h_{<t}, a))$.

Note that we can form a *chain* of inequalities by repeatedly applying the lemma to δ_∞ on the right hand side, as follows:

$$\begin{aligned}
\delta_\infty(h_{<t}) &\le \gamma \sum_{e'_t} \nu(e'_t \mid h_{<t}a'_t)\, \delta_\infty(h'_{<t+1}) + \delta_1(h_{<t}) \\
&\le \gamma^2 \sum_{e'_{t:t+1}} \nu(e'_{t:t+1} \mid h_{<t} \parallel a'_{t:t+1})\, \delta_\infty(h'_{<t+2}) \\
&\quad + \gamma \sum_{e'_t} \nu(e'_t \mid h_{<t}a'_t)\, \delta_1(h'_{<t+1}) + \delta_1(h_{<t}) \\
&\le \cdots \\
&\le \gamma^L \sum_{e'_{t:t+L-1}} \nu(e'_{t:t+L-1} \mid h_{<t} \parallel a'_{t:t+L-1})\, \delta_\infty(h'_{<t+L}) \\
&\quad + \sum_{l=0}^{L-1} \gamma^l \Big[\sum_{e'_{t:t+l-1}} \nu(e'_{t:t+l-1} \mid h_{<t} \parallel a'_{t:t+l-1})\, \delta_1(h'_{<t+l}) \Big]
\end{aligned}$$

where $a'_i = \arg\max_a |Q^*_\nu(h'_{<i}, a) - Q^\pi_\nu(h'_{<i}, a)|$.

Furthermore, the first term in the last inequality is smaller than or equal to γ^L, since $\delta_\infty(h'_{<t+L}) \le 1$. Thus $\delta_\infty(h_{<t})$ is smaller than or equal to

$$\gamma^L + \sum_{l=0}^{L-1} \gamma^l \sum_{e'_{t:t+l-1}} \nu(e'_{t:t+l-1} | h_{<t} \| a'_{t:t+l-1})\, \delta_1(h'_{<t+l}). \quad (8)$$

We have bound the global optimality gap $\delta_\infty(h_{<t})$ using *only* one-step optimality gaps δ_1.

We are now ready to prove the central result of our paper, that AIQI is strong asymptotically ε-optimal.

Theorem 4.6 (AIQI is strong asymptotically ε-optimal)**.** *Fix an environment class $\mathcal{M}$ and tolerance $\varepsilon > 0$. Let $\tau \le \frac{\varepsilon(1-\gamma)}{10}$, $M \ge \frac{10}{\varepsilon(1-\gamma)}$, $H = H(\eta)$ with $\eta \le \frac{\varepsilon(1-\gamma)}{10}$, and $N \ge H + \log_\gamma \frac{\varepsilon}{5}$. Suppose ψ is a unified predictor that has a grain of truth with respect to all environments $\nu \in \mathcal{M}$ and the above parameters for AIQI. Then, the AIQI policy $\hat{\pi}_\psi^{H,M,N,\tau}$ is strong asymptotically ε-optimal in $\mathcal{M}$.*

Proof. We simply write π for the AIQI policy $\hat{\pi}_\psi^{H,M,N,\tau}$. Fix an environment $\nu \in \mathcal{M}$, and let $L = N - H + 1$. Fix some $\beta > 0$, and with it, choose $S \subseteq \Omega$ as in Lemma 4.2. For any $h \in S$, choose t_0, again as in Lemma 4.2. Then for all $t \ge t_0$, substituting $m = t + l$ in Eq. 8 and noting that

$$a'_i = \arg\max_a |Q^*_\nu(h'_{<i}, a) - Q^\pi_\nu(h'_{<i}, a)|$$

is the output of a *deterministic policy*, we apply Lemma 4.4 to obtain

$$\delta_\infty(h_{<t}) < \gamma^L + \sum_{l=0}^{L-1} \gamma^l \Big[2(2\beta \left(\frac{\tau}{|\mathcal{A}|}\right)^{-(l+1)} + M^{-1} + \eta) + 2\tau \Big].$$

Recall that $L \geq \log_\gamma \frac{\varepsilon}{5}$, $\tau \leq \frac{\varepsilon(1-\gamma)}{10}$, $M \geq \frac{10}{\varepsilon(1-\gamma)}$, and $\eta \leq \frac{\varepsilon(1-\gamma)}{10}$. If we use $\beta = \frac{\varepsilon\tau}{20}|\mathcal{A}|^{-1}(\sum_{l=0}^{L-1}(\gamma|\mathcal{A}|/\tau)^l)^{-1}$ in choosing S and t_0, the right hand side is smaller than or equal to ε. Thus for $h \in S$ and $t \geq t_0$, $\delta_\infty(h_{<t}) \leq \varepsilon$. This concludes our proof that

$$\limsup_{t\to\infty} \delta_\infty(h_{<t}) \leq \varepsilon, \quad \nu^\pi\text{-a.s.} \qquad \square$$

We can also show that strong asymptotic ε-optimality in the environment class $\mathcal{M}$ implies asymptotic optimality in the mixture environment ξ.

Lemma 4.7 (Optimality in $\mathcal{M}$ implies optimality in ξ)**.** *Given a mixture ξ of environment class $\mathcal{M}$, and a policy π that is strong asymptotically ε-optimal with respect to $\mathcal{M}$, the same policy π is asymptotically ε-optimal in ξ:*

$$\limsup_{t\to\infty} V_\xi^*(h_{<t}) - V_\xi^\pi(h_{<t}) \leq \varepsilon$$

holds both ξ^π-a.s. and, for all $\nu \in \mathcal{M}$, ν^π-a.s.

As a consequence, AIQI is also asymptotically optimal in ξ. In other words, AIQI approximates *Bayes-optimality*, which is the defining feature of AIXI.

Theorem 4.8 (AIQI is asymptotically ε-Bayes-optimal)**.** *For any mixture ξ of environment class $\mathcal{M}$, the AIQI policy $\hat{\pi}_\psi^{H,M,N,\tau}$ with the parameters in Theorem 4.6 is asymptotically ε-optimal in ξ.*

Proof. Theorem 4.6 and Lemma 4.7. $\square$

The single-agent result extends directly to multi-agent interaction by viewing each agent i as acting in the subjective environment σ_i induced by the other agents' policies $\pi_{\neq i}$ [Leike et al., 2016b, Section 4.1]. Thus, if each AIQI agent $\pi_i = \hat{\pi}_{\psi_i}^{H_i,M_i,N_i,\tau_i}$ satisfies the assumptions of Theorem 4.6, every agent is eventually an ε-best response to the others. Hence the joint policy profile $\pi_{1:n}$ converges to an ε-Nash equilibrium, $\sigma^{\pi_{1:n}}$-almost surely. We detail this extension in § D.

4.4 OFF-POLICY BEHAVIOR

Until now we have solely discussed the on-policy setting, where the policy that interacts with the environment is the same as the policy that we are optimizing. In the *off-policy setting*, we must infer an optimal policy from the interactions generated by some "historic policy" π' that we do not have control over. We can formalize this notion of optimality with the *self-optimizing* property.

Definition 4.9 (Self-optimizing policy)**.** *A policy $\bar{\pi}$ is called self-optimizing for a class of environments $\mathcal{M}$ and a historic policy π', if for every $\nu \in \mathcal{M}$,*

$$\lim_{t\to\infty} V_\nu^*(h_{<t}) - V_\nu^{\bar{\pi}}(h_{<t}) = 0, \quad \nu^{\pi'}\text{-a.s.}$$

Hutter [2002] showed a remarkable property of AIXI, that if there *exists* a self-optimizing policy for $\mathcal{M}$ and π', then AIXI is one of those self-optimizing policies. On the other hand, we can expect that AIQI does not satisfy this property, since it is an on-policy MC control algorithm.

In fact, if the predictor ψ supports the return-predictor $\phi^* = \nu^{\pi'}$, AIQI will eventually choose actions with the highest value under the historic policy π', and not its own value—the optimality of AIQI is thus no longer guaranteed. Even if we relax the self-optimizing property to the ε-self-optimizing property by replacing $\lim = 0$ with $\limsup \leq \varepsilon$, the following can be shown.

Theorem 4.10 (AIQI is not self-optimizing)**.** *Suppose the unified predictor ψ is built from classes $\mathcal{P}_n$ that contain all computable variable-order Markov models. For any period N, exploration rate τ, horizon $H \geq 2$, tolerance ε small enough (depending on H), and discretization level M large enough (depending on H, ε, τ), there exist an environment class $\mathcal{M}$, a historic policy π', and a discount factor $\gamma \in (0,1)$, for which there exists a self-optimizing policy, but the AIQI policy $\hat{\pi}_\psi^{H,M,N,\tau}$ is not even ε-self-optimizing.*

This completes our investigation of the theoretical properties of AIQI. We have assumed a geometric discount setting for simplicity, but we can extend all our results to the general discount setting, where returns are defined as $R_t = \sum_{k=t}^\infty \gamma_k r_k / \sum_{k=t}^\infty \gamma_k$ for a general discount sequence γ_k. Specifically, our results can be generalized to cases where the discount sequence decays faster than a geometric sequence. We detail the generalization in § E.

5 APPLICATION TO SELF-AIXI

Self-AIXI [Catt et al., 2023] is similar to AIQI in that it makes predictions of its own action values, and chooses the action a_t that maximizes the predicted value. As noted by Wyeth [2025], Catt et al. [2023] make unjustified assumptions in their attempt to prove the asymptotic optimality of Self-AIXI. We suggest adding ε-greedy exploration to Self-AIXI and proving its asymptotic ε-optimality without any ad-hoc assumptions. Fortunately, our proof techniques can be applied with minimal changes. The key is in proving an analogue of Lemma 4.3. We present the key results below, and provide their proofs in § C. First, we define the ε-greedy Self-AIXI.

Definition 5.1 (ε-greedy Self-AIXI)**.** *Let ξ be a mixture environment constructed from some environment class $\mathcal{M}$, and ζ be a mixture policy constructed from some policy class $\mathcal{P}$. Then, ε-greedy Self-AIXI, denoted π_S, is defined as*

$$\pi_S(a \mid h_{<t}) := (1-\tau)\mathbb{1}\left[a = a^*\right] + \tau/|\mathcal{A}|,$$

where $a^* = \arg\max_{a_t} Q_\xi^\zeta(h_{<t}, a_t)$. We allow for stochastic tie-breaking when multiple actions maximize Q_ξ^ζ.

The following lemmas serve as a bridge between Self-AIXI and our proof technique.

Lemma 5.2 (Convergence of ξ^ζ to ν^π)**.** *For any environment $\nu \in \mathcal{M}$ and policy $\pi \in P$,*

$$\lim_{t\to\infty} D\big(\xi^\zeta, \nu^\pi \mid h_{<t}\big) = 0, \quad \nu^\pi\text{-a.s.}$$

Lemma 5.3 (Bounds on Q-value difference)**.** *For any two policies π_1, π_2 and two environments ν_1, ν_2,*

$$\big|Q^{\pi_1}_{\nu_1}(h_{<t}, a_t) - Q^{\pi_2}_{\nu_2}(h_{<t}, a_t)\big| \leq D(\nu_1^{\pi_1}, \nu_2^{\pi_2} \mid h_{<t}\, a_t)$$

Proof. See Leike [2016, Lemma 4.17]. □

Lemma 5.4 (Average conditional TV bound)**.** *Let P and Q be probability measures on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is finite. Let P_X and Q_X denote the marginals on $\mathcal{X}$. Then*

$$\sum_{x\in\mathcal{X}} P_X(x)\, D\big(P, Q \mid X = x\big) \;\leq\; 2D(P, Q).$$

Now we can prove an analogue of Lemma 4.3 for π_S. Let

$$\delta_Q(h_{<t}a_t) := |Q^\zeta_\xi(h_{<t}a_t) - Q^\pi_\nu(h_{<t}a_t)|,$$

and define $\mathcal{T}$ as in Eq. 7.

Lemma 5.5 (Convergence of Q-value prediction for ε-greedy Self-AIXI)**.** *Let $\pi = \pi_S$. For every $\beta > 0$, there exists a ν^π-probability-one set $S \subseteq \Omega$ where, for every outcome $h \in S$, there exists t_0 such that: for $t \geq t_0$ and $m \geq t$,*

$$\sum_{h'_{<m}\in\mathcal{T}, a'_m\in\mathcal{A}} \nu(e'_{t:m-1} \mid h_{<t} \parallel a'_{t:m-1}) \cdot \delta_Q(h'_{<m}a'_m) < 2\beta\,(\tau/|\mathcal{A}|)^{-(m-t+1)}.$$

We can then show analogues of Lemmas 4.4 and 4.5 and Theorem 4.6 for ε-greedy Self-AIXI, with our proof for AIQI repeated almost verbatim. Note that the error term 2τ would be introduced by the analogue of Lemma 4.4, and thus π_S can only be asymptotically ε-optimal, where ε depends on the exploration rate τ.

6 RELATED WORK

Compress and Control [CNC; Veness et al., 2015] is an algorithm that shares many similarities to AIQI such as return prediction, but has only been analyzed for fully observable MDPs. Both CNC and AIQI are on-policy Monte Carlo control methods with a distributional flavour [Bellemare et al., 2017], since they predict the return distribution instead of the average return. CNC uses a model-based approach for MC evaluation, unlike AIQI.

Feature reinforcement learning [Daswani, 2015] converts a general environment to a finite state MDP through a learned feature map, and then performs planning [Hutter, 2009] or model-free RL [Nguyen et al., 2011, Daswani et al., 2013] inside the MDP. Hutter [2014] shows that any general environment can be mapped to an MDP where the optimal policy is still ε-optimal in the general environment. However, these theoretical results do not tell us about the learning of such mappings, and results that are concerned with the learning [Sunehag and Hutter, 2010] do not concern themselves with control.

Self-AIXI [Catt et al., 2023] is a variant of AIXI that learns both a model of the environment and a model of itself. At every step, it uses both models to estimate its own action-values, by simulating the model of itself interacting with the model of the environment. It then chooses the action with the highest estimated value. Although Self-AIXI gets rid of planning, it is clearly model-based in that it learns and uses an environment model. Our work was inspired by Self-AIXI, and can be seen as a modification that further gets rid of environment models by directly predicting action-values.

Finally, Optimal Direct Policy Search [ODPS; Glasmachers and Schmidhuber, 2011] is a policy search method that enumerates computable policies (or more generally policies in some arbitrary class) and evaluates them by repeated interaction with the environment. Like AIQI, ODPS is model-free and avoids planning. However, ODPS assumes an episodic POMDP setting with bounded episode lengths, whereas our results apply to general environments.

7 DISCUSSION

7.1 PROOF TECHNIQUE

To prove that AIQI is strong asymptotically ε-optimal, we first showed that the action-value estimates converge to the true ν^π action-values (Lemma 4.3), and that choosing the action with the highest (estimated) ν^π value makes the policy converge to optimal actions (Theorem 4.6). There are two key novelties in our proofs compared to existing work in universal AI. First, we had to deal with the problem of delayed feedback using periodic augmentations. This was not needed in previous works which model policies or environments, since actions and percepts are never delayed, unlike H-step returns. Second, we introduced a chain of inequalities using Lemma 4.5, to make the δ_∞ term in the upper bound vanish with γ^L. This allowed us to bypass ad-hoc assumptions such as off-policy sensibility [Catt et al., 2023, Theorem 16]. In fact, we show in § 5 that our technique can be applied to Self-AIXI to show analogous results without off-policy sensibility. Overall, our work provides a blueprint for the analysis of policy iteration algorithms in general environments.

7.2 CONTINUAL REINFORCEMENT LEARNING

The class of general environments is equivalent to the class of infinite-state POMDPs, which might seem unnecessarily large. However, there are nontrivial settings that cannot be captured by finite-state POMDPs, one such example being continual RL [Ring, 1994]. Continual RL studies environments that never stop changing [Khetarpal et al., 2022], and where agents must keep learning in order to act optimally [Abel et al., 2023]. The theory of universal AI sets aside problems with continual *learning*, e.g., catastrophic forgetting, by assuming a perfect Bayesian learner with a broad prior. This lets it focus on a different question: how to design principled *objectives* for continual RL, assuming continual learning is approximately solved. Our work shows that learning distributional action-values is one such objective.

7.3 AGENT FOUNDATIONS

The theory of universal AI is central to agent foundations research [Soares and Fallenstein, 2017], since it provides formal models of powerful goal-directed agents under minimal assumptions about the environment. As such, it has been used to analyze important issues in AI safety in a theoretically grounded way [Everitt et al., 2016, Orseau and Armstrong, 2016, Majha et al., 2019, Everitt and Hutter, 2018a, Cohen et al., 2021]. Our results show that asymptotic optimality in general reinforcement learning does not require an explicit world model. We therefore view AIQI as a new theoretical object that may help clarify which ingredients of general-agent theory matter for safety.

Although our paper set out to primarily show that AIQI and AIXI are similar, they might differ in other ways, e.g., the off-policy behavior as we discussed in § 4.4. Another interesting difference is that, unlike AIXI, AIQI needs to deal with self-reference (Definition 3.4), which is in turn related to embedded agency [Demski and Garrabrant, 2019, Orseau and Ring, 2012b].

8 CONCLUSION

We introduced Universal AI with Q-Induction (AIQI), the first model-free universal agent. AIQI performs universal induction over distributional action-value functions—rather than over environments or policies as in prior work. It is essentially a Monte Carlo control algorithm, the most primitive form of model-free RL. We prove that AIQI is strong asymptotically ε-optimal, asymptotically ε-Bayes-optimal, and, as expected from an on-policy MC control algorithm, not self-optimizing. Remarkably, the only assumption we make is grain of truth, which is standard in universal AI and has a known resolution via reflective oracles.

Our findings significantly broaden the known landscape of universal agents. We hope this provides both a theoretical foundation for understanding MC control with sequence models, and a blueprint for the analysis of other policy iteration algorithms in general environments. Promising directions for future work include analyzing the properties of AIQI with exploration strategies that are more sophisticated than ε-greedy, such as knowledge-seeking agents [Orseau, 2014], investigating similarities/differences between AIQI and AIXI, and exploring other possible forms of model-free universal AI such as those that involve policy search.

Acknowledgements

This work was supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2019-II190075, Artificial Intelligence Graduate School Program(KAIST); RS-2024-00509279, Fundamental Research in Artificial Intelligence).

We thank Cole Wyeth for carefully reading the paper and discussing ways to improve it. We also thank the Universal Algorithmic Intelligence community for inviting us for a talk and providing insightful questions/feedback.

References

David Abel, André Barreto, Benjamin Van Roy, Doina Precup, Hado P van Hasselt, and Satinder Singh. A definition of continual reinforcement learning. *Advances in Neural Information Processing Systems*, 36:50377–50407, 2023.

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International conference on machine learning*, pages 449–458. Pmlr, 2017.

Richard Ernest Bellman. *Dynamic Programming*. Courier Dover Publications, Mineola, NY, 1957.

David Blackwell and Lester Dubins. Merging of opinions with increasing information. *The Annals of Mathematical Statistics*, 33(3):882–886, 1962.

Elliot Catt. *On the Foundations of Universal Artificial Intelligence*. PhD thesis, The Australian National University, Canberra, Australia, 2022.

Elliot Catt, Jordi Grau-Moya, Marcus Hutter, Matthew Aitchison, Tim Genewein, Gregoire Deletang, Kevin Li, and Joel Veness. Self-predictive universal ai. *Advances in Neural Information Processing Systems*, 36:27181–27198, 2023.

Michael K Cohen, Elliot Catt, and Marcus Hutter. A strongly asymptotically optimal agent in general environments. *arXiv preprint arXiv:1903.01021*, 2019.

Michael K Cohen, Elliot Catt, and Marcus Hutter. Curiosity killed or incapacitated the cat and the asymptotically optimal agent. *IEEE Journal on Selected Areas in Information Theory*, 2(2):665–677, 2021.

Rémi Coulom. Efficient selectivity and backup operators in monte-carlo tree search. In *International conference on computers and games*, pages 72–83. Springer, 2006.

Mayank Daswani. *Generic Reinforcement Learning Beyond Small MDPs*. PhD thesis, Australian National University, 2015.

Mayank Daswani, Peter Sunehag, and Marcus Hutter. Q-learning for history-based reinforcement learning. In *Asian Conference on Machine Learning*, pages 213–228. PMLR, 2013.

Abram Demski and Scott Garrabrant. Embedded agency. *arXiv preprint arXiv:1902.09469*, 2019.

Tom Everitt and Marcus Hutter. The alignment problem for bayesian history-based reinforcement learners. *Under submission*, 2018a.

Tom Everitt and Marcus Hutter. Universal artificial intelligence: Practical agents and fundamental challenges. In *Foundations of trusted autonomy*, pages 15–46. Springer, 2018b.

Tom Everitt, Daniel Filan, Mayank Daswani, and Marcus Hutter. Self-modification of policy and utility function in rational agents. In *International conference on artificial general intelligence*, pages 1–11. Springer, 2016.

Benja Fallenstein, Jessica Taylor, and Paul F Christiano. Reflective oracles: A foundation for game theory in artificial intelligence. In *International Workshop on Logic, Rationality and Interaction*, pages 411–415. Springer, 2015.

Tobias Glasmachers and Jürgen Schmidhuber. Optimal direct policy search. In *International Conference on Artificial General Intelligence*, pages 52–61. Springer, 2011.

Marcus Hutter. Self-optimizing and pareto-optimal policies in general environments based on bayes-mixtures. In *International Conference on Computational Learning Theory*, pages 364–379. Springer, 2002.

Marcus Hutter. *Universal artificial intelligence: Sequential decisions based on algorithmic probability*, volume 300. Springer, 2005.

Marcus Hutter. Feature reinforcement learning: Part i. unstructured mdps. *arXiv preprint arXiv:0906.1713*, 2009.

Marcus Hutter. Extreme state aggregation beyond mdps. In *International Conference on Algorithmic Learning Theory*, pages 185–199. Springer, 2014.

Marcus Hutter, David Quarel, and Elliot Catt. *An introduction to universal artificial intelligence*. Chapman and Hall/CRC, 2024.

Ehud Kalai and Ehud Lehrer. Rational learning leads to nash equilibrium. *Econometrica: Journal of the Econometric Society*, pages 1019–1045, 1993.

Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup. Towards continual reinforcement learning: A review and perspectives. *Journal of Artificial Intelligence Research*, 75:1401–1476, 2022.

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

Tor Lattimore. *Theory of General Reinforcement Learning*. PhD thesis, Australian National University, 2014.

Tor Lattimore and Marcus Hutter. Asymptotically optimal agents. In *International Conference on Algorithmic Learning Theory*, pages 368–382. Springer, 2011.

Tor Lattimore and Marcus Hutter. Bayesian reinforcement learning with exploration. In *International conference on algorithmic learning theory*, pages 170–184. Springer, 2014.

Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. *Minds and machines*, 17(4):391–444, 2007.

Jan Leike. *Nonparametric general reinforcement learning*. The Australian National University (Australia), 2016.

Jan Leike, Tor Lattimore, Laurent Orseau, and Marcus Hutter. Thompson sampling is asymptotically optimal in general environments. *arXiv preprint arXiv:1602.07905*, 2016a.

Jan Leike, Jessica Taylor, and Benya Fallenstein. A formal solution to the grain of truth problem. *arXiv preprint arXiv:1609.05058*, 2016b.

Arushi Majha, Sayan Sarkar, and Davide Zagami. Categorizing wireheading in partially embedded agents. *arXiv preprint arXiv:1906.09136*, 2019.

Alexander Meulemans, Rajai Nasser, Maciej Wołczyk, Marissa A Weis, Seijin Kobayashi, Blake Richards, Guillaume Lajoie, Angelika Steger, Marcus Hutter, James Manyika, et al. Embedded universal predictive intelligence: a coherent framework for multi-agent learning. *arXiv preprint arXiv:2511.22226*, 2025.

Phuong Nguyen, Peter Sunehag, and Marcus Hutter. Feature reinforcement learning in practice. In *European Workshop on Reinforcement Learning*, pages 66–77. Springer, 2011.

Laurent Orseau. Universal knowledge-seeking agents. *Theoretical Computer Science*, 519:127–139, 2014.

Laurent Orseau and Stuart Armstrong. Safely interruptible agents. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, pages 557–566, 2016.

Laurent Orseau and Mark Ring. Memory issues of intelligent agents. In *International Conference on Artificial General Intelligence*, pages 219–231. Springer, 2012a.

Laurent Orseau and Mark Ring. Space-time embedded intelligence. In *International Conference on Artificial General Intelligence*, pages 209–218. Springer, 2012b.

Mark Bishop Ring. *Continual learning in reinforcement environments*. The University of Texas at Austin, 1994.

Gavin A Rummery and Mahesan Niranjan. *On-line Q-learning using connectionist systems*, volume 37. University of Cambridge, Department of Engineering Cambridge, UK, 1994.

Nate Soares and Benya Fallenstein. Agent foundations for aligning machine intelligence with human interests: a technical research agenda. In *The technological singularity: Managing the journey*, pages 103–125. Springer, 2017.

Ray J Solomonoff. A formal theory of inductive inference. part i. *Information and control*, 7(1):1–22, 1964a.

Ray J Solomonoff. A formal theory of inductive inference. part ii. *Information and control*, 7(2):224–254, 1964b.

Peter Sunehag and Marcus Hutter. Consistency of feature markov processes. In *International Conference on Algorithmic Learning Theory*, pages 360–374. Springer, 2010.

Joel Veness, Kee Siong Ng, Marcus Hutter, William Uther, and David Silver. A monte-carlo aixi approximation. *Journal of Artificial Intelligence Research*, 40:95–142, 2011.

Joel Veness, Marc Bellemare, Marcus Hutter, Alvin Chua, and Guillaume Desjardins. Compress and control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

John Von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, 1944.

Christopher John Cornish Hellaby Watkins. *Learning from delayed rewards*. PhD thesis, King's College, University of Cambridge, Cambridge, United Kingdom, 1989.

FMJ Willems, Yu M Shtarkov, and TJ Tjalkens. The context-tree weighting method: basic properties. *IEEE Transactions on Information Theory*, 41(3):653–664, 1995.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Cole Wyeth. Unbounded embedded agency: AEDT w.r.t. rOSI. `https://www.lesswrong.com/posts/B6gumHyuxzR5yn5tH/unbounded-embedded-agency-aedt-w-r-t-rosi`, July 2025. LessWrong blog post. Accessed: 2026-06-02.

Cole Wyeth, Marcus Hutter, Jan Leike, and Jessica Taylor. Limit-computable grains of truth for arbitrary computable extensive-form (un) known games. *arXiv preprint arXiv:2508.16245*, 2025.

A Model-Free Universal AI (Supplementary Material)

Yegon Kim[1] **Juho Lee**[1]

[1] Graduate School of AI, KAIST, Seoul, South Korea

A NOTATION

Table 1: Summary of Notations

Symbol	Description
ϵ	The empty string, e.g., initial history $h_{<1}$
$\mathbb{N}$	Set of natural numbers $\{0, 1, 2, \dots\}$
$\mathbb{Z}_+$	Set of positive integers $\{1, 2, 3, \dots\}$
$\mathcal{A}, \mathcal{O}, \mathcal{R}$	Sets of actions, observations, and rewards
$\mathcal{E}$	Percept space, defined as $\mathcal{O} \times \mathcal{R}$
$\mathcal{H}$	The set of all finite histories $(\mathcal{A} \times \mathcal{E})^*$
$h_{<t}$	History sequence $a_1 e_1 \dots a_{t-1} e_{t-1}$
π, ν	Policy $\pi : \mathcal{H} \to \Delta\mathcal{A}$ and environment $\nu : \mathcal{H} \times \mathcal{A} \to \Delta\mathcal{E}$
V_ν^π, Q_ν^π	Value function and Q-value function
$H(\eta)$	η-effective horizon
M	Discretization level for returns
$\mathcal{Z}$	Alphabet of discretized returns $\{0, \frac{1}{M}, \dots, \frac{M-1}{M}\}$
z_t	M-discretized H-step return at time t
$\tilde{z}_t$	Augmented return at time t in an augmented outcome
N	Augmentation period ($N \geq H$)
n	Phase of periodic augmentation ($n \in \{0, \dots, N-1\}$)
$\tilde{\Omega}^{(n)}, \tilde{\mathcal{H}}^{(n)}$	Phase n augmented sample space and histories
aug_n	Mapping from outcomes to phase n-augmented outcomes
ψ_n	Mixture return-predictor for phase n
ψ	Unified predictor that selects ψ_n where $n = t \bmod N$
$\hat{Q}(h_{<t}, a_t)$	AIQI estimated Q-value (expected discretized return)
τ	Exploration probability
$D(P, Q)$	Total Variation (TV) distance between measures P and Q
L	Effective lookahead length ($N - H + 1$)

B PROOFS

Lemma 4.1 (Convergence in TV distance)**.** *Let π be the AIQI policy $\hat{\pi}_\psi^{H,M,N,\tau}$ where ψ has a grain of truth w.r.t. an environment ν. There exists a ν^π-probability-one set $S \subseteq \Omega$ such that for all $h \in S$ and $n \in [0, N-1]$,*

$$\lim_{t \to \infty} D(q_n, \tilde{\nu}_n^\pi \mid \text{aug}_n(h)_{<t}) = 0.$$

Proof. Recall that

$$\lim_{t\to\infty} D(q_n, \tilde{\nu}_n^\pi \mid \tilde{h}_{<t}^{(n)}) = 0, \quad \tilde{\nu}_n^\pi\text{-a.s.}$$

In other words, there exists an event $\tilde{S}_n \subseteq \tilde{\Omega}^{(n)}$, with $\tilde{\nu}_n^\pi(\tilde{S}_n) = 1$, where $D(q_n, \tilde{\nu}_n^\pi \mid \tilde{h}_{<t}^{(n)})$ converges to 0 pointwise. Let F be the set of "faulty" outcomes in $\tilde{S}_n$ whose augmented returns $\tilde{z}_{i\equiv n}$ do not agree with the true returns $z_{i\equiv n}$. Recall the definition of $\tilde{\nu}_n^\pi$ as the pushforward of ν^π by aug_n. Since F is disjoint from the image of aug_n, we get $\tilde{\nu}_n^\pi(F) = 0$.

Let S_n be the preimage of $\tilde{S}_n - F$ under aug_n. Then, by the injectivity of aug_n, we see that $\nu^\pi(S_n) = \tilde{\nu}_n^\pi(\tilde{S}_n - F) = \tilde{\nu}_n^\pi(\tilde{S}_n) = 1$. Since pointwise convergence on $\tilde{S}_n$ implies pointwise convergence on the subset $\tilde{S}_n - F$, we see that $D(q_n, \tilde{\nu}_n^\pi \mid \mathrm{aug}_n(h)_{<t})$ converges pointwise to 0 on S_n. Construct such S_n for every n, and let $S = \bigcap_n S_n$. Then, $\nu^\pi(S) = 1$, and $D(q_n, \tilde{\nu}_n^\pi \mid \mathrm{aug}_n(h)_{<t})$ converges pointwise to 0 on S for all n. □

Lemma 4.2 (Convergence of return-predictor)**.** *Let π be the AIQI policy $\hat{\pi}_\psi^{H,M,N,\tau}$ where ψ has a grain of truth w.r.t. an environment ν. For every $\beta > 0$, there exists a ν^π-probability-one set $S \subseteq \Omega$ where, for every outcome $h \in S$, there exists t_0 such that: for $t \geq t_0$ and $m \in [t, t+N-H]$,*

$$\sum_{h'_{<m}\in\mathcal{T}, a'_m\in\mathcal{A}} \nu(e'_{t:m-1} \mid h_{<t} \parallel a'_{t:m-1}) \cdot \delta_\psi(h'_{<m}a'_m) < 2\beta\,(\tau/|\mathcal{A}|)^{-(m-t+1)}.$$

Proof. Fix some $\beta > 0$. By Lemma 4.1, we can choose some $S \subseteq \Omega$ with $\nu^\pi(S) = 1$ such that $\forall h \in S, \exists t_0, \forall t \geq t_0, \forall n \in [0, N-1]$,

$$D(q_n, \tilde{\nu}_n^\pi \mid \mathrm{aug}_n(h)_{<t}) < \beta. \tag{9}$$

Consider some $h \in S$ and $t \geq t_0$. For all $m \in [t, t+N-H]$ and $n = m \bmod N$, we see that Eq. 2 is satisfied for both t and m, so we can apply aug_n on histories of lengths $t-1$ and $m-1$. Then,

$$\begin{aligned}
&\sum_{\substack{h'_{<m}\in\mathcal{T},\, a'_m\in\mathcal{A},\\ \tilde{z}_m\in\mathcal{Z}}} \nu^\pi\left(h'_{t:m-1}a'_m \mid h_{<t}\right) \cdot \left|\psi(\tilde{z}_m \mid h'_{<m}a'_m) - \nu^\pi(\tilde{z}_m \mid h'_{<m}a'_m)\right| \\
&\overset{\text{(a)}}{=} \sum_{\substack{h'_{<m}\in\mathcal{T},\, a'_m\in\mathcal{A},\\ \tilde{z}_m\in\mathcal{Z}}} \nu^\pi\left(h'_{t:m-1}a'_m \mid h_{<t}\right) \cdot \left|\psi_n(\tilde{z}_m \mid \mathrm{aug}_n(h'_{<m})\, a'_m) - \tilde{\nu}_n^\pi(\tilde{z}_m \mid \mathrm{aug}_n(h'_{<m})\, a'_m)\right| \\
&\overset{\text{(b)}}{=} \sum_{\substack{h'_{<m}\in\mathcal{T},\, a'_m\in\mathcal{A},\\ \tilde{z}_m\in\mathcal{Z}}} \tilde{\nu}_n^\pi\left(h'_{t:m-1}a'_m \mid \mathrm{aug}_n(h_{<t})\right) \cdot \left|\psi_n(\tilde{z}_m \mid \mathrm{aug}_n(h_{<t})\, h'_{t:m-1}a'_m) - \tilde{\nu}_n^\pi(\tilde{z}_m \mid \mathrm{aug}_n(h_{<t})\, h'_{t:m-1}a'_m)\right| \\
&\overset{\text{(c)}}{=} \sum_{\substack{h'_{<m}\in\mathcal{T},\, a'_m\in\mathcal{A},\\ \tilde{z}_m\in\mathcal{Z}}} \left|q_n\left(h'_{t:m-1}a'_m\tilde{z}_m \mid \mathrm{aug}_n(h_{<t})\right) - \tilde{\nu}_n^\pi\left(h'_{t:m-1}a'_m\tilde{z}_m \mid \mathrm{aug}_n(h_{<t})\right)\right| \\
&\overset{\text{(d)}}{\leq} 2 \cdot D(q_n, \tilde{\nu}_n^\pi \mid \mathrm{aug}_n(h)_{<t}).
\end{aligned} \tag{10}$$

Equality (a) is just rewriting the unified predictor ψ as the phase n predictor ψ_n (and likewise rewriting the true conditional using $\tilde{\nu}_n^\pi$) via the definitions of ψ and $\tilde{\nu}_n^\pi$. Equality (b) uses that $\mathrm{aug}_n(h_{<t})$ is a deterministic function of $h_{<t}$, and thus conditioning under $h_{<t}$ is the same as conditioning under $\mathrm{aug}_n(h_{<t})$. Equality (c) uses the definitions of q_n and $\tilde{\nu}_n^\pi$, noting that there are no augmented returns between indices t and m other than $\tilde{z}_m$. Inequality (d) is Eq. 3 applied to the disjoint cylinder sets formed by $h'_{<m}a'_m\tilde{z}_m$.

By the definition of AIQI, each action in $a'_{t:m}$ receives a probability of at least $\tau/|\mathcal{A}|$ from the AIQI policy π. Thus,

$$\begin{aligned}
\nu^\pi\left(h'_{t:m-1}a'_m \mid h_{<t}\right) &= \prod_{i=t}^{m} \pi(a'_i \mid h'_{<i}) \prod_{i=t}^{m-1} \nu(e'_i \mid h'_{<i}a'_i) \\
&\geq (\tau/|\mathcal{A}|)^{m-t+1}\, \nu(e'_{t:m-1} \mid h_{<t} \parallel a'_{t:m-1}).
\end{aligned} \tag{11}$$

By combining Eqs. 9 to 11, we obtain the desired result. □

Lemma 4.3 (Convergence of Q-value prediction)**.** *Under the conditions established in Lemma 4.2,*

$$\sum_{h'_{<m}\in\mathcal{T},a'_m\in\mathcal{A}} \nu(e'_{t:m-1} \mid h_{<t} \parallel a'_{t:m-1}) \cdot \delta_Q(h'_{<m}a'_m) < 2\beta\left(\tau/|\mathcal{A}|\right)^{-(m-t+1)} + M^{-1} + \eta.$$

Proof. Let us define the lower-approximate Q-value as

$$\underline{Q}^\pi_\nu(h_{<t}, a_t) := \mathbb{E}_{\nu^\pi}[z_t \mid h_{<t}a_t] = \sum_{z_t\in\mathcal{Z}} z_t \cdot \nu^\pi(z_t \mid h_{<t}, a_t).$$

Recall that the H-step Q-value is $Q^\pi_{\nu,H}(h_{<t}, a_t) = \mathbb{E}_{\nu^\pi}[R_{t,H} \mid h_{<t}a_t]$. Since $|z_t - R_{t,H}| < 1/M$, we see that $|\underline{Q}^\pi_\nu(\cdot) - Q^\pi_{\nu,H}(\cdot)| < 1/M$. Since we set H to be the η-effective horizon $H(\eta)$, $|R_{t,H} - R_t| \le \eta$, so that $|Q^\pi_{\nu,H}(\cdot) - Q^\pi_\nu(\cdot)| \le \eta$. Thus we obtain $|\underline{Q}^\pi_\nu(\cdot) - Q^\pi_\nu(\cdot)| < M^{-1} + \eta$, and

$$\sum_{h'_{<m}\in\mathcal{T}} \nu(e'_{t:m-1} \mid h_{<t} \parallel a'_{t:m-1}) \cdot |\underline{Q}^\pi_\nu(h'_{<m}a'_m) - Q^\pi_\nu(h'_{<m}a'_m)| < M^{-1} + \eta, \tag{12}$$

where we used Eq. 1.

Recall the definition of $\hat{Q}$ in Definition 3.3.

$$\begin{aligned}
&\sum_{\substack{h'_{<m}\in\mathcal{T},\\ a'_m\in\mathcal{A}}} \nu(e'_{t:m-1} \mid h_{<t} \parallel a'_{t:m-1}) \cdot |\hat{Q}(h'_{<m}\, a'_m) - \underline{Q}^\pi_\nu(h'_{<m}\, a'_m)| \\
&\overset{\text{(a)}}{=} \sum_{\substack{h'_{<m}\in\mathcal{T},\\ a'_m\in\mathcal{A}}} \nu(e'_{t:m-1} \mid h_{<t} \parallel a'_{t:m-1}) \cdot \Big|\sum_{\tilde{z}_m\in\mathcal{Z}} \tilde{z}_m \left[\psi(\tilde{z}_m \mid h'_{<m}\, a'_m) - \nu^\pi(\tilde{z}_m \mid h'_{<m}\, a'_m)\right]\Big| \\
&\overset{\text{(b)}}{\le} \sum_{\substack{h'_{<m}\in\mathcal{T},\\ a'_m\in\mathcal{A},\, \tilde{z}_m\in\mathcal{Z}}} \nu(e'_{t:m-1} \mid h_{<t} \parallel a'_{t:m-1}) \cdot \tilde{z}_m \cdot |\psi(\tilde{z}_m \mid h'_{<m}\, a'_m) - \nu^\pi(\tilde{z}_m \mid h'_{<m}\, a'_m)| \\
&\overset{\text{(c)}}{\le} \sum_{\substack{h'_{<m}\in\mathcal{T},\\ a'_m\in\mathcal{A},\, \tilde{z}_m\in\mathcal{Z}}} \nu(e'_{t:m-1} \mid h_{<t} \parallel a'_{t:m-1}) \cdot |\psi(\tilde{z}_m \mid h'_{<m}\, a'_m) - \nu^\pi(\tilde{z}_m \mid h'_{<m}\, a'_m)| \\
&= \sum_{\substack{h'_{<m}\in\mathcal{T},\\ a'_m\in\mathcal{A}}} \nu(e'_{t:m-1} \mid h_{<t} \parallel a'_{t:m-1}) \cdot \delta_\psi(h'_{<m}a'_m) \\
&\overset{\text{(d)}}{<} 2\beta\left(\tau/|\mathcal{A}|\right)^{-(m-t+1)}.
\end{aligned} \tag{13}$$

Equality (a) expands $\hat{Q}$ and $\underline{Q}^\pi_\nu$ as expectations over $\tilde{z}_m$ (by definition). Inequality (b) is the triangle inequality. Inequality (c) uses $\tilde{z}_m \in [0, 1]$. Finally, (d) is an application of Lemma 4.2.

Combining Eqs. 12 and 13 with triangle inequality, we obtain the desired result. □

Lemma 4.4 (Convergence of one-step optimality gap)**.** *Under the conditions established in Lemma 4.2,*

$$\sum_{h'_{<m}\in\mathcal{T}} \nu(e'_{t:m-1} \mid h_{<t} \parallel a'_{t:m-1}) \cdot \delta_1(h'_{<m}) < 2 \cdot (2\beta\left(\tau/|\mathcal{A}|\right)^{-(m-t+1)} + M^{-1} + \eta) + 2\tau.$$

Proof.

$$
\begin{aligned}
\delta_1(h'_{<m}) &= |\max_a Q(h'_{<m}, a) - V(h'_{<m})| \\
&\overset{\text{(a)}}{=} \Big|\max_a Q(h'_{<m}, a) - \big[(1-\tau)Q(h'_{<m}, \arg\max_a \hat{Q}) + (\tau/|\mathcal{A}|)\sum_a Q(h'_{<m}, a)\big]\Big| \\
&\overset{\text{(b)}}{\le} |\max_a Q(h'_{<m}, a) - Q(h'_{<m}, \arg\max_a \hat{Q})| + 2\tau \\
&\le |\max_a Q(h'_{<m}, a) - \max_a \hat{Q}(h'_{<m}, a)| + |\max_a \hat{Q}(h'_{<m}, a) - Q(h'_{<m}, \arg\max_a \hat{Q})| + 2\tau \\
&\overset{\text{(c)}}{\le} \max_a |Q(h'_{<m}, a) - \hat{Q}(h'_{<m}, a)| + |\hat{Q}(h'_{<m}, a^\ddagger) - Q(h'_{<m}, a^\ddagger)| + 2\tau \\
&= |Q(h'_{<m}, a^\dagger) - \hat{Q}(h'_{<m}, a^\dagger)| + |\hat{Q}(h'_{<m}, a^\ddagger) - Q(h'_{<m}, a^\ddagger)| + 2\tau,
\end{aligned}
$$

where $a^\dagger := \arg\max_a |Q(h'_{<m}, a) - \hat{Q}(h'_{<m}, a)|$ and $a^\ddagger := \arg\max_a \hat{Q}(h'_{<m}, a)$ are functions of $h'_{<m}$.

Equality (a) uses the Bellman equation $V(\cdot) = \sum_a \pi(a \mid \cdot)\, Q(\cdot, a)$ together with the AIQI policy definition (greedy with probability $1-\tau$ and uniform exploration with probability τ). Inequality (b) upper-bounds the effect of the exploration mixture; since $Q \in [0,1]$, the mixture term can change the value by at most 2τ. Inequality (c) uses the standard bound $|\max f - \max g| \le \max |f - g|$ and that $a^\ddagger = \arg\max_a \hat{Q}(h'_{<m}, a)$.

Multiplying both sides by $\nu(e'_{t:m-1} \mid h_{<t} \parallel a'_{t:m-1})$ and summing, we obtain

$$
\begin{aligned}
&\sum_{h'_{<m} \in \mathcal{T}} \nu(e'_{t:m-1} \mid h_{<t} \parallel a'_{t:m-1}) \cdot \delta_1(h'_{<m}) \\
&\le \sum_{h'_{<m} \in \mathcal{T}} \nu(e'_{t:m-1} \mid h_{<t} \parallel a'_{t:m-1}) \Big(|Q(h'_{<m}, a^\dagger) - \hat{Q}(h'_{<m}, a^\dagger)| + |\hat{Q}(h'_{<m}, a^\ddagger) - Q(h'_{<m}, a^\ddagger)|\Big) \\
&\quad + \sum_{h'_{<m} \in \mathcal{T}} \nu(e'_{t:m-1} \mid h_{<t} \parallel a'_{t:m-1}) \cdot 2\tau \\
&\overset{\text{(a)}}{\le} 2 \cdot \sum_{\substack{h'_{<m} \in \mathcal{T}, \\ a'_m \in \mathcal{A}}} \nu(e'_{t:m-1} \mid h_{<t} \parallel a'_{t:m-1}) \cdot |\hat{Q}(h'_{<m} a'_m) - Q^\pi_\nu(h'_{<m} a'_m)| + 2\tau \\
&\overset{\text{(b)}}{<} 2 \cdot \left(2\beta\,(\tau/|\mathcal{A}|)^{-(m-t+1)} + M^{-1} + \eta\right) + 2\tau.
\end{aligned}
$$

Inequality (a) uses that $a^\dagger(h'_{<m})$ and $a^\ddagger(h'_{<m})$ select particular actions, so summing over all $a'_m \in \mathcal{A}$ upper-bounds each choice. Finally, (b) applies Lemma 4.3. □

Lemma 4.5 (Bound on global optimality gap)**.** *For any environment ν, policy π, and history $h_{<t}$,*

$$
\delta_\infty(h_{<t}) \le \gamma \sum_{e'_t} \nu(e'_t \mid h_{<t}\, a'_t)\, \delta_\infty(h_{<t}\, a'_t\, e'_t) + \delta_1(h_{<t}),
$$

where $a'_t = \arg\max_a \left(Q^*_\nu(h_{<t}, a) - Q^\pi_\nu(h_{<t}, a)\right)$.

Proof. We omit the subscript ν under Q and V for brevity.

$$\begin{aligned}
\delta_\infty(h_{<t}) &= |\max_a Q^*(h_{<t},a) - V^\pi(h_{<t})| \\
&\overset{\text{(a)}}{\leq} |\max_a Q^*(h_{<t},a) - \max_a Q^\pi(h_{<t},a)| + |\max_a Q^\pi(h_{<t},a) - V^\pi(h_{<t})| \\
&\overset{\text{(b)}}{\leq} \max_a |Q^*(h_{<t},a) - Q^\pi(h_{<t},a)| + \delta_1(h_{<t}) \\
&\overset{\text{(c)}}{=} |Q^*(h_{<t},a'_t) - Q^\pi(h_{<t},a'_t)| + \delta_1(h_{<t}) \\
&\overset{\text{(d)}}{=} \left| \sum_{e'_t} \nu(e_t|h_{<t}a'_t) \left[((1-\gamma)r_t + \gamma V^*(h_{<t}a'_te'_t)) - ((1-\gamma)r_t + \gamma V^\pi(h_{<t}a'_te'_t))\right] \right| + \delta_1(h_{<t}) \\
&= \left| \gamma \sum_{e'_t} \nu(e_t|h_{<t}a'_t) \left[V^*(h_{<t}a'_te'_t) - V^\pi(h_{<t}a'_te'_t)\right] \right| + \delta_1(h_{<t}) \\
&\overset{\text{(e)}}{\leq} \gamma \sum_{e'_t} \nu(e_t \mid h_{<t}a'_t)\delta_\infty(h_{<t}a'_te'_t) + \delta_1(h_{<t})
\end{aligned}$$

Inequality (a) is the triangle inequality. Inequality (b) uses $|\max f - \max g| \leq \max |f-g|$ and the definition of δ_1. Equality (c) is by the definition of $a'_t = \arg\max_a |Q^*(h_{<t},a) - Q^\pi(h_{<t},a)|$ as stated in the lemma. Equality (d) is the Bellman expansion of Q^* and Q^π. Finally, (e) uses $|\sum_x p_x u_x| \leq \sum_x p_x|u_x|$ and $\delta_\infty = V^* - V^\pi$. □

Lemma 4.7 (Optimality in $\mathcal{M}$ implies optimality in ξ)**.** *Given a mixture ξ of environment class $\mathcal{M}$, and a policy π that is strong asymptotically ε-optimal with respect to $\mathcal{M}$, the same policy π is asymptotically ε-optimal in ξ:*

$$\limsup_{t\to\infty} V^*_\xi(h_{<t}) - V^\pi_\xi(h_{<t}) \leq \varepsilon$$

holds both ξ^π-a.s. and, for all $\nu \in \mathcal{M}$, ν^π-a.s.

Proof. Proving that the inequality in our theorem holds ξ^π-almost surely implies that the inequality holds ν^π-almost surely for any $\nu \in \mathcal{M}$. This is because, if E is the set of all outcomes h where the inequality *doesn't* hold, $\nu^\pi(E) \leq \xi^\pi(E)/w(\nu) = 0$. Thus it suffices to prove that the inequality holds ξ^π-almost surely.

By the definition of the mixture ξ and the value function V, we see that

$$V^\pi_\xi(h_{<t}) = \sum_{\nu\in\mathcal{M}} w(\nu \mid h_{<t})\, V^\pi_\nu(h_{<t}).$$

Let π^*_ξ be an optimal policy in ξ. Then

$$V^*_\xi(h_{<t}) = V^{\pi^*_\xi}_\xi(h_{<t}) = \sum_\nu w(\nu \mid h_{<t})\, V^{\pi^*_\xi}_\nu(h_{<t}) \leq \sum_\nu w(\nu \mid h_{<t})\, V^*_\nu(h_{<t}),$$

so that

$$0 \leq V^*_\xi(h_{<t}) - V^\pi_\xi(h_{<t}) \leq \sum_\nu w(\nu \mid h_{<t}) \left(V^*_\nu(h_{<t}) - V^\pi_\nu(h_{<t})\right). \tag{14}$$

Thus it suffices to show that the right-hand side has $\limsup \leq \varepsilon$, ξ^π-almost surely.

For each environment $\nu \in \mathcal{M}$, by Theorem 4.6 there exists an event $S_\nu \subseteq \Omega$ with $\nu^\pi(S_\nu) = 1$ such that for all outcomes $h \in S_\nu$,

$$\limsup_{t\to\infty} \left(V^*_\nu(h_{<t}) - V^\pi_\nu(h_{<t})\right) \leq \varepsilon. \tag{15}$$

Define the posterior process $w_t(\nu) := w(\nu \mid h_{<t})$ and its limit $w_\infty(\nu) := \lim_{t\to\infty} w_t(\nu)$, which exists ξ^π-a.s. since $(w_t(\nu))_{t\geq 1}$ is a bounded martingale. Let us consider the joint probability measure on $\mathcal{M}\times\Omega$, where we first draw $\boldsymbol{\nu} \sim w$, then draw the outcome $h \sim \boldsymbol{\nu}^\pi$. Then ξ^π is its marginal over Ω. A standard identity is the following:

$$\mathbb{E}[w_\infty(\nu)\mathbb{1}_E] = \mathbb{P}(\boldsymbol{\nu} = \nu, h \in E) = w(\nu)\nu^\pi(E). \tag{16}$$

Due to Eq. 16 and $\nu^\pi(S_\nu^\complement) = 0$, $\mathbb{E}_{\xi^\pi}[w_\infty(\nu)\mathbb{1}_{S_\nu^\complement}] = 0$. Since the integrand is nonnegative,

$$w_\infty(\nu)\mathbb{1}_{S_\nu^\complement} = 0, \quad \xi^\pi\text{-a.s.} \tag{17}$$

Hence, there exists a ξ^π-probability-one set, such that for any outcome h in this set, Eq. 15 holds for all ν where $w_\infty(\nu) > 0$.

Now define the per-environment gap

$$\delta_{t,\nu} := V_\nu^*(h_{<t}) - V_\nu^\pi(h_{<t}) \in [0,1],$$

and the posterior-weighted gap

$$G_t := \sum_\nu w_t(\nu)\delta_{t,\nu}.$$

By Eq. 14, it is enough to show $\limsup_{t\to\infty} G_t \le \varepsilon$, ξ^π-a.s.

Fix an outcome h in a ξ^π-probability-one set on which $w_t(\nu) \to w_\infty(\nu)$ for all ν and Eq. 17 holds for all ν. Let $\alpha > 0$. Since $w_\infty(\cdot)$ is a probability mass function on a countable set, there exists a finite $F \subseteq \mathcal{M}$ such that $w_\infty(\nu) > 0$ for all $\nu \in F$, and

$$\sum_{\nu\in F} w_\infty(\nu) \ge 1 - \alpha. \tag{18}$$

Since $w_\infty(\nu) > 0$, we have $\limsup_{t\to\infty} \delta_{t,\nu} \le \varepsilon$ by Eq. 17. Therefore, for every such ν there exists t_ν such that for all $t \ge t_\nu$, $\delta_{t,\nu} \le \varepsilon + \alpha$. Let $t_0 := \max_{\nu\in F} t_\nu$. Then for all $t \ge t_0$,

$$\sum_{\nu\in F} w_t(\nu)\delta_{t,\nu} \le (\varepsilon+\alpha)\sum_{\nu\in F} w_t(\nu) \le \varepsilon + \alpha. \tag{19}$$

Moreover, since $\delta_{t,\nu} \le 1$, $\sum_{\nu\notin F} w_t(\nu)\delta_{t,\nu} \le \sum_{\nu\notin F} w_t(\nu)$. Because F is finite and $w_t(\nu) \to w_\infty(\nu)$ pointwise, we have $\sum_{\nu\in F} w_t(\nu) \to \sum_{\nu\in F} w_\infty(\nu)$, so by Eq. 18, $\sum_{\nu\notin F} w_t(\nu) = 1 - \sum_{\nu\in F} w_t(\nu) \to 1 - \sum_{\nu\in F} w_\infty(\nu) \le \alpha$. Hence there exists t_1 such that for all $t \ge t_1$,

$$\sum_{\nu\notin F} w_t(\nu) \le 2\alpha. \tag{20}$$

Combining Eqs. 19 and 20, we obtain that for all $t \ge \max\{t_0, t_1\}$,

$$G_t = \sum_{\nu\in F} w_t(\nu)\delta_{t,\nu} + \sum_{\nu\notin F} w_t(\nu)\delta_{t,\nu} \le (\varepsilon+\alpha) + 2\alpha.$$

Since $\alpha > 0$ is arbitrary, we conclude that

$$\limsup_{t\to\infty} G_t \le \varepsilon, \quad \xi^\pi\text{-a.s.} \qquad \square$$

Theorem 4.10 (AIQI is not self-optimizing)**.** *Suppose the unified predictor ψ is built from classes $\mathcal{P}_n$ that contain all computable variable-order Markov models. For any period N, exploration rate τ, horizon $H \ge 2$, tolerance ε small enough (depending on H), and discretization level M large enough (depending on H, ε, τ), there exist an environment class $\mathcal{M}$, a historic policy π', and a discount factor $\gamma \in (0,1)$, for which there exists a self-optimizing policy, but the AIQI policy $\hat{\pi}_\psi^{H,M,N,\tau}$ is not even ε-self-optimizing.*

Proof. Given $H \ge 2$, define

$$c_H := \max_{\gamma\in(0,1)} (1-\gamma)\gamma^{H-1} = \frac{1}{H}\Big(\frac{H-1}{H}\Big)^{H-1}.$$

Fix a tolerance $\varepsilon \in (0, c_H)$.

Let

$$\gamma := \frac{H-1}{H} \in (0,1), \qquad c := (1-\gamma)\gamma^{H-1} = c_H.$$

Choose

$$\delta := \frac{1-\varepsilon/c}{3} \in (0,1), \qquad K \in \mathbb{N} \text{ s.t. } \frac{\tau}{K} \le \frac{\delta}{2}.$$

Finally, assume M is large enough so that discretization cannot collapse the strict value gap:

$$\frac{1}{M} < c\frac{\delta}{2} \qquad \text{equivalently} \qquad M > \frac{2}{c\delta}. \tag{21}$$

Environment. Let $\mathcal{A} = \{1, 2, \ldots, K\}$, $\mathcal{O} = \{\cdot, \star\}$, and $\mathcal{R} = \{0, \delta, 1\}$. The environment μ consists of episodes of length H. For the first $H-1$ steps of each episode, output $(o, r) = (\cdot, 0)$. On the last step output $(o, r) = (\star, r)$ where r depends on the H actions taken in that episode:

- if all H actions are 1, then $r = 1$;
- if the first action is 1 but not all H actions are 1, then $r = 0$;
- if the first action is not 1, then $r = \delta$.

Then a new episode starts and the construction repeats forever.

Historic policy. Define π' episode-wise as follows. At the first step of an episode, sample a uniformly from $\mathcal{A}$. If this first action is not 1, sample the remaining $H-1$ actions uniformly from $\mathcal{A}$. If the first action is 1, then for each of the remaining $H-1$ steps play 1 with probability

$$p \;:=\; \Big(\frac{\delta}{2}\Big)^{1/(H-1)},$$

and otherwise sample uniformly from $\{2, \ldots, K\}$. Then conditioned on starting an episode with action 1, the probability of playing 1 for all remaining $H-1$ steps is $p^{H-1} = \delta/2$.

Existence of a self-optimizing policy. Since $\mathcal{M}$ is a singleton, the μ-optimal policy π^*_μ that always plays action 1 is trivially self-optimizing for $(\mathcal{M}, \pi')$.

AIQI learns the π'-value. Under $\mu^{\pi'}$, the (periodically augmented) discretized return process is a computable variable-order Markov model of order $O(H)$ (episode boundaries are marked by $\star$ and terminal rewards depend only on the H actions in the current episode). Because ψ is a Bayesian mixture over a class containing the true return-predictor for $\mu^{\pi'}$, Blackwell and Dubins [1962] implies that ψ's posterior predictive distribution converges to the true conditional return distribution under $\mu^{\pi'}$. In particular, AIQI's value estimate

$$\hat{Q}(h_{<t}, a) = \sum_{\tilde{z} \in \mathcal{Z}} \tilde{z}\; \psi(\tilde{z} \mid h_{<t} a)$$

converges to the true conditional expectation of the discretized target z_t under $\mu^{\pi'}$:

$$\underline{Q}^{\pi'}_{\mu}(h_{<t}, a) := \mathbb{E}^{\pi'}_{\mu}[z_t \mid h_{<t}, a_t = a].$$

Now fix an episode-start time t (immediately after observing $\star$). In this environment, the only nonzero reward within the next H steps is the terminal reward, so the H-step discounted return from t satisfies

$$R_{t,H} = (1-\gamma) \sum_{k=0}^{H-1} \gamma^k r_{t+k} = (1-\gamma)\gamma^{H-1} r_{t+H-1} = c\, r_{t+H-1}.$$

If we force $a_t \neq 1$ and then follow π', the terminal reward is deterministically δ, hence

$$Q^{\pi'}_{\mu,H}(h_{<t}, a \neq 1) = c\delta.$$

If we force $a_t = 1$ and then follow π', the terminal reward is 1 with probability $\delta/2$ and 0 otherwise, hence

$$Q^{\pi'}_{\mu,H}(h_{<t}, a = 1) = c \cdot \frac{\delta}{2}.$$

Because AIQI optimizes the discretized target $z_t = \lfloor M R_{t,H} \rfloor / M$ and $0 \leq R_{t,H} - z_t < 1/M$ pointwise, we have for each action a:

$$Q^{\pi'}_{\mu,H}(h_{<t}, a) - \frac{1}{M} < \underline{Q}^{\pi'}_{\mu}(h_{<t}, a) \leq Q^{\pi'}_{\mu,H}(h_{<t}, a).$$

Therefore at episode starts,

$$\underline{Q}^{\pi'}_{\mu}(h_{<t}, a \neq 1) - \underline{Q}^{\pi'}_{\mu}(h_{<t}, a = 1) \;\geq\; c\frac{\delta}{2} - \frac{1}{M} \;>\; 0$$

by (21). So action 1 is strictly suboptimal under the discretized π'-value at episode starts.

By $\mu^{\pi'}$-a.s. convergence of $\hat{Q}$ to $\underline{Q}^{\pi'}_{\mu}$, there exists a t_0 such that for all episode-start times $t \geq t_0$, AIQI's exploitation choice $a^*(h_{<t}) \neq 1$.

Lower-bound on the self-optimization gap. Fix any episode-start $t \geq t_0$. Since $a^*(h_{<t}) \neq 1$, AIQI chooses action 1 at time t with probability at most τ/K (it can only happen by uniform exploration). Hence the expected terminal reward of the *next* episode satisfies

$$\mathbb{E}_\mu^{\hat{\pi}}[r_{t+H-1} \mid h_{<t}] \;\leq\; \delta\Big(1 - \frac{\tau}{K}\Big) + 1 \cdot \frac{\tau}{K} \;\leq\; \delta + \frac{\tau}{K} \;\leq\; \frac{3\delta}{2}.$$

Let π^* denote the policy that always outputs action 1. Then π^* achieves terminal reward 1 at the end of *every* episode, in particular $\mathbb{E}_\mu^{\pi^*}[r_{t+H-1} \mid h_{<t}] = 1$. Because rewards are in $[0,1]$, the termwise differences at episode-terminal times are nonnegative, so

$$\begin{aligned}
V_\mu^*(h_{<t}) - V_\mu^{\hat{\pi}}(h_{<t}) &\geq V_\mu^{\pi^*}(h_{<t}) - V_\mu^{\hat{\pi}}(h_{<t}) \\
&= (1-\gamma) \sum_{j=0}^{\infty} \gamma^{jH+H-1} \Big(1 - \mathbb{E}_\mu^{\hat{\pi}}[r_{t+jH+H-1} \mid h_{<t}]\Big) \\
&\geq (1-\gamma)\gamma^{H-1} \Big(1 - \mathbb{E}_\mu^{\hat{\pi}}[r_{t+H-1} \mid h_{<t}]\Big) \\
&\geq c\Big(1 - \frac{3\delta}{2}\Big).
\end{aligned}$$

With our choice $\delta = (1 - \varepsilon/c)/3$, we have

$$1 - \frac{3\delta}{2} = 1 - \frac{1 - \varepsilon/c}{2} = \frac{1 + \varepsilon/c}{2},$$

so

$$V_\mu^*(h_{<t}) - V_\mu^{\hat{\pi}}(h_{<t}) \;\geq\; c \cdot \frac{1 + \varepsilon/c}{2} \;=\; \frac{c + \varepsilon}{2} \;>\; \varepsilon.$$

Episode starts occur infinitely often, hence

$$\limsup_{t \to \infty} \big(V_\mu^*(h_{<t}) - V_\mu^{\hat{\pi}}(h_{<t})\big) > \varepsilon \qquad \mu^{\pi'}\text{-a.s.}$$

Thus AIQI is not ε-self-optimizing for $(\mathcal{M}, \pi')$. □

C PROOFS FOR SELF-AIXI

Lemma 5.2 (Convergence of ξ^ζ to ν^π)**.** *For any environment $\nu \in \mathcal{M}$ and policy $\pi \in P$,*

$$\lim_{t \to \infty} D\big(\xi^\zeta, \nu^\pi \mid h_{<t}\big) = 0, \quad \nu^\pi\text{-a.s.}$$

Proof. Since ν^π is absolutely continuous w.r.t. ξ^π, we obtain by Blackwell and Dubins [1962] that

$$\lim_{t \to \infty} D(\xi^\pi, \nu^\pi \mid h_{<t}) = 0, \quad \nu^\pi\text{-a.s.}$$

Similarly, we obtain that

$$\lim_{t \to \infty} D\big(\xi^\zeta, \xi^\pi \mid h_{<t}\big) = 0, \quad \xi^\pi\text{-a.s.}$$

ξ^π-almost sure convergence implies ν^π-almost sure convergence, since ν^π is absolutely continuous w.r.t. ξ^π. By triangle inequality for total variation distance, we obtain the desired result. □

Lemma 5.4 (Average conditional TV bound)**.** *Let P and Q be probability measures on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is finite. Let P_X and Q_X denote the marginals on $\mathcal{X}$. Then*

$$\sum_{x \in \mathcal{X}} P_X(x)\, D\big(P, Q \mid X = x\big) \;\leq\; 2D(P, Q).$$

Proof. Define a probability measure R on $\mathcal{X} \times \mathcal{Y}$ by

$$R(x, dy) := P_X(x)\, Q(dy \mid x).$$

Thus, R has the same marginal on X as P, but uses the conditional law of $Y \mid X$ from Q. We can show that

$$D(P, R) = \sum_{x \in \mathcal{X}} P_X(x)\, D\big(P, Q \mid X = x\big),$$

and that

$$D(R, Q) = D(P_X, Q_X).$$

Now apply the triangle inequality:

$$D(P, R) \le D(P, Q) + D(R, Q).$$

Substituting the identities above gives

$$\sum_{x \in \mathcal{X}} P_X(x)\, D\big(P, Q \mid X = x\big) \;\le\; D(P, Q) + D(P_X, Q_X).$$

Finally, total variation distance is non-increasing under measurable maps. Applying this to the projection $(x, y) \mapsto x$, we get $D(P_X, Q_X) \le D(P, Q)$. Therefore,

$$D(P, Q) + D(P_X, Q_X) \le 2D(P, Q),$$

which completes the proof.

□

Lemma 5.5 (Convergence of Q-value prediction for ε-greedy Self-AIXI)**.** *Let $\pi = \pi_S$. For every $\beta > 0$, there exists a ν^π-probability-one set $S \subseteq \Omega$ where, for every outcome $h \in S$, there exists t_0 such that: for $t \ge t_0$ and $m \ge t$,*

$$\sum_{h'_{<m} \in \mathcal{T}, a'_m \in \mathcal{A}} \nu(e'_{t:m-1} \mid h_{<t} \parallel a'_{t:m-1}) \cdot \delta_Q(h'_{<m} a'_m) < 2\beta\, (\tau/|\mathcal{A}|)^{-(m-t+1)}.$$

Proof. Fix $\beta > 0$. By Lemma 5.2, there exists a ν^π-probability-one set $S \subseteq \Omega$ such that for every $h \in S$ there is some t_0 with

$$D(\nu^\pi, \xi^\zeta \mid h_{<t}) < \beta \qquad \text{for all } t \ge t_0.$$

Now fix $h \in S$, $t \ge t_0$, and $m \ge t$. By Lemma 5.4,

$$\sum_{\substack{h'_{<m} \in \mathcal{T}, \\ a'_m \in \mathcal{A}}} \nu^\pi\left(h'_{t:m-1} a'_m \mid h_{<t}\right) \cdot D(\nu^\pi, \xi^\zeta \mid h'_{<m} a'_m) \le 2D(\nu^\pi, \xi^\zeta \mid h_{<t}),$$

and by Lemma 5.3,

$$\delta_Q(h'_{<m} a'_m) = \left|Q^\zeta_\xi(h'_{<m}, a'_m) - Q^\pi_\nu(h'_{<m}, a'_m)\right| \le D(\nu^\pi, \xi^\zeta \mid h'_{<m} a'_m).$$

Hence

$$\sum_{\substack{h'_{<m} \in \mathcal{T}, \\ a'_m \in \mathcal{A}}} \nu^\pi\left(h'_{t:m-1} a'_m \mid h_{<t}\right) \cdot \delta_Q(h'_{<m} a'_m) \le 2D(\nu^\pi, \xi^\zeta \mid h_{<t}) < 2\beta.$$

Since π_S assigns probability of at least $\tau/|\mathcal{A}|$ to each action,

$$\begin{aligned} \nu^\pi\left(h'_{t:m-1} a'_m \mid h_{<t}\right) &= \prod_{i=t}^{m} \pi(a'_i \mid h'_{<i}) \prod_{i=t}^{m-1} \nu(e'_i \mid h'_{<i} a'_i) \\ &\ge (\tau/|\mathcal{A}|)^{m-t+1}\, \nu(e'_{t:m-1} \mid h_{<t} \parallel a'_{t:m-1}), \end{aligned}$$

and we obtain the desired result:

$$\sum_{\substack{h'_{<m} \in \mathcal{T}, \\ a'_m \in \mathcal{A}}} \nu(e'_{t:m-1} \mid h_{<t} \parallel a'_{t:m-1}) \cdot \delta_Q(h'_{<m} a'_m) < 2\beta\, (\tau/|\mathcal{A}|)^{-(m-t+1)}.$$

□

D MULTI-AGENT ENVIRONMENT

The single-agent results in § 4.3 extend directly to multi-agent sequential interaction by viewing each agent as acting in a *subjective environment* induced by the other agents [Leike et al., 2016b, Section 4.1]. Consider n agents interacting in a multi-agent environment

$$\sigma : (\mathcal{A}^n \times \mathcal{E}^n)^* \times \mathcal{A}^n \to \Delta(\mathcal{E}^n),$$

where at time t the joint action and percept are $\mathbf{a}_t := (a_t^1, \dots, a_t^n)$ and $\mathbf{e}_t := (e_t^1, \dots, e_t^n)$. Agent i only observes its own action-percept history

$$h_{<t}^i := a_1^i e_1^i \dots a_{t-1}^i e_{t-1}^i.$$

Given n policies $\pi_1, \dots, \pi_n$, let σ_i denote the subjective environment of agent i, obtained by combining σ with the policies of the other agents and marginalizing out the components not observed by agent i. Together with π_i, this defines an ordinary environment-policy pair, so the formalism of the previous sections applies unchanged.

We say that π_i is an ε-best response at history $h_{<t}^i$ if

$$V_{\sigma_i}^*(h_{<t}^i) - V_{\sigma_i}^{\pi_i}(h_{<t}^i) \le \varepsilon.$$

If this holds for every $i \in \{1, \dots, n\}$, then the policies $\pi_{1:n}$ is an ε-Nash equilibrium at time t.

As in Definition 3.4, the key self-referential requirement is that each agent's unified predictor ψ_i has a grain of truth with respect to its own subjective environment σ_i. A nontrivial instantiation can again be obtained by letting $\mathcal{P}_n^i$ be the class of all O-computable return-predictors for every i and n, and assuming that the environment σ is O-computable.

Theorem D.1 (AIQI converges to ε-Nash equilibria)**.** *Fix a multi-agent environment σ and a tolerance $\varepsilon > 0$. For each agent $i \in \{1, \dots, n\}$, let*

$$\pi_i := \hat{\pi}_{\psi_i}^{H_i, M_i, N_i, \tau_i},$$

where the parameters H_i, M_i, N_i, τ_i satisfy the conditions of Theorem 4.6 with tolerance $\varepsilon/2$, and suppose that ψ_i has a grain of truth with respect to the subjective environment σ_i induced by σ and the policies $\pi_{\neq i}$. Then, the policies $\pi_{1:n}$ converge to an ε-Nash equilibrium, $\sigma^{\pi_{1:n}}$-almost surely.

Proof. Fix an agent i. Since σ_i is an ordinary environment for agent i, Theorem 4.6 applied with tolerance $\varepsilon/2$ gives

$$\limsup_{t \to \infty} \left(V_{\sigma_i}^*(h_{<t}^i) - V_{\sigma_i}^{\pi_i}(h_{<t}^i) \right) \le \varepsilon/2, \qquad \sigma_i^{\pi_i}\text{-a.s.}$$

Because $\sigma_i^{\pi_i}$ is exactly the marginal of the joint history distribution $\sigma^{\pi_{1:n}}$ on agent i's coordinates, the same statement holds $\sigma^{\pi_{1:n}}$-almost surely. Hence, for $\sigma^{\pi_{1:n}}$-almost every outcome h, there exists t_i such that

$$V_{\sigma_i}^*(h_{<t}^i) - V_{\sigma_i}^{\pi_i}(h_{<t}^i) \le \varepsilon \qquad \text{for all } t \ge t_i.$$

Since there are finitely many agents, intersecting these probability-one events over i still yields a probability-one event. Taking $t_0 := \max_i t_i$ proves that all agents are simultaneously ε-best responses for all $t \ge t_0$, which is exactly the claim. □

E GENERAL DISCOUNT SEQUENCES

We can extend all our results to general time-consistent discount sequences [Hutter et al., 2024, p. 245] that decay faster than a geometric sequence.

First, we should introduce the relevant concepts. A time-consistent discount sequence is a sequence $\{\gamma_k\}_{k=1}^\infty$ with $\gamma_k \ge 0$ and $\sum_{k=1}^\infty \gamma_k < \infty$. An example is the geometric discount sequence $\gamma_k = \gamma^k$. The discount normalization factor is defined as $\Gamma_t = \sum_{k=t}^\infty \gamma_k$. The full return at time t is defined as

$$R_t = \frac{1}{\Gamma_t} \sum_{k=t}^\infty \gamma_k r_k,$$

and the value functions V and Q are defined with R_t as in § 2.1. The value functions satisfy the general discount Bellman equations:

$$V_\nu^\pi(h_{<t}) = \sum_{a_t' \in \mathcal{A}} \pi(a_t' \mid h_{<t})\, Q_\nu^\pi(h_{<t}, a_t'),$$

$$Q^\pi_\nu(h_{<t}, a_t) = \frac{1}{\Gamma_t} \sum_{e'_t \in \mathcal{E}} \nu(e'_t \mid h_{<t} a_t) \big[\gamma_t r_t + \Gamma_{t+1} V^\pi_\nu(h_{<t} a_t e'_t)\big].$$

H-step returns and values are defined as in § 2.1. We define the η-effective horizon at time t as

$$H_t(\eta) := \min \left\{ H \in \mathbb{Z}_+ \;\middle|\; \frac{\Gamma_{t+H}}{\Gamma_t} \leq \eta \right\}.$$

One can easily show that the $H_t(\eta)$-step return at t differs from the full return at t by at most η. If a discount sequence decays faster than a geometric sequence, i.e.,

$$\limsup_{t \to \infty} \frac{\gamma_{t+1}}{\gamma_t} < \lambda$$

for some $\lambda < 1$, the effective horizon $H_t(\eta)$ for a fixed η has an upper bound across all t.

We only need to modify our proofs and results in several places to accommodate the generalization.

- The statement of Lemma 4.5 should change to

  $$\delta_\infty(h_{<t}) \leq \frac{\Gamma_{t+1}}{\Gamma_t} \sum_{e'_t} \nu(e'_t \mid h_{<t}\, a'_t)\, \delta_\infty(h_{<t}\, a'_t\, e'_t) + \delta_1(h_{<t}),$$

  which is just the original statement with γ replaced by Γ_{t+1}/Γ_t. The logic of its proof remains the same, except that we should use the general discount Bellman equation instead of the geometric discount Bellman equation, in one of its steps.
- The chain of inequalities resulting from Lemma 4.5 should be slightly modified: γ^l turns into Γ_{t+l}/Γ_t.
- In the proof of Theorem 4.6, the core inequality should be changed to

  $$\delta_\infty(h_{<t}) < \frac{\Gamma_{t+L}}{\Gamma_t} + \sum_{l=0}^{L-1} \frac{\Gamma_{t+l}}{\Gamma_t} \big[2(2\beta \left(\frac{\tau}{|\mathcal{A}|}\right)^{-(l+1)} + M^{-1} + \eta) + 2\tau\big],$$

  so that for large enough t,

  $$\delta_\infty(h_{<t}) < \lambda^L + \sum_{l=0}^{L-1} \lambda^l \big[2(2\beta \left(\frac{\tau}{|\mathcal{A}|}\right)^{-(l+1)} + M^{-1} + \eta) + 2\tau\big].$$

- Theorem 4.6 thus generalizes to general discount sequences that decay faster than a geometric sequence. It suffices to change γ with λ in the criterion for parameters, and use $H = \limsup_t H_t(\eta) + 1$.
- Theorem 4.8 and Theorem D.1 generalize accordingly.

F AIQI-CTW

We introduce AIQI-CTW, a computable instantiation of AIQI with the context tree weighting [CTW; Willems et al., 1995] algorithm. We also provide experiments that compare AIQI-CTW and MC-AIXI-CTW [Veness et al., 2011]. The source code for experiments can be found at `https://github.com/yegonkim/aiqi`.

Context tree weighting. Context Tree Weighting (CTW) is a fast, principled method for doing Bayesian sequence prediction over binary strings. It builds a variable-order Markov model by maintaining a context tree up to some maximum depth. Rather than committing to a single context length, CTW mixes predictions from all context depths using an efficient recursive weighting scheme, which effectively performs Bayesian model averaging over a large family of context trees. Remarkably, it makes online predictions by updating in time *linear* in the context depth, assigns non-zero probability to all sequences, and tends to quickly exploit repeated structure.

Parameter	Biased Rock-Paper-Scissors		Kuhn Poker		4×4 Grid	
	MC-AIXI-CTW	AIQI-CTW	MC-AIXI-CTW	AIQI-CTW	MC-AIXI-CTW	AIQI-CTW
Horizon H	—	4	—	2	—	12
Period N	—	4	—	2	—	12
Discretization M	—	9	—	9	—	13
Baseline exploration τ	—	0.01	—	0.01	—	0.01
Exploration (initial)	0.999	0.999	0.99	0.999	0.999	0.999
Explore decay rate	0.99999	0.9999	0.9999	0.9999	0.9999	0.9999
CTW depth	32	32	42	42	96	96
MCTS horizon	4	—	2	—	12	—
MCTS simulations	200	—	200	—	40	—
Learning period	5000	100000	5000	100000	5000	100000
Terminating age	10000	100000	10000	100000	10000	100000

Table 2: Hyperparameters for MC-AIXI-CTW and AIQI-CTW across environments

Implementation. AIQI-CTW is an instantiation of AIQI that uses CTW as its return-predictor ψ_n. MC-AIXI-CTW is similarly an approximation of AIXI that uses CTW as the environment model, and additionally a Monte Carlo Tree Search [MCTS; Coulom, 2006] algorithm that replaces exact planning. AIQI-CTW uses N CTW models, each corresponding to one of the N return-predictors. They receive the augmented sequences as described in § 3, to predict the M-discretized H-step returns. All rewards, observations, actions, and discretized returns are represented as blocks of bits. Note that the weighted probabilities of CTW are updated only with the bits corresponding to returns. This is in line with Definition 3.1. Expectation over returns is taken in an exact manner by computing all the probabilities. The implementation of MC-AIXI-CTW was taken from the `pyaixi` repository[1], and AIQI-CTW was adapted from this implementation.

Experimental setup. We tested the algorithms on 3 environments: "Biased Rock-Paper-Scissor", "Kuhn Poker", and "4×4 Grid". Biased Rock-Paper-Scissor is a repeated rock paper scissors game against an opponent with a simple exploitable bias: if it won the previous round by playing rock it plays rock again, otherwise it plays uniformly at random. Kuhn Poker is a simplified two-player, zero-sum poker game with a three-card deck (K, Q, J) in which players alternately choose between pass and bet under hidden information, capturing core phenomena such as bluffing and slow-playing. The 4×4 Grid task is a small gridworld environment in which an agent moves on a 4-by-4 lattice using the primitive actions up, down, left, right (with boundary constraints) to reach a designated goal state and receive reward. More details on the environments and experimental setup can be found in Veness et al. [2011].

Hyperparameters. For MC-AIXI-CTW, we use the default hyperparameter used in Veness et al. [2011]. One exception is the terminating age, which was reduced to keep the runtime manageable. The number of MCTS simulations in 4×4 Grid was also reduced for the same reason. For AIQI-CTW, we use a decaying ε-greedy exploration and a minimum baseline exploration rate τ. The CTW depth and horizon H were chosen to equal those of MC-AIXI-CTW. Note that we choose $N = H$, since the buffer period $N - H$ was introduced in § 4 solely as a theoretical device for proofs. We report all the important hyperparameters in Table 2.

Results. We performed each experiment on 8 seeds. Fig. 1 is a plot of the exponential moving average (EMA) of reward (with $\alpha = 10^{-3}$) against wall clock time. AIQI-CTW spends much less time in deciding an action than MC-AIXI-CTW, since it doesn't involve planning with MCTS. In all experiments we find that AIQI-CTW has an advantage over MC-AIXI-CTW given the restricted computational budget. Note that optimal performance can be obtained with MC-AIXI-CTW, albeit with orders of magnitude more compute, as described in Veness et al. [2011].

[1] https://github.com/sgkasselau/pyaixi

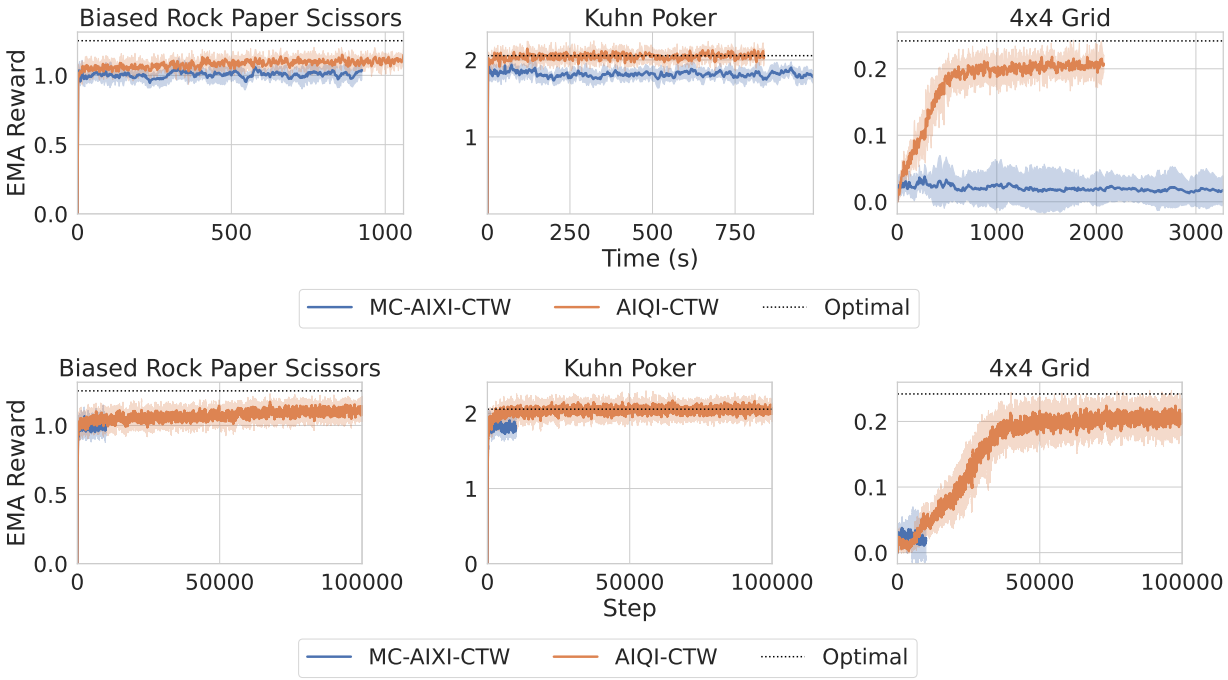

Figure 1: Plots of EMA reward vs wall clock time (in seconds) and environment steps on three environments.