Strong Asymptotic Optimality in General Environments

Michael K. Cohen\textsuperscript{1*}, Elliot Catt\textsuperscript{1} and Marcus Hutter\textsuperscript{1}

\textsuperscript{1}Australian National University
{michael.cohen, elliot.carpentercatt, marcus.hutter}@anu.edu.au

Abstract

Reinforcement Learning agents are expected to eventually perform well. Typically, this takes the form of a guarantee about the asymptotic behavior of an algorithm given some assumptions about the environment. We present an algorithm for a policy whose value approaches the optimal value with probability 1 in all computable probabilistic environments, provided the agent has a bounded horizon. This is known as strong asymptotic optimality, and it was previously unknown whether it was possible for a policy to be strongly asymptotically optimal in the class of all computable probabilistic environments. Our agent, Inquisitive Reinforcement Learner (Inq), is more likely to explore the more it expects an exploratory action to reduce its uncertainty about which environment it is in, hence the term inquisitive. Exploring inquisitively is a strategy that can be applied generally; for more manageable environment classes, inquisitiveness is tractable. We conducted experiments in "grid-worlds" to compare the Inquisitive Reinforcement Learner to other weakly asymptotically optimal agents.

1 Introduction

"Efforts to solve [an instance of the exploration-exploitation problem] so sapped the energies and minds of Allied analysts that the suggestion was made that the problem be dropped over Germany, as the ultimate instrument of intellectual sabotage." –Peter Whittle [Whittle, 1979]

The Allied analysts were considering the simplest possible problem in which there is a trade-off to be made between exploiting, taking the apparently best option, and exploring, choosing a different option to learn more. We tackle what we consider the most difficult instance of the exploration-exploitation trade-off problem: when the environment could be any computable probability distribution, not just a multi-armed bandit, how can one achieve optimal performance in the limit?

Our work is within the Reinforcement Learning (RL) paradigm: an agent selects an action, and the environment responds with an observation and a reward. The interaction may end, or it may continue forever. Each interaction cycle is called a timestep. The agent has a discount function that weights its relative concern for the reward it achieves in various future timesteps. The agent’s job is to select actions that maximize the total expected discounted reward it achieves in its lifetime. The “value” of an agent’s policy at a certain point in time is the expected total discounted reward it achieves after that time if it follows that policy. One formal specification of the exploration-exploitation problem is: what policy can an agent follow so that the policy’s value approaches the value of the optimal informed policy with probability 1, even when the agent doesn’t start out knowing the true dynamics of its environment?

Most work in RL makes strong assumptions about the environment—that the environment is Markov, for instance. Impressive recent development in the field of reinforcement learning often makes use of the Markov assumption, including Deep Q Networks [Mnih et al., 2015], A3C [Mnih et al., 2016], Rainbow [Hessel et al., 2018], and AlphaZero [Silver et al., 2017]. Another example of making strong assumptions in RL comes from some model-based algorithms that implicitly assume that the environment is representable by, for example, a fixed-size neural network, or whatever construct is used to model the environment.

Many recent developments in RL are largely about tractably learning to exploit; how to explore intelligently is a separate problem. We address the latter problem. Our approach, inquisitiveness, is based on Orseau et al.’s [2013] Knowledge Seeking Agent for Stochastic Environments, which selects the actions that best inform the agent about what environment it is in. Our Inquisitive Reinforcement Learner (Inq) explores like a knowledge seeking agent, and is more likely to explore when there is apparently (according to its current beliefs) more to be learned. Sometimes exploring well requires “expeditions,” or many consecutive exploratory actions. Inq entertains expeditions of all lengths, although it follows the longer ones less often, and it doesn’t resolutely commit in advance to seeing the expedition through.

This is a very human approach to information acquisition. When we spot an opportunity to learn something about our
natural environment, we feel inquisitive. We get distracted. We are inclined to check it out, even if we don’t see directly in advance how this information might help us better achieve our goals. Moreover, if we can tell that the opportunity to learn something requires a longer term project, we may find ourselves less inquisitive.

For the class of computable environments (stochastic environments that follow a computable probability distribution), it was previously unknown whether any policy could achieve strong asymptotic optimality (convergence of the value to optimality with probability 1). Lattimore et al. [2011] showed that no deterministic policy could achieve this. The key advantage that stochastic policies have is that they can let the exploration probability go to 0 while still exploring infinitely often.

There is a weaker notion of optimality “weak asymptotic optimality” for which positive results already exist; this condition requires that the average value over the agent’s lifetime approach optimality. Lattimore et al. [2011] identified a weakly asymptotically optimal agent for deterministic computable environments; the agent maintains a list of environments consistent with its observations, exploiting as if it is in the first such one, and exploring in bursts. A recent algorithm for a Thompson Sampling Bayesian agent was shown, with an elegant proof, to be weakly asymptotically optimal in all computable environments, but not strongly asymptotically optimal [Leike et al., 2016].

In Section 2, we formally describe the RL setup and present notation. In Section 3, we present the algorithm for Inq. In Section 4, we prove our main result: that Inq is weakly asymptotically optimal. In Section 5, we present experimental results comparing Inq to weakly asymptotically optimal algorithms. Finally, we discuss the relevance of this exploration regime to tractable algorithms. Appendix A collates notation and definitions for quick reference. Appendix B contains the proofs of the lemmas.

2 Notation

We follow the notation in [Orseau et al., 2013]. The reinforcement learning setup is as follows: \( \mathcal{A} \) is a finite set of actions available to the agent; \( \mathcal{O} \) is a finite set of observations it might observe, and \( \mathcal{R} = [0, 1] \cap \mathbb{Q} \) is the set of possible rewards. The set of all possible interactions in a timestep is \( \mathcal{H} := \mathcal{A} \times \mathcal{O} \times \mathcal{R} \). A reinforcement learner’s policy \( \pi \) is a stochastic function which outputs an action given an interaction history, denoted by \( \pi : \mathcal{H}^* \rightarrow \mathcal{A} \). (\( \mathcal{X}^* := \bigcup_{t=0}^{\infty} \mathcal{X}^t \)) represents all finite strings from an alphabet \( \mathcal{X} \). An environment is a stochastic function which outputs an observation and reward given an interaction history and an action: \( \nu : \mathcal{H}^* \times \mathcal{A} \rightarrow \mathcal{O} \times \mathcal{R} \).

A policy and an environment induce a probability measure over \( \mathcal{H}^\infty \), the set of all possible infinite histories. For \( h \in \mathcal{H}^\infty \), \( P_\nu^\pi (h) \) denotes the probability that an infinite history begins with \( h \) when actions are sampled from the policy \( \pi \), and observations and rewards are sampled from the environment \( \nu \). Formally, we define this inductively: \( P_\nu^\pi (\epsilon) \mapsto 1 \), where \( \epsilon \) is the empty history, and for \( h \in \mathcal{H}^\infty \), \( a \in \mathcal{A}, o \in \mathcal{O}, r \in \mathcal{R} \), we define \( P_\nu^\pi (hao^r) \mapsto P_\nu^\pi (h)\pi (a|h)\nu (or|ha) \). In an infinite history \( h_{t:}\infty \in \mathcal{H}^\infty \), \( o_t, a_t, r_t \) refer to the \( t \)th action, observation and reward, and \( h_t \) refers to the \( t \)th timestep; \( a_o r_o h_o \) refers to the first \( t - 1 \) timesteps, and \( h_{t:k} \) refers to the string of timesteps \( t \) through \( k \) (inclusive). Strings of actions, observations, and rewards are notated similarly.

A Bayesian agent considers a class of environments a priori feasible. Its “beliefs” take the form of a probability distribution over which environment is the true one. We call this the agent’s belief distribution. In our formulation, Inq considers any computable environment feasible, and starts with a prior belief distribution based on the environments’ Kolmogorov complexities: that is, the length of the shortest program that computes the environment on some reference machine. However, all our results hold as long as the true environment is contained in the class of environments that are considered feasible, and as long as the prior belief distribution assigns nonzero probability to each environment in the class. We take \( \mathcal{M} \) to be the class of all computable environments, and \( w(\nu) := 2^{-K(\nu)} \) to be the prior probability of the environment \( \nu \), where \( K \) is the Kolmogorov complexity. A smaller class and a different prior probability could easily be substituted for \( \mathcal{M} \) and \( w(\nu) \).

We use \( \xi \) to denote the agent’s beliefs about future observations. Together with a policy \( \pi \) it defines a Bayesian mixture measure: \( P_\xi^\pi (\cdot) := \sum_{\nu \in \mathcal{M}} w(\nu)P_\nu^\pi (\cdot) \). The posterior belief distribution of the agent after observing a history \( h \in \mathcal{H}^* \) is \( w(\nu|h) := w(\nu)P_\nu^\pi (h) / P_\xi^\pi (h) \). This definition is independent of the choice of \( \pi \)’ as long as \( P_\xi^\pi (h) > 0 \); we can fix a reference policy \( \pi' \) just for this definition if we like. We sometimes also refer to the conditional distribution \( \xi(or|ha):=\sum_{\nu \in \mathcal{M}} w(\nu|h)\nu(or|ha) \).

The agent’s discount at a timestep is denoted \( \gamma_t \). To normalize the agent’s value to \( [0, 1] \), we introduce \( \Gamma_t := \sum_{k=1}^{\infty} \gamma_k \). We consider an agent with a bounded horizon: \( \forall \varepsilon > 0 \exists m \forall t : \Gamma_{t+m}/\Gamma_t \leq \varepsilon \). Intuitively, this means that the agent does not become more and more farsighted over time. A classic discount function giving a bounded horizon is a geometric one: \( 0 \leq \gamma < 1, \gamma = \gamma^t \). The value of a policy \( \pi \) in an environment \( \nu \), given a history \( h_{\leq t} \in \mathcal{H}^{t-1} \), is

\[
V_\nu^\pi (h_{\leq t}) := \frac{1}{\Gamma_t} \mathbb{E}_\nu \left[ \sum_{k=t}^{\infty} \gamma_k r_k \mid h_{\leq t} \right] \tag{1}
\]

Here, the expectation is with respect to the probability measure \( P_\nu^\pi \). Reinforcement Learning is the attempt to find a policy that makes this value high, without access to \( \nu \).

3 Inquisitive Reinforcement Learner

We first describe how Inq exploits, then how it explores. It exploits by maximizing the discounted sum of its reward in expectation over its current beliefs, and it explores by following maximally informative “exploratory expeditions” of various lengths.

The optimal policy with respect to an environment \( \nu \) is the policy that maximizes the value.

\[
\pi^*_\nu (\cdot) := \arg\max_{\pi \in \Pi} V_\nu^\pi (\cdot) \tag{2}
\]
where $\Pi = \mathcal{H}^* \sim \mathcal{A}$ is the space of all policies. An optimal deterministic policy always exists [Lattimore and Hutter, 2014b]. When exploiting, Inq simply maximizes the value according to its belief distribution $\xi$. Since this policy is deterministic, we write $\alpha^*(h_{<t})$ to mean the unique action at time $t$ for which $\pi_\xi(a|h_{<t}) = 1$. That is the exploitative action.

The most interesting feature of Inq is how it gets distracted by the opportunity to explore. Inq explores to learn. An agent has learned from an observation if its belief distribution $w$ changes significantly after making that observation. If the belief distribution has hardly changed, then the observation was not very informative. The typical information-theoretic measure for how well a distribution $Q$ approximates a distribution $P$ is the KL-divergence, $\text{KL}(P||Q)$. Thus, a principled way to quantify the information that an agent gains in a timestep is the KL-divergence from the belief distribution at time $t + 1$ to the belief distribution at time $t$. This is the rationale behind the construction of Orseau, et al.’s [2013] Knowledge Seeking Agent, which maximizes this expected information gain.

Letting $h_{<t} \in \mathcal{H}^{t-1}$ and $h' \in \mathcal{H}^*$, the information gain at time $t$ is defined:

$$\text{IG}(h'|h_{<t}) := \sum_{\nu \in \mathcal{A}} w(\nu|h_{<t}h') \log \frac{w(\nu|h_{<t}h')}{w(\nu|h_{<t})} \tag{3}$$

Recall that $w(\nu|h)$ is the posterior probability assigned to $\nu$ after observing $h$.

An $m$-step expedition, denoted $\alpha^m$, represents all contingencies for how an agent will act for the next $m$ timesteps. It is a deterministic policy that takes history-fragments of length less than $m$ and returns an action:

$$\alpha^m : \bigcup_{i=0}^{m-1} \mathcal{H}^i \to \mathcal{A} \tag{4}$$

$P^m_\xi(h_{<t+k}|h_{<t})$ is a conditional distribution defined for $0 \leq k \leq m$, which represents the conditional probability of observing $h_{<t+k}$ if the expedition $\alpha^m$ is followed starting at time $t$, after observing $h_{<t}$. Now we can consider the information-gain value of an $m$-step expedition. It is the expected information gain upon following that expedition:

$$V^\text{IG}(\alpha^m, h_{<t}) := \sum_{h_{t:t+m-1} \in \mathcal{H}^m} P^m_\xi(h_{<t+m}|h_{<t}) \text{IG}(h_{t:t+m-1}|h_{<t}) \tag{5}$$

At a time $t$, one might consider many expeditions: the one-step expedition which maximizes expected information gain, the two-step expedition doing the same, etc. Or one might consider carrying on with an expedition that began three timesteps ago.

**Definition 1.** At time $t$, the $m$-$k$ expedition is the $m$-step expedition beginning at time $t-k$ which maximized the expected information gain from that point.\(^1\)

$$\alpha^\text{IG}_{m,k}(h_{<t}) := \arg \max_{\alpha^m \in \mathcal{A}^{t+m-1}} V^\text{IG}(\alpha^m, h_{<t-k}) \tag{6}$$

\(^1\)Ties in the argmax are broken arbitrarily.

\(^2\)This algorithm is written in a simplified way that does not halt, but if a real number in $[0, 1]$ is sampled first, the actions can be assigned to disjoint intervals successively until the sampled real number lands in one of them.

Figure 1: **Example Expeditions.** Expeditions maximize the expected KL-divergence from the posterior at the end to the posterior at the beginning.

Example expeditions are diagrammed in Figure 1.

Expeditions are functions which return an action given what’s been seen so far on the expedition. The $m$-$k$ exploratory action is the action to take at time $t$ according to the $m$-$k$ expedition:

$$a^\text{IG}_{m,k}(h_{<t}) := a^\text{IG}_{m,k}(h_{<t})(h_{t-k:t-1}) \tag{7}$$

Naturally, this is only defined for $k < m$, since the expedition function can’t accept a history fragment of length $\geq m$, and $t-k$ must be positive. Note also that if $k = 0$, $h_{t-k:t-1}$ evaluates to the empty string, $\epsilon$.

**Definition 2.** Let $\rho(h_{<t}, m, k)$ is the probability of taking the $m$-$k$ exploratory action after observing a history $h_{<t}$.

$$\rho(h_{<t}, m, k) := \min \left\{ \frac{1}{m^2(m + 1)}, \eta V^\text{IG}(\alpha^\text{IG}_{m,k}(h_{<t}), h_{<t-k}) \right\} \tag{8}$$

where $\eta$ is an exploration constant.

This makes the total probability of exploration:

$$\beta(h_{<t}) := \sum_{m \in \mathcal{N}} \sum_{k < m} \rho(h_{<t}, m, k) \leq 1 \tag{9}$$

The feature that makes Inq inquisitive is that $\rho(h_{<t}, m, k)$ is proportional to the expected information gain from the $m$-$k$ expedition, $V^\text{IG}(\alpha^\text{IG}_{m,k}(h_{<t}), h_{<t-k})$. Note that completing an $m$-$k$ expedition requires randomly deciding to explore in that way on $m$ separate occasions. While this may seem inefficient, if the agent always got boxed into long expeditions, the value of its policy would plummet infinitely often.

Finally, Inq’s policy $\pi\dagger$, defined in Algorithm 1, takes the $m$-$k$ exploratory action with probability $\rho(\cdot, m, k)$, and takes the exploitative action otherwise.\(^2\)
Algorithm 1 Inquisitive Reinforcement Learner’s Policy $\pi^*$

1: while True do
2:     calculate $\rho(h_{c,t}, m, k)$ for all $m$ and for all $k < \min\{m, t\}$
3:     take action $a_{m,k}^\pi(h_{c,t})$ with probability $\rho(h_{c,t}, m, k)$
4:     take action $a^*(h_{c,t})$ with probability $1 - \beta(h_{c,t})$

4 Strong Asymptotic Optimality

Here we present our central result: that the value of $\pi^*$ approaches the optimal value. We present the theorem, motivate the result, and proceed to the proof. We recommend the reader have Appendix A and hand for quickly looking up definitions and notation.

Before presenting the theorem, we clarify an assumption, and define the optimal value. We call the true environment $\mu$, and we assume that $\mu \in \mathcal{M}$. For $\mathcal{M}$ the class of computable environments, this is a very unassumming assumption.

The optimal value is simply the value of the optimal policy with respect to the true environment:

$$V^*(h_{c,t}) := \sup_{\pi} V^\pi(h_{c,t}) = V_{\mu}^\pi(h_{c,t})$$  \hspace{1cm} (10)

Recall also that we have assumed the agent has a bounded horizon in the sense that $\forall \in \exists m \forall t : \Gamma_{t+m}/\Gamma_t \leq \varepsilon$. The Strong Asymptotic Optimality theorem is that under these conditions, the value of $\text{Inq}$’s policy approaches the optimal value with probability 1, when actions are sampled from $\text{Inq}$’s policy and observations and rewards are sampled from the true environment $\mu$.

**Theorem 3** (Strong Asymptotic Optimality). As $t \to \infty$, $V^\mu(h_{c,t}) - V^\pi^*(h_{c,t}) \to 0$ with $P_{\mu}^\pi$-prob. 1 where $\mu \in \mathcal{M}$ is the true environment.

For a Bayesian agent, uncertainty about on-policy observations goes to 0. Since “on-policy” for $\text{Inq}$ includes, with some probability, all maximally informative expeditions, $\text{Inq}$ eventually has little uncertainty about the result of any course of action, and can therefore successfully select the optimal course. For any fixed horizon, $\text{Inq}$’s mixture measure $\xi$ approaches the true environment $\mu$.

We use the following notation for a particular KL-divergence that plays a central role in the proof:

$$KL(P_{\mu}^\pi || P_{\mu}^\nu) := \sum_{h' \in \mathcal{H}^n} P_{\nu}^\pi(h'|h_{c,t}) \log \frac{P_{\nu}^\pi(h'|h_{c,t})}{P_{\mu}^\nu(h'|h_{c,t})}$$  \hspace{1cm} (11)

This quantifies the difference between the expected observations of two different environments that would arise in the next $n$ timesteps when following policy $\pi$. $KL_{h_{c,t},\infty}$ denotes the limit of the above as $n \to \infty$, which exists by [Orseau et al., 2013, proof of Theorem 3].

In dealing with the KL-divergence, we simplify matters by asserting that $0 \log 0 := 0$, and $0 \log \frac{0}{0} := 0$.

We begin with a lemma that equates the information gain value of an expedition with the expected prediction error. The KL-divergence on the right hand side represents how different $\nu$ and $\xi$ appear when following the expedition in question.

**Lemma 4.**

$$V^IG(a^m, h_{c,t}) = \sum_{\nu \in \mathcal{M}} w(\nu|h_{c,t}) KL(P_{\nu}^\pi || P_{\xi}^{a_{\infty}^m})$$

**Proof.** This result is shown in Orseau et al. (2013), Eq. 4. \hfill $\square$

We would like to use the fact that Bayesians become good at on-policy prediction to bound $\text{Inq}$’s uncertainty regarding potential expeditions. Thus we have,

**Lemma 5.**

$$KL_{h_{c,t},m}(P_{\nu}^\pi || P_{\xi}^*) \geq \rho(h_{c,t}, m, 0)^m KL_{h_{c,t},m}(P_{\nu}^{IG,a_0(h_{c,t})} || P_{\xi}^{IG,a_0(h_{c,t})})$$

The left-hand side is the on-policy prediction error, and will be pushed down to 0, and right-hand side is the prediction error for a given expedition times the probability of seeing that expedition through. The proofs of the lemmas are in Appendix B.

Recall that $w(\nu|h_{c,k})$ is the posterior weight that $\text{Inq}$ assigns to the environment $\nu$ after observing $h_{c,k}$. Another piece of notation we need is the infimum of $\text{Inq}$’s belief that the $\mu$ is the true environment:

$$w_{\text{inf}}(\mu|h_{c,\infty}) := \inf_{k \in \mathbb{N}} w(\mu|h_{c,k})$$

We show that this is strictly positive with probability 1.

**Lemma 6.** $w_{\text{inf}}(\mu|h_{c,\infty}) > 0$ w.$P_{\mu}^\pi$-prob. 1

Next we show that the exploration probability goes to 0. From here, all “w.p.1” statements mean with $P_{\mu}^\pi$-probability 1, if not otherwise specified.

**Lemma 7.**

$$\beta(h_{c,t}) \to 0 \text{ w.p.1}$$

The next lemma states that the expectation of a certain sum is finite: the sum over time of the inaccuracy (measured in a certain way) of $\xi$’s forecasts regarding the outcome of the $m$-step expedition that begins at that time. This holds even though most of those expeditions are not followed. As a first pass reading the lemma, ignore the quantifiers and the conditional.

**Lemma 8.** $\forall m \in \mathbb{N} \exists c > 0 \forall \varepsilon \in (0, c)$:

$$\mathbb{E}_{P_{\mu}^\pi} \left[ \sum_{t=1}^{\infty} \max_{a^m_{h_{c,t},m}} KL(P_{\mu}^\pi || P_{\xi}^{a_{\infty}^m}) \left| w_{\text{inf}}(\mu|h_{c,\infty}) > \varepsilon \right| \right] < \infty$$

The proof roughly follows the following argument: if all exploration probabilities go to 0, then the informativeness of the maximally informative expeditions goes to 0, so the informativeness of all expeditions goes to 0, meaning the prediction error as measured by the KL-divergence goes to 0.

**Lemma 9.** The probabilities assigned by $\xi$ converge to those of $\mu$. Lemma 8 showed this as well in a way, but in a less usable form.
Lemma 9. As $t \to \infty$, $\forall m \in \mathbb{N}$, $\alpha_m \in \mathcal{A}^{\infty}_{t;0} \mathcal{H}^t$, $h_{t:t+m-1} \in \mathcal{H}^m$:

$$P_{\mu}^\alpha(h_{t:t+m}|h_t) - P_{\xi}^\alpha(h_{t:t+m}|h_t) \to 0 \text{ w.p.1}$$

Finally, we prove the Strong Asymptotic Optimality Theorem: $V_{\mu}^\gamma(h_t) - V_{\mu}^\gamma(h_t) \to 0$ with $P_{\mu}^\gamma$-prob. 1.

**Proof of Theorem 1.** Let $\varepsilon > 0$. Since the agent has a bounded horizon, there exists an $m$ such that for all $t$, $\Gamma_{t,m} \leq \varepsilon$. Recall

$$V_{\mu}^\gamma(h_t) = \frac{1}{\Gamma_t} \mathbb{E}_{\mu}^\gamma \left[ \sum_{k=t}^{\infty} \gamma_k r_k \mid h_t \right]$$

Let

$$V_{\mu}^\gamma(h_t) := \frac{1}{\Gamma_t} \mathbb{E}_{\mu}^\gamma \left[ \sum_{k=t}^{\infty} \gamma_k r_k \mid h_t \right]$$

Since $r_t \in [0,1]$, we have

$$|V_{\mu}^\gamma(h_t) - V_{\mu}^\gamma(h_t)| \leq \frac{\Gamma_{t+m}}{\Gamma_t} \leq \varepsilon$$

From that we get,

$$V_{\mu}^\gamma(h_t) \leq V_{\mu}^\gamma(h_t)^\gamma + \varepsilon$$

$$= \frac{1}{\Gamma_t} \sum_{h_{t:t+m} \in \mathcal{H}^m} P_{\mu}^\gamma(h_{t:t+m}|h_t) \sum_{k=t}^{t+m-1} \gamma_k r_k + \varepsilon$$

$$\leq \frac{1}{\Gamma_t} \sum_{h_{t:t+m} \in \mathcal{H}^m} P_{\mu}^\gamma(h_{t:t+m}|h_t) \sum_{k=t}^{t+m-1} \gamma_k r_k + 2\varepsilon$$

$$\leq \frac{1}{\Gamma_t} \sum_{h_{t:t+m} \in \mathcal{H}^m} P_{\mu}^\gamma(h_{t:t+m}|h_t) \sum_{k=t}^{t+m-1} \gamma_k r_k + 3\varepsilon$$

Strong Asymptotic Optimality is not a guarantee of efficacy; consider an agent that “commits suicide” on the first timestep, and thereafter receives a reward of 0 no matter what it does. This agent is asymptotically optimal, but not very useful. In general, when considering many environments with many different “traps,” bounded regret is impossible to guarantee [Hutter, 2005], but one can still demand from a reinforcement learner that it make the best of whatever situation it finds itself in by correctly identifying (in the limit) the optimal policy.

We suspect that strong asymptotic optimality would not hold if Inq had an unbounded horizon, since its horizon of concern may grow faster than it can learn about progressively more long-term dynamics of the environment. Indeed, we tentuously suspect that it is impossible for an agent with an
unbounded time horizon to be strongly asymptotically optimal in the class of all computable environments. If that is true, then the assumptions that our result relies on (namely that the true environment is computable, and the agent has a bounded horizon) are the bare minimum for strong asymptotic optimality to be possible.

Inq is not computable; in fact, no computable policy can be strongly asymptotically optimal in the class of all computable environments (Lattimore, et al. [2011] show this for deterministic policies, but a simple modification extends this to stochastic policies). For many smaller environment classes, however, Inq would be computable, for example if \( \mathcal{M} \) is finite, and perhaps for decidable \( \mathcal{M} \) in general. The central result, that inquisitiveness is an effective exploration strategy, applies to any Bayesian agent.

5 Experimental Results

We compared Inq with other known weakly asymptotically optimal agents, Thompson sampling and BayesExp [Lattimore and Hutter, 2014a], in the grid-world environment using AIXIjs [Aslanides, 2017] which has previously been used to compare asymptotically optimal agents [Aslanides et al., 2017]. We tested in 10 \times 10 grid-worlds, and 20 \times 20 grid-worlds, both with a single dispenser with probability of dispensing reward 0.75. Following the conventions of [Aslanides et al., 2017] we averaged over 50 simulations, used discount factor \( \gamma = 0.99 \), 600 MCTS samples, and planning horizon of 6. The code used for this experiment is available online at https://github.com/ejcatt/aixijs, and this version of Inq can be run in the browser at https://ejcatt.github.io/aixijs/demo.html#inq.

In the 10 \times 10 grid-worlds Inq performed comparably to both BayesExp and Thompson sampling. However in the 20 \times 20 grid-worlds Inq performed comparably to BayesExp, and out performed Thompson sampling.

6 Conclusion

We have shown that it is possible for an agent with a bounded horizon to be strongly asymptotically optimal in the class of all computable environments. No existing RL agent has as strong an optimality guarantee as Inq. The nature of the exploration regime that accomplishes this is perhaps of wider interest. We formalize an agent that gets distracted from reward maximization by its inquisitiveness: the more it expects to learn from an expedition, the more inclined it is to take it.

We have confirmed experimentally that inquisitiveness is a practical and effective exploration strategy for Bayesian agents with manageable model classes.

There are two main avenues for future work we would like to see. The first regards possible extensions of inquisitiveness: we have defined inquisitiveness for Bayesian agents with countable model-classes, but inquisitiveness could also be defined for a Bayesian agent with a continuous model class, such as a Q-learner using a Bayesian Neural Network. The second avenue regards the theory of strong asymptotic optimality itself: is Inq strongly asymptotically optimal for more farsighted discounters? If not, can it be modified to accomplish that? Or is it indeed impossible for an agent with an unbounded horizon to be strongly asymptotically optimal in the class of computable environments? Answers to these questions, besides being interesting in their own right, will likely inform the design of tractable exploration strategies, in the same way that this work has done.
References

[Aslanides et al., 2017] John Aslanides, Jan Leike, and Marcus Hutter. Universal reinforcement learning algorithms: survey and experiments. In Proceedings of the 26th International Joint Conference on Artificial Intelligence, pages 1403–1410. AAAI Press, 2017.

[Aslanides, 2017] John Aslanides. Aixijs: A software demo for general reinforcement learning. arXiv preprint arXiv:1705.07615, 2017.

[Hessel et al., 2018] M. Hessel, J. Modayil, H. van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, and D. Silver. Rainbow: Combining improvements in deep reinforcement learning. In Proc. of AAAI Conference on Artificial Intelligence, 2018.

[Hutter, 2005] Marcus Hutter. Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability. Springer, Berlin, 2005.

[Lattimore and Hutter, 2011] Tor Lattimore and Marcus Hutter. Asymptotically optimal agents. In Proc. 22nd International Conf. on Algorithmic Learning Theory (ALT’11), volume 6925 of LNAI, pages 368–382, Espoo, Finland, 2011. Springer.

[Lattimore and Hutter, 2014a] Tor Lattimore and Marcus Hutter. Bayesian reinforcement learning with exploration. In International Conference on Algorithmic Learning Theory, pages 170–184. Springer, 2014.

[Lattimore and Hutter, 2014b] Tor Lattimore and Marcus Hutter. General time consistent discounting. Theoretical Computer Science, 519:140–154, 2014.

[Leike et al., 2016] Jan Leike, Tor Lattimore, Laurent Orseau, and Marcus Hutter. Thompson sampling is asymptotically optimal in general environments. In Proc. 32nd International Conf. on Uncertainty in Artificial Intelligence (UAI’16), pages 417–426, New Jersey, USA, 2016. AUAI Press.

[Mnih et al., 2015] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. Nature, 518(7540):529, 2015.

[Mnih et al., 2016] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In International conference on machine learning, pages 1928–1937, 2016.

[Orseau et al., 2013] Laurent Orseau, Tor Lattimore, and Marcus Hutter. Universal knowledge-seeking agents for stochastic environments. In Proc. 24th International Conf. on Algorithmic Learning Theory (ALT’13), volume 8139 of LNAI, pages 158–172, Singapore, 2013. Springer.

[Silver et al., 2017] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. arXiv preprint arXiv:1712.01815, 2017.

[Whittle, 1979] Peter Whittle. Discussion of Dr Gittins’ paper. Journal of the Royal Statistical Society, 41:164–177, 1979.
Appendix A: Definitions and Notation – Quick Reference

\( \mathcal{H} := A \times O \times R \)
\( h_{<t} \in \mathcal{H}^{t-1} \)
\( h_{t:k} \in \mathcal{H}^{k-t+1} \)
\( \mu, \nu \in \mathcal{M} \)
\( \mu, \nu : \mathcal{H}^* \times A \rightarrow O \times R \)
\( \pi : \mathcal{H}^* \rightarrow A \)

\( P^\pi_\nu(e) := 1; \quad P^\pi_\nu(haor) := P^\pi_\nu(h)\pi(a|h)\nu(or|ha) \)
\( w(\nu) := 2^{-K(\nu)} \)

\( P^\pi_\xi(\cdot) := \sum_{\nu \in \mathcal{M}} w(\nu) P^\pi_\nu(\cdot) \)
\( w(\nu|h) := w(\nu) P^\pi_\nu(h) P^\pi_\xi(h) \)
\( w_{\inf}(\mu|h_{1:\infty}) := \inf_{k \in \mathbb{N}} w(\mu|h_{<k}) \)

\( \pi^\nu(\cdot) := 1 \)
\( \pi^\nu(h | a_{or}) := P^\nu_\pi(h | a_{or}) \)

\( w(\nu|h) := w(\nu) P^\pi_\nu(h) \)
\( w_{\inf}(\mu|h_{1:\infty}) := \inf_{k \in \mathbb{N}} w(\mu|h_{<k}) \)
\( \Xi(\nu|ha) := \sum_{\nu \in \mathcal{M}} w(\nu|h)\nu(or|ha) \)

\( \text{IG}(h_{t:t+k-1}|h_{<t}) := \sum_{\nu \in \mathcal{M}} w(\nu|h_{<t+k-1}) \log \frac{w(\nu|h_{<t+k-1})}{w(\nu|h_{<t})} \quad \text{for } h_{<t} \in \mathcal{H}^{t-1}, h' \in \mathcal{H}^{k} \)

\( V^\pi(\alpha_m, h_{<t}) := \sum_{h_{t:t+m-1} \in \mathcal{H}^{m}} P^\pi_\xi(h_{t:t+m-1}|h_{<t}) \text{IG}(h_{t:t+m-1}|h_{<t}) \)
\( a^\pi_{m,k}(h_{<t}) := \arg\max_{a \in A} \quad V^\pi_{\alpha_m, h_{<t}} \)

\( V^\pi_{\alpha_m, h_{<t}} := \sup_{\pi \in \Pi} V^\pi_{\nu}(h_{<t}) = V^\pi_{\alpha^\nu}(h_{<t}) \)

\( \rho(h_{<t}, m, k) := \max\{ \frac{1}{m^2(m+1)}, \eta V^\pi_{\alpha_m, h_{<t-k}} \} \)

\( \pi^\nu(a|h_{<t}) := \sum_{m \in \mathbb{N}} \sum_{k < m, t} \rho(h_{<t}, m, k)[[a = a^\pi_{m,k}(h_{<t})]] + (1 - \beta(h_{<t}, m, k))[[a = a^*(h_{<t})]] \)

\( \text{KL} \left( P^\pi_{v_1} || P^\pi_{v_2} \right) := \sum_{h' \in \mathcal{H}^{n}} P^\pi_{v_1}(h'|h_{<t}) \log \frac{P^\pi_{v_1}(h'|h_{<t})}{P^\pi_{v_2}(h'|h_{<t})} \)
Appendix B: Proofs of Lemmas

Lemma 5.

\[
\text{KL}_{h_{t:t+m-1}} \left( P^{\pi^+}_\nu \| P^{\pi^+}_\xi \right) = \sum_{h_{t:t+m-1} \in \mathcal{H}^m} P^{\pi^+}_\nu (h_{t:t+m} | h_t) \log \frac{P^{\pi^+}_\nu (h_{t:t+m} | h_t)}{P^{\pi^+}_\xi (h_{t:t+m} | h_t)}
\]

Proof. (a) follows because the KL-divergence for all nonzero terms in the sum. Note (a) is the definition of this KL-divergence. (b) follows from conditional probabilities \( \pi^+ \) canceling, which is nonzero for all nonzero terms in the sum. (c) follows from expanding \( P^{\pi^+}_\nu \) into the probabilities of the actions and probabilities of the observations and rewards. (d) follows because the KL-divergence is non-negative when the first term is a measure and the second a (semi)measure, and the probability of following the action sequence \( \alpha_{m,0}^{\pi^+} (h_{t:t+m}) \) under \( \pi^+ \) is at least the probability of exploring that way for \( m \) timesteps. Note that \( \xi \) is a semimeasure even though it is a mixture over measures because \( \text{KL}_{\mathcal{H}^m} w(\nu) < 1 \). (e) follows from \( \rho(h_{t:t+k},m) = \rho(h_{t:t+k},m,0) \). (f) applies the definition of the KL-divergence.

Lemma 6. \( w_{\text{inf}}(\mu | h_{1:1}) > 0 \) w.p.1 (with \( P^{\pi^+}_\mu \)-probability 1)

Proof. We consider a general \( \pi \), inducing the measure \( P^{\pi}_\mu \) over \( \mathcal{H}^\infty \). Suppose \( w_{\text{inf}}(\mu | h_{1:1}) = 0 \). \( w(\mu | h_{1:t}) > 0 \) for all histories generated by \( P^{\pi}_\mu \). Therefore, \( w_{\text{inf}}(\mu | h_{1:1}) = 0 \) \( \implies \lim_{t \to \infty} w(\mu | h_{1:t}) = 0 \), and \( \lim_{t \to \infty} w(\mu | h_{1:t})^{-1} = \infty \). We show that this has probability 0.

Let \( z_t := w(\mu | h_{1:t})^{-1} = \frac{P^{\pi}_\mu (h_{1:t})}{P^{\pi}_\mu (h_{1:t})} w(\mu)^{-1} \). We first show that \( z_t \) is a \( \mu \)-martingale.

\[
\mathbb{E}^{\pi}[z_t | h_{1:t}] = w(\mu)^{-1} \sum_{h_{1:t} \in \mathcal{H}} P^{\pi}_\mu (h_{1:t}) \frac{P^{\pi}_\mu (h_{1:t})}{P^{\pi}_\mu (h_{1:t})} = w(\mu)^{-1} \frac{\sum_{h_{1:t} \in \mathcal{H}} P^{\pi}_\mu (h_{1:t})}{P^{\pi}_\mu (h_{1:t})} = 1 \]

By the martingale convergence theorem \( z_t \to f(\omega) < \infty \) w.p.1, for \( \omega \in \Omega \), the sample space, and some \( f : \Omega \to \mathbb{R} \). Therefore, \( w_{\text{inf}}(\mu | h_{1:1}) > 0 \) w.p.1.

Lemma 7. \( \beta(h_{1:t}) \to 0 \) w.p.1
Proof.

\[ 0 \overset{(a)}{=} \lim_{t \to \infty} \frac{1}{w(\mu)} \sum_{\nu \in M} w(\nu) m E^\pi_{\nu} (KL(P^\pi_{\nu} || P^\xi_{\nu})) - KL(P^\xi_{\nu} || P^\xi_{\nu}) \]

\[ \overset{(b)}{=} \lim_{t \to \infty} \frac{1}{w(\mu)} \sum_{\nu \in M} w(\nu) m E^\pi_{\nu} \KL(h_{<t}, \infty) (P^\pi_{\nu} || P^\xi_{\nu}) \]

\[ \overset{(c)}{=} \lim_{t \to \infty} \frac{1}{w(\mu)} \sum_{\nu \in M} w(\nu) \sum_{i=0}^{m-1} E^\pi_{\nu} \KL(h_{<t+i}, \infty) (P^\pi_{\nu} || P^\xi_{\nu}) \]

\[ \overset{(d)}{=} \lim_{t \to \infty} \frac{1}{w(\mu)} \sum_{\nu \in M} w(\nu) \sum_{k=k}^{\infty} E^\pi_{\nu} \KL(h_{<t,k,m}) (P^\pi_{\nu} || P^\xi_{\nu}) \]

\[ \overset{(e)}{=} \lim_{t \to \infty} \frac{1}{w(\mu)} \sum_{\nu \in M} \sum_{k=k}^{\infty} E^\pi_{\nu} \KL(h_{<t,k,m}) (P^\pi_{\nu} || P^\xi_{\nu}) \]

Steps (a), (b), (f), and (g) are borrowed almost verbatim from Orseau et al. [2013]. Recalling that \( \varepsilon \) is the empty string, (a) follows from the well-known \( KL_{\varepsilon, \infty}(P^\nu_{\nu} || P^\xi_{\nu}) \leq \log \frac{1}{w(\nu)} < \infty \), which renders the monotone convergence theorem applicable, (b) from Orseau et al., 2013, Lem. 6, the chain rule for KL-divergence, (c) by \( E^\pi_{\nu} \KL(h_{<t}, \infty) (P^\pi_{\nu} || P^\xi_{\nu}) \) = \( E^\pi_{\nu} \KL(h_{<t}, \nu) (P^\pi_{\nu} || P^\xi_{\nu}) \) ≥ 0, (d) by the telescoping property of the KL-divergence, and dividing the infinite part into blocks of length \( m \), (e) by reindexing the sums, (f) by exchanging the sum and expectation, then using the definition of \( w(\nu)h_{<k} \) and the definition of the \( \xi \)-expectation, (g) by the definition of \( w(\nu)h_{<k} \) and the definition of the expectation, (h) by Lemma 5, (i) by pulling \( \rho(h_{<k}, m, 0)^m \) out of the sum, and Lemma 4, (j) by the definition of \( \rho(h_{<k}, m, 0) \), and (k) by the non-negativity of all terms involved.

This gives us

\[ \lim_{t \to \infty} \sum_{u=t}^{\infty} E^\pi_{\mu} \frac{1}{w(\mu)h_{<u}} \rho(h_{<u}, m, 0)^{m+1} = 0 \]

If the limit of the tail of the sum goes to 0, the sum is finite:

\[ \sum_{i=1}^{\infty} E^\pi_{\mu} \frac{1}{w(\mu)h_{<i}} \rho(h_{<i}, m, 0)^{m+1} < \infty \]

The expectation and sum can be exchanged, and \( \frac{1}{w(\mu)h_{<i}} \geq 1 \), so

\[ E^\pi_{\mu} \sum_{i=1}^{\infty} \rho(h_{<i}, m, 0)^{m+1} < \infty \]

Suppose \( P^\pi_{\mu} (|\rho(h_{<t}, m, 0)| > \delta) \) i.o. > 0.

\[ |\rho(h_{<t}, m, 0)| > \delta \text{ i.o.} \implies \sum_{i=1}^{\infty} \rho(h_{<i}, m, 0)^{m+1} = \infty \]
so

\[ P^\pi_\mu \left( \sum_{t=1}^{\infty} \rho(h_{<t}, m, 0)^{m+1} \right) = \infty > 0 \]

and

\[ E^\pi_\mu \left( \sum_{t=1}^{\infty} \rho(h_{<t}, m, 0)^{m+1} \right) = \infty \]

contradicting the result above. Thus, \( P^\pi_\mu (|\rho(h_{<t}, m, 0)| > \delta \ i.o.) = 0 \), that is,

\[ \rho(h_{<t}, m, 0) \rightarrow 0 \ w.p.1 \]

\( \rho(h_{<t}, m, 0) = \rho(h_{<t-k}, m, 0) \), so

\[ \rho(h_{<t}, m, k) \rightarrow 0 \ w.p.1 \]

Now we show that \( \beta(h_{<t}) = \sum_{m=1}^{\infty} \sum_{k\leq m} \rho(h_{<t}, m, k) \) goes to 0 w.p.1 as well. Recall that \( \rho(\cdot, m, k) \leq 1/(m^2(m+1)) \). Thus, for all \( \varepsilon > 0 \),

\[ \exists M : \sum_{m=M}^{\infty} \sum_{k\leq m} \rho(\cdot, m, k) < \frac{\varepsilon}{2} \]

because \( \rho(\cdot, m, k) \leq 1/(m^2(m+1)) \). Suppose by contradiction that \( \beta(h_{<t}) > \varepsilon \) infinitely often for some \( \varepsilon > 0 \). Then it must be the case that \( \sum_{m=1}^{M} \sum_{k\leq m} \rho(h_{<t}, m, k) > \varepsilon/2 \) infinitely often, since the remainder of the sum cannot sum to more than \( \varepsilon/2 \). Since this is a finite sum with \( M(M+1)/2 \) terms, at least one of these terms must exceed \( \varepsilon/(M(M+1)) \) infinitely often, which contradicts that \( \rho(h_{<t}, m, k) \rightarrow 0 \). Therefore,

\[ \beta(h_{<t}) \rightarrow 0 \ w.p.1 \]

**Lemma 8.**

\[ \forall m \in \mathbb{N} \exists c > 0 \forall \varepsilon \in (0, c] : E_\mu^{\pi_1} \left[ \max_{h_{<t}, m} \text{KL} (P_{\mu}^{\alpha_{m}} || P_{\xi}^{\alpha_{m}})^{m+1} | \inf_{h_{1}\rightarrow \infty} (\mu|h_{1}\rightarrow \infty) > \varepsilon \right] < \infty \]

**Proof.** We begin with

\[ \lim_{k \to \infty} \sum_{m=1}^{\infty} \frac{1}{w(\mu|h_{<k})} V^{IG}(\alpha_{m,0}(h_{<k}), h_{<k}) = 0 \]

from the proof of Lemma 7, adding in a factor of \( \eta^m \). From the definition of \( \rho(h_{<k}, m, k), \rho(h_{<k}, m, 0) \leq \frac{1}{m^2(m+1)} \Rightarrow \rho(h_{<k}, m, 0) = \eta V^{IG}(\alpha_{m,0}(h_{<k}), h_{<k}) \). From Lemma 7, \( \rho(h_{<k}, m, 0) \rightarrow 0 \ w.p.1 \), so as \( t \to \infty \), the probability of the event \( \forall k \geq t : \rho(h_{<k}, m, 0) \leq \frac{1}{w^{(m+1)}(k) \rightarrow 1} \) goes to 1. Likewise, as \( t \to \infty \), the probability of the event \( \forall k > t : \rho(h_{<k}, m, 0) = \eta V^{IG}(\alpha_{m,0}(h_{<k}), h_{<k}) \) goes to 1. Recall these are all \( P^{\pi_1}_\mu \)-probabilities. This gives us

\[ \lim_{k \to \infty} \sum_{m=1}^{\infty} \frac{1}{w(\mu|h_{<k})} V^{IG}(\alpha_{m,0}(h_{<k}), h_{<k})^{m+1} = 0 \]

\[ 0 = \lim_{k \to \infty} \sum_{m=1}^{\infty} \frac{1}{w(\mu|h_{<k})} V^{IG}(\alpha_{m,0}(h_{<k}), h_{<k})^{m+1} \]

\[ \begin{align*}
  &\leq \lim_{k \to \infty} \sum_{m=1}^{\infty} \frac{1}{w(\mu|h_{<k})} \max_{\alpha_{m}} V^{IG}(\alpha_{m}, h_{<k})^{m+1} \\
  &\leq \lim_{k \to \infty} \sum_{m=1}^{\infty} \frac{1}{w(\mu|h_{<k})} \max_{\alpha_{m}} \sum_{h_{<k}, m} w(\mu|h_{<k}) \text{KL} (P_{\mu}^{\alpha_{m}} || P_{\xi}^{\alpha_{m}})^{m+1} \\
  &\geq \lim_{k \to \infty} \sum_{m=1}^{\infty} w(\mu|h_{<k}) \max_{\alpha_{m}} \text{KL} (P_{\mu}^{\alpha_{m}} || P_{\xi}^{\alpha_{m}})^{m+1} \\
  &\geq \lim_{k \to \infty} \sum_{m=1}^{\infty} w(\mu|h_{1}\rightarrow \infty) \max_{\alpha_{m}} \text{KL} (P_{\mu}^{\alpha_{m}} || P_{\xi}^{\alpha_{m}})^{m+1}
\end{align*} \]
Lemma 6. by exchanging the sum and expectation, by the fact that limit of the tail of the sum goes to 0, expectation is finite, is ε, by factoring out δ > ε > 0, so for such ε, by exchanging the sum and expectation back.

(a) follows from the entropy inequality, which holds when the first term is a measure, and the second a (semi)measure, and (b) follows from the definition of α, and the non-negativity of all other terms. P(μ(h1:∞) = 0) = 0 by Lemma 6, so there exists 0 : P(μ(h1:∞) > c) > 0, and ∀ε ∈ (0, c) : P(μ(h1:∞) > ε) > 0, so for such ε, ε

by factoring out ε and P(μ(h1:∞) > ε) from above, and

by exchanging the sum and the expectation, by the fact that limit of the tail of the sum goes to 0 only if the sum is finite, and by exchanging the sum and expectation back.

Lemma 9. As t → ∞,

∀m ∈ N, α ∈ A, μ : h1:t+1 → H : P(μ(h1:t+1|h1:t) − P(μ(h1:t+1|h1:t)) → 0 w.p.1

Proof.

(a) follows from the entropy inequality, which holds when the first term is a measure, and the second a (semi)measure, and (b) follows from Lemma 8 for ε ∈ (0, c), for some c > 0.

We show that the following event has probability 0 when h1:∞ ∼ P : |P(μ(h1:t+1|h1:t) − P(μ(h1:t+1|h1:t))| > δ i.o., for δ > 0. Note that this event implies |P(μ(h1:t+1|h1:t) − P(μ(h1:t+1|h1:t))| > δ i.o. ∨ w(μ(h1:∞)) > ε) = 0.

Suppose w(μ(h1:∞)) > ε. By Markov’s Inequality on the result above, the probability of the sum being infinite when the expectation is finite, is 0, so

Therefore, ∀ε ∈ (0, c) : P(μ(h1:t+1|h1:t) > δ i.o.) ≤ P(0 ≤ w(μ(h1:∞)) ≤ ε). From Lemma 6, w(μ(h1:∞)) > 0 w.p.1, so

∀ε′ > 0 ∃ε > 0 : P(0 ≤ w(μ(h1:∞)) ≤ ε) < ε′

∀ε′ > 0 : P(0 ≤ w(μ(h1:∞)) ≤ ε) < ε′

P(μ(h1:t+1|h1:t) > δ i.o.) = 0

12
Equivalently, as $t \to \infty$:

$$\forall \alpha^m \in A^{\cup_{i=0}^{m-1} H^i}, h_{t:t+m-1} \in H^m : P_{\mu}^{\alpha^m}(h_{<t+m}| h_{<t}) - P_{\xi}^{\alpha^m}(h_{<t+m}| h_{<t}) \to 0 \text{ w.p.1}$$