Optimistic AIXI

Peter Sunehag and Marcus Hutter

Research School of Computer Science, Australian National University
Canberra Australia
{Peter.Sunehag,Marcus.Hutter}@anu.edu.au

Abstract. We consider extending the AIXI agent by using multiple (or even a compact class of) priors. This has the benefit of weakening the conditions on the true environment that we need to prove asymptotic optimality. Furthermore, it decreases the arbitrariness of picking the prior or reference machine. We connect this to removing symmetry between accepting and rejecting bets in the rationality axiomatization of AIXI and replacing it with optimism. Optimism is often used to encourage exploration in the more restrictive Markov Decision Process setting and it alleviates the problem that AIXI (with geometric discounting) stops exploring prematurely.

Keywords: AIXI, Reinforcement Learning, Optimism, Optimality.

1 Introduction

In this article, we aim to define agents that adapt to asymptotically act optimally for as large a class of environments as possible. This task is fundamental for Artificial General Intelligence with many authors [LH07] using it as a definition of intelligence. In [Hut05] the AIXI agent is defined as a Bayesian reinforcement learning agent with particular attention being put on using the class of all computable environments as the hypothesis class. This agent has some interesting optimality properties. Besides maximizing expected utility with respect to the a-priori distribution by design, it is also Pareto optimal and self-optimizing when this is possible for the considered class. It was, however, shown in [Ors10] that at least with computable horizons, AIXI is not guaranteed to be asymptotically optimal for all computable (deterministic) environments. Furthermore, [LH11] shows that no agent can be.

Here we use multiple priors (or more generally multiple a-priori environments) and the principle of optimism to define more explorative extensions of the AIXI agent with the aim of being able to prove asymptotic optimality under weaker conditions on the true environment. In other words, the agent can adapt successfully to a larger class of environments. The more priors used the more explorative the agent will be; indeed we can even define the agent for all priors though the convergence results will not apply and the agent can end up having no preference between any of the actions in any situation. The meaningful cases include having a compact class of strictly positive weight sequences $w_\nu, \nu \in \mathcal{M}$ for a countable

J. Bach, B. Goertzel, and M. Iklé (Eds.): AGI 2012, LNAI 7716, pp. 312–321, 2012.

hypothesis class $\mathcal{M}$. We can, for example, consider a sequence $\alpha_\nu > 0$ and the set of mixtures with weights satisfying $\omega_\nu \geq \alpha_\nu$ and $\sum_\nu w_\nu = 1$.

In Section 2, we discuss the rational betting theory that has recently been used to derive AIXI [SH11a] and in Section 3, after introducing the reinforcement learning agent setting, we describe how the betting theory leads to active agents. Furthermore, in Section 2, we weaken the assumptions to introduce (in Section 3) our extended AIXI agent. In [SH11a], rationality axioms were presented that lead to the AIXI agent. Here we are going to extend AIXI by breaking the symmetry between accepting and rejecting bets in an optimistic fashion and as a consequence get a multiple-prior model. In the active AI setting where decisions affect the environment, the optimism makes the agent more explorative, which improves its chances of finding an optimal policy. Optimism has previously been used to encourage exploration in the more restrictive setting of Markov Decision Processes [SLL09]. Here we study general countable classes of environments.

In Section 3.2, we present our main results on asymptotic optimality under two conditions on how the a priori environment(s) relate to the true environment. If the a-priori environment ξ dominates an environment ν in the sense that $\xi(\cdot) \geq w_\nu \nu(\cdot)$, then we know from the Blackwell-Dubins theorem [BD62] that ξ will almost surely merge with ν in total variation distance under the followed policy. This is, however, not enough for achieving asymptotic optimality. We will say that ξ is optimistic for ν, if the expected value of following an optimal policy in ξ is always higher than it is in ν. If ξ is both dominating ν and optimistic for ν, then almost surely AIXI asymptotically achieves optimality. In this article, we extend the class of environments that we can prove optimality for by replacing ξ with a compact class of a-priori environments Ξ and decisions are taken according to the policy that maximizes the expected value for the environment in Ξ that is the most optimistic in the current situation. To guarantee asymptotic optimality we only need to assume that the optimistic environment is also optimistic relative the true environment. In a separate article [SH12] we remove those two conditions and replace them with the condition that the true environment lies in the class of a-priori environments, which then essentially serves as a hypothesis class.

2 Optimistic Rational Choice

In [SH11a, SH11b], AIXI was derived from rationality axioms inspired by the traditional literature [NM44, Ram31, Sav54, deF37] on decision making under uncertainty. Here we suggest replacing a symmetry condition between accepting and rejecting bets with optimism. The new weaker condition says that if we reject one side of a bet we must be prepared to accept the other side. The principle of optimism results in a more explorative agent and leads to multiple-prior models. Multiple-priors are also used in an approach sometimes called imprecise probability [Wal00], though our work is distinguished from the imprecise probability approach by actually making a choice among the priors. Axiomatics of multiple-prior models has been studied by [GS89, CMKO00]. These models can be understood as quantifying the uncertainty in estimated probabilities by assigning a whole set (or range) of probabilities. In the passive prediction case when

the decisions do not affect the environment, one often combines the multiple-prior model with caution to achieve more risk averse decisions [CMKO00]. In the active case, we need to take risk to generate experience that one can learn successful behavior from and, therefore, optimism is appropriate.

2.1 Bets

The basic setting used in [SH11a] was inspired by the betting approach of [Ram31, deF37]. In this setting we are about to observe an event from a finite (or countable) alphabet and we are offered a bet (contract) $x = (x_1, ..., x_n)$ where $x_i \in \mathbb{R}$ is the reward received for the outcome i. We first introduce the setting of [SH11a] and its main theorem for the finite case.

Definition 1 (Bet). *Suppose that we are going to observe an event whose outcome is represented by a symbol from an alphabet with m elements. A bet for such an event is an element $x = (x_1, ..., x_m)$ in $\mathbb{R}^m$ and x_j is the reward received if the outcome of the event is the j:th symbol.*

Definition 2 (Decision Maker, Decision). *A decision maker is a pair of sets $Z, \tilde{Z} \subset \mathbb{R}^m$ which defines exactly the bets that are acceptable Z and those that are rejectable $\tilde{Z}$. In other words, a decision maker is a function from $\mathbb{R}^m$ to $\{$accepted, rejected, either, neither$\}$. The function value is called the decision.*

Next we present the axioms and representation theorem from [SH11a].

Definition 3 (Rationality). *We say that the decision maker $(Z, \tilde{Z})$ is rational if*

1. $Z \cup \tilde{Z} = \mathbb{R}^m$
2. $x \in Z \iff -x \in \tilde{Z}$
3. $x, y \in Z,\ \lambda, \gamma \geq 0 \Rightarrow \lambda x + \gamma y \in Z$
4. $\forall k\ x_k > 0\ \Rightarrow x \in Z \setminus \tilde{Z}$

Theorem 1 (Existence of Probabilities, Sunehag&Hutter 2011). *Given a rational decision maker, there are numbers $p_i \geq 0$ that satisfy*

$$\{x \mid \sum x_i p_i > 0\} \subseteq Z \subseteq \{x \mid \sum x_i p_i \geq 0\}. \tag{1}$$

Assuming $\sum_i p_i = 1$ makes the numbers unique and we will use the notation $Pr(i) = p_i$.

Axiom 1 in Definition 3 is really describing the setting rather than an assumption. It says that we must always choose at least one of accept or reject. Axioms $3-4$ were motivated as follows in [SH11a]. If $x \in Z$ and $\lambda \geq 0$ then we want $\lambda x \in Z$ since it is simply a multiple of the same bet. We also want the sum of two acceptable bets to be acceptable. If we are guaranteed to win money we accept the bet and we are not prepared to reject it. Axiom 2 is a symmetry condition between accepting and rejecting which we are going to break in the optimistic setting. In the optimistic setting we will still demand that if we reject x we must accept $-x$ but not the other way around.

2.2 Rational Optimism

We present four axioms for rational optimism. They state properties that the set of accepted and the set of rejected bets must satisfy. The first two relate to optimism. The first one says that if a bet is not rejected it is accepted. The second says that if x is rejected then $-x$ must be accepted. In other words, if we reject one side of a bet we must accept the opposite. This was also argued for in the first set of axioms in the previous setting but in the optimistic setting we do not have the opposite direction. Namely we do *not* say that if x is accepted then $-x$ is rejected. The other two axioms are about rational rejection. If we reject two bets x and y, we reject $\lambda x + \gamma y$ if $\lambda \geq 0$ and $\gamma \geq 0$. The final axiom says that if the reward is guaranteed to be strictly negative we reject the bet. If the $\Rightarrow$ in Axiom 2 was instead an $\iff$ we would have the same axioms as before, just slightly differently expressed.

Definition 4 (Rational Optimism). *We say that the decision maker $Z, \tilde{Z} \subset \mathbb{R}^m$ is a rational optimist if*

1. $x \notin \tilde{Z} \Rightarrow x \in Z$
2. $x \in \tilde{Z} \Rightarrow -x \notin \tilde{Z}$
3. $x, y \in \tilde{Z}$ *and* $\lambda, \gamma \geq 0 \Rightarrow \lambda x + \gamma y \in \tilde{Z}$
4. $x_k < 0 \ \forall k \ \Rightarrow x \in \tilde{Z} \setminus Z$

Theorem 2 (Existence of a set of probabilities). *Given a rational optimist, there is a set $\mathcal{P}$ of probability vectors (p_i), that satisfy*

$$\{x \mid \exists (q_i) \in \mathcal{P} : \sum x_i q_i > 0\} \subseteq Z \subseteq \{x \mid \exists (q_i) \in \mathcal{P} : \sum x_i q_i \geq 0\}. \quad (2)$$

One can always replace $\mathcal{P}$ with an extreme set the size of the alphabet.

Proof. Properties 2 and 3 tell us that the closure $\bar{\tilde{Z}}$ of $\tilde{Z}$ is a (one sided) convex cone. Let $\mathcal{P} = \{(p_i) \in \mathbb{R}^m \mid \sum p_i x_i \leq 0 \ \forall (x_i) \in \bar{\tilde{Z}}\}$. Then, it follows from convexity that $\bar{\tilde{Z}} = \{(x_i) \mid \sum x_i p_i \leq 0 \ \forall (p_i) \in \mathcal{P}\}$. Property 4 tells us that it contains all the elements of only strictly negative coefficients and this implies that for all $(p_i) \in \mathcal{P}$, $p_i \geq 0$ for all i. We can directly conclude that $Z \subseteq \{x \mid \exists (q_i) \in \mathcal{P} : \sum x_i q_i \geq 0\}$ and furthermore, it follows from property 2 that $\{x \mid \sum x_i p_i > 0\} \subseteq Z$ for all $(p_i) \in \mathcal{P}$. Normalizing to $\sum p_i = 1$ does not change anything. Property 1 tells us that $Z \subseteq \{x \mid \exists (q_i) \in \mathcal{P} : \sum x_i q_i \geq 0\}$. □

2.3 Making Choices

If we want to go from decisions on accepting or rejecting bets to a setting where we choose between different bets $x^j, j = 1, 2, 3, ...$, we define preferences by saying that x is better or equal (as in equally good) than y if $x - y \in \bar{Z}$ (the closure of Z), while it is worse or equal if $x - y$ is rejectable. For the first form of rationality stated in Definition 3, the consequence is that one chooses the option with the highest expected utility. If we instead consider optimistic rationality, and if there

is $(p_i) \in \mathcal{P}$ such that $\sum x_i p_i \geq \sum y_i q_i \ \forall (q_i) \in \mathcal{P}$ then $\sum p_i(x_i - y_i) \geq 0$ and, therefore, $x - y \in \bar{Z}$. Therefore, if we choose the bet x^j by

$$\arg\max_j \max_{p \in \mathcal{P}} \sum x_i^j p_i$$

we are guaranteed that this bet is preferable to all other bets but not necessarily strictly so, even if $\max_{p \in \mathcal{P}} \sum x_i^j p_i$ is strictly larger than all competitors.

3 Intelligent Agents

We will consider an agent [RN10, Hut05] that interacts with an environment through performing actions a_t from a finite set $\mathcal{A}$ and receives observations o_t from a finite set $\mathcal{O}$ and rewards r_t from a finite set $\mathcal{R} \subset [0,1]$. Let $\mathcal{H} = \cup_n (\mathcal{A} \times \mathcal{O} \times R)^n$ be the set of histories and let ϵ be the empty history. A function $\nu : \mathcal{H} \times \mathcal{A} \to \mathcal{O} \times \mathcal{R}$ is called a deterministic environment. A function $\pi : \mathcal{H} \to \mathcal{A}$ is called a (deterministic) policy or an agent. We define the value function V based on geometric discounting by $V_\nu^\pi(h_{t-1}) = \sum_{i=t}^\infty \gamma^{i-t} r_i$ where the sequence r_i are the rewards achieved by following π from time step t onwards in the environment ν after having seen h_{t-1}.

Instead of viewing the environment as a function $\mathcal{H} \times \mathcal{A} \to \mathcal{O} \times \mathcal{R}$ we can equivalently write it as a function $\nu : \mathcal{H} \times \mathcal{A} \times \mathcal{O} \times \mathcal{R} \to \{0,1\}$ where we write $\nu(o, r|h, a)$ for the function value of (h, a, o, r). It equals zero if in the first formulation (h, a) is not sent to (o, r) and 1 if it is. In the case of stochastic environments we instead have a function $\nu : \mathcal{H} \times \mathcal{A} \times \mathcal{O} \times \mathcal{R} \to [0,1]$ such that $\sum_{o,r} \nu(o, r|h, a) = 1 \ \forall h, a$. Furthermore, we define $\nu(h_t|\pi) := \nu(o r_{1:t}|\pi) := \Pi_{i=1}^t \nu(o_i r_i | a_i, h_{i-1})$ where $a_i = \pi(h_{i-1})$. $\nu(\cdot|\pi)$ is a probability measure over strings or sequences as will be discussed in the next section and we can define $\nu(\cdot|\pi, h_{t-1})$ by conditioning $\nu(\cdot|\pi)$ on h_{t-1}. We define $V_\nu^\pi(h_{t-1}) := \mathbb{E}_{\nu(\cdot|\pi, h_{t-1})} \sum_{i=t}^\infty \gamma^{i-t} r_i$ and $V_\nu^*(h_{t-1}) := \max_\pi V_\nu^\pi(h_{t-1})$. Given a countable class of environments $\mathcal{M}$ and strictly positive prior weights w_ν for all $\nu \in \mathcal{M}$, we define the a-priori environment ξ by letting $\xi(\cdot) = \sum w_\nu \nu(\cdot)$ and the AIXI agent is defined by following the policy

$$\pi^* := \arg\max_\pi V_\xi^\pi(\epsilon).$$

3.1 Rational Optimistic Sequential Decisions

There are some extensions to the results from Section 2 needed to reach the full AI (generic reinforcement learning) case we have in mind, but the procedure for doing this has already been outlined in [SH11a]. The first extension is to reactive environments where the outcome is affected by the choice made. One then chooses between different actions to take. It was concluded that it follows from the rationality axioms that there is a probability (p_i^j) for the outcome i given action j, and the action given a bet $x = (x_i)$ is chosen by

$$\arg\max_j \sum x_i p_i^j.$$

The extension to finitely many sequential decisions is simply about considering the choice to be made to be a choice of policy π (previously j). The discounted value $\sum r_t\gamma^t$ achieved then plays the role of the bet x_i and the decision on what policy to follow is taken according to

$$\arg\max_{\pi} V_{\xi}^{\pi}$$

where ξ is the probabilistic a priori belief (the p_i^j) and $V_{\xi}^{\pi} = \sum p_i^j(\sum r_t^i\gamma^t)$ where r_t^i is the reward achieved at time t in outcome sequence i in an enumeration of all the possible histories. The rational optimist takes the decision

$$\pi^{\circ} := \arg\max_{\pi}\max_{\xi\in\Xi} V_{\xi}^{\pi}$$

for a set of beliefs (environments) Ξ (corresponds to $\mathcal{P}$ before) which we will assume is compact in the metric topology of the total variation distance as in [SH12].

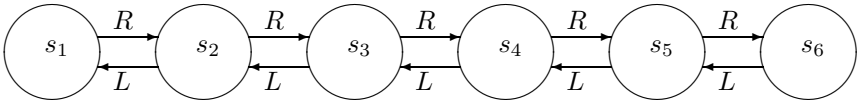
3.2 Asymptotic Optimality

In this section we will first prove that AIXI is asymptotically optimal if its a-priori environment ξ is both dominating the true environment μ in the sense of $\xi(\cdot) \geq c\mu(\cdot)$ and optimistic in the sense that $V_{\xi}^*(h_t) \geq V_{\mu}^*(h_t)$ (for large t). We extend this by replacing ξ with a compact (with respect to the total variation distance) set Ξ and prove that we then only need there to be, for each h_t (for t large), some $\xi \in \Xi$ such that $V_{\xi}^*(h_t) \geq V_{\mu}^*(h_t)$. The first domination property is most easily satisfied for $\xi(\cdot) = \sum_{\nu\in\mathcal{M}} w_\nu\nu(\cdot)$ with $w_\nu > 0$ where $\mathcal{M}$ is a countable class of environments with $\mu \in \mathcal{M}$. We are going to provide one simple example for the first theorem to illustrate what it is saying in a simple setting while after the second theorem we discuss the example that we really have in mind. This example addresses the AIXI agent as it was introduced in [Hut05] with a Solomonoff prior and the problem of defining a natural Universal Turing Machine [Mül10].

Theorem 3. *Suppose that $\xi(\cdot) \geq c\mu(\cdot)$ for some $c > 0$ and μ is the true environment. Also suppose that there almost surely is $T_1 < \infty$ such that $V_{\xi}^*(h_t) \geq V_{\mu}^*(h_t)$ $\forall t \geq T_1$. Suppose that the policy π^* acts according to the AIXI agent based on ξ in μ. Then there is almost surely, for every $\varepsilon > 0$, a time $T < \infty$ such that $V_{\mu}^{\pi^*}(h_t) \geq V_{\mu}^*(h_t) - \varepsilon$ $\forall t \geq T$.*

Proof. Due to the dominance we can (using the Blackwell-Dubins merging of opinions theorem [BD62]) say that almost surely there is for every $\varepsilon' > 0$, a $T < \infty$ such that $d(\xi(\cdot|h_t, \pi^*), \mu(\cdot|h_t, \pi^*)) < \varepsilon$ where d is the total variation distance. This implies that $|V_{\xi}^{\pi^*}(h_t) - V_{\mu}^{\pi^*}(h_t)| < \frac{\varepsilon'}{1-\gamma} := \varepsilon$ which means that, if $T \geq T_1$, $V_{\mu}^{\pi^*}(h_t) \geq V_{\xi}^*(h_t) - \varepsilon \geq V_{\mu}^*(h_t) - \varepsilon$. □

Example 1 (Line Environment). Consider an agent who, when given a class of environments, will always choose its prior based on simplicity which is in accordance with Occam's razor [Hut05]. First let us look at a class $\mathcal{M}$ of two environments which both have six states $s_1, ..., s_6$ and two actions L (left) and R (right). Action R changes s_k to s_{k+1}, L to s_{k-1}. Also L in s_1 or R in s_6 result in staying. We start at s_1. Being at s_1 yields a reward of 0, while s_2, s_3, s_4, s_5 give reward -1 and the reward in s_6 depends on the environment.



In one of the environments ν_1, this reward is -1 while in ν_2 it is 1. Since ν_2 is not simpler than ν_1 it will not have higher weight and if γ is only modestly high we will not explore along the line despite that in ν_2 it would be optimal to do so. However, if we define another environment ν_3 by letting the reward at s_6 be really high, then when including ν_3 in the mixture, the agent will end up with an a priori environment ξ that is optimistic for ν_1 and ν_2 and we can guarantee optimality for any γ.

Note that the example above is only supposed to show how the optimism condition can be satisfied for a subclass of the class one has a prior over. It will almost never be satisfied for the whole class. In the next theorem we prove that for the extended agent with a class of priors, only one of them needs to be optimistic at a time while we need all to be dominant.

Theorem 4. *Suppose that Ξ is a compact set for the total variation topology (maximized over all policies and histories) of a-priori environments such that for each $\xi \in \Xi$ there is $c_{\xi,\mu} > 0$ such that $\xi(\cdot) \geq c_{\xi,\mu}\mu(\cdot)$ where μ is the true environment. Also suppose that there almost surely is $T_1 < \infty$ such that for $t \geq T_1$ there is $\xi \in \Xi$ such that $V^*_\xi(h_t) \geq V^*_\mu(h_t)$. Suppose that the policy π° acts according to the rational optimistic agent based on Ξ in μ. Then there is almost surely, for every $\varepsilon > 0$, a time $T < \infty$ such that $V^{\pi^\circ}_\mu(h_t) \geq V^*_\mu(h_t) - \varepsilon \ \forall t \geq T$.*

The theorem is proven by combining the proof technique from the previous theorem with the following lemma. We have made this lemma easier by formulating it for time $t = 0$ (when the history is the empty string ϵ), though when proving Theorem 4 it is used for a later time point when the environments in the class have merged sufficiently in the sense of total variation diameter.

Lemma 1 (Optimism is nearly optimal). *Suppose that an infinite history h has been generated by running π° in the environment μ. Given $\varepsilon > 0$ there is $\tilde{\varepsilon} > 0$ such that $V^{\pi^\circ}_\mu(\epsilon) \geq \max_\pi V^\pi_\mu(\epsilon) - \varepsilon$ if*

$$|V^{\pi^\circ}_{\nu_1}(h_t) - V^{\pi^\circ}_{\nu_2}(h_t)| < \tilde{\varepsilon} \ \forall t, \forall \nu_1, \nu_2 \in \Xi.$$

Proof. Let $\nu^*_{h_t}$ be the environment in $\arg\max_\nu \max_\pi V^\pi_\nu(h_t)$ that π° use to choose the next action a_{t+1} after experiencing h_t. Define $\hat{\nu}$ by letting

$$\hat{\nu}(o_t r_t | h_{t-1}, a) = \nu^*_{h_{t-1}}(o_t r_t | h_{t-1}, a).$$

We will show that this implies that $V_{\hat{\nu}}^{\pi^\circ} \geq \max_{\nu\in\mathcal{M},\pi} V_\nu^\pi$ where V_ν^π denotes $V_\nu^\pi(\epsilon)$. Let

$$\hat{\nu}_s(o_t r_t|h_{t-1},a) = \begin{cases} \hat{\nu}(o_t r_t|h_{t-1},a)\ \forall h_{t-1}, \text{ for } t \leq s \\ \hat{\nu}_s(o_t r_t|h_{t-1},a) = \nu^*_{h_s}(o_t r_t|h_{t-1},a)\ \forall h_{t-1}, \text{ for } t > s. \end{cases}$$

$\hat{\nu}_1$ equals ν^*_ε at all time points and thus $V^\pi_{\hat{\nu}_1} = V^\pi_{\nu^*_\varepsilon}$. Let $\hat{R}_t^\nu$ be the expected accumulated (discounted) reward ($\mathbb{E}\sum_{i=1}^t \gamma^{i-1} r_i$) when following π° in environment ν up to time t.

$$\max_{\pi_{2:\infty}} V^{\pi^\circ_{0:1}\pi_{2:\infty}}_{\hat{\nu}_2} = \max_{\pi_{1:\infty}} (\hat{R}_1^{\nu^*_\epsilon} + \gamma \mathbb{E}_{h_1|\nu^*_\epsilon,\pi^\circ} V^{\pi_{1:\infty}}_{\nu^*_{h_1}}(h_1)) \geq$$

$$\max_{\pi_{1:\infty}} (\hat{R}_1^{\nu^*_\epsilon} + \gamma \mathbb{E}_{h_1|\nu^*_\epsilon,\pi^\circ} V^{\pi_{1:\infty}}_{\nu^*_\epsilon}(h_1)) = \max_\pi V^\pi_{\hat{\nu}_1}$$

since $\max_\pi V^\pi_{\nu^*_{h_1}}(h_1)) \geq \max_\pi V^\pi_\nu(h_1))\ \forall \nu \in \mathcal{M}$. In the same way,

$$\max_{\pi_{k:\infty}} V^{\pi^\circ_{0:k-1}\pi_{k:\infty}}_{\hat{\nu}_k} \geq \max_{\pi_{k-1:\infty}} V^{\pi^\circ_{0:k-2}\pi_{k-1:\infty}}_{\hat{\nu}_{k-1}} \forall k$$

and it follows that $V_{\hat{\nu}}^{\pi^\circ} \geq \max_{\pi,\nu\in\mathcal{M}} V_\nu^\pi$. To conclude the proof, we show that if $\tilde{\varepsilon}$ is small enough, then

$$|V_{\hat{\nu}}^{\pi^\circ} - V_\mu^{\pi^\circ}| < \varepsilon \tag{3}$$

where μ is the true environment. That (3) is true is shown by induction. $\hat{\nu}_1 \in \mathcal{M}$ and, therefore, (3) holds with $\hat{\nu}_1$ instead of $\hat{\nu}$ if $\tilde{\varepsilon} \leq \varepsilon$. $\hat{\nu}_k$ and $\hat{\nu}_{k+1}$ are identical for the first k time step so $|V_{\hat{\nu}_k}^{\pi^\circ} - V_{\hat{\nu}_{k+1}}^{\pi^\circ}| < \gamma^k\tilde{\varepsilon}$. We conclude that

$$|V_{\hat{\nu}_1}^{\pi^\circ} - V_{\hat{\nu}}^{\pi^\circ}| < \frac{\tilde{\varepsilon}}{1-\gamma}$$

and if $\tilde{\varepsilon} + \frac{\tilde{\varepsilon}}{1-\gamma} \leq \varepsilon$ then (3) holds and the proof is complete. □

Proof. **of Theorem** 4. Due to the compactness, there is almost surely for every ε', a $T < \infty$ such that $d(\xi(\cdot|h_t,\pi^\circ),\mu(\cdot|h_t,\pi^\circ)) < \varepsilon\ \forall \xi \in \Xi\ \forall t \geq T$. This means that $|V_\xi^\pi(h_t) - V_\mu^\pi(h_t)| < \frac{\varepsilon'}{1-\gamma} := \varepsilon\ \forall \xi \in \Xi$. Applying Lemma 1 to the ξ that is optimistic at time T proves the result.

Example 2. For any Universal Turing Machine (UTM) U the corresponding Solomonoff distribution ξ_U, (see [LV93] for details) is dominant for any lower semi-computable semi-measure over infinite sequences. [Hut05] extends these constructions to the active case and defines (for each U) an environment that is dominant for all lower semi-computable environments and defines the AIXI agent based on it. The AIXI agent would have uniquely defined the most intelligent agent according to the underlying sense of intelligence (maximizing expected reward), if the choice of UTM was clear. Many have without success tried to find a single "natural" Turing machine and there might in fact be no such machine [Mül10]. With the approach that we introduce in this article one can pick

finitely many machines that one considers to be natural. Though this does not fully resolve the issue of having to make arbitrary choices, it alleviates it by no longer demanding a unique choice of UTM. We can consider an enumeration of all UTMs U_i and let the agent $Agent_n$ be based on the first n machines. $Agent_n$ has better guarantees than $Agent_m$ (in the sense of Theorem 4) if $n > m$. The conclusion does, however, not carry through to a limiting case. Note, that if we instead combine finitely many machines into one by letting the first few bits represent a choice of machine, the resulting environment will not be optimistic for all the environments that we achieve optimism for with the multiple-prior approach.

4 Conclusions

We extended AIXI to a multiple-prior setting using the principle of optimism. This decreases the arbitrariness of picking an a-priori environment or a reference machine to base a Solomonoff prior on. Furthermore, we show that this leads to asymptotic optimality guarantees for more environments. We also explain that this extension is related to replacing symmetry with optimism in the recently introduced axiomatization of AIXI.

In a separate article [SH12], we perform a different sort of analysis where it is not assumed that all the environments in Ξ are dominating the true environment μ. The analysis, however, adds the assumption that the true environment is a member of this class of environments. The a priori environments are then naturally thought of as a hypothesis class rather than mixtures over some hypothesis class. In this article we note, that there is no mathematical difference between a class of environments that is considered a hypothesis class and one that is considered a class of a priori environments. However, in the case where we consider Ξ to be a hypothesis class, Ξ has to be very large to yield an agent that is guaranteed asymptotic optimality for many environments (the environments in Ξ), while in the case when it represents a mixture over a hypothesis class, a singleton Ξ (the AIXI case) is already a powerful agent. Another distinction is that in the case studied in [SH12], we need a mechanism for excluding environments from the class as they become inconsistent with experience.

A practical agent that builds upon the ideas of this article and the companion article [SH12], is a variation of a Bayesian reinforcement learning agent. A common way of implementing a practical Bayesian agent is that one samples several environments from the posterior and then act for a period of time according to what would give the highest expected value when averaging the expected value over the sampled environments. Instead we here suggest acting optimistically with respect to those sampled environments who are then, for a period of time, basically treated as a restricted hypothesis class. In the MDP case this is close to what the BOSS algorithm [ALL09] is doing.

Acknowledgement. This work was supported by ARC grant DP120100950. The authors are grateful for feedback from Tor Lattimore.

References

[ALL09] Asmuth, J., Li, L., Littman, M.L., Nouri, A., Wingate, D.: Pac-mdp reinforcement learning with bayesian priors (2009)

[BD62] Blackwell, D., Dubins, L.: Merging of Opinions with Increasing Information. The Annals of Mathematical Statistics 33(3), 882–886 (1962)

[CMKO00] Casadesus-Masanell, R., Klibanoff, P., Ozdenoren, E.: Maxmin Expected Utility over Savage Acts with a Set of Priors. Journal of Economic Theory 92(1), 35–65 (2000)

[deF37] deFinetti, B.: La prévision: Ses lois logiques, ses sources subjectives. In: Annales de l'Institut Henri Poincaré 7, Paris, pp. 1–68 (1937)

[GS89] Gilboa, I., Schmeidler, D.: Maxmin expected utility with non-unique prior. Journal of Mathematical Economics 18(2), 141–153 (1989)

[Hut05] Hutter, M.: Universal Articial Intelligence: Sequential Decisions based on Algorithmic Probability. Springer, Berlin (2005)

[LH07] Legg, S., Hutter, M.: Universal Intelligence: A defintion of machine intelligence. Mind and Machine 17, 391–444 (2007)

[LH11] Lattimore, T., Hutter, M.: Asymptotically Optimal Agents. In: Kivinen, J., Szepesvári, C., Ukkonen, E., Zeugmann, T. (eds.) ALT 2011. LNCS, vol. 6925, pp. 368–382. Springer, Heidelberg (2011)

[LV93] Li, M., Vitany, P.: An Introduction to Kolmogov Complexity and Its Applications. Springer (1993)

[Mül10] Müller, M.: Stationary algorithmic probability. Theor. Comput. Sci. 411(1), 113–130 (2010)

[NM44] Neumann, J., Morgenstern, O.: Theory of Games and Economic Behavior. Princeton University Press (1944)

[Ors10] Orseau, L.: Optimality Issues of Universal Greedy Agents with Static Priors. In: Hutter, M., Stephan, F., Vovk, V., Zeugmann, T. (eds.) ALT 2010. LNCS, vol. 6331, pp. 345–359. Springer, Heidelberg (2010)

[Ram31] Ramsey, F.: Truth and probability. In: Braithwaite, R.B. (ed.) The Foundations of Mathematics and other Logical Essays, ch. 7, pp. 156–198. Brace & Co. (1931)

[RN10] Russell, S.J., Norvig, P.: Artificial Intelligence: A Modern Approach, 3rd edn. Prentice Hall, Englewood Cliffs (2010)

[Sav54] Savage, L.: The Foundations of Statistics. Wiley, New York (1954)

[SH11a] Sunehag, P., Hutter, M.: Axioms for Rational Reinforcement Learning. In: Kivinen, J., Szepesvári, C., Ukkonen, E., Zeugmann, T. (eds.) ALT 2011. LNCS, vol. 6925, pp. 338–352. Springer, Heidelberg (2011)

[SH11b] Sunehag, P., Hutter, M.: Principles of Solomonoff induction and AIXI. In: Solomonoff Memorial Conference, Melbourne, Australia (2011)

[SH12] Sunehag, P., Hutter, M.: Optimistic Agents Are Asymptotically Optimal. In: Thielscher, M., Zhang, D. (eds.) AI 2012. LNCS, vol. 7691, pp. 15–26. Springer, Heidelberg (2012)

[SLL09] Strehl, A.L., Li, L., Littman, M.L.: Reinforcement learning in finite MDPs: PAC analysis. Journal of Machine Learing Research 10, 2413–2444 (2009)

[Wal00] Walley, P.: Towards a unified theory of imprecise probability. Int. J. Approx. Reasoning, 125–148 (2000)