Journal of Machine Learning Research 16 (2015) 1345-1390 Submitted 12/13; Revised 2/15; Published 8/15

Rationality, Optimism and Guarantees in General Reinforcement Learning

Peter Sunehag* sunehag@google.com
Marcus Hutter Marcus.Hutter@anu.edu.au
Research School of Computer Science (RSISE BLD 115)
The Australian National University, ACT 0200, Canberra Australia

Editor: Laurent Orseau

Abstract

In this article,[1] we present a top-down theoretical study of general reinforcement learning agents. We begin with rational agents with unlimited resources and then move to a setting where an agent can only maintain a limited number of hypotheses and optimizes plans over a horizon much shorter than what the agent designer actually wants. We axiomatize what is rational in such a setting in a manner that enables optimism, which is important to achieve systematic explorative behavior. Then, within the class of agents deemed rational, we achieve convergence and finite-error bounds. Such results are desirable since they imply that the agent learns well from its experiences, but the bounds do not directly guarantee good performance and can be achieved by agents doing things one should obviously not. Good performance cannot in fact be guaranteed for any agent in fully general settings. Our approach is to design agents that learn well from experience and act rationally. We introduce a framework for general reinforcement learning agents based on rationality axioms for a decision function and an hypothesis-generating function designed so as to achieve guarantees on the number errors. We will consistently use an optimistic decision function but the hypothesis-generating function needs to change depending on what is known/assumed. We investigate a number of natural situations having either a frequentist or Bayesian flavor, deterministic or stochastic environments and either finite or countable hypothesis class. Further, to achieve sufficiently good bounds as to hold promise for practical success we introduce a notion of a class of environments being generated by a set of laws. None of the above has previously been done for fully general reinforcement learning environments.

Keywords: reinforcement learning, rationality, optimism, optimality, error bounds

1. Introduction

A general reinforcement learning environment returns observations and rewards in cycles to an agent that feeds actions to the environment. An agent designer's aim is to construct an agent that accumulates as much reward as possible. Ideally, the agent should maximize a given quality measure like e.g., expected accumulated reward or the maximum accumulated reward that is guaranteed with a certain given probability. The probabilities and expectation should be the actual, i.e., with respect to the true environment. Performing this task

∗. The first author is now at Google - DeepMind, London UK

1. This article combines and extends our conference articles (Sunehag and Hutter, 2011, 2012a,b, 2013, 2014) and is further extended by (Sunehag and Hutter, 2015) covering stochastic laws.

well in an unknown environment is an extremely challenging problem (Hutter, 2005). Hutter (2005) advocated a Bayesian approach to this problem while we here introduce optimistic agents as an alternative.

The Bayesian approach to the above task is to design an agent that approximately maximizes the quality measure with respect to an a priori environment chosen by the designer. There are two immediate problems with this approach. The first problem is that the arbitrary choice of a priori environment, e.g., through a prior defining a mixture of a hypothesis class, substantially influences the outcome. The defined policy is optimal by definition in the sense of achieving the highest quality with respect to the a priori environment, but its quality with respect to other environments like the true one or a different mixture, might be much lower. The second problem is that computing the maximizing actions is typically too hard, even approximately. We will below explain how a recent line of work attempts to address these problems and see that the first problem is partially resolved by using information-theoretic principles to make a "universal" choice of prior, while the second is not resolved. Then we will discuss another way in which Bayesian methods are motivated which is through rational choice theory (Savage, 1954).

The optimistic agents that we introduce in this article have the advantage that they satisfy guarantees that hold regardless of which environment from a given class is the true one. We introduce the concept of a class being generated by a set of laws and improve our bounds from being linear in the number of environments to linear in the number of laws. Since the number of environments can be exponentially larger than the number of laws this is of vital importance and practically useful environment classes should be such that its size is exponential in the number of laws. We will discuss such guarantees below as well as the mild modification of the classical rationality framework required to deem an optimistic agent rational. We also explain why such a modification makes sense when the choice to be made by an agent is one in a long sequence of such choices in an unknown environment.

Information-theoretic priors and limited horizons. Hutter (2005) and Veness et al. (2011) choose the prior, which can never be fully objective (Leike and Hutter, 2015), through an information-theoretic approach based on the code length of an environment by letting environments with shorter implementations be more likely. Hutter (2005) does this for the universal though impractical class of all lower semi-computable environments while Veness et al. (2011) use a limited but useful class based on context trees. For the latter, the context tree weighting (CTW) algorithm (Willems et al., 1995) allows for efficient calculation of the posterior. However, to optimize even approximately the quality measure used to evaluate the algorithm for the actual time-horizon (e.g., a million time steps), is impossible in complex domains. The MC-AIXI-CTW agent in Veness et al. (2011), which we employ to illustrate the point, uses a Monte-Carlo tree search method to optimize a geometrically discounted objective. Given a discount factor close to 1 (e.g., 0.99999) the effective horizon becomes large (100000). However, the tree search is only played out until the end of episode in the tasks considered in Veness et al. (2011). Playing it out for 100000 time steps for each simulation at each time step would be completely infeasible. When an agent maximizes the return from a much shorter horizon than the actual, e.g., one game instead of a 1000 games of PacMan, the exploration versus exploitation dilemma shows up. If the environment is fully known, then maximizing the return for one episode is perfect. In an unknown environment such a strategy can be a fatal mistake. If the expected return

is maximized for a shorter working horizon, i.e., the agent always exploits, then it is likely to keep a severely sub-optimal policy due to insufficient exploration. Veness et al. (2011) addressed this heuristically through random exploration moves.

Our agent framework. In Section 3, we introduce a framework that combines notions of what is considered desirable in decision theory with optimality concepts from reinforcement learning. In this framework, an agent is defined by the choice of a decision function and a hypothesis-generating function. The hypothesis-generating function feeds the decision function a finite class of environments at every time step and the decision function chooses an action/policy given such a class. The decision-theoretic analysis of rationality is used to restrict the choice of the decision function, while we consider guarantees for asymptotic properties and error bounds when designing the hypothesis-generating function.

All the agents we study can be expressed with an optimistic decision function but we study many different hypothesis-generating functions which are suitable under different assumptions. For example, with a domination assumption there is no need to remove environments, it would only worsen the guarantees. Hence a constant hypothesis-generating function is used. If we know that the environment is in a certain finite class of deterministic environments, then a hypothesis-generating function that removes contradicted environments but does not add any is appropriate. Similarly, when we have a finite class of stochastic but non-dominant environments that we assume the truth belongs to, the hypothesis-generating function should not add to the class but needs to exclude those environments that have become implausible.

If we only know that the true environment is in a countable class and we choose an optimistic decision function, the agent needs to have a growing finite class. In the countable case, a Bayesian agent can still work with the whole countable class at once (Lattimore, 2014), though to satisfy the desired guarantees that agent (BayesExp) was adjusted in a manner we here deem irrational. Another alternative adjustment of a Bayesian agent that is closer to fitting our framework is the Best of Sampled Set (BOSS) algorithm (Asmuth et al., 2009). This agent samples a finite set of environments (i.e., hypothesis-generation) from the posterior and then constructs an optimistic environment by combining transition dynamics from all those environments in the most optimistic manner and then optimize for this new environment (optimistic decision). This is an example of an agent that uses what we refer to as environments constructed by combining laws, though BOSS belongs in the narrow Markov Decision Process setting, while we here aim for full generality.

Rationality. In the foundations of decision theory, the focus is on axioms for rational preferences (Neumann and Morgenstern, 1944; Savage, 1954) and on making a single decision that does not affect the event in question but only its utility. The single decision setting can actually be understood as incorporating sequential decision-making since the one choice can be for a policy to follow for a period of time. This latter perspective is called normal form in game theory. We extend rational choice theory to the full reinforcement learning problem. It follows from the strictest version of the axioms we present that the agent must be a Bayesian agent. These axioms are appropriate when an agent is capable of optimizing the plan for its entire life. Then we loosen the axioms in a way that is analogous to the multiple-prior setting by Gilboa and Schmeidler (1989), except that ours enable optimism instead of pessimism and are based on a given utility function. These more permissive

axioms are suitable for a setting where the agent must actually make the decisions in a sequence due to not being able to optimize over the full horizon. We prove that optimism allows for asymptotic optimality guarantees and finite error bounds not enjoyed by a realist (expected utility maximizer).

Guarantees. In the field of reinforcement learning, there has been much work dedicated to designing agents for which one can prove asymptotic optimality or sample complexity bounds. The latter are high probability bounds on the number of time steps where the agent does not make a near optimal decision (Strehl et al., 2009). However, a weakness with sample complexity bounds is that they do not directly guarantee good performance for the agent. For example, an agent who has the opportunity to self-destruct can achieve subsequent optimality by choosing this option. Hence, aiming only for the best sample complexity can be a very bad idea in general reinforcement learning. If the environment is an ergodic MDPs or value-stable environment (Ryabko and Hutter, 2008) where the agent can always recover, these bounds are more directly meaningful. However, optimizing them blindly is still not necessarily good. Methods that during explicit exploration phases, aim at minimizing uncertainty by exploring the relatively unknown, can make very bad decisions. If one has an option offering return in the interval $[0, 0.3]$ and another option has return in the interval $[0.7, 0.8]$ one should have no interest in the first option since its best case scenario is worse than the worst case scenario of the other option. Nevertheless, some devised algorithms have phases of pure exploration where the most uncertain option is chosen. On the other hand, we will argue that one can rationally choose an option with return known to be in $[0.2, 0.85]$ over either. Assuming uniform belief over those intervals, the latter option is, however, not strictly rational under the classical axioms that are equivalent to choosing according to maximum subjective expected utility. We will sometimes use the term weakly rational for the less strict version of rationality considered below.

Here we consider agents that are rational in a certain decision-theoretic sense and within this class we design agents that make few errors. Examples of irrational agents, as discussed above, are agents that rely on explicit phases of pure exploration that aim directly at excluding environments while a category of prominent agents instead rely on optimism (Szita and Lörincz, 2008; Strehl et al., 2009; Lattimore and Hutter, 2012). Optimistic agents investigate whether a policy is as good as the hypothesis class says it might be but not whether something is bad or very bad. We extend these kinds of agents from MDP to general reinforcement learning and we deem them rational according to axioms presented here in Section 2.

The bounds presented here, like discussed above, are of a sort that the agent is guaranteed to eventually act nearly as well as possible given the history that has been generated. Since the risk of having all prospects destroyed cannot be avoided in the fully general setting, we have above argued that the bounds should be complemented with a demand for acting rationally. This does of course not prevent disaster, since nothing can. Hutter (2005) brings up a heaven and hell example where either action a_1 takes the agent to hell (min reward forever) and a_2 to heaven (max reward forever) or the other way around with a_2 to hell and a_1 to heaven. If one assumes that the true environment is safe (Ryabko and Hutter, 2008) as in always having the same optimal value from all histories that can occur, this kind of bounds are directly meaningful. Otherwise, one can consider an agent that is first pessimistic and rules out all actions that would lead to disaster for some environment

in its class and then takes an optimistic decision among the remaining actions. The bounds then apply to the environment class that remains after the pessimist has ruled out some actions. The resulting environments might not have as good prospects anymore due to the best action being ruled out, and in the heaven and hell example both actions would be ruled out and one would have to consider both. However, we repeat: there are no agents that can guarantee good outcomes in general reinforcement learning (Hutter, 2005).

The bounds given in Section 5 have a linear dependence on the number of environments in the class. While this rate is easily seen to be the best one can do in general (Lattimore et al., 2013a), it is exponentially worse than what we are used to from Markov Decision Processes (MDPs) (Lattimore and Hutter, 2012) where the linear (up to logarithms) dependence is on the size of the state space instead. In Section 5.2 we introduce the concept of laws and environments generated by sets of laws and we achieve bounds that are linear in the number of laws instead of the number of environments. All environment classes are trivially generated by sets of laws that equal the environments but some can also be represented as generated by exponentially fewer laws than there are environments. Such environment classes have key elements in common with an approach that has been heuristically developed for a long time, namely collaborative multi-agent systems called Learning Classifier Systems (LCS) (Holland, 1986; Hutter, 1991; Drugowitsch, 2007) or artificial economies (Baum and Durdanovic, 2001; Kwee et al., 2001). Such systems combine sub-agents that make recommendations and predictions in limited contexts (localization), sometimes combined with other sub-agents' predictions for the same single decision (factorization). The LCS family of approaches are primarily model-free by predicting the return and not future observations while what we introduce here is model-based and has a dual interpretation as an optimistic agent, which allows for theoretical guarantees.

Related work. Besides the work mentioned above, which all use discounted reward sums, Maillard et al. (2011); Nguyen et al. (2013); Maillard et al. (2013) extend the UCRL algorithm and regret bounds (Auer and Ortner, 2006) from undiscounted MDPs to problems where the environments are defined by combining maps from histories to states with MDP parameters as in Hutter (2009b); Sunehag and Hutter (2010). Though Maillard et al. (2011, 2013) study finite classes, Nguyen et al. (2013) extend their results by incrementally adding maps. Their algorithms use undiscounted reward sums and are, therefore, in theory not focused on a shorter horizon but on average reward over an infinite horizon. However, to optimize performance over long horizons is practically impossible in general. The online MDP with bandit feedback work (Neu et al., 2010; Abbasi-Yadkori et al., 2013) aims at general environments but limited to finitely many policies called experts to choose between. We instead limit the environment class in size, but consider any policies.

Outline. We start below with notation and background for general reinforcement learning and then in Section 2 we introduce the axioms for rational and rational optimistic agents. In Section 3 we introduce an agent framework that fits all the agents studied in this article and we make the philosophy fully explicit. It consists of two main parts, rational decision functions (Section 3.1) and hypothesis-generating functions (Section 3.2) that given a history delivers a class of environments to the decision function. In Section 4 we show the importance of optimism for asymptotic optimality for a generic Bayesian reinforcement learning agent called AIXI and we extend this agent to an optimistic multiple-prior agent

with stronger asymptotic guarantees. The required assumption is the a priori environments' dominance over the true environment and that at least one a priori environment is optimistic for the true environment.

In Section 5 and Section 6 we continue to study optimistic agents that pick an optimistic hypothesis instead of an optimistic a priori distribution. This is actually the very same mathematical formula for how to optimistically make a decision given a hypothesis class. However, in this case we do not assume that the environments in the class dominate the truth and the agent, therefore, needs to exclude environments which are not aligned with observations received. Instead of assuming dominance as in the previous section, we here assume that the truth is a member of the class. It is interesting to notice that the only difference between the two sections, despite their very different interpretations, is the assumptions used for the mathematical analysis. In Section 5.2 we also show that understanding environment classes as being generated by finite sets of partial environments that we call laws, allows for error bounds that are linear in the number of laws instead of in the number of environments. This can be an exponential improvement.

In earlier sections the hypothesis-generating functions either deliver the exact same class (except for conditioning the environments on the past) at all times or just remove implausible environments from an initial class while in Section 7 we consider hypothesis-generating functions that also add new environments and exhaust a countable class in the limit. We prove error bounds that depend on how fast new environments are introduced. Section 8 contains the conclusions. The appendix contains extensions of various results.

We summarize our contributions and where they can be found in the following list:

- Axiomatic treatment of rationality and optimism: Section 2.
- Agent framework: Section 3
- Asymptotic results for AIXI (rational) and optimistic agents using finite classes of dominant stochastic environments: Section 4
- Asymptotic and finite error bounds for optimistic agents with finite classes of deterministic (non-dominant) environments containing the truth, as well as improved error rates for environment classes based on laws: Section 5
- Asymptotic results for optimistic agents with finite classes of stochastic non-dominant environments containing the truth: Section 6
- Extensions to countable classes: Section 7.
- Extending deterministic results from smaller class of conservative optimistic agents to larger class of liberal optimistic agents: Appendix A
- Extending axioms for rationality to countable case: Appendix B
- A list of important notation can be found in Appendix C

General reinforcement learning: notation and background. We will consider an agent (Russell and Norvig, 2010; Hutter, 2005) 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 $\mathcal{R} \subset [0,1]$ resulting in a history $h_t := a_0 o_1 r_1 a_1, ..., o_t r_t$. These sets can be allowed to depend on time or context but we do not write this out explicitly. Let $\mathcal{H} := \epsilon \cup (\mathcal{A} \times \cup_n (\mathcal{O} \times R \times \mathcal{A})^n \times (\mathcal{O} \times \mathcal{R}))$ be the set of histories where ϵ is the empty history and $\mathcal{A} \times (\mathcal{O} \times R \times \mathcal{A})^0 \times (\mathcal{O} \times \mathcal{R}) := \mathcal{A} \times \mathcal{O} \times \mathcal{R}$. 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 $\mathcal{H} \times \mathcal{A} \times \mathcal{O} \times \mathcal{R} \to \{0,1\}$ where we write $\nu(o, r|h, a)$ for the function value. 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 $\sum_{o,r} \nu(o, r|h, a) = 1 \ \forall h, a$. The deterministic environments are then just a degenerate special case. Furthermore, we define $\nu(h_t|\pi) := \Pi_{i=1}^t \nu(o_i r_i | a_i, h_{i-1})$ where $a_i = \pi(h_{i-1})$. $\nu(\cdot|\pi)$ is a probability measure over strings, actually one measure for each string length with the corresponding power set as the σ-algebra. We define $\nu(\cdot|\pi, h_{t-1})$ by conditioning $\nu(\cdot|\pi)$ on h_{t-1} and we let $V_\nu^\pi(h_{t-1}) := \mathbb{E}_{\nu(\cdot|\pi,h_{t-1})} \sum_{i=t}^\infty \gamma^{i-t} r_i$ and $V_\nu^*(h_{t-1}) := \max_\pi V_\nu^\pi(h_{t-1})$.

Examples of agents: AIXI and Optimist. Suppose we are 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

$$\pi^* := \arg\max_\pi V_\xi^\pi(\epsilon) \tag{1}$$

which is its general form. Sometimes AIXI refers to the case of a certain universal class and a Solomonoff style prior (Hutter, 2005). The above agent, and only agents of that form, satisfies the strict rationality axioms presented first in Section 2 while the slightly looser version we present afterwards enables optimism. The optimist chooses its next action after history h based on

$$\pi^\circ := \arg\max_\pi \max_{\xi \in \Xi} V_\xi^\pi(h) \tag{2}$$

for a set of environments (beliefs) Ξ which we in the rest of the article will assume to be finite, though results can be extended further.

2. Rationality in Sequential Decision-Making

In this section, we first derive the above introduced AIXI agent from rationality axioms inspired by the traditional literature (Neumann and Morgenstern, 1944; Ramsey, 1931; Savage, 1954; deFinetti, 1937) on decision-making under uncertainty. Then we suggest weakening a symmetry condition between accepting and rejecting bets. The weaker condition only says that if an agent considers one side of a bet to be rejectable, it must be prepared to accept the other side but it can accept either. Since the conditions are meant for sequential decision and one does not accept several bets at a time, considering both sides of a bet to

be acceptable is not necessarily vulnerable to combinations of bets that would otherwise cause our agent a sure loss. Further, if an outcome is only revealed when a bet is accepted, one can only learn about the world by accepting bets. What is learned early on can lead to higher earnings later. The principle of optimism results in a more explorative agent and leads to multiple-prior models or the imprecise probability by Walley (2000). Axiomatics of multiple-prior models has been studied by Gilboa and Schmeidler (1989); Casadesus-Masanell et al. (2000). These models can be understood as quantifying the uncertainty in estimated probabilities by assigning a whole set of probabilities. In the passive prediction case, one typically combines the multiple-prior model with caution to achieve more risk averse decisions (Casadesus-Masanell et al., 2000). In the active case, agents need to take risk to generate experience that they can learn successful behavior from and, therefore, optimism is useful.

Bets. The basic setting we use is inspired by the betting approach of Ramsey (1931); deFinetti (1937). In this setting, the agent is about to observe a symbol from a finite alphabet and is offered a bet $x = (x_1, ..., x_n)$ where $x_i \in \mathbb{R}$ is the reward received for the outcome i.

Definition 1 (Bet) *Suppose that we have an unknown symbol from an alphabet with m elements, say $\{1, ..., m\}$. A bet (or contract) is a vector $x = (x_1, ..., x_m)$ in $\mathbb{R}^m$ where x_j is the reward received if the symbol is j.*

In our definition of decision maker we allow for choosing neither accept nor reject, while when we move on to axiomatize rational decision makers we will no longer allow for neither. In the case of a strictly rational decision maker it will only be the zero bet that can, and actually must, be both acceptable and rejectable. For the rational optimist the zero bet is always accepted and all bets are exactly one of acceptable or rejectable.

Definition 2 (Decision maker, Decision) *A decision maker (for bets regarding an unknown symbol) is a pair of sets $(Z, \tilde{Z}) \subset \mathbb{R}^m \times \mathbb{R}^m$ 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 stricter version of the axioms and a representation theorem.

Definition 3 (Strict rationality) *We say that $(Z, \tilde{Z})$ is strictly rational if it has the following properties:*

1. **Completeness**: $Z \cup \tilde{Z} = \mathbb{R}^m$
2. **Symmetry**: $x \in Z \iff -x \in \tilde{Z}$
3. **Convexity of accepting**: $x, y \in Z,\ \lambda, \gamma > 0 \Rightarrow \lambda x + \gamma y \in Z$
4. **Accepting sure profits**: $\forall k\ x_k > 0\ \Rightarrow x \in Z \setminus \tilde{Z}$

Axiom 1 in Definition 3 is really describing the setting rather than an assumption. It says that the agent must always choose at least one of accept or reject. Axiom 2 is a symmetry condition between accepting and rejecting that we will replace in the optimistic setting. In the optimistic setting we still demand that if the agent rejects x, then it must accept $-x$ but not the other way around. Axiom 3 is motivated as follows: If $x \in Z$ and $\lambda \geq 0$, then $\lambda x \in Z$ since it is simply a multiple of the same bet. Also, the sum of two acceptable bets should be acceptable. Axiom 4 says that if the agent is guaranteed to win money it must accept the bet and cannot reject it.

The following representation theorem says that a strictly rational decision maker can be represented as choosing bets to accept based on if they have positive expected utility for some probability vector. The same probabilities are consistently used for all decisions. Hence, the decision maker can be understood as a Bayesian agent with an a priori environment distribution. In Sunehag and Hutter (2011) we derived Bayes rule by showing how the concepts of marginal and conditional probabilities also come out of the same rational decision-making framework.

Theorem 4 (Existence of probabilities, Sunehag&Hutter 2011) *Given a rational decision maker, there are numbers $p_i \geq 0$ that satisfy*

$$\{x \mid \sum x_i p_i > 0\} \subseteq Z \subseteq \{x \mid \sum x_i p_i \geq 0\}. \tag{3}$$

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

Proof The third property tells us that Z and $-Z$ ($= \tilde{Z}$ according to the second property) are convex cones. The second and fourth property tells us that $Z \neq \mathbb{R}^m$. Suppose that there is a point x that lies in both the interior of Z and of $-Z$. Then, the same is true for $-x$ according to the second property and for the origin according to the third property. That a ball around the origin lies in Z means that $Z = \mathbb{R}^m$ which is not true. Thus the interiors of Z and $-Z$ are disjoint open convex sets and can, therefore, according to the Hahn-Banach Theorem be separated by a hyperplane which goes through the origin since according to the first and second property the origin is both acceptable and rejectable. The first two properties tell us that $Z \cup -Z = \mathbb{R}^m$. Given a separating hyperplane between the interiors of Z and $-Z$, Z must contain everything on one side. This means that Z is a half space whose boundary is a hyperplane that goes through the origin and the closure $\bar{Z}$ of Z is a closed half space and can be written as

$$\bar{Z} = \{x \mid \sum x_i p_i \geq 0\}$$

for some vector $p = (p_i)$ such that not every p_i is 0. The fourth property tells us that $p_i \geq 0 \ \forall i$. ■

In Appendix B we extend the above results to the countable case with Banach sequence spaces as the spaces of bets. Sunehag and Hutter (2011) showed how one can derive basic probability-theoretic concepts like marginalization and conditionalization from rationality.

Rational optimism. We now present four axioms for rational optimism. They state properties that the set of accepted and the set of rejected bets must satisfy.

Definition 5 (Rational optimism, Weak rationality) *We say that the decision maker $(Z, \tilde{Z}) \subset \mathbb{R}^m \times \mathbb{R}^m$ is a rational optimist or weakly rational if it satisfies the following:*

1. **Disjoint Completeness***: $x \notin \tilde{Z} \iff x \in Z$*
2. **Optimism***: $x \in \tilde{Z} \Rightarrow -x \notin \tilde{Z}$*
3. **Convexity of rejecting***: $x, y \in \tilde{Z}$ and $\lambda, \gamma > 0 \Rightarrow \lambda x + \gamma y \in \tilde{Z}$*
4. **Rejecting sure losses***: $\forall k \; x_k < 0 \quad \Rightarrow x \in \tilde{Z} \setminus Z$*

The first axiom is again a completeness axiom where we here demand that each contract is either accepted or rejected but not both. We introduce this stronger disjoint completeness assumption since the other axioms now concern the set of rejected bets, while we want to conclude something about what is accepted. The following three axioms concern rational rejection. The second says that if x is rejected then $-x$ must not be rejected. Hence, if the agent rejects one side of a bet it must, due to the first property, accept its negation. This was also argued for in the first set of axioms in the previous setting but in the optimistic set we do not have the opposite direction. In other words, if x is accepted then $-x$ can also be accepted. The agent is strictly rational about how it rejects bets. Rational rejection also means that if the agent rejects two bets x and y, it also rejects $\lambda x + \gamma y$ if $\lambda \geq 0$ and $\gamma \geq 0$. The final axiom says that if the reward is guaranteed to be strictly negative the bet must be rejected.

The representation theorem for rational optimism differs from that of strict rationality by not having a single unique environment distribution. Instead the agent has a set of such and if the bet has positive expected utility for any of them, the bet is accepted.

Theorem 6 (Existence of a set of probabilities) *Given a rational optimist, there is a set $\mathcal{P} \subset \mathbb{R}^m$ that satisfies*

$$\{x \mid \exists p \in \mathcal{P} : \sum x_i p_i > 0\} \subseteq Z \subseteq \{x \mid \exists p \in \mathcal{P} : \sum x_i p_i \geq 0\}. \tag{4}$$

One can always replace $\mathcal{P}$ with an extreme set the size of the alphabet. Also, one can demand that all the vectors in $\mathcal{P}$ be probability vectors, i.e., $\sum p_i = 1$ and $\forall i \; p_i \geq 0$.

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

Remark 7 (Pessimism) *If one wants an axiomatic system for rational pessimism, one can reverse the roles of Z and $\tilde{Z}$ in the definition of rational optimism and the theorem applies with a similar reversal: The conclusion could be rewritten by replacing $\exists$ with $\forall$ in the conclusion of Theorem 6.*

Making choices. To go from agents making decisions on accepting or rejecting bets to agents choosing between different bets $x^j,\ j = 1, 2, 3, ...$, we define preferences by saying that x is better than or equal to 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 the agent chooses the option with the highest expected utility. If we instead consider optimistic rationality, and if there is $p \in \mathcal{P}$ such that $\sum x_i p_i \geq \sum y_i q_i\ \forall q \in \mathcal{P}$ then $\sum p_i(x_i - y_i) \geq 0$ and, therefore, $x - y \in \bar{Z}$. Therefore, if the agent chooses the bet $x^j = (x_i^j)_i$ by

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

it is guaranteed that this bet is preferable to all other bets. We call this the optimistic decision or the rational optimistic decision. If the environment is reactive, i.e., if the probabilities for the outcome depends on the action, then p_i is above replaced by p_i^j. We discussed this in more detail in Sunehag and Hutter (2011).

Rational sequential decisions. For the general reinforcement learning setting we consider the choice of policy to use for the next T time steps. After one chooses a policy to use for those T steps the result is a history h_T and the value/return $\sum_{t=1}^{T} r_t \gamma^t$. There are finitely many possible h_T, each of them containing a specific return. If we enumerate all the possible h_T using i and the possible policies by j then for each policy and history there is a probability p_i^j for that history to be the result when policy j is used. Further we will denote the return achieved in history i by x_i. The bet x_i does depend on j since the rewards are part of the history.

By considering the choice to be for a policy π (previously j), an extension to finitely many sequential decisions is directly achieved. 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

$$\pi^* \in \arg\max_\pi V_\xi^\pi$$

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 chooses the next action based on a policy

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

for a finite set of environments Ξ ($\mathcal{P}$ before) and recalculates this at every time step.

3. Our Agent Framework

In this section, we introduce an agent framework that all agents we study in this paper can be fitted into by a choice of what we call a decision function and a hypothesis-generating function.

3.1 Decision Functions

The primary component of our agent framework is a decision function $f : \mathbb{M} \to \mathcal{A}$ where $\mathbb{M}$ is the class of all finite sets $\mathcal{M}$ of environments. The function value only depends on

the class of environments $\mathcal{M}$ that is the argument. The decision function is independent of the history, however, the class $\mathcal{M}$ fed to the decision function introduces an indirect dependence. For example, the environments at time $t+1$ can be the environments at time t, conditioned on the new observation. Therefore, we will in this section often write the value function without an argument: $V_{\nu_t}^{\tilde{\pi}} = V_{\nu_0}^{\pi}(h_t)$ if $\nu_t = \nu_0(\cdot|h_t)$ where the policy $\tilde{\pi}$ on the left hand side is the same as the policy π on the right, just after h_t have been seen. It starts at a later stage, meaning $\tilde{\pi}(h) = \pi(h_t h)$, where $h_t h$ is a concatenation.

Definition 8 (Rational decision function) *Given alphabets $\mathcal{A}$, $\mathcal{O}$ and $\mathcal{R}$ we say that a decision function $f : \mathbb{M} \to \mathcal{A}$ is a function $f(\mathcal{M}) = a$ that for any class of environments $\mathcal{M}$ based on those alphabets produces an action $a \in \mathcal{A}$. We say that f is strictly rational for the class $\mathcal{M}$ if there are $\omega_\nu \geq 0,\ \nu \in \mathcal{M}, \sum_{\nu\in\mathcal{M}} w_\nu = 1$ and there is a policy*

$$\pi \in \arg\max_{\tilde{\pi}} \sum_{\nu\in\mathcal{M}} \omega_\nu V_\nu^{\tilde{\pi}} \tag{5}$$

such that $a = \pi(\epsilon)$.

Agents as in Definition 8 are also called admissible if $w_\nu > 0\ \forall \nu \in \mathcal{M}$ since then they are Pareto optimal (Hutter, 2005). Being Pareto optimal means that if another agent (of this form or not) is strictly better (higher expected value) than a particular agent of this form in one environment, then it is strictly worse in another. A special case is when $|\mathcal{M}| = 1$ and (5) becomes

$$\pi \in \arg\max_{\tilde{\pi}} V_\nu^{\tilde{\pi}}$$

where ν is the environment in $\mathcal{M}$. The more general case connects to this by letting $\tilde{\nu}(\cdot) := \sum_{\nu\in\mathcal{M}} w_\nu \nu(\cdot)$ since then $V_{\tilde{\nu}}^{\pi} = \sum w_\nu V_\nu^{\pi}$ (Hutter, 2005). The next definition defines optimistic decision functions. They only coincide with strictly rational ones for the case $|\mathcal{M}| = 1$, however agents based on such decision functions satisfy the looser axioms that define a weaker form of rationality as presented in Section 2.

Definition 9 (Optimistic decision function) *We call a decision function f optimistic if $f(\mathcal{M}) = a$ implies that $a = \pi(\epsilon)$ for an optimistic policy π, i.e., for*

$$\pi \in \arg\max_{\tilde{\pi}} \max_{\nu\in\mathcal{M}} V_\nu^{\tilde{\pi}}. \tag{6}$$

3.2 Hypothesis-Generating Functions

Given a decision function, what remains to create a complete agent is a hypothesis-generating function $\mathcal{G}(h) = \mathcal{M}$ that for any history $h \in \mathcal{H}$ produces a set of environments $\mathcal{M}$. A special form of hypothesis-generating function is defined by combining the initial class $\mathcal{G}(\epsilon) = \mathcal{M}_0$ with an update function $\psi(\mathcal{M}_{t-1}, h_t) = \mathcal{M}_t$. An agent is defined from a hypothesis-generating function $\mathcal{G}$ and a decision function f by choosing action $a = f(\mathcal{G}(h))$ after seeing history h. We discuss a number of examples below to elucidate the framework and as a basis for the results we later present.

Example 10 (Bayesian agent) *Suppose that ν is a stochastic environment and $\mathcal{G}(h) = \{\nu(\cdot|h)\}$ for all h and let f be a strictly rational decision function. The agent formed by combining f and $\mathcal{G}$ is a rational agent in the stricter sense . Also, if $\mathcal{M}$ is a finite or countable class of environments and $\mathcal{G}(h) = \{\nu(\cdot|h) \mid \nu \in \mathcal{M}\}$ for all $h \in \mathcal{H}$ (same $\mathcal{M}$ for all h) and there are $\omega_\nu > 0,\ \nu \in \mathcal{M}, \sum_{\nu\in\mathcal{M}} w_\nu = 1$ such that $a = \pi(\epsilon)$ for a policy*

$$\pi \in \arg\max_{\tilde{\pi}} \sum_{\nu\in\mathcal{G}(h)} \omega_\nu V_\nu^{\tilde{\pi}}, \tag{7}$$

then we say that the agent is Bayesian and it can be represented more simply in the first way by $\mathcal{G}(h) = \{\sum w_\nu \nu(\cdot|h)\}$ due to linearity of the value function (Hutter, 2005)

Example 11 (Optimist deterministic case) *Suppose that $\mathcal{M}$ is a finite class of deterministic environments and let $\mathcal{G}(h) = \{\nu(\cdot|h) \mid \nu \in \mathcal{M}$ consistent with $h\}$. If we combine $\mathcal{G}$ with the optimistic decision function we have defined the optimistic agents for classes of deterministic environments (Algorithm 1) from Section 4. In Section 7 we extend these agents to infinite classes by letting $\mathcal{G}(h_t)$ contain new environments that were not in $\mathcal{G}(h_{t-1})$.*

Example 12 (Optimistic AIXI) *Suppose that $\mathcal{M}$ is a finite class of stochastic environments and that $\mathcal{G}(h) = \{\nu(\cdot|h) \mid \nu \in \mathcal{M}\}$. If we combine $\mathcal{G}$ with the optimistic decision function we have defined the optimistic AIXI agent (Equation 2 with $\Xi = \mathcal{M}$).*

Example 13 (MBIE) *The Model Based Interval Estimation (MBIE) (Strehl et al., 2009) method for Markov Decision Processes (MDPs) defines $\mathcal{G}(h)$ as a set of MDPs (for a given state space) with transition probabilities in confidence intervals calculated from h. This is combined with the optimistic decision function. MBIE satisfies strong sample complexity guarantees for MDPs and is, therefore, an example of what we want but in a narrower setting.*

Example 14 (Optimist stochastic case) *Suppose that $\mathcal{M}$ is a finite class of stochastic environments and that $\mathcal{G}(h) = \{\nu(\cdot|h) \mid \nu \in \mathcal{M} : \nu(h) \geq z \max_{\tilde{\nu}\in\mathcal{M}} \tilde{\nu}(h)\}$ for some $z \in (0, 1)$. If we combine $\mathcal{G}$ with the optimistic decision function we have defined the optimistic agent with stochastic environments from Section 5.*

Example 15 (MERL and BayesExp) *Agents that switch explicitly between exploration and exploitation are typically* **not** *satisfying even our weak rationality demand. An example is Lattimore et al. (2013a) where the introduced Maximum Exploration Reinforcement Learning (MERL) agent performs certain tests when the remaining candidate environments are disagreeing sufficiently. This decision function is not satisfying rationality while our Algorithm 3, which uses the exclusion criteria of MERL but with an optimistic decision function, does satisfy our notion of rationality. Another example of an explicitly exploring irrational agent is BayesExp (Lattimore, 2014).*

4. Finite Classes of Dominant A Priori Environments

In this section, we study convergence results for optimistic agents with finite classes of dominant environments. In terms of the agent framework we here use an optimistic decision

function and a hypothesis-generating function that neither adds to nor removes from the initial class but just updates the environments through conditioning. Such agents were previously described in Example 12. In the next section we consider a setting where we instead of domination assume that one of the environments in the class is the true environment. The first setting is natural for Bayesian approaches, while the second is more frequentist in flavor. If we assume that all uncertainty is epistemic, i.e., caused by the agent's lack of knowledge, and that the true environment is deterministic, then for the first (Bayesian) setting the assumption means that the environments assign strictly positive probability to the truth. In the second (frequentist) setting, the assumption says that the environment class must contain this deterministic environment. In Section 6, we also consider a stochastic version of the second setting where the true environment is potentially stochastic in itself.

We first prove that AIXI is asymptotically optimal if the a priori environment ξ both dominates the true environment μ in the sense that $\exists c > 0 : \xi(\cdot) \geq c\mu(\cdot)$ and optimistic in the sense that $\forall h_t \; V^*_\xi(h_t) \geq V^*_\mu(h_t)$ (for large t). We extend this by replacing ξ with a finite 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 $V^*_\xi(h_t) \geq V^*_\mu(h_t)$. We refer to this second domination property as optimism. The first domination property, which we simply refer to as domination, 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 provide a simple illustrative example for the first theorem and a more interesting one after the second theorem. First, we introduce some definitions related to the purpose of domination, namely it implies absolute continuity which according to the Blackwell-Dubins Theorem (Blackwell and Dubins, 1962) implies merging in total variation.

Definition 16 (Total variation distance, Merging, Absolute continuity)
i) The total variation distance between two (non-negative) measures P and Q is defined to be

$$d(P, Q) \; = \; \sup_A |P(A) - Q(A)|$$

where A ranges over the σ-algebra of the relevant measure space.
ii) P and Q are said to merge iff $d(P(\cdot|\omega_{1:t}), Q(\cdot|\omega_{1:t})) \to 0$ P-a.s. as $t \to \infty$, i.e., almost surely if the sequence ω is generated by P. The environments ν_1 and ν_2 merge under π if $\nu_1(\cdot|h_t, \pi)$ and $\nu_2(\cdot|h_t, \pi)$ merge.
iii) P is absolutely continuous with respect to Q if $Q(A) = 0$ implies that $P(A) = 0$.

We will make ample use of the classical Blackwell-Dubins Theorem (Blackwell and Dubins, 1962) so we state it explicitly.

Theorem 17 (Blackwell-Dubins Theorem) *If P is absolutely continuous with respect to Q, then P and Q merge P-almost surely.*

Lemma 18 (Value convergence for merging environments) *Given a policy π and environments μ and ν it follows that for all h*

$$|V^\pi_\mu(h) - V^\pi_\nu(h)| \; \leq \; \frac{1}{1-\gamma} d(\mu(\cdot|h, \pi), \nu(\cdot|h, \pi)).$$

Proof The lemma follows from the general inequality

$$\left|\mathbb{E}_P(f) - \mathbb{E}_Q(f)\right| \;\leq\; \sup|f| \cdot \sup_A \left|P(A) - Q(A)\right|$$

by letting f be the return in the history and $P = \mu(\cdot|h,\pi)$ and $Q = \nu(\cdot|h,\pi)$, and using $0 \leq f \leq 1/(1-\gamma)$ that follows from the rewards being in $[0,1]$. ■

The next theorem is the first of the two convergence theorems in this section. It relates to a strictly rational agent and imposes two conditions. The domination condition is a standard assumption that a Bayesian agent satisfies if it has strictly positive prior weight for the truth. The other assumption, the optimism assumption, is restrictive but the convergence result does not hold if only domination is assumed and the known alternative (Hutter, 2005) of demanding that a Bayesian agent's hypothesis class is self-optimizing is only satisfied for environments of very particular form such as ergodic Markov Decision Processes.

Algorithm 1: Optimistic-AIXI Agent (π°)

Require: Finite class of dominant a priori environments Ξ

$t = 1, h_0 = \epsilon$
repeat
  $(\pi^*, \xi^*) \in \arg\max_{\pi\in\Pi, \xi\in\Xi} V_\xi^\pi(h_{t-1})$
  $a_{t-1} = \pi^*(h_{t-1})$
  Perceive $o_t r_t$ from environment μ
  $h_t \leftarrow h_{t-1} a_{t-1} o_t r_t$
  $t \leftarrow t+1$
until end of time

Theorem 19 (AIXI convergence) *Suppose that $\xi(\cdot) \geq c\mu(\cdot)$ for some $c > 0$ and μ is the true environment. Also suppose that there μ-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 in μ according to the AIXI agent based on ξ, i.e.,*

$$\pi^* \in \arg\max_\pi V_\xi^\pi(\epsilon)$$

or equivalently Algorithm 1 with $\Xi = \{\xi\}$. Then there is μ-almost surely, i.e., almost surely if the sequence h_t is generated by π^* acting in μ, 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 (Blackwell and Dubins, 1962)) say that μ-almost surely there is for every $\varepsilon' > 0$, a $T < \infty$ such that $\forall t \geq T$ $d(\xi(\cdot|h_t,\pi^*), \mu(\cdot|h_t,\pi^*)) < \varepsilon'$ 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 $t \geq T$, $V_\mu^{\pi^*}(h_t) \geq V_\xi^*(h_t) - \varepsilon \geq V_\mu^*(h_t) - \varepsilon$. ■

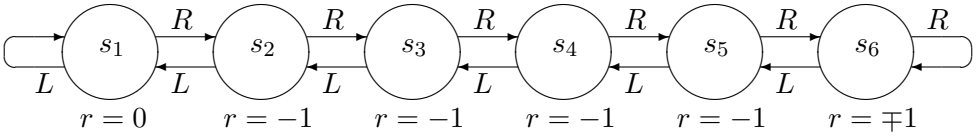

Figure 1: Line environment

Example 20 (Line environment) *We consider an agent who, when given a class of environments, will choose its prior based on simplicity in accordance with Occam's razor (Hutter, 2005). First let us look at a class $\mathcal{M}$ of two environments which both have six states (Figure 1) $s_1, ..., s_6$ 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 has a reward of 0, s_2, s_3, s_4, s_5 have reward -1 while 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 the agent 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 that is optimistic for ν_1 and ν_2 and we can guarantee optimality for any γ.*

In the next theorem we prove that for the optimistic agent with a class of a priori environments, only one of them needs to be optimistic at a time while all are assumed to be dominant. As before, domination is achieved if the a priori environments are of the form of a mixture over a hypothesis class containing the truth. The optimism is in this case milder and is e.g., trivially satisfied if the truth is one of the a priori environments. Since the optimistic agent is guaranteed convergence under milder assumptions we believe that it would succeed in a broader range of environments than the single-prior rational agent.

Theorem 21 (Multiple-prior convergence) *Suppose that Ξ is a finite set 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 μ-almost surely is $T_1 < \infty$ such that for $t \geq T_1$ there is $\xi_t \in \Xi$ such that $V^*_{\xi_t}(h_t) \geq V^*_\mu(h_t)$. Suppose that the policy π°, defined as in (2) or equivalently Algorithm 1, acts according to the rational optimistic agent based on Ξ in μ. Then there is μ-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 to formulate by formulating it for time $t = 0$ (when the history is the empty string ϵ), though when proving Theorem 21 it is used for a later time point when the environments in the class have merged sufficiently under π° in the sense of total variation diameter. The lemma simply says that if the environments are sufficiently close under π°, then π° must be nearly optimal. This follows from optimism since it means that the value function that π° maximizes is the highest among the value functions for the environments in the class and it is also close to the actual value by the

assumption. The only thing that makes the proof non-trivial is that π° might maximize for different environments at each step but since they are all close, the conclusion that would otherwise have been trivial is still true. One can simply construct a new environment that combines the dynamics of the environments that are optimistic at different times. Then, the policy maximizes value for this environment at each times step and this new environment is also close to all the environments in the class. We let ν_h^* be an environment in $\arg\max_\nu \max_\pi V_\nu^\pi(h)$ that π° uses to choose the next action after experiencing h.

Definition 22 (Environment used by π°) *Suppose that Ξ is a finite set of environments and that π° is the optimistic agent. Let ν_h^* be an environment in $\arg\max_\nu \max_\pi V_\nu^\pi(h)$ that π° uses to choose the next action after experiencing h, i.e., ν_h^* is such that $V_{\nu_h^*}^*(h) = \max_{\nu,\pi} V_\nu^\pi(h)$ and $\pi^\circ(h) = \tilde{\pi}(h)$ for some $\tilde{\pi} \in \arg\max_\pi V_{\nu_h^*}^\pi(h)$. Note, the choice might not be unique.*

The next definition introduces the concept of constructing an environment that is consistently used.

Definition 23 (Constructed environment) *Define $\hat{\nu}$ by $\hat{\nu}(o,r|h,a) = \nu_h^*(o,r|h,a)$.*

The following lemma is intuitively obvious. It says that if at each time step we define an environment by using the dynamics of the environment in the class that promises the most value, then the resulting environment will always be optimistic relative to any environment in the class. The proof is only complicated by the cumbersome notation required due to studying fully general reinforcement learning. The key tool is the Bellman equation that for general reinforcement learning is

$$V_\nu^\pi(h) = \sum_{o,r} \nu(o,r|h,\pi(h))[r + \gamma V_\nu^\pi(h')]$$

where $h' = h\pi(h)or$. Together with induction this will be used to prove the next lemma.

Lemma 24 $V_{\hat{\nu}}^{\pi^\circ} \geq \max_{\nu\in\mathcal{M},\pi} V_\nu^\pi(\epsilon)$

Proof Let V_ν^π denote $V_\nu^\pi(\epsilon)$. We reason by induction using a sequence of environments approaching $\hat{\nu}$. Let

$$\hat{\nu}_s(o_t r_t|h_{t-1},a) = \hat{\nu}(o_t r_t|h_{t-1},a)\ \forall h_{t-1}\forall a,\ t \leq s$$

and

$$\hat{\nu}_s(o_t r_t|h_{t-1},a) = \nu_{h_s}^*(o_t r_t|h_{t-1},a),\ \forall h_{t-1}\forall a, t > s.$$

$\hat{\nu}_1$ equals ν_ε^* at all time points and thus $V_{\hat{\nu}_1}^\pi = V_{\nu_\varepsilon^*}^\pi$. Let $\hat{R}_t^\nu$ be the expected accumulated (discounted) reward ($\mathbb{E}\sum_{i=1}^t \gamma^{i-1} r_i$) up to time t when following π° up until that time in the environment ν. We first do the base case $t = 1$.

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

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

The middle inequality is due to $\max_\pi V^\pi_{\nu^*_{h_1}}(h_1) \geq \max_\pi V^\pi_\nu(h_1)\ \forall \nu \in \Xi$. The first equality is the Bellman equation together with the fact that π° makes a first action that optimize for ν^*_ϵ. The second is due to $\hat{\nu}_1 = \nu^*_\epsilon$ and the Bellman equation. In the same way,

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

and it follows by induction that $V^{\pi^\circ}_{\hat{\nu}} \geq \max_{\pi,\nu\in\mathcal{M}} V^\pi_\nu \geq V^*_\mu$.

■

Lemma 25 (Optimism is nearly optimal) *Suppose that the assumptions of Theorem 21 hold and that we denote the optimistic agent again by (π°). Then for each $\varepsilon > 0$ there exists $\tilde{\varepsilon} > 0$ such that $V^{\pi^\circ}_\mu(\epsilon) \geq \max_\pi V^\pi_\mu(\epsilon) - \varepsilon$ whenever*

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

Proof We will show that if we choose $\tilde{\varepsilon}$ small enough, then

$$|V^{\pi^\circ}_{\hat{\nu}} - V^{\pi^\circ}_\mu| < \varepsilon \tag{8}$$

where μ is the true environment. Equation (8), when proven to hold when $\tilde{\varepsilon}$ is chosen small enough, concludes the proof since then $|V^*_\mu - V^{\pi^\circ}_\mu| < \varepsilon$, due to $V^{\pi^\circ}_{\hat{\nu}} \geq V^*_\mu \geq V^{\pi^\circ}_\mu$. This is easy since

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

and if $\tilde{\varepsilon} + \frac{\tilde{\varepsilon}}{1-\gamma} \leq \varepsilon$ then (8) holds and the proof is complete as we concluded above since $|V^{\pi^\circ}_{\hat{\nu}_\varepsilon} - V^{\pi^\circ}_\mu| < \tilde{\varepsilon}$. ■

Proof of Theorem 21. Since Ξ is finite and by using Theorem 17 (Blackwell-Dubins), there is for every ε', a $T < \infty$ when $\forall \xi \in \Xi\ \forall t \geq T,\ d(\xi(\cdot|h_t,\pi^\circ), \mu(\cdot|h_t,\pi^\circ)) < \varepsilon'$. This implies that $\forall \xi \in \Xi\ |V^{\pi^\circ}_\xi(h_t) - V^{\pi^\circ}_\mu(h_t)| < \frac{\varepsilon'}{1-\gamma}$ by Lemma 18. Choose ε' such that $\frac{\varepsilon'}{1-\gamma} = \varepsilon$. Applying Lemma 25 with class $\tilde{\Xi} = \{\xi(\cdot|\ h_T) : \xi \in \Xi\}$ now directly proves the result. The application of Lemma 25 is viewing time T from this proof as time zero and the ϵ context. ■

Example 26 (Multiple-prior AIXI) *For any Universal Turing Machine (UTM) U the corresponding Solomonoff distribution ξ_U is defined by putting coin flips on the input tape (see Li and Vitani (2008); Hutter (2005) for details). ξ_U is dominant for any lower semi-computable semi-measure over infinite sequences. Hutter (2005) extends these constructions and introduces an environment ξ_U that is dominant for all reactive lower semi-computable reactive environments and defines the AIXI agent based on it as in Theorem 19. A difficulty*

is to choose the UTM to use. Many have without success tried to find a single "natural" Turing machine and it might in fact be impossible (Müller, 2010). Examples includes defining a machine from a programming language like C or Haskell and another possibility is to use Lambda calculus. 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, and the issue might not be fully resolvable, it alleviates it.

5. Finite Classes of Deterministic (Non-Dominant) A Priori Environments

In this section, we perform a different sort of analysis where it is not assumed that all the environments in Ξ dominate the true environment μ. We instead rely on the assumption that the true environment is a member of the agent's class of environments. The a priori environments are then naturally thought of as a hypothesis class rather than mixtures over some hypothesis class and we will write $\mathcal{M}$ instead of Ξ to mark this difference. We begin with the deterministic case, where one could not have introduced the domination assumption, in this section and look at stochastic non-dominant a priori environments in the next. The agent in this section can be described, as was done in Example 11 as having an optimistic decision function and a hypothesis-generating function that begins with an initial class and removes excluded environments.

5.1 Optimistic Agents for Deterministic Environments

Given a finite class of deterministic environments $\mathcal{M} = \{\nu_1, ..., \nu_m\}$, we define an algorithm that for any unknown environment from $\mathcal{M}$ eventually achieves optimal behavior in the sense that there exists T such that maximum reward is achieved from time T onwards. The algorithm chooses an optimistic hypothesis from $\mathcal{M}$ in the sense that it picks the environment in which one can achieve the highest reward and then the policy that is optimal for this environment is followed. If this hypothesis is contradicted by the feedback from the environment, a new optimistic hypothesis is picked from the environments that are still consistent with h. This technique has the important consequence that if the hypothesis is not contradicted, the agent acts optimally even when optimizing for an incorrect hypothesis.

Let $h_t^{\pi,\nu}$ be the history up to time t generated by policy π in environment ν. In particular let $h^\circ := h^{\pi^\circ,\mu}$ be the history generated by Algorithm 2 (policy π°) interacting with the actual "true" environment μ. At the end of cycle t we know $h_t^\circ = h_t$. An environment ν is called consistent with h_t if $h_t^{\pi^\circ,\nu} = h_t$. Let $\mathcal{M}_t$ be the environments consistent with h_t. The algorithm only needs to check whether $o_t^{\pi^\circ,\nu} = o_t$ and $r_t^{\pi^\circ,\nu} = r_t$ for each $\nu \in \mathcal{M}_{t-1}$, since previous cycles ensure $h_{t-1}^{\pi^\circ,\nu} = h_{t-1}$ and trivially $a_t^{\pi^\circ,\nu} = a_t$. The maximization in Algorithm 2 that defines optimism at time t is performed over $\nu \in \mathcal{M}_{t-1}$, the set of consistent hypotheses at time t, and $\pi \in \Pi = \Pi^{all}$ is the class of *all* deterministic policies. In Example 11, we described the same agent by saying that it combines an optimistic decision function with a hypothesis generating function that begins with an initial finite class of deterministic environments and excludes those that are contradicted. More precisely, we have here first narrowed down the optimistic decision function further by saying that it needs to stick to hypothesis until contradicted, while we will below further discuss not

Algorithm 2: Optimistic Agent (π°) for Deterministic Environments

Require: Finite class of deterministic environments $\mathcal{M}_0 \equiv \mathcal{M}$
$t = 1$
repeat
  $(\pi^*, \nu^*) \in \arg\max_{\pi\in\Pi,\nu\in\mathcal{M}_{t-1}} V_\nu^\pi(h_{t-1})$
  repeat
    $a_{t-1} = \pi^*(h_{t-1})$
    Perceive $o_t r_t$ from environment μ
    $h_t \leftarrow h_{t-1}a_{t-1}o_t r_t$
    Remove all inconsistent ν from $\mathcal{M}_t$ ($\mathcal{M}_t := \{\nu \in \mathcal{M}_{t-1} : h_t^{\pi^\circ,\nu} = h_t\}$)
    $t \leftarrow t + 1$
  until $\nu^* \notin \mathcal{M}_{t-1}$
until $\mathcal{M}$ is empty

making this simplifying extra specification. Its an important fact, proven below, that an optimistic hypothesis does not cease to be optimistic until contradicted. The guarantees we prove for this agent are stronger than in the previous chapter where only dominance was assumed while here we assume that the truth belongs to the given finite class of deterministic environments.

Theorem 27 (Optimality, Finite deterministic class) *Suppose $\mathcal{M}$ is a finite class of deterministic environments. If we use Algorithm 2 (π°) in an environment $\mu \in \mathcal{M}$, then there is $T < \infty$ such that*

$$V_\mu^{\pi^\circ}(h_t) \;=\; \max_\pi V_\mu^\pi(h_t) \;\forall t \geq T.$$

A key to proving Theorem 27 is time-consistency (Lattimore and Hutter, 2011b) of geometric discounting. The following lemma tells us that if the agent acts optimally with respect to a chosen optimistic hypothesis, this hypothesis remains optimistic until contradicted.

Lemma 28 (Time-consistency) *Suppose $(\pi^*, \nu^*) \in \arg\max_{\pi\in\Pi,\nu\in\mathcal{M}_{t-1}} V_\nu^\pi(h_{t-1})$ and that an agent acts according to π^* from a time point t to another time point $\tilde{t} - 1$, i.e., $a_s = \pi^*(h_{s-1})$ for $t \leq s \leq \tilde{t} - 1$. For any choice of $t < \tilde{t}$ such that ν^* is still consistent at time $\tilde{t}$, it holds that $(\pi^*, \nu^*) \in \arg\max_{\pi\in\Pi,\nu\in\mathcal{M}_{\tilde{t}}} V_\nu^\pi(h_{\tilde{t}})$.*

Proof Suppose that $V_{\nu^*}^{\pi^*}(h_{\tilde{t}}) < V_{\tilde{\nu}}^{\tilde{\pi}}(h_{\tilde{t}})$ for some $\tilde{\pi}$, $\tilde{\nu}$. It holds that $V_{\nu^*}^{\pi^*}(h_t) = C + \gamma^{\tilde{t}-t} V_{\nu^*}^{\pi^*}(h_{\tilde{t}})$ where C is the accumulated reward between t and $\tilde{t} - 1$. Let $\hat{\pi}$ be a policy that equals π^* from t to $\tilde{t} - 1$ and then equals $\tilde{\pi}$. It follows that $V_{\tilde{\nu}}^{\hat{\pi}}(h_t) = C + \gamma^{\tilde{t}-t} V_{\tilde{\nu}}^{\hat{\pi}}(h_{\tilde{t}}) > C + \gamma^{\tilde{t}-t} V_{\nu^*}^{\pi^*}(h_{\tilde{t}}) = V_{\nu^*}^{\pi^*}(h_t)$ which contradicts the assumption $(\pi^*, \nu^*) \in \arg\max_{\pi\in\Pi,\nu\in\mathcal{M}_t} V_\nu^\pi(h_t)$. Therefore, $V_{\nu^*}^{\pi^*}(h_{\tilde{t}}) \geq V_{\tilde{\nu}}^{\tilde{\pi}}(h_{\tilde{t}})$ for all $\tilde{\pi}$, $\tilde{\nu}$. ■

Proof (Theorem 27) At time t we know h_t. If some $\nu \in \mathcal{M}_{t-1}$ is inconsistent with h_t, i.e., $h_t^{\pi^\circ,\nu} \neq h_t$, it gets removed, i.e., is not in $\mathcal{M}_{t'}$ for all $t' \geq t$.

Since $\mathcal{M}_0 = \mathcal{M}$ is finite, such inconsistencies can only happen finitely often, i.e., from some T onwards we have $\mathcal{M}_t = \mathcal{M}_\infty$ for all $t \geq T$. Since $h_t^{\pi^\circ,\mu} = h_t \ \forall t$, we know that $\mu \in \mathcal{M}_t \ \forall t$.

Assume $t \geq T$ henceforth. The optimistic hypothesis will not change after this point. If the optimistic hypothesis is the true environment μ, the agent has obviously chosen a truly optimal policy.

In general, the optimistic hypothesis ν^* is such that it will never be contradicted while actions are taken according to π°, hence (π^*, ν^*) do not change anymore. This implies

$$V_\mu^{\pi^\circ}(h_t) \;=\; V_\mu^{\pi^*}(h_t) \;=\; V_{\nu^*}^{\pi^*}(h_t) \;=\; \max_{\nu\in\mathcal{M}_t}\max_{\pi\in\Pi} V_\nu^\pi(h_t) \;\geq\; \max_{\pi\in\Pi} V_\mu^\pi(h_t)$$

for all $t \geq T$. The first equality follows from π° equals π^* from $t \geq T$ onwards. The second equality follows from consistency of ν^* with $h_{1:\infty}^\circ$. The third equality follows from optimism, the constancy of π^*, ν^*, and $\mathcal{M}_t$ for $t \geq T$, and time-consistency of geometric discounting (Lemma 28). The last inequality follows from $\mu \in \mathcal{M}_t$. The reverse inequality $V_\mu^{\pi^*}(h_t) \leq \max_\pi V_\mu^\pi(h_t)$ follows from $\pi^* \in \Pi$. Therefore π° is acting optimally at all times $t \geq T$. ■

Besides the eventual optimality guarantee above, we also provide a bound on the number of time steps for which the value of following Algorithm 2 is more than a certain $\varepsilon > 0$ less than optimal. The reason this bound is true is that we only have such suboptimality for a certain number of time steps immediately before the current hypothesis becomes inconsistent and the number of such inconsistency points are bounded by the number of environments. Note that the bound tends to infinity as $\varepsilon \to 0$, hence we need Theorem 27 with its distinct proof technique for the $\varepsilon = 0$ case.

Theorem 29 (Finite error bound) *Following π° (Algorithm 2),*

$$V_\mu^{\pi^\circ}(h_t) \geq \max_{\pi\in\Pi} V_\mu^\pi(h_t) - \varepsilon,\ 0 < \varepsilon < 1/(1-\gamma)$$

for all but at most $K\frac{-\log\varepsilon(1-\gamma)}{1-\gamma} \leq |\mathcal{M}-1|\frac{-\log\varepsilon(1-\gamma)}{1-\gamma}$ time steps t where K is the number of times that some environment is contradicted.

Proof Consider the ℓ-truncated value

$$V_{\nu,\ell}^\pi(h_t) \;:=\; \sum_{i=t+1}^{t+\ell} \gamma^{i-t-1} r_i$$

where the sequence r_i are the rewards achieved by following π from time $t+1$ to $t+\ell$ in ν after seeing h_t. By letting $\ell = \frac{\log\varepsilon(1-\gamma)}{\log\gamma}$ (which is positive due to negativity of both numerator and denominator) we achieve $|V_{\nu,\ell}^\pi(h_t) - V_\nu^\pi(h_t)| \leq \frac{\gamma^l}{1-\gamma} = \epsilon$. Let (π_t^*, ν_t^*) be the policy-environment pair selected by Algorithm 2 in cycle t.

Let us first assume $h_{t+1:t+\ell}^{\pi^\circ,\mu} = h_{t+1:t+\ell}^{\pi^\circ,\nu_t^*}$, i.e., ν_t^* is consistent with $h_{t+1:t+\ell}^\circ$, and hence π_t^* and ν_t^* do not change from $t+1, ..., t+\ell$ (inner loop of Algorithm 2). Then

$$V_\mu^{\pi^\circ}(h_t) \overset{\text{drop terms,}}{\geq} V_{\mu,\ell}^{\pi^\circ}(h_t) \overset{\text{same } h_{t+1:t+\ell},}{=} V_{\nu_t^*,\ell}^{\pi^\circ}(h_t) \overset{\pi^\circ=\pi_t^* \text{ on } h_{t+1:t+\ell},}{=} V_{\nu_t^*,\ell}^{\pi_t^*}(h_t)$$

$$\underset{\text{bound extra terms}}{\overset{\uparrow}{\geq}} V_{\nu_t^*}^{\pi_t^*}(h_t) - \tfrac{\gamma^\ell}{1-\gamma} \quad \underset{\text{def. of } (\pi_t^*, \nu_t^*) \text{ and } \varepsilon := \frac{\gamma^\ell}{1-\gamma}}{\overset{\uparrow}{=}} \max_{\nu \in \mathcal{M}_t} \max_{\pi \in \Pi} V_\nu^\pi(h_t) - \varepsilon \quad \underset{\mu \in \mathcal{M}_t}{\overset{\uparrow}{\geq}} \max_{\pi \in \Pi} V_\mu^\pi(h_t) - \varepsilon.$$

Now let $t_1, ..., t_K$ be the times t at which the currently selected ν_t^* becomes inconsistent with h_t, i.e., $\{t_1, ..., t_K\} = \{t : \nu_t^* \notin \mathcal{M}_t\}$.

Therefore $h^\circ_{t+1:t+\ell} \neq h^{\pi^\circ, \nu_t^*}_{t+1:t+\ell}$ (only) at times $t \in \mathcal{T}_\times := \bigcup_{i=1}^K \{t_i - \ell, ..., t_i - 1\}$, which implies $V_\mu^{\pi^\circ}(h_t) \geq \max_{\pi \in \Pi} V_\mu^\pi(h_t) - \varepsilon$ except possibly for $t \in \mathcal{T}_\times$. Finally

$$|\mathcal{T}_\times| \; = \; \ell \cdot K \; = \; \frac{\log \varepsilon(1-\gamma)}{\log \gamma} K \; \leq \; K \frac{\log \varepsilon(1-\gamma)}{\gamma - 1} \; \leq \; |\mathcal{M} - 1| \frac{\log \varepsilon(1-\gamma)}{\gamma - 1}$$

■

Conservative or liberal optimistic agents. We refer to the algorithm above as the conservative agent since it keeps its hypothesis for as long as it can. We can define a more *liberal* agent that re-evaluates its optimistic hypothesis at every time step and can switch between different optimistic policies at any time. Algorithm 2 is actually a special case of this as shown by Lemma 28. The liberal agent is really a class of algorithms and this larger class of algorithms consists of exactly the algorithms that are optimistic at every time step without further restrictions. The conservative agent is the subclass of algorithms that only switch hypothesis when the previous is contradicted. The results for the conservative agent can be extended to the liberal one. We do this for Theorem 27 in Appendix A together with analyzing further subtleties about the conservative case. It is worth noting that the liberal agent can also be understood as a conservative agent but for an extended class of environments where one creates a new environment by letting it have, at each time step, the dynamics of the chosen optimistic environment. Contradiction of such an environment will then always coincide with contradiction of the chosen optimistic environment and there will be no extra contradictions due to these new environments. Hence, the finite-error bound can also be extended to the liberal case. In the stochastic case below, we have to use a liberal agent. Note that both the conservative and liberal agents are based on an optimistic decision function and the same hypothesis-generating function. There can be several optimistic decision functions due to ties.

5.2 Environments and Laws

The bounds given above have a linear dependence on the number of environments in the class and though this is the best one can do in general (Lattimore et al., 2013a), it is bad compared to what we are used to from Markov Decision Processes (Lattimore and Hutter, 2012) where the linear (up to logarithms) dependence is on the size of the state space instead. Markov Decision Processes are finitely generated in a sense that makes it possible to exclude whole parts of the environment class together, e.g., all environments for which a state s_2 is likely to follow the state s_1 if action a_1 is taken. Unfortunately, the Markov assumption is very restrictive.

In this section we will improve the bounds above by introducing the concept of laws and of an environment being generated by a set of laws. Any environment class can be described this way and the linear dependence on the size of the environment class in the bounds is

replaced by a linear dependence on the size of the smallest set of laws that can generate the class. Since any class is trivially generated by the laws that simply equal an environment from the class each, we are not making further restrictions compared to previous results. However, in the worst situations the bounds presented here equal the previous bounds, while for other environment classes the bounds in this section are exponentially better. The latter classes with good bounds are the only option for practical generic agents. Classes of such form have the property that one can exclude laws and thereby exclude whole classes of environments simultaneously like when one learns about a state transition for an MDP.

Environments defined by laws. We consider observations of the form of a feature vector $o = \vec{x} = (x_j)_{j=1}^m \in \mathcal{O} = \times_{j=1}^m \mathcal{O}_j$ including the reward as one coefficient where x_j is an element of some finite alphabet $\mathcal{O}_i$. Let $\mathcal{O}_\perp = \times_{j=1}^m (\mathcal{O}_j \cup \{\perp\})$, i.e., $\mathcal{O}_\perp$ consists of the feature vectors from $\mathcal{O}$ but where some elements are replaced by a special letter $\perp$. The meaning of $\perp$ is that there is no prediction for this feature. We first consider deterministic laws.

Definition 30 (Deterministic laws) *A law is a function $\tau : \mathcal{H} \times \mathcal{A} \to \mathcal{O}_\perp$.*

Using a feature vector representation of the observations and saying that a law predicts some of the features is a convenient special case of saying that the law predicts that the next observation will belong to a certain subset of the observation space. Each law τ predicts, given the history and a new action, some or none but not necessarily all of the features x_j at the next time point. We first consider sets of laws such that for any given history and action, and for every feature, there is at least one law that makes a prediction of this feature. Such sets are said to be complete.

Definition 31 (Complete set of laws) *A set of laws $\tilde{\mathcal{T}}$ is complete if*

$$\forall h, a \forall j \in \{1, ..., m\} \; \exists \tau \in \tilde{\mathcal{T}} : \tau(h, a)_j \neq \perp.$$

We will only consider combining deterministic laws that never contradict each other and we call such sets of laws coherent. The main reason for this restriction is that one can then exclude a law when it is contradicted. If one does not demand coherence, an environment might only sometimes be consistent with a certain law and the agent can then only exclude the contradicted environment, not the contradicted law which is key to achieving better bounds.

Definition 32 (Coherent set of laws) *We say that $\tilde{\mathcal{T}}$ is coherent if for all $\tau \in \tilde{\mathcal{T}}, h, a$ and j*

$$\tau(h, a)_j \neq \perp \; \Rightarrow \; \tilde{\tau}(h, a)_j \in \{\perp, \tau(h, a)_j\} \; \forall \tilde{\tau} \in \tilde{\mathcal{T}}.$$

Definition 33 (Environment from a complete and coherent set of laws) *Given a complete and coherent set of laws $\tilde{\mathcal{T}}$, $\nu(\tilde{\mathcal{T}})$ is the unique environment ν which is such that*

$$\forall h, a \forall j \in \{1, ..., m\} \exists \tau \in \tilde{\mathcal{T}} : \nu(h, a)_j = \tau(h, a)_j.$$

The existence of $\nu(\tilde{\mathcal{T}})$ follows from completeness of $\tilde{\mathcal{T}}$ and uniqueness is due to coherence.

Definition 34 (Environment class from deterministic laws) *Given a set of laws $\mathcal{T}$, let $\mathcal{C}(\mathcal{T})$ denote the complete and coherent subsets of $\mathcal{T}$. Given a set of laws $\mathcal{T}$, we define the class of environments generated by $\mathcal{T}$ through*

$$\mathcal{M}(\mathcal{T}) := \{\nu(\tilde{\mathcal{T}}) \; | \tilde{\mathcal{T}} \in \mathcal{C}(\mathcal{T})\}.$$

Example 35 (Deterministic laws for fixed vector) *Consider an environment with a constant binary feature vector of length m. There are 2^m such environments. Every such environment can be defined by combining m out of a class of $2m$ laws. Each law says what the value of one of the features is, one law for 0 and one for 1. In this example, a coherent set of laws is simply one feature for each coefficient. The generated environment is the constant vector defined by that vector and the set of all the generated environments is the full set of 2^m environments.*

Error analysis. Every contradiction of an environment is a contradiction of at least one law and there are finitely many laws. This is what is needed for the finite error result from Section 4 to hold but with $|\mathcal{M}|$ replaced by $|\mathcal{T}|$ (see Theorem 36 below) which can be exponentially smaller. Furthermore, the extension to countable $\mathcal{T}$ works the same as in Theorem 45.

Theorem 36 (Finite error bound when using laws) *Suppose that $\mathcal{T}$ is a finite class of deterministic laws and let $\mathcal{G}(h) = \{\nu(\cdot|h) \mid \nu \in \mathcal{M}(\{\tau| \; \tau \in \mathcal{T} \text{ consistent with } h\})\}$. We define $\bar{\pi}$ by combining $\mathcal{G}$ with the optimistic decision function. Following $\bar{\pi}$ for a finite class of deterministic laws $\mathcal{T}$ in an environment $\mu \in \mathcal{M}(\mathcal{T})$, we have for any $0 < \varepsilon < \frac{1}{1-\gamma}$ that*

$$V_\mu^{\bar{\pi}}(h_t) \geq \max_\pi V_\mu^\pi(h_t) - \varepsilon \tag{9}$$

for all but at most $|\mathcal{T} - l| \frac{-\log \varepsilon(1-\gamma)}{1-\gamma}$ time steps t where l is the minimum number of laws from $\mathcal{T}$ needed to define a complete environment.

Proof This theorem follows from Theorem 29 since there are at most $K = |\mathcal{T} - l|$ time steps with a contradiction.

■

6. Finite Classes of Stochastic Non-Dominant A Priori Environments

A stochastic hypothesis may never become completely inconsistent in the sense of assigning zero probability to the observed sequence while still assigning very different probabilities than the true environment. Therefore, we exclude based on a threshold for the probability assigned to the generated history proportional to the highest probability assigned by some environment in the remaining class. An obvious alternative is to instead compare to a weighted average of all the remaining environments as done by Lattimore et al. (2013b) for the BayesExp algorithm. This latter alternative means that one can interpret the criterion as excluding environments of low posterior probability where the weights define the prior. The alternatives differ only by a constant factor depending on the weights.

Unlike in the deterministic case, a hypothesis can cease to be optimistic without having been excluded. We, therefore, only consider an algorithm that re-evaluates its optimistic hypothesis at every time step. Algorithm 3 specifies the procedure and Theorem 37 states that it is asymptotically optimal. We previously introduced the agent described in Algorithm 3, in Example 14 by saying it has an optimistic decision function and by describing the hypothesis-generating function based on a criterion for excluding environments from an initial class. We also consider a different exclusion criterion, i.e., a different hypothesis-generating function, for an optimistic agent to be able to present sample complexity bounds that we believe also holds for the first agent. The criterion used to achieve near-optimal sample complexity has previously been used in the MERL algorithm (Lattimore et al., 2013a), which has a decision function that we deem irrational according to our theory. Our agent instead uses an optimistic decision function but the same hypothesis-generating function as MERL. A very similar agent and bound can also be achieved as an optimistically acting realization of the adaptive k-meteorologists' algorithm by Diuk et al. (2009) and its bound. This agent would only have a slightly different exclusion criterion compared to MERL. A further step that we do not take here would be to improve the bounds dramatically by using stochastic laws (Sunehag and Hutter, 2015) as we did with deterministic laws previously.

Algorithm 3: Optimistic Agent (π°) with Stochastic Finite Class

Require: Finite class of stochastic environments $\mathcal{M}_1 \equiv \mathcal{M}$, threshold $z \in (0,1)$
1: $t = 1$
2: **repeat**
3: $(\pi^*, \nu^*) = \arg\max_{\pi,\nu\in\mathcal{M}_t} V_\nu^\pi(h_{t-1})$
4: $a_{t-1} = \pi^*(h_{t-1})$
5: Perceive $o_t r_t$ from environment μ
6: $h_t \leftarrow h_{t-1}a_{t-1}o_t r_t$
7: $t \leftarrow t+1$
8: $\mathcal{M}_t := \{\nu \in \mathcal{M}_{t-1} : \frac{\nu(h_t|a_{1:t})}{\max_{\tilde{\nu}\in\mathcal{M}} \tilde{\nu}(h_t|a_{1:t})} > z\}$
9: **until** the end of time

Theorem 37 (Optimality, Finite stochastic class) *Define π° by using Algorithm 3 with any threshold $z \in (0,1)$ and a finite class $\mathcal{M}$ of stochastic environments containing the true environment μ, then with probability $1 - z|\mathcal{M}-1|$ there exists, for every $\varepsilon > 0$, a number $T < \infty$ such that*

$$V_\mu^{\pi^\circ}(h_t) \; > \; \max_\pi V_\mu^\pi(h_t) - \varepsilon \;\; \forall t \geq T.$$

We borrow some techniques from Hutter (2009a) that introduced a "merging of opinions" result that generalized the classical theorem by Blackwell and Dubins (1962), restated here as Theorem 17. The classical result says that it is sufficient that the true measure (over infinite sequences) is absolutely continuous with respect to a chosen a priori distribution to guarantee that they will almost surely merge in the sense of total variation distance. The generalized version is given in Lemma 38. When we combine a policy π with an environment ν by letting the actions be taken by the policy, we have defined a measure, denoted by $\nu(\cdot|\pi)$, on the space of infinite sequences from a finite alphabet. We denote such a sample

sequence by ω and the a:th to b:th elements of ω by $\omega_{a:b}$. The σ-algebra is generated by the cylinder sets $\Gamma_{y_{1:t}} := \{\omega|\omega_{1:t} = y_{1:t}\}$ and a measure is determined by its values on those sets. To simplify notation in the next lemmas we will write $P(\cdot) = \nu(\cdot|\pi)$, meaning that $P(\omega_{1:t}) = \nu(h_t|a_{1:t})$ where $\omega_j = o_j r_j$ and $a_j = \pi(h_{j-1})$. Furthermore, $\nu(\cdot|h_t, \pi) = P(\cdot|h_t)$.

The results from Hutter (2009a) are based on the fact that $Z_t = \frac{Q(\omega_{1:t})}{P(\omega_{1:t})}$ is a martingale sequence if P is the true measure and therefore converges with P probability 1 (Doob, 1953). The crucial question is if the limit is strictly positive or not. The following lemma shows that with P probability 1 we are either in the case where the limit is 0 or in the case where $d(P(\cdot|\omega_{1:t}), Q(\cdot|\omega_{1:t})) \to 0$.

Lemma 38 (Generalized merging of opinions Hutter (2009a)) *For any measures P and Q it holds that $P(\Omega^\circ \cup \bar{\Omega}) = 1$ where*

$$\Omega^\circ := \Big\{\omega : \frac{Q(\omega_{1:t})}{P(\omega_{1:t})} \to 0\Big\} \quad \textit{and} \quad \bar{\Omega} := \Big\{\omega : d(P(\cdot|\omega_{1:t}), Q(\cdot|\omega_{1:t})) \to 0\Big\}$$

The following lemma replaces the property for deterministic environments that either they are consistent indefinitely or the probability of the generated history becomes 0.

Lemma 39 (Merging of environments) *Suppose we are given two environments μ (the true one) and ν and a policy π (defined e.g., by Algorithm 3). Let $P(\cdot) = \mu(\cdot|\pi)$ and $Q(\cdot) = \nu(\cdot|\pi)$. Then with P probability 1 we have that*

$$\lim_{t\to\infty} \frac{Q(\omega_{1:t})}{P(\omega_{1:t})} = 0 \quad \textit{or} \quad \lim_{t\to\infty} |V_\mu^\pi(h_t) - V_\nu^\pi(h_t)| = 0.$$

Proof This follows from a combination of Lemma 38 and Lemma 18. ■

Proof (Theorem 37) Given a policy π, let $P(\cdot) = \mu(\cdot|\pi)$ where $\mu \in \mathcal{M}$ is the true environment and $Q = \nu(\cdot|\pi)$ where $\nu \in \mathcal{M}$. Let the outcome sequence $(o_1r_1), (o_2r_2), ...$ be denoted by ω. It follows from Doob's martingale inequality (Doob, 1953) that for all $z \in (0, 1)$

$$P\Big(\sup_t \frac{Q(\omega_{1:t})}{P(\omega_{1:t})} \geq 1/z\Big) \leq z, \qquad \text{which implies} \quad P\Big(\inf_t \frac{P(\omega_{1:t})}{Q(\omega_{1:t})} \leq z\Big) \leq z.$$

This implies, using a union bound, that the probability of Algorithm 3 ever excluding the true environment is less than $z|\mathcal{M} - 1|$.

The limits $\frac{\nu(h_t|\pi^\circ)}{\mu(h_t|\pi^\circ)}$ converge μ-almost surely as argued before using the martingale convergence theorem. Lemma 39 tells us that any given environment (with probability one) is eventually excluded or is permanently included and merges with the true one under π°. Hence, the remaining environments do merge with the true environment, according to and in the sense of Lemma 39. Lemma 18 tells us that the difference between value functions (for the same policy) of merging environments converges to zero. Since there are finitely many environments and the ones that remain indefinitely in $\mathcal{M}_t$ merge with the true environment under π°, there is for every $\tilde{\varepsilon} > 0$ a T such that for all continuations h of h_T, it holds that

$$|V_{\nu_1}^{\pi^\circ}(h) - V_{\nu_2}^{\pi^\circ}(h)| < \tilde{\varepsilon} \ \forall \nu_1, \nu_2 \in \mathcal{M}_{\ell(h)}.$$

The proof is concluded by Lemma 25 (applied to $\Xi = \mathcal{M}_t$) in the case where the true environment remains indefinitely included which happens with probability $z|\mathcal{M}-1|$. ■

Optimal sample complexity for optimistic agent. We state the below results for $\gamma = 0$ even if some of the results referred to are more general, both for simplicity and because we can only prove that our new agent is optimal for this myopic case and only conjecture that the result extends to $0 < \gamma < 1$. For $\gamma = 0$ we can replace π by a in e.g., V^π because the value then only depends on the immediate action.

Definition 40 (ε-error) *Given $0 \leq \varepsilon < 1$, we define the number of ε-errors for $\gamma = 0$ in history h to be*

$$m(h,\varepsilon) = |\{t \leq \ell(h) \mid V_\mu^{a_t}(h_t) < V_\mu^*(h_t) - \varepsilon\}|$$

where μ is the true environment, $\ell(h)$ is the length of h, a_t is the t:th action of an agent π and $V_\mu^*(h) = \max_a V_\mu^a(h)$. Each such time point t where $V_\mu^{a_t}(h_t) < V_\mu^*(h_t) - \varepsilon$ is called an ε-error.

In Lattimore et al. (2013a), an agent (MERL) that achieves optimal sample complexity for general finite classes of stochastic environments was presented and we provided a high-level description of it in Example 15 in terms of an irrational decision function and a hypothesis-generating function. Here we point out that one can take the hypothesis-generating function of MERL and combine it with an optimistic decision function and still satisfy optimal sample complexity for the case $\gamma = 0$. We conjecture that our optimistic agent also satisfies MERL's bound for $0 < \gamma < 1$, but it is even harder to prove than the difficult analysis of MERL, which was designed to enable the proof. Our resulting optimistic agent is described in Algorithm 4. Lattimore et al. (2013a) proves the matching lower bound $O(\frac{M}{\varepsilon^2(1-\gamma)^3}\log\frac{1}{\delta})$. We conjecture that the optimistic agent just like MERL satisfies an upper bound matching the generic lower up to logarithmic factors for all $\gamma < 1$ and not just for $\gamma = 0$, which we can prove it for.

The advantage of the optimistic agent is that its exploration is not irrationally exploring an option with values in e.g., the interval $[0, 0.3]$ if there is an option with guaranteed value of 0.9. MERL does this because it looks for the maximum discrepancy in values, which is why it is called Maximum Exploration Reinforcement Learning. The agent eliminates all wrong environments regardless if this is useful or not. The exclusion criterion is based on what return is predicted by the remaining environments. If the most optimistic and the most pessimistic differ substantially one of them will turn out to be wrong and the plausibility of it being the truth decreases. When an environment becomes sufficiently implausible it is excluded. The technical difficulty is about both making sure that the truth is with high probability not excluded while also not keeping an environment unnecessarily long which would cause excess exploration. Investigating this particular technical difficulty, while important, is not among the main conceptual issues this article is focused on.

Theorem 41 (Sample complexity for optimistic agent) *Suppose we have a finite class of M (stochastic) environments $\mathcal{M}$. Letting $\alpha = 1 + (4\sqrt{M}-1)^{-1}$ and $\delta_1 = \delta(32(3 + \log_2 1/\varepsilon)M^{3/2})^{-1}$ in Algorithm 4, the number of ε-errors, i.e., time points t such that*

Algorithm 4: Optimistic agent with hypothesis-generation from Lattimore et al. (2013a)

Require: $\varepsilon, \delta_1, \alpha, \mathcal{M} = \{\nu_1, ..., \nu_M\}$
Ensure: $t = 1$, $h = \epsilon$, $\alpha_j = \lceil \alpha^j \rceil$, $n_{\nu,\kappa} := 0 \ \forall \nu \in \mathcal{M}, \kappa \in \mathbb{N}$
  while True **do**
    $(\overline{\nu}, a_t) := \arg\max_{\nu \in \mathcal{M}, a \in \mathcal{A}} V_\nu^a(h)$ # Choosing the optimistic action.
    Take action a_t, receive r_t,o_t # h is not appended until the end of the loop
    $\underline{\nu} := \arg\min_{\nu \in \mathcal{M}} V_\nu^{a_t}(h)$ # Find the pessimistic environment for a_t
    $\Delta = V_{\overline{\nu}}^{a_t}(h) - V_{\underline{\nu}}^{a_t}(h)$ # Difference between optimistic and pessimistic
    if $\Delta > \varepsilon/4$ # If large, one of them is significantly off
    # and we got an effective test
   then
      $\kappa = \max\{k \in \mathbb{N} : \Delta > \varepsilon 2^{k-2}\}$
      $n_{\overline{\nu},\kappa} = n_{\overline{\nu},\kappa} + 1, n_{\underline{\nu},\kappa} = n_{\underline{\nu},\kappa} + 1$
      $X_{\overline{\nu},\kappa}^{n_{\overline{\nu},\kappa}} = V_{\overline{\nu}}^{a_t}(h) - r_t$
      $X_{\underline{\nu},\kappa}^{n_{\underline{\nu},\kappa}} = r_t - V_{\underline{\nu}}^{a_t}(h)$
      if $\exists j, \kappa : n_{\overline{\nu},\kappa} = \alpha_j$ and $\sum_{i=1}^{n_{\overline{\nu},\kappa}} X_{\overline{\nu},\kappa}^i \geq \sqrt{2 n_{\overline{\nu},\kappa} \log \frac{n_{\overline{\nu},\kappa}}{\delta_1}}$ **then**
        $\mathcal{M} = \mathcal{M} \setminus \{\overline{\nu}\}$
      end if
      if $\exists j, \kappa : n_{\underline{\nu},\kappa} = \alpha_j$ and $\sum_{i=1}^{n_{\overline{\nu},\kappa}} X_{\underline{\nu},\kappa}^i \geq \sqrt{2 n_{\underline{\nu},\kappa} \log \frac{n_{\underline{\nu},\kappa}}{\delta_1}}$ **then**
        $\mathcal{M} = \mathcal{M} \setminus \{\underline{\nu}\}$
      end if
      $t := t + 1$, $h := h a_t o_t r_t$
    end if
  end while

$V_\mu^*(h_t) - V_\mu^\pi(h_t) > \varepsilon$ where π is Algorithm 4, resulting from running it on any environment in $\mathcal{M}$ is with probability $1 - \delta$ less than

$$\tilde{\mathcal{O}}(\frac{M}{\varepsilon^2} \log^2 \frac{1}{\delta})$$

where $\tilde{O}$ means O but up to logarithmic factors.

Proof The result follows from the analysis in Lattimore et al. (2013a) and we only provide an overview here. More precisely, the claim follows from the proofs of Lemma 2 and 4 in Lattimore et al. (2013a) which are both based on Azuma's inequality. Lemma 2 proves that the true environment will not be excluded with high probability (we need this to be at least $1 - \delta/2$). Lemma 4 shows that the number of exploration phases will not be larger than $\tilde{O}(\frac{M}{\varepsilon^2} \log^2 \frac{1}{\delta})$ with high probability, at least $1 - \delta/2$. The proof shows that before we reach that many we will with at least that probability have excluded all but the true environment. However, all environments do not have to be excluded and some environments might remain indefinitely by offering just slightly less reward for the optimal action than the true environment. For our agent, unlike MERL, an environment might also remain by differing arbitrarily much on actions that will never optimistically be taken. ■

For a reader that is familiar with MERL we explain why the bound for our agent should naturally be expected to be the same as for the MERL agent for $\gamma = 0$. To ensure that it can be guaranteed that no ε-errors are made during exploitation, MERL checks the maximum distance between environments for any policy and decides based on this if it needs to explore. Our agent, however, will still have this guarantee in the case $\gamma = 0$ and we can, from the analysis of MERL in Lattimore et al. (2013a), conclude that it makes, with probability $1 - \delta$, at most $\tilde{O}(\frac{M}{\varepsilon^2} \log^2 \frac{1}{\delta})$ ε-errors. In fact, for $\gamma = 0$ we only need to know that the maximum difference between any two environments' values under the optimistic action is less than $\varepsilon/2$, to guarantee that the agent does not make an ε-error.

Model-free vs model-based. We will here discuss our two main ways of excluding environments, namely exclusion by accuracy of return predictions (Algorithm 4 and MERL) and plausibility given observations and rewards (Algorithm 3 and BayesExp). Algorithm 4 above is essentially a model-free algorithm since what is used from each environment are two things; a recommended policy and a predicted return (its value in the given environment). Algorithm 4 evaluates the plausibility of an environment based on its predicted return. Hence, for each time step it only needs pairs of policy and return prediction and not complete environments. Such pairs are exactly what is considered in the Learning Classifier Systems (LCS) approach as mentioned in the introduction and as will be discussed in Section 5.2.

We will primarily consider a model-based situation where predictions are made also for future observations. Also, including the observations in the evaluation of one's hypotheses makes better use of available data. However, Hutter (2009b) argues that observations can be extremely complex and that focusing on reward prediction for selecting a model, may still be preferable due its more discriminative nature. We do not here take a definite position.

Lattimore et al. (2013b) studied confidence and concentration in sequence prediction and used exclusion based on a probability ratio, in that case with a weighted average instead of the max in our Algorithm 3. This alternative expression, which is closely related to the one used by Algorithm 3, differing only by a constant factor, can be interpreted as the posterior probability for the hypothesis and hypotheses with low posterior probability are excluded. Lattimore (2014) extended this work to a reinforcement learning algorithm BayesExp that like MERL above switches between phases of exploitation and pure exploration. When the remaining environments are sufficiently concentrated, one can guarantee that an agent does not make a mistake and the agent exploits this. The exploitation in BayesExp is performed by maximizing value for a weighted average, although one can also use optimism and not make a mistake. We deem both behaviors rational based on the definitions in Section 2. However, when the environments are not close enough, BayesExp explores by maximizing Bayesian information gain or by acting greedily with respect to the policy with the largest Hellinger distance to the Bayes mixture. Pure exploration is in this article not deemed rational and we suggest replacing it with acting greedily with respect to the most optimistic environment, i.e., being optimistic. This results again in an always optimistic agent with a criterion for when to exclude environments and we conjecture that this agent satisfies near optimal sample-complexity.

Compact classes. One can extend our results for finite classes to classes that are compact in a suitable topology, e.g., defined by the pseudo-metric

$$\tilde{d}(\nu_1, \nu_2) = \sup_{h,\pi} |V^{\pi}_{\nu_1}(h) - V^{\pi}_{\nu_2}(h)|$$

used by Lattimore et al. (2013a) or a variant based on total variation distance used for the same purpose in Sunehag and Hutter (2012a). If one wants accuracy of $\varepsilon > 0$ one can cover the compact space with finitely many $\tilde{d}$-balls of radius $\varepsilon/2$ and then apply an algorithm for finite classes to the finite class of ball centers to achieve accuracy $\varepsilon/2$. This adds up to accuracy ε for the actual compact class. The number of environments in the finite class is then equal to the number of balls. This number also feature prominently in the theory of supervised learning using reproducing kernel Hilbert spaces (Cucker and Smale, 2002).

Feature Markov decision processes. One can define interesting compact classes of environments using the feature Markov decision process framework (ϕMDP) (Hutter, 2009b; Sunehag and Hutter, 2010). The main idea in this framework is to reduce an environment to an MDP through applying a function ϕ to the history h_t and define a state $s_t = \phi(h_t)$. Given a class of functions of this sort, Sunehag and Hutter (2010) define a class of environments that consists of those that can be exactly represented as an MDP using a function from the class. The class of feature Markov decision processes defined from a finite set of maps is a compact continuously parameterized class. Given a map ϕ from histories to a finite state set $\mathcal{S}$, a sequence of actions, observations, rewards is transformed into a sequence of states $s_1, ..., s_n$ where $s_t = \phi(h_t)$. Defining probability distributions $Pr(or|s, a)$ leads to having defined an environment. In other words, a combination of a map from histories to states with probability parameters stating, for each state-action pair (s, a) the probability of each possible perception $or \in \mathcal{O} \times \mathcal{R}$, is a fully specified environment. Furthermore,

$$Pr(s_{t+1}, r_{t+1}|s_t, a_{t+1}) = \sum_{o_{t+1}r_{t+1}|\phi(h_t a_{t+1} o_{t+1} r_{t+1})=s_{t+1}} Pr(o_{t+1}r_{t+1}|s_t, a_{t+1})$$

and we have, therefore, also defined a Markov Decision Process based on the states defined by the map ϕ. When considering an environment's optimal policy, this means that we can restrict our study to policies that are functions from the states of the environment to actions. Finding the best such *stationary policy* becomes the goal in this setting. Considering a finite class of maps, each map gives us a compact class of environments and we can embed all of them into $\mathbb{R}^d$ for some d. Since a finite union of compact sets is compact, we have defined a compact class. Hence, one can cover the space with finite many balls regardless of how small positive radius one chooses. However, the bounds are linear in the number of balls which can be very large. This is because those bounds are worst case bounds for fully general environments. In the feature MDP case we learn simultaneously about large subsets of environments and one should be able to have bounds that are linear in the size of a maximal state space (see Section 5.2).

Example 42 (Automata) *A special form of maps are those that can be defined by a deterministic function (a table) $\tau(s, a, o, r) = s'$. Maps of this sort have been considered by Mahmud (2010) for the class of Probabilistic-Deterministic Finite Automata.*

7. Countable and Growing Classes

In this section, we extend the agents and analysis from the previous section to arbitrary countable environment classes.

Properties of hypothesis-generating functions. After seeing examples of decision functions and hypothesis generating functions above, we will discuss what properties are desirable in a hypothesis-generating function. We discussed what a decision function should be like in Section 3.1 based on decision-theoretic axioms defining rationality. In the choice of hypothesis-generating functions we focus on what kind of performance can be guaranteed in terms of how many suboptimal decisions will be taken. First, however, we want to restrict our study to hypothesis-generating functions that are following Epicurus' principle that says that one should keep all consistent hypotheses. In the case of deterministic environments it is clear what it means to have a contradiction between a hypothesis and an observation while in the stochastic case it is not. One can typically only say that the data makes a hypothesis unlikely as in Example 14. We say that a hypothesis generating function satisfies Epicurus if the update function is such that it might add new environments in any way while removing environments if a hypothesis becomes implausible (likely to be false) in light of the observations made. Aside from satisfying Epicurus' principle, we design hypothesis generating functions based mainly on wanting few mistakes to be made. For this purpose we first define the term ε-(in)confidence. We are going to formulate the rest of the definitions and results in this section for $\gamma = 0$, while explaining also how the general $0 < \gamma < 1$ works. We choose to formulate the formal results for this case ($\gamma = 0$) to clarify the reasoning and conceptual issues that apply to endless variations of the setting.

Since the true environment is unknown, an agent cannot know if it has made an ε-error or not. However, if one assumes that the true environment is in the class $\mathcal{G}(h_t)$, or more generally that the class contains an environment that is optimistic with respect to the true environment, and if the class is narrow in total variation distance in the sense (of Lemma 25) that the distance between any pair of environments in the class is small, then one can conclude that an error is not made. Since we do not know if this extra assumption holds for $\mathcal{G}(h_t)$, we will use the terms ε-confident and ε-inconfident.

If the value functions in the class $\mathcal{G}(h_t)$ differ in their predicted value by more than $\varepsilon > 0$, then we cannot be sure not to make an ε-error even if we knew that the true environment is in $\mathcal{G}(h_t)$. We call such points ε-inconfidence points.

Definition 43 (ε-(in)confidence) *Given $0 < \varepsilon < 1$, we define the number of ε-inconfidence points in the history h to be*

$$n(h, \varepsilon) := |\{t \leq \ell(h) \mid \max_{\nu_1, \nu_2 \in \mathcal{G}(h_t)} |V_{\nu_1}^{\pi^*} - V_{\nu_2}^{\pi^*}| > \varepsilon\}|$$

where $\pi^* := \arg\max_\pi \max_{\nu \in \mathcal{G}(h_t)} V_\nu^\pi$. In the $\gamma = 0$ case studied here, we can equivalently write $a^* := \arg\max_a \max_{\nu \in \mathcal{G}(h_t)} V_\nu^a$ instead of π^*. The individual time points where $\max_{\nu_1, \nu_2 \in \mathcal{G}(h_t)} |V_{\nu_1}^{\pi^*} - V_{\nu_2}^{\pi^*}| > \varepsilon$ are the points of ε-inconfidence and the other points are the points of ε-confidence.

Hypothesis-generating functions with budget. We suggest defining a hypothesis-generating function from a countable enumerated class $\mathcal{M}$ based on a *budget function* for ε-inconfidence.

The budget function $N : \mathbb{N} \to \mathbb{N}$ is always such that $N(t) \to \infty$ as $t \to \infty$. The idea is simply that when the number of ε-inconfidence points is below budget the next environment is introduced into the class. The intuition is that if the current hypotheses are frequently contradictory, then the agent should resolve these contradictions before adding more. The definition is also mathematically convenient for proving bounds on ε-errors. Besides the budget function we also require a criterion for excluding environments. An exclusion function (criterion) is here a function $\psi(\tilde{\mathcal{M}}, h) = \mathcal{M}'i$ for $\tilde{\mathcal{M}} \subset \mathcal{M}$ and $h \in \mathcal{H}$ such that $\mathcal{M}' \subset \tilde{\mathcal{M}}$. We will use the trivial $\psi(\tilde{\mathcal{M}}, h) = \tilde{\mathcal{M}}$ when the class of environments is guaranteed to asymptotically merge with the truth. The definitions below are slightly complicated by the fact that the hypothesis class $\mathcal{G}(h)$ consists of environments $\tilde{\nu}(\cdot) = \nu(\cdot|h)$ for ν in a subset of $\mathcal{M}$ that can be described as $\{\nu \in \mathcal{M} \mid \nu(\cdot|h) \in \mathcal{G}(h)\}$.

Definition 44 (Hypothesis generation with budget and exclusion function) *The hypothesis-generating function $\mathcal{G}$ with class $\mathcal{M}$, initial class $\mathcal{M}^0 \subset \mathcal{M}$, accuracy $\varepsilon \geq 0$, budget N and exclusion criterion ψ, is defined recursively: First, let $\mathcal{G}(\epsilon) := \mathcal{M}^0$. If $n(h_t, \varepsilon) \geq N(t)$, then*

$$\mathcal{G}(h_t) := \{\nu(\cdot|h_t) \mid \nu \in \psi(\{\nu \in \mathcal{M} \mid \nu(\cdot|h_{t-1}) \in \mathcal{G}(h_{t-1})\}, h_t)\}$$

while if $n(h_t, \varepsilon) < N(t)$, let $\tilde{\nu}$ be the environment in $\mathcal{M}$ with the lowest index that is not in $\bigcup_{i=1}^{t-1}\{\nu \in \mathcal{M} \mid \nu(\cdot|h_i) \in \mathcal{G}(h_i)\}$ (i.e., the next environment to introduce) and let

$$\mathcal{G}(h_t) := \{\nu(\cdot|h_t) \mid \nu \in \{\tilde{\nu} \cup \psi(\{\nu \in \mathcal{M} \mid \nu(\cdot|h_{t-1}) \in \mathcal{G}(h_{t-1})\}, h_t)\}\}.$$

7.1 Error Analysis

We now extend the agents described in Example 11 and Example 12 by removing the demand for the class $\mathcal{M}$ to be finite and analyze the effect on the number of ε-errors made. We still use the optimistic decision function and apply it to finite classes but incrementally add environments from the full class to the finite working class of environments. The resulting agent differs from agents such as the one in Example 15 by (among other things) instead of having exploration phases as part of the decision function, it has a hypothesis-generating function that sometimes adds an environment. This may cause new explorative behavior if it becomes the optimistic hypothesis and it deviates significantly from the other environments. A point to note about our results is that the agent designer chooses the asymptotic error rate but a constant term gets higher for slower rates. This trade-off is due to the fact that if new environments are included at a slower rate, then it takes longer until the right environment is introduced while the error rate afterwards is better. If the agent knew that the true environment had been found, then it could stop introducing more but this is typically impossible to know.

Deterministic environments. We first extend the agent for finite classes of deterministic environments in Example 11 to the countable case. In the finite case with a fixed class, the proof of the finite error bound builds on the fact that every ε-error must be within $\frac{-\log(\varepsilon(1-\gamma))}{1-\gamma}$ time steps before a contradiction and the bound followed immediately because there are at most $|\mathcal{M} - 1|$ contradictions. In the case where environments are being added, errors occur either before the truth is added or within that many time steps before a

contradiction or that many time steps before the addition of a new environment. The addition of a new environment can change the optimistic policy without a contradiction, because the event temporarily breaks time-consistency. Hence, every added environment after the truth has been included can add at most $2\frac{-\log(\varepsilon(1-\gamma))}{1-\gamma}$ ε-errors. In the $\gamma = 0$ case it is only at contradictions and when the truth has not been added that errors occur.

Theorem 45 (Countable deterministic class) *Suppose we have a countable class of deterministic environments $\mathcal{M}$ with a chosen enumeration and containing the true environment. Also suppose we have a hypothesis-generating function $\mathcal{G}$ with a finite initial class $\mathcal{G}(\epsilon) := \mathcal{M}^0 \subset \mathcal{M}$, budget function $N : \mathbb{N} \to \mathbb{N}$, accuracy $\varepsilon = 0$ and exclusion function $\psi(\tilde{\mathcal{M}}, h) := \{\nu \in \tilde{\mathcal{M}} \mid \nu \text{ consistent with } h\}$. π° is defined by combining $\mathcal{G}$ with an optimistic decision function. It follows that*
i) The number of 0-errors $m(h_t, 0)$ is for all t at most $n(h_t, 0) + C$ for some constant $C \geq 0$ (the time steps until the true environment is introduced) dependent on choice of budget function N but not on t.
ii) $\forall i \in \mathbb{N}$ there is $t_i \in \mathbb{N}$ such that $t_i < t_{i+1}$ and $n(h_{t_i}, 0) < N(t_i)$.
Further, if we modify the hypothesis-generating function above by delaying a new environment from being introduced if more than $N(t)$ environments (including the initial class) have been introduced at time t, then
iii) $\forall t: \ n(h_t, 0) < N(t)$
iv) $m(h_t, 0)/t \to 0$ if $N(t)/t \to 0$, i.e., π° satisfies weak asymptotic optimality

In the theorem above, ii) says that we will always see the number of errors fall within the budget $N(t)$ again (except for a constant term) even if it can be temporarily above. This means that we will always introduce more environments and exhaust the class in the limit. The final conclusion (iv) is that π° satisfies weak asymptotic optimality as defined by Lattimore and Hutter (2011a) and previously considered by Orseau (2010) who showed that AIXI does not achieve this for the class of all computable environments. An agent with explicit exploration phases that achieved such weak asymptotic optimality was presented by Lattimore and Hutter (2011a) where it was also showed that for the specific countable class of all computable environments, no agent can achieve strong asymptotic optimality, i.e., convergence to optimal performance without averaging.

Comparing to the previous results on finite deterministic environments, we then assumed that the truth was already in that initial class and, therefore, $C = 0$. Further, one will in that case have at most have $|\mathcal{M} - 1|$ inconfidence points as argued in the proof of Theorem 29. Hence, $m(h_t, 0) \leq n(h_t, 0) + C$ says that we will at most have $|\mathcal{M} - 1|$ errors as stated also by Theorem 29 with $\gamma = 0$. The second part of the conclusion of Theorem 45 does not mean anything for the finite case since it relates to an indefinitely increasing budget and environments being continually added. Therefore, the case with a finite fixed class is more cleanly studied first by itself to then reuse the techniques adapted to the setting of growing classes in this section.

Proof Suppose that at time t, the true environment μ is in $\mathcal{G}(h_t)$. Then, if we do not have a 0-inconfidence point, it follows from optimism that

$$V_\mu^{\pi^\circ}(h_t) = \max_a V_\mu^a(h_t) \tag{10}$$

since all the environments in $\mathcal{G}(h_t)$ agree on the reward for the optimistic action. Hence $m(h_t, 0) \leq n(h_t, 0) + C$ where C is the time the true environment is introduced.

However, we need to show that the truth will be introduced by proving that the class will be exhausted in the limit. If this was not the case, then there is T such that $n(0, h_t) \geq N(t)\ \forall t \geq T$. Since we have 0-inconfidence points exactly when a contradiction is guaranteed, $n(0, h_t)$ is then bounded by the number of environments that have been introduced up to time t if we include the number of environments in the initial class. Hence $n(0, h_t)$ is bounded by a finite number while (by the definition of budget function) $N(t) \to \infty$ which contradicts the assumption. iii) follows because if there are at most $N(t)$ environments, and if the truth has been introduced, then one cannot have had more than $N(t)$ contradictions. iv) follows directly from iii). ■

Stochastic environments. We continue by also performing the extension of the agent in Example 12 from finite to countable classes of stochastic environments. The absolute continuity assumption (Definition 16) is best understood in a Bayesian setting but with multiple priors. That is, the environment class can arise as different mixtures of the environments in a hypothesis class that the true environment is assumed to belong to. An alternative stochastic setting is the one in Example 14 where one does not make this assumption but instead assumes that the true environment is in the class and the agent needs to have an exclusion criterion. In this section no exclusion is necessary but we instead rely on the merging of environments guaranteed by Theorem 17. As for the deterministic setting, one can derive the corresponding finite class result, Theorem 21, from the inequality $m(h_t, \varepsilon) \leq n(h_t, \varepsilon) + C$ but it requires some of the reasoning of its proof.

Theorem 46 (Countable stochastic class) *Suppose we have a enumerated countable class of stochastic environments $\mathcal{M}$ such that the true environment μ is absolutely continuous with respect to every environment in $\mathcal{M}$, a hypothesis-generating function $\mathcal{G}$ with a finite initial class $\mathcal{G}(\epsilon) = \mathcal{M}^0 \subset \mathcal{M}$, a budget function $N : \mathbb{N} \to \mathbb{N}$ and accuracy $\varepsilon > 0$ and exclusion function $\psi(\tilde{\mathcal{M}}, h) := \tilde{\mathcal{M}}$. The agent is defined by combining $\mathcal{G}$ with an optimistic decision function. If for all h, there is $\nu_h \in \mathcal{M}$ that is optimistic in the sense that*

$$\max_a V^a_{\nu_h}(h) \geq \max_a V^a_{\mu}(h),$$

then there is
i) μ-almost surely a $C \geq 0$ such that

$$\forall t\ m(h_t, \varepsilon) \leq n(h_t, \varepsilon) + C$$

ii) μ-almost surely a sequence $t_i \to \infty$ such that $n(h_{t_i}, \varepsilon) < N(t_i)$ and, therefore, any environment in $\mathcal{M}$ is eventually included in $\mathcal{G}(h_t)$ for sufficiently large t.

Proof Suppose we have a finite class Ξ of stochastic environments such that the true environment μ is absolutely continuous with respect to all of them. Suppose that $\varepsilon > 0$ and that π is defined by letting $\mathcal{G}(h_t) = \{\nu(\cdot|h_t) \mid \nu \in \Xi\}$ for all t and letting the decision

function be optimistic. If we act according to π then we will first show that there will μ-almost surely only be finitely many ε-inconfidence points. Furthermore, if Ξ contains an environment that is optimistic relative to μ then only ε -inconfidence points can be ε-errors so there are only finitely many of those.

By Theorem 17 and the finiteness of the class, there is (μ-almost surely) for any $\varepsilon > 0$ and policy π, a $T < \infty$ such that $d(\xi(\cdot|h_t, \pi), \mu(\cdot|h_t, \pi)) < \varepsilon \ \forall \xi \in \Xi \ \forall t \geq T$. We cannot know for sure when the environments in Ξ have merged with the truth (under policy π) in this sense but we do know when the environments have merged with each other. That they will merge with each other follows from the fact that they all merge with μ under π. More precisely, for all $\varepsilon' > 0$ there is $T < \infty$ such that $d(\xi_1(\cdot|h_t, \pi), \xi_2(\cdot|h_t, \pi)) < \varepsilon' \ \forall \xi_1, \xi_2 \in \Xi \ \forall t \geq T$. It follows that then $|V^\pi_{\xi_1}(h_t) - V^\pi_{\xi_2}(h_t)| < \frac{\varepsilon'}{1-\gamma} \ \forall \xi_1, \xi_2 \in \Xi \ \forall t \geq T$ by Lemma 18. Hence, for any $\varepsilon > 0$ there are only finitely many ε-inconfidence points.

Now, let $t_i, i = 1, 2, 3, ...$ be the points where $n(\varepsilon, h_t) < N(t)$, i.e., where new environments are added. $t_i < t_{i+1}$ by definition. One of the aims is to show that $t_i \to \infty$ as $i \to \infty$. Before that we do not know if there is a t_i defined for each i. Suppose that i is such that t_i is defined and suppose that there is no t_{i+1}, i.e., that $n(h_t, \varepsilon) \geq N(t) \ \forall t > t_i$. Let $\Xi := \mathcal{G}(h_{t_i+1})$. Then the argument above shows that there are only finitely many ε-inconfidence points which contradicts the assumption that $n(h_t, \varepsilon) \geq N(t) \ \forall t > t_i$ since $N(t) \to \infty$. Hence t_i is defined for all i and since $t_i < t_{i+1}$, $t_i \to \infty$ as $i \to \infty$.

Finally, ε-errors can only occur at time points before there always is an optimistic environment for μ in $\mathcal{G}(h_t)$, before an environment in the class has merged sufficiently with μ or at points of ε-inconfidence and this proves the claims. ■

Remark 47 (Extensions: $\gamma > 0$, Separable classes) *As in the deterministic case, the difference between the $\gamma = 0$ case and the $0 < \gamma < 1$ case is that ε-errors can then also occur within $\frac{-\log(\varepsilon(1-\gamma))}{1-\gamma}$ time steps before a new environment is introduced, hence the Theorem still holds. Further, one can extend our algorithms for countable classes to separable classes since they can by definition be covered by countably many balls of arbitrarily small radius.*

Discussion and future plans. The agent studied above has the behaviour that after its current class merges it could remain confident for such a long time that its average number of points of inconfidence gets close to zero, but then when a new environments is introduced a finite but potentially very long stretch of inconfidence sets in before we are back to a stretch of confidence. Since we do not have bound on how long the inconfidence will last, we can not set the budget function such as to guarantee convergence to zero for the average number of errors.

If we want to achieve such convergence, extending the agent that excludes implausible stochastic environments is more promising. The reasoning is closer to the deterministic case. In particular if we look at the adaptive k-meteorologist algorithm, when two environments have disagreed sufficiently much m times, one of them is excluded. The number m depends on the desired confidence. In the deterministic case $m = 1$ and the confidence is complete. Having an environment excluded after m disagreements, bounds the amount of inconfidence caused by adding a new environment. If one wants asymptotic optimality in average, the agent also needs to decrease ε when a new environment is introduced. We intend in

the future to pursue the investigation into asymptotic optimality for countable classes of stochastic environments, which together with stochastic laws (Sunehag and Hutter, 2015) and practical implementation constitute important questions not addressed here.

8. Conclusions

We studied sequential decision-making in general reinforcement learning. Our starting point was decision-theoretic axiomatic systems of rational behavior and a framework to define agents within. We wanted to axiomatically exclude agents that are doing things that one clearly should not, before considering achieving good performance guarantees. This is important because if the guarantees are for a relatively short horizon they can sometimes be achieved by highly undesirable strategies. The guarantees only imply that the agent learns well from its experiences.

After introducing two sets of rationality axioms, one for agents with a full horizon and one for agents with a limited horizon that required optimism, we then introduced a framework using hypothesis-generating functions and decision functions to define rational general reinforcement learning agents. Further, we designed optimistic agents within this framework for different kinds of environment classes and proved error bounds and asymptotic properties. This was first done for finite classes and then extended to arbitrary countable classes. Along the way we introduced the concept of deterministic environments defined by combining partial laws and showed that the studied optimistic agents satisfy more desirable, potentially exponentially better, guarantees in such a setting. A further step would be to also apply that strategy in the stochastic setting.

Acknowledgments

This work was supported by ARC grant DP120100950. The first author was affiliated with the Australian National University during most of this work. The authors are also grateful to the helpful reviewers.

References

Y. Abbasi-Yadkori, P. Bartlett, V. Kanade, Y. Seldin, and C. Szepesvári. Online learning in Markov decision processes with adversarially chosen transition probability distributions. In *Advances in Neural Information Processing Systems 26 (NIPS)*, pages 2508–2516, 2013.

J. Asmuth, L. Li, M. Littman, A. Nouri, and D. Wingate. A Bayesian sampling approach to exploration in reinforcement learning. In *Uncertainty in Artificial Intelligence (UAI)*, pages 19–26, 2009.

P. Auer and R. Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. In *Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems(NIPS'2006)*, pages 49–56, 2006.

E. Baum and I. Durdanovic. Evolution of cooperative problem solving in an artificial economy. *Neural Computation*, 12(12):2743–2775, 2001.

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

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

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39:1–49, 2002.

B. deFinetti. La prevision: Ses lois logiques, ses sources subjectives. In *Annales de l'Institut Henri Poincare 7*, pages 1–68. Paris, 1937.

J. Diestel. *Sequences and series in Banach spaces*. Springer-Verlag, 1984.

C. Diuk, L. Li, and B. Leffler. The adaptive k-meteorologists problem and its application to structure learning and feature selection in reinforcement learning. In *Proceedings of the 26th International Conference on Machine Learning, ICML 2009*, pages 249–256, 2009.

J. Doob. *Stochastic processes*. Wiley, New York, NY, 1953.

J. Drugowitsch. *Learning Classifier Systems from First Principles: A Probabilistic Reformulation of Learning Classifier Systems from the Perspective of Machine Learning*. Technical report (University of Bath. Dept. of Computer Science). University of Bath, Department of Computer Science, 2007.

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

J.H. Holland. Escaping brittleness: The possibilities of general purpose learning algorithms applied to parallel rule-based systems. In R.S. Michalski, J.G. Carbonell, and T.M. Mitchell, editors, *Machine learning: An artificial intelligence approach*, volume 2, chapter 20, pages 593–623. Morgan Kaufmann, Los Altos, CA, 1986.

M. Hutter. Implementierung eines Klassifizierungs-Systems. Master's thesis, Theoretische Informatik, TU München, 1991.

M. Hutter. *Universal Articial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2005.

M. Hutter. Discrete MDL predicts in total variation. In *Advances in Neural Information Processing Systems 22: (NIPS'2009)*, pages 817–825, 2009a.

M. Hutter. Feature reinforcement learning: Part I: Unstructured MDPs. *Journal of Artificial General Intelligence*, 1:3–24, 2009b.

E. Kreyszig. *Introductory Functional Analysis With Applications*. Wiley, 1989.

I. Kwee, M. Hutter, and J. Schmidhuber. Market-based reinforcement learning in partially observable worlds. *Proceedings of the International Conference on Artificial Neural Networks (ICANN-2001)*, pages 865–873, 2001.

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

T. Lattimore and M. Hutter. Asymptotically optimal agents. In *Proc. of Algorithmic Learning Theory, (ALT'2011)*, volume 6925 of *Lecture Notes in Computer Science*, pages 368–382. Springer, 2011a.

T. Lattimore and M. Hutter. Time consistent discounting. In *Proc. 22nd International Conf. on Algorithmic Learning Theory (ALT'11)*, volume 6925 of *LNAI*, pages 383–397, Espoo, Finland, 2011b. Springer, Berlin.

T. Lattimore and M. Hutter. PAC Bounds for Discounted MDPs. In Nader H. Bshouty, Gilles Stoltz, Nicolas Vayatis, and Thomas Zeugmann, editors, *ALT*, volume 7568 of *Lecture Notes in Computer Science*, pages 320–334. Springer, 2012. ISBN 978-3-642-34105-2.

T. Lattimore, M. Hutter, and P. Sunehag. The sample-complexity of general reinforcement learning. *Journal of Machine Learning Research, W&CP: ICML*, 28(3):28–36, 2013a.

T. Lattimore, M. Hutter, and P. Sunehag. Concentration and confidence for discrete Bayesian sequence predictors. In *Proc. 24th International Conf. on Algorithmic Learning Theory (ALT'13)*, volume 8139 of *LNAI*, pages 324–338, Singapore, 2013b. Springer, Berlin.

J. Leike and M. Hutter. Bad Universal Priors and Notions of Optimality. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015*, pages 1244–1259, 2015.

M. Li and P. Vitani. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, 2008.

M. M. Mahmud. Constructing states for reinforcement learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 727–734, 2010.

O.-A. Maillard, R. Munos, and D. Ryabko. Selecting the state-representation in reinforcement learning. In *Advances in Neural Information Processing Systems 24 (NIPS'2011)*, pages 2627–2635, 2011.

O.-A. Maillard, P. Nguyen, R. Ortner, and D. Ryabko. Optimal regret bounds for selecting the state representation in reinforcement learning. In *International Conference on Machine Learning (ICML'2013)*, 2013.

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

L. Naricia and E. Beckenstein. The Hahn-Banach theorem: the life and times. *Topology and its Applications*, 77(2):193–211, 1997.

G. Neu, A. György, C. Szepesvári, and A. Antos. Online Markov decision processes under bandit feedback. In *Advances in Neural Information Processing Systems 23: 2010.*, pages 1804–1812, 2010.

J. Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

P. Nguyen, O.-A. Maillard, D. Ryabko, and Ronald Ortner. Competing with an infinite set of models in reinforcement learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS'2013).*, 2013.

L. Orseau. Optimality issues of universal greedy agents with static priors. In *Proc. of Algorithmic Learning Theory, 21st International Conference, (ALT'2010)*, volume 6331 of *Lecture Notes in Computer Science*, pages 345–359. Springer, 2010.

F. Ramsey. Truth and probability. In R. B. Braithwaite, editor, *The Foundations of Mathematics and other Logical Essays*, chapter 7, pages 156–198. Brace & Co., 1931.

S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, Englewood Cliffs, NJ, 3^{nd} edition, 2010.

D. Ryabko and M. Hutter. On the possibility of learning in reactive environments with arbitrary dependence. *Theoretical Computer Science*, 405(3):274–284, 2008.

L. Savage. *The Foundations of Statistics*. Wiley, New York, 1954.

A. L. Strehl, L. Li, and M. L. Littman. Reinforcement learning in finite MDPs: PAC analysis. *J. of Machine Learning Research*, 10:2413–2444, 2009.

P. Sunehag and M. Hutter. Consistency of feature Markov processes. In *Proc. 21st International Conf. on Algorithmic Learning Theory (ALT'10)*, volume 6331 of *LNAI*, pages 360–374, Canberra, 2010. Springer, Berlin.

P. Sunehag and M. Hutter. Axioms for rational reinforcement learning. In *Algorithmic Learning Theory, (ALT'2011)*, volume 6925 of *Lecture Notes in Computer Science*, pages 338–352. Springer, 2011.

P. Sunehag and M. Hutter. Optimistic agents are asymptotically optimal. In *Proc. 25th Australasian Joint Conference on Artificial Intelligence (AusAI'12)*, volume 7691 of *LNAI*, pages 15–26, Sydney, Australia, 2012a. Springer.

P. Sunehag and M. Hutter. Optimistic AIXI. In *Proc. 5th Conf. on Artificial General Intelligence (AGI'12)*, volume 7716 of *LNAI*, pages 312–321. Springer, Heidelberg, 2012b.

P. Sunehag and M. Hutter. Learning agents with evolving hypothesis classes. In *Proc. 6th Conf. on Artificial General Intelligence (AGI'13)*, volume 7999 of *LNAI*, pages 150–159. Springer, Heidelberg, 2013.

P. Sunehag and M. Hutter. A dual process theory of optimistic cognition. In *Proc. 36th Annual Meeting of the Cognitive Science Society (CogSci'14)*, pages 2949–2954, 2014.

P. Sunehag and M. Hutter. Using Localization and Factorization to Reduce the Complexity of Reinforcement Learning In *Proc. 8th Conf. on Artificial General Intelligence (AGI'15)*, volume 9205 of *LNAI*, pages 177–186. Springer, Heidelberg, 2015.

I. Szita and A. Lörincz. The many faces of optimism: a unifying approach. In *Proceedings of the* 20^{th} *International Conference on Machine Learning*, pages 1048–1055, 2008.

J. Veness, K. S. Ng, M. Hutter, W. Uther, and D. Silver. A Monte-Carlo AIXI approximation. *Journal of Artificial Intelligence Research*, 40(1):95–142, 2011.

P. Walley. Towards a unified theory of imprecise probability. *Int. J. Approx. Reasoning*, pages 125–148, 2000.

F. Willems, Y. Shtarkov, and T. Tjalkens. The context tree weighting method: Basic properties. *IEEE Transactions on Information Theory*, 41:653–664, 1995.

Appendix A. Asymptotic Optimality of the Liberal Agent

This section contains a proof of the asymptotic optimality Theorem 27 for the liberal version of Algorithm 1 called Algorithm 1', which can (but does not have to) leave the inner loop even when $\nu^* \in \mathcal{M}_{t-1}$. We are also more explicit and provide some intuition behind the subtleties hidden in the conservative case. The notation used here is somewhat different to the main paper. The fact that environments and policies are deterministic is heavily exploited in notation and proof technique.

Policies versus action sequences. A deterministic policy $\pi : \mathcal{H} \to \mathcal{A}$ in some fixed deterministic environment $\nu : \mathcal{H} \times \mathcal{A} \to \mathcal{O} \times \mathcal{R}$ induces a unique history $h^{\pi,\nu}$, and in particular an action an sequence $a_{1:\infty}$. Conversely, an action sequence $a_{1:\infty}$ defines a policy in a *fixed* environment ν. Given ν, a policy and an action sequence are therefore equivalent. But a policy applied to multiple environments is more than just an action sequence. More on this later. For now we only consider action sequences $a_{1:\infty}$ rather than policies.

Definitions. Let

$$\begin{aligned}
\mathcal{M}_\infty &= \text{finite class of environments} \\
r_t^\nu(a_{1:t}) &= \text{reward at time } t \text{ when performing actions } a_{1:t} \text{ in environment } \nu \\
V_\nu^{a_{t:\infty}}(a_{<t}) &= \textstyle\sum_{k=t}^\infty r_k^\nu(a_{1:k})\gamma^{k-t} = \text{value of } \nu \text{ and } a_{1:\infty} \text{ from time } t \text{ on} \\
V_*^{a_{t:\infty}}(a_{<t}) &= \max_{\nu\in\mathcal{M}_\infty} V_\nu^{a_{t:\infty}}(a_{<t}) = \text{optimistic value from time } t \text{ on} \\
a^*_{1:\infty} \in \mathcal{A}^*_{1:\infty} &= \{\arg\max_{a_{1:\infty}} V_*^{a_{1:\infty}}(\epsilon)\} = \text{set of optimistic action sequences} \\
h_t^\circ &= h_t^{\pi^\circ,\mu} = \dot{a}_1\dot{o}_1\dot{r}_1 ... \dot{a}_t\dot{o}_t\dot{r}_t = \text{actually realized history} \\
&\quad \text{by Algorithm } \pi^\circ \text{ in true environment } \mu \\
&\quad \text{generated via } \mu(h_{t-1}^\circ, \dot{a}_t) = \dot{o}_t\dot{r}_t \text{ and } \pi^\circ(h_{t-1}^\circ) = \dot{a}_t
\end{aligned}$$

Consistency. There is a finite initial phase during which environments ν can become inconsistent with h_t° in the sense of $h_t^{\pi^\circ,\nu} \neq h_t^\circ$. Algorithm 1 eliminates environments as soon as they become known to be inconsistent. Since here we are interested in asymptotic optimality only, we can ignore this finite initial phase $1, ..., T-1$ and shift time T back to 1. This simplifies notation considerable. We hence assume that *all* environments in $\mathcal{M}_\infty$ are from the outset and forever consistent, i.e., $h_\infty^{\pi^\circ,\nu} = h_\infty^\circ \ \forall \nu \in \mathcal{M}_\infty$. This implies that

$$\dot{r}_t \ = \ r_t^\nu(\dot{a}_{1:t}) \ \text{ is independent of } \nu \in \mathcal{M}_\infty \text{ for all } t \quad (\infty\text{-consistency}) \tag{11}$$

It does *not* imply that all environments in $\mathcal{M}_\infty$ are the same, they only *look* the same on the one chosen action path $\dot{a}_{1:\infty}$, but different actions, e.g., $\tilde{a}_t$ =*look-left* instead of $\dot{a}_t$ =*look-right* could reveal that ν differs from μ, and $\tilde{a}_t$ =*go-left* instead of $\dot{a}_t$ =*go-right* can probe completely different futures. This is relevant and complicates analysis and actually foils many naively plausible conjectures, since an action $\dot{a}_t$ is only optimal if alternative actions are not better, and this depends on how the environment looks off the trodden path, and there the environments in $\mathcal{M}_\infty$ can differ.

Optimistic liberal algorithm $\boldsymbol{\pi^\circ}$. At time t, given $\dot{a}_{<t}$, Algorithm π° chooses action $\dot{a}_t$ optimistically, i.e., among

$$\dot{a}_t \in \{\arg\max_{a_t} \max_{a_{t+1:\infty}} V_*^{a_{t:\infty}}(\dot{a}_{<t})\} \tag{12}$$

More precisely, we define Algorithm 1' properly with using $\mathcal{M}_{t-1}$ at time t generating action sequence $\dot{a}_{1:\infty}$. After $t > T$, we can use $\mathcal{M}_\infty = \mathcal{M}_{t-1}$, i.e., (12) is equivalent to Algorithm 1' for $t > T$. Now we shift back $T \rightsquigarrow 1$, and (12), which uses $\mathcal{M}_\infty$, is a correct formalization of Algorithm 1'. Note that $\mathcal{M}_\infty$ depends on the choice of $\dot{a}_{1:\infty}$ the algorithm actually makes in case of ambiguities. From now on $\dot{a}_{1:\infty}$ will be a single fixed sequence, chosen by some particular deterministic optimistic algorithm.

Lemma 48 (Optimistic actions) $\dot{a}_{1:\infty} \in \mathcal{A}^*_{1:\infty}$ *i.e.,* $V_*^{\dot{a}_{1:\infty}}(\epsilon) = \max_{a_{1:\infty}} V_*^{a_{1:\infty}}(\epsilon)$.

Proof For $|\mathcal{M}_\infty| = 1$, this follows from the well-known fact in planning that optimal action trees lead to optimal policies and vice versa (under time-consistency (Lattimore and Hutter, 2011b)). For general $|\mathcal{M}_\infty| \geq 1$, ∞-consistency (11) is crucial. Using the value recursion

$$\begin{aligned}
V_\nu^{a_{1:\infty}}(\epsilon) &= \sum_{k=1}^{t-1} r_k^\nu(a_{1:k})\,\gamma^{k-1} + \gamma^t V_\nu^{a_{t:\infty}}(a_{<t}), \quad \text{we get:} \\
\gamma^t \max_{a_{t:\infty}} V_*^{a_{t:\infty}}(\dot{a}_{<t}) &= \max_{a_{t:\infty}} \max_{\nu\in\mathcal{M}_\infty} \Big[V_\nu^{\dot{a}_{<t}a_{t:\infty}}(\epsilon) - \underbrace{\textstyle\sum_{k=1}^{t-1} \overbrace{r_k^\nu(\dot{a}_{1:k})}^{=\dot{r}_k} \gamma^{k-1}}_{\text{independent } \nu \text{ and } a_{t:\infty}} \Big] \\
&= \max_{a_{t:\infty}} V_*^{\dot{a}_{<t}a_{t:\infty}}(\epsilon) - \text{const.}
\end{aligned}$$

Replacing $\max_{a_t}$ by $\arg\max_{a_t}$ we get

$$\arg\max_{a_t} \max_{a_{t+1:\infty}} V_*^{a_{t:\infty}}(\dot{a}_{<t}) \;=\; \arg\max_{a_t} \max_{a_{t+1:\infty}} V_*^{\dot{a}_{<t}a_{t:\infty}}(\epsilon) \tag{13}$$

We can define the set of optimistic action sequences $\mathcal{A}^*_{1:\infty} = \{\arg\max_{a_{1:\infty}} V_*^{a_{1:\infty}}(\epsilon)\}$ recursively as

$$\begin{aligned}
\mathcal{A}^*_{1:t} &:= \{\arg\max_{a_{1:t}} \max_{a_{t+1:\infty}} V_*^{a_{1:\infty}}(\epsilon))\} \\
&= \{(a^*_{<t}, \arg\max_{a_t} \max_{a_{t+1:\infty}} V_*^{a^*_{<t}a_{t:\infty}}(\epsilon)) : a^*_{<t} \in \mathcal{A}^*_{<t}\}, \\
\mathcal{A}^*_{1:\infty} &= \{a_{1:\infty} : a_{1:t} \in \mathcal{A}^*_{1:t}\, \forall t\}
\end{aligned}$$

This shows that any sequence $\tilde{a}_{1:\infty}$ that satisfies the recursion

$$\tilde{a}_t \in \{\arg\max_{a_t} \max_{a_{t+1:\infty}} V_*^{\tilde{a}_{<t}a_{t:\infty}}(\epsilon))\} \tag{14}$$

is in $\mathcal{A}^*_{1:\infty}$. Plugging (13) into (12) shows that $\tilde{a}_{1:\infty} = \dot{a}_{1:\infty}$ satisfies recursion (14), hence $\dot{a}_{1:\infty} \in \mathcal{A}^*_{1:\infty}$. ■

Lemma 49 (Optimism is optimal) $V_\mu^{\dot{a}_{1:\infty}}(\epsilon) = \max_{a_{1:\infty}} V_\mu^{a_{1:\infty}}(\epsilon)$.

Note that by construction and Lemma 48, $\dot{a}_{1:\infty}$ maximizes the (known) optimistic value $V_*^{a_{1:\infty}}$ and by Lemma 49 also the (unknown) true value $V_\mu^{a_{1:\infty}}$; a consequence of the strong asymptotic consistency condition (11). Also note that $V_\mu^{\dot{a}_{1:\infty}} = V_*^{\dot{a}_{1:\infty}}$ but $V_\mu^{a_{1:\infty}} \neq V_*^{a_{1:\infty}}$ for $a_{1:\infty} \neq \dot{a}_{1:\infty}$ is possible and common.

Proof The $\leq$ direction is trivial (since maximization is over all action sequences. For limited policy spaces $\Pi \neq \Pi^{all}$ this may no longer be true). The following chain of (in)equalities proves the $\geq$ direction

$$\max_{a_{1:\infty}} V_\mu^{a_{1:\infty}}(\epsilon) \;\leq\; \max_{a_{1:\infty}} V_*^{a_{1:\infty}}(\epsilon) \;=\; V_*^{\dot{a}_{1:\infty}}(\epsilon) \;=\; \max_{\nu\in\mathcal{M}_\infty} \sum_{k=1}^{\infty} r_k^\nu(\dot{a}_{1:k})\gamma^{k-1}$$
$$= \max_{\nu\in\mathcal{M}_\infty} \sum_{k=1}^{\infty} \underbrace{\dot{r}_k\gamma^{k-1}}_{\text{indep.}\nu} \;=\; \sum_{k=1}^{\infty} \dot{r}_k\gamma^{k-1} \;=\; \sum_{k=1}^{\infty} r_k^\mu(\dot{a}_{1:k})\gamma^{k-1} \;=\; V_\mu^{\dot{a}_{1:\infty}}(\epsilon)$$

where we used in order: definition, Lemma 48, definition, consistency of $\nu \in \mathcal{M}_\infty$, independence of ν, $\mu \in \mathcal{M}_\infty$ and consistency again, and definition. ■

Proof of Theorem 27 for liberal Algorithm 1.

As mentioned, for a fixed deterministic environment ν, policies and action sequences are interchangeable. In particular $\max_\pi V_\nu^\pi(\epsilon) = \max_{a_{1:\infty}} V_\nu^{a_{1:\infty}}(\epsilon)$. This is no longer true for V_*: There are π such that for all $a_{1:\infty}$, $V_*^\pi \neq V_*^{a_{1:\infty}}$, since π may depend on ν but $a_{1:\infty}$ not. This causes us no problems, since still $\max_\pi V_*^\pi = \max_{a_{1:\infty}} V_*^{a_{1:\infty}}$, since

$$\max_\pi \max_\nu V_\nu^\pi(\epsilon) \;=\; \max_\nu \max_\pi V_\nu^\pi(\epsilon) \;=\; \max_\nu \max_{a_{1:\infty}} V_\nu^{a_{1:\infty}}(\epsilon) \;=\; \max_{a_{1:\infty}} \max_\nu V_\nu^{a_{1:\infty}}(\epsilon)$$

Similar (non)equalities hold for $V(h_t)$. Hence Lemmas 48 and 49 imply $V_*^{\pi^\circ} = \max_\pi V_*^\pi$ and $V_\mu^{\pi^\circ} = \max_\pi V_\mu^\pi$.

Now if we undo the shift $T \rightsquigarrow 1$, actually shift $T \rightsquigarrow t$, Lemma 49 implies $V_\mu^{\pi^\circ}(h_t^\circ) = \max_\pi V_\mu^\pi(h_t^\circ)$ for all $t \geq T$. This is just Theorem 1 for the liberal algorithm. ■

Appendix B. Countable Sets of Events

Instead of a finite set of possible outcomes, we will in this section assume a countable set. We suppose that the set of bets is a vector space of sequences $x_k, k = 0, 1, 2, ...$ where we use point-wise addition and multiplication with a scalar. We will define a space by choosing a norm and let the space consist of the sequences that have finite norm as is common in Banach space theory. If the norm makes the space complete it is called a Banach sequence space (Diestel, 1984). Interesting examples are ℓ^∞ of bounded sequences with the maximum norm

$\|(\alpha_k)\|_\infty = \max |\alpha_k|$, c_0 of sequence that converges to 0 equipped with the same maximum norm and ℓ^p which for $1 \leq p < \infty$ is defined by the norm

$$\|(\alpha_k)\|_p = (\sum |\alpha_k|^p)^{1/p}.$$

For all of these spaces we can consider weighted versions ($w_k > 0$) where

$$\|(\alpha_k)\|_{p,w_k} = \|(\alpha_k w_k)\|_p.$$

This means that $\alpha \in \ell^p(w)$ iff $(\alpha_k w_k) \in \ell^p$, e.g., $\alpha \in \ell^\infty(w)$ iff $\sup_k |\alpha_k w_k| < \infty$. Given a Banach (sequence) space X we use X' to denote the dual space that consists of all continuous linear functionals $f : X \to \mathbb{R}$. It is well known that a linear functional on a Banach space is continuous if and only if it is bounded, i.e that there is $C < \infty$ such that $\frac{|f(x)|}{\|x\|} \leq C \; \forall x \in X$. Equipping X' with the norm $\|f\| = \sup \frac{|f(x)|}{\|x\|}$ makes it into a Banach space. Some examples are $(\ell^1)' = \ell^\infty$, $c_0' = \ell^1$ and for $1 < p < \infty$ we have that $(\ell^p)' = \ell^q$ where $1/p + 1/q = 1$. These identifications are all based on formulas of the form

$$f(x) = \sum x_i p_i$$

where the dual space is the space that (p_i) must lie in to make the functional both well defined and bounded. It is clear that $\ell^1 \subset (\ell^\infty)'$ but $(\ell^\infty)'$ also contains "stranger" objects.

The existence of these other objects can be deduced from the Hahn-Banach theorem (see e.g., Kreyszig (1989) or Naricia and Beckenstein (1997)) that says that if we have a linear function defined on a subspace $Y \in X$ and if it is bounded on Y then there is an extension to a bounded linear functional on X. If Y is dense in X the extension is unique but in general it is not. One can use this Theorem by first looking at the subspace of all sequences in ℓ^∞ that converge and let $f(\alpha) = \lim_{k\to\infty} \alpha_k$. The Hahn-Banach theorem guarantees the existence of extensions to bounded linear functionals that are defined on all of ℓ^∞. These are called Banach limits. The space $(\ell^\infty)'$ can be identified with the so called ba space of bounded and finitely additive measures with the variation norm $\|\nu\| = |\nu|(A)$ where A is the underlying set. Note that ℓ^1 can be identified with the smaller space of countably additive bounded measures with the same norm. The Hahn-Banach Theorem has several equivalent forms. One of these identifies the hyper-planes with the bounded linear functionals (Naricia and Beckenstein, 1997).

Definition 50 (Rationality (countable case)) *Given a Banach sequence space X of bets, we say that the decision maker (subset Z of X defining acceptable bets and $\tilde{Z}$ the rejectable bets) is rational if*

1. *Every bet $x \in X$ is either acceptable or rejectable or both*
2. *x is acceptable if and only if $-x$ is rejectable.*
3. *$x, y \in Z$, $\lambda, \gamma \geq 0$ then $\lambda x + \gamma y \in Z$*
4. *If $x_k > 0 \; \forall k$ then x is acceptable and not rejectable*

In the case of a finite dimensional space X, the above definition reduces to Definition 8.

Theorem 51 (Linear separation) *Suppose that we have a space of bets X that is a Banach sequence space. Given a rational decision maker there is a positive continuous linear functional $f : X \to \mathbb{R}$ such that*

$$\{x \mid f(x) > 0\} \subseteq Z \subseteq \{x \mid f(x) \geq 0\}. \tag{15}$$

Proof The third property tells us that Z and $-Z$ are convex cones. The second and fourth property tells us that $Z \neq X$. Suppose that there is a point x that lies in both the interior of Z and of $-Z$. Then the same is true for $-x$ according to the second property and for the origin. That a ball around the origin lies in Z means that $Z = X$ which is not true. Thus the interiors of Z and $-Z$ are disjoint open convex sets and can, therefore, be separated by a hyperplane (according to the Hahn-Banach theorem) which goes through the origin (since according to the second and fourth property the origin is both acceptable and rejectable). The first two properties tell us that $Z \cup -Z = X$. Given a separating hyperplane (between the interiors of Z and $-Z$), Z must contain everything on one side. This means that Z is a half space whose boundary is a hyperplane that goes through the origin and the closure $\bar{Z}$ of Z is a closed half space and can be written as $\{x \mid f(x) \geq 0\}$ for some $f \in X'$. The fourth property tells us that f is positive. ■

Corollary 52 (Additivity) *1. If $X = c_0$ then a rational decision maker is described by a countably additive (probability) measure.*
2. If $X = \ell^\infty$ then a rational decision maker is described by a finitely additive (probability) measure.

It seems from Corollary 52 that we pay the price of losing countable additivity for expanding the space of bets from c_0 to ℓ^∞ but we can expand the space even more by looking at $c_0(w)$ where $w_k \to 0$ which contains ℓ^∞ and X' is then $\ell^1((1/w_k))$. This means that we get countable additivity back but we instead have a restriction on how fast the probabilities p_k must tend to 0. Note that a bounded linear functional on c_0 can always be extended to a bounded linear functional on ℓ^∞ by the formula $f(x) = \sum p_i x_i$ but that is not the only extension. Note also that every bounded linear functional on ℓ^∞ can be restricted to c_0 and there be represented as $f(x) = \sum p_i x_i$. Therefore, a rational decision maker for ℓ^∞-bets has probabilistic beliefs (unless $p_i = 0\ \forall i$), though it might also take asymptotic behavior of a bet into account. For example the decision maker that makes decisions based on asymptotic averages $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} x_i$ when they exist. This strategy can be extended to all of ℓ^∞ and is then called a Banach limit. The following proposition will help us decide which decision maker on ℓ^∞ is endowed with countably additive probabilities.

Proposition 53 *Suppose that $f \in (\ell^\infty)'$. For any $x \in \ell^\infty$, let $x_i^j = x_i$ if $i \leq j$ and $x_i^j = 0$ otherwise. If for any x,*

$$\lim_{j\to\infty} f(x^j) = f(x),$$

then f can be written as $f(x) = \sum p_i x_i$ where $p_i \geq 0$ and $\sum_{i=1}^{\infty} p_i < \infty$.

Proof The restriction of f to c_0 gives us numbers $p_i \geq 0$ such that $\sum_{i=1}^{\infty} p_i < \infty$ and $f(x) = \sum p_i x_i$ for $x \in c_0$. This means that $f(x^j) = \sum_{i=1}^{j} p_i x_i$ for any $x \in \ell^\infty$ and $j < \infty$. Thus $\lim_{j\to\infty} f(x^j) = \sum_{i=1}^{\infty} p_i x_i$. ■

Definition 54 (Monotone decisions) *We define the concept of a* monotone decision maker in the following way. Suppose that for every $x \in \ell^\infty$ there is $N < \infty$ such that the decision is the same for all (as defined above) $x^j,\ j \geq N$ as for x. Then we say that the decision maker is monotone.

Example 55 *Let $f \in \ell^\infty$ be such that if $\lim \alpha_k \to L$ then $f(\alpha) = L$ (i.e., f is a Banach limit). Furthermore define a rational decision maker by letting the set of acceptable bets be $Z = \{x \mid f(x) \geq 0\}$. Then $f(x^j) = 0$ (where we use notation from Proposition 53) for all $j < \infty$ and regardless of which x we define x^j from. Therefore, all sequences that are eventually zero are acceptable bets. This means that this decision maker is not monotone since there are bets that are not acceptable.*

Theorem 56 (Monotone rationality) *Given a monotone rational decision maker for ℓ^∞ bets, there are $p_i \geq 0$ such that $\sum p_i < \infty$ and*

$$\{x \mid \sum x_i p_i > 0\} \subseteq Z \subset \{x \mid \sum x_i p_i \geq 0\}. \tag{16}$$

Proof According to Theorem 51 there is $f \in (\ell^\infty)'$ such that (the closure of Z) $\bar{Z} = \{x|\ f(x) \geq 0\}$. Let $p_i \geq 0$ be such that $\sum p_i < \infty$ and such that $f(x) = \sum x_i p_i$ for $x \in c_0$. Remember that x^j (notation as in Proposition 53) is always in c_0. Suppose that there is x such that x is accepted but $\sum x_i p_i < 0$. This violate monotonicity since there exists $N < \infty$ such that $\sum_{i=1}^{n} x_i p_i < 0$ for all $n \geq N$ and, therefore, x^j is not accepted for $j \geq N$ but x is accepted. We conclude that if x is accepted then $\sum p_i x_i \geq 0$ and if $\sum p_i x_i > 0$ then x is accepted. ■

Appendix C. List of important notation

t	generic time point
T	special time point
$\mathcal{A}, \mathcal{O}, \mathcal{R}$	action/observation/reward sets
h_t	$= a_1 o_1 r_1 ... a_t o_t r_t$ = (action,observation,reward) history
$h_0 = \epsilon$	empty history/string
$\varepsilon \geq 0$	accuracy
δ	probability/confidence
$0 \leq \gamma < 1$	discount factor
$\mathcal{O}_j$	set for the j:th feature
$\vec{x} = (x_i) \in \mathcal{O} = \times_{j=1}^{m} \mathcal{O}_j$	feature vector in Section 5

$\perp$	not predicted feature
$\mathcal{O}_\perp = \times_{j=1}^m (\mathcal{O}_j \cup \{\perp\})$	observation set enhanced by $\perp$
$\pi : \mathcal{H} \to \mathcal{A}$	generic policy $\pi \in \Pi$
$\tilde{\pi}$	some specific policy π
π°	optimistic policy actually followed.
(π_t^*, ν_t^*)	optimistic (policy,environment) (used only) at time t
$V_\nu^\pi(h_t)$	future value of π interacting with ν given h_t
$\mathcal{M}, \tilde{\mathcal{M}}, \hat{\mathcal{M}}$	finite or countable class of environments
$\mathcal{M}^0$	initial class of environments
$m(h, \varepsilon)$	number of ε-errors during h
$n(h, \varepsilon)$	number of ε-inconfidence points
Ξ	finite class of dominant environments
$\nu \in \mathcal{M}$	generic environment
$\xi \in \Xi$	dominant environment
μ	true environment
$\mathcal{T}$	finite class of laws
$\tau \in \mathcal{T}$	generic law
$q_1(\tau, h, a)$	features not predicted by τ in context h, a
$q_2(\tau, h, a)$	features predicted by τ in context h, a
$\mathcal{M}(\mathcal{T})$	environments generated by deterministic laws
$\Xi(\mathcal{T})$	environments generated by stochastic laws
$\bar{\mathcal{M}}(P, \mathcal{T})$	semi-deterministic environments from background and laws
ω	elementary random outcome from some sample space
ω_t	$= o_t r_t =$ perception at time t
$x = (x_i)$	bet in Section 2
$y = (y_i)$	bet in Section 2
$p = (p_i)$	probability vector
f	decision function
$\mathcal{G}$	hypothesis-generating function