Universal Knowledge-Seeking Agents for Stochastic Environments

Laurent Orseau[1], Tor Lattimore[2], and Marcus Hutter[2]

[1] AgroParisTech, UMR 518 MIA, F-75005 Paris, France
INRA, UMR 518 MIA, F-75005 Paris, France
[2] RSCS, Australian National University
Canberra, ACT, 0200, Australia

Abstract. We define an optimal Bayesian knowledge-seeking agent, KL-KSA, designed for countable hypothesis classes of stochastic environments and whose goal is to gather as much information about the unknown world as possible. Although this agent works for arbitrary countable classes and priors, we focus on the especially interesting case where all stochastic computable environments are considered and the prior is based on Solomonoff's universal prior. Among other properties, we show that KL-KSA learns the true environment in the sense that it learns to predict the consequences of actions it does not take. We show that it does not consider noise to be information and avoids taking actions leading to inescapable traps. We also present a variety of toy experiments demonstrating that KL-KSA behaves according to expectation.

Keywords: Universal artificial intelligence; exploration; reinforcement learning; algorithmic information theory; Solomonoff induction.

1 Introduction

The goal of scientists is to acquire knowledge about the universe in which we reside. To this end, they must explore the world while designing experiments to test, discard and refine hypotheses. At the core of science lies the problem of induction that is arguably solved by Solomonoff induction, which uses algorithmic information theory to obtain a universal[3] semi-computable prior and Bayes theorem to perform induction. This approach learns to predict (fast) in any stochastically computable environment and has numerous attractive properties both theoretical [Hut05] and philosophical [RH11]. Its (in)famous incomputability is an unavoidable consequence of its generality.

The main difficulty with applying Solomonoff induction to construct an optimal scientist – which we call a knowledge-seeking agent – is that, although it defines how to predict, it gives no guidance on how to choose actions so as to maximise the acquisition of knowledge to make better predictions. The extension of Solomonoff induction to the reinforcement learning framework [SB98] has

[3] Universal in the sense that it dominates all lower-semi-computable priors [LV08].

been done by Hutter [Hut05]. An optimal reinforcement learner is different from an optimal scientist because it is rewarded extrinsically by the environment, rather than intrinsically by information gain.

Defining strategies to explore the environment optimally is not a new idea with a number of researchers having previously tackled this problem, especially Schmidhuber; see [Sch06] and references therein. Storck et al. [SHS95] use various information gain criteria in a frequentist setting to explore non-deterministic Markov environments, bending the reinforcement learning framework to turn information gain into rewards. The beauty of this approach is that exploration is not a means to the ends of getting more rewards, but is the goal per se [BO13, Sch06]. In this context, exploration *is* exploitation, thus making the old [SB98] and persisting [Ors13, LH11a] exploration/exploitation problem collapse into a unified objective.

Generalising the previous approach and placing it in a Bayesian setting, Sun et al. [SGS11] construct a policy that explores by maximising the discounted expected information gain in the class of finite-state Markov decision processes. The choice of a continuous parametric class introduces some challenging problems because the expected information gain when observing statistics depending on a continuous parameter is typically infinite. The authors side-step these problems by introducing a geometric discount factor, but this is unattractive for a universal algorithm, especially when environments are non-Markovian and may have unbounded diameter. In this work we prove most results for both the discounted and undiscounted settings, resorting to discounting only when necessary.

In 2011, Orseau presented two universal knowledge-seeking agents, Square-KSA and Shannon-KSA, designed for the class of all deterministic computable environments [Ors11]. Both agents maximise a version of the Bayes-expected entropy of their future observations, which is equivalent to maximising expected information gain with respect to the prior. Unfortunately, neither Square-KSA nor Shannon-KSA perform well when environments are permitted to be stochastic with both agents preferring to observe coin flips rather than explore a more informative part of their environment. The reason for this is that these agents mistake stochastic outcomes for complex information. In the present paper, we define a new universal knowledge-seeking agent designed for arbitrary countable classes of stochastic environments. An especially interesting case is when the class of environments is chosen to be the set of all stochastic computable environments. The new agent has a natural definition, is resistant to noise and behaves as expected in a variety of toy examples. The main idea is to choose a policy maximising the (un)discounted Bayes-expected information gain.

First we give some basic notation (Section 2). We then present the definitions of the knowledge-seeking agent and prove that it learns to predict all possible futures (Section 3). The special case where the hypothesis class is chosen to be the class of all stochastic computable in environments is then considered (Section 4). Finally, we demonstrate the agent in action on a number of toy examples to further motivate the definitions and show that the new agent performs as expected (Section 5) and conclude (Section 6).

2 Notation

Sequences. Let $\mathcal{A}$ be the finite set of all possible actions, and $\mathcal{O}$ the finite set of all possible observations. Let $\mathcal{H} := \mathcal{A} \times \mathcal{O}$ be the finite set of interaction tuples containing action/observation pairs. The sets $\mathcal{H}^t$, $\mathcal{H}^*$ and $\mathcal{H}^\infty$ are defined to contain all histories of length t, finite length and infinite length respectively. The empty history of length 0 is denoted by ϵ. We write $a_{n:m} \in \mathcal{A}^{m-n+1}$ to denote the (ordered) sequence of actions $a_n a_{n+1} \dots a_m$ and $a_{<n} := a_{1:n-1}$ and similarly for observations o and histories h, and length $\ell(a_{n:m}) := m - n + 1$.

Environments and policies. A policy is a stochastic function $\pi : \mathcal{H}^* \rightsquigarrow \mathcal{A}$ while an environment is a stochastic function $\mu : \mathcal{H}^* \times \mathcal{A} \rightsquigarrow \mathcal{O}$. We write $\pi(a|h)$ for the π-probability that policy π takes action $a \in \mathcal{A}$ in history $h \in \mathcal{H}^*$ and similarly $\nu(o|ha)$ is the ν-probability that ν outputs observation $o \in \mathcal{O}$ having observed history $h \in \mathcal{H}^*$ and action $a \in \mathcal{A}$. A policy π and environment ν interact sequentially to induce a measure P^π_ν on the space of infinite histories with $P^\pi_\nu(\epsilon) := 1$ and $P^\pi_\nu(hao)$ defined inductively by $P^\pi_\nu(hao) := \nu(o|ha)\pi(a|h)P^\pi_\nu(h)$ where $o \in \mathcal{O}$ and $a \in \mathcal{A}$. From now on, unless otherwise mentioned, all policies are assumed to be deterministic with Π being the set of all such policies. For finite history h we define $\Pi(h) \subseteq \Pi$ to be the set of policies consistent with h, so $\pi \in \Pi(h)$ if $\pi(a_t|h_{<t}) = 1$ for all $t \leq \ell(h)$, where a_t is the tth action in history h.

Bayesian mixture. Let $\mathcal{M}$ be a countable set of environments and $w : \mathcal{M} \to (0, 1]$ satisfy $\sum_{\nu\in\mathcal{M}} w_\nu \leq 1$. Given a policy π, the Bayesian mixture measure is defined by $P^\pi_\xi(h) := \sum_{\nu\in\mathcal{M}} w_\nu P^\pi_\nu(h)$, for all histories $h \in \mathcal{H}^*$. The posterior of an environment ν having observed h is $w_\nu(h) := w_\nu P^\pi_\nu(h)/P^\pi_\xi(h)$ where π is some policy consistent with h. We also take $0 \log 0 := 0$. All logarithms are in base 2. The entropy of prior w is defined by $\mathrm{Ent}(w) := \sum_{\nu\in\mathcal{M}} w_\nu \log \frac{1}{w_\nu}$. Note that P^π_ξ may only be a semimeasure in the case when $\sum_{\nu\in\mathcal{M}} w_\nu < 1$. This detail is inconsequential throughout and may be ignored by the reader unfamiliar with semimeasures.

Discounting. A discount vector is a function $\gamma : \mathbb{N} \to [0, 1]$. It is summable if $\sum_{t=1}^\infty \gamma_t < \infty$ and asymptotically non-trivial if for all $t \in \mathbb{N}$ there exists a $\tau > t$ such that $\gamma_\tau > 0$. For summable γ we define $\Gamma_t := \sum_{k=t}^\infty \gamma_k$ and otherwise $\Gamma_t := 1$. The undiscounted case fits in the framework by letting $\boldsymbol{\infty}$ be the discount vector with $\boldsymbol{\infty}_k = 1$ for all k. The finite horizon discount vector is $\mathbf{n}$ with $\mathbf{n}_k = [\![k \leq n]\!]$.

3 Knowledge-Seeking Agent

Distances between measures. The goal of the knowledge-seeking agent is to gain as much information about its environment as possible. An important quantity in information theory is the Kullback-Leibler divergence or relative entropy, which measures the expected difference in code lengths between two measures.

Let ν be an environment and π a policy. The 1-step generalized distance between measures P^π_ν and P^π_ξ having observed history h of length $t-1$ is defined as

$$\underset{h,1}{\mathrm{D}}\big(P^\pi_\nu \| P^\pi_\xi\big) := \sum_{h' \in \mathcal{H}} d(P^\pi_\nu(h'|h), P^\pi_\xi(h'|h)) = \sum_{h' \in \mathcal{H}} P^\pi_\nu(h'|h)\; f\!\left(\frac{P^\pi_\xi(h'|h)}{P^\pi_\nu(h'|h)}\right) .$$

Classical choices are given in the table on the right and are discussed in [Hut05, Sec.3.2.5]. The most interesting distance for us is the KL-divergence, but various sub-results hold for more general D. A distance D is called an f-divergence if it can be expressed via a convex f with $f(1) = 0$. All distances in the table are f-divergences with the exception of Square. Also, all but Absolute are upper bounded by KL. Therefore, besides KL itself, only Hellinger possesses both important properties simultaneously. A natural generalisation of $\mathrm{D}_{h,1}$ is the ∞-step discounted version. If $h \in \mathcal{H}^{t-1}$,

D	KL	Absolute	Square	Hellinger
$d(a,b)$	$a \log \frac{a}{b}$	$\lvert a-b\rvert$	$(a-b)^2$	$(\sqrt{a}-\sqrt{b})^2$
$f(x)$	$-\log x$	$\lvert x-1\rvert$	no f	$(\sqrt{x}-1)^2$

$$\underset{h,\gamma}{\mathrm{D}}\big(P^\pi_\nu \| P^\pi_\xi\big) := \sum_{k=t}^{\infty} \gamma_k \sum_{h' \in \mathcal{H}^{k-t}} P^\pi_\nu(h'|h)\, \underset{hh',1}{\mathrm{D}}\big(P^\pi_\nu \| P^\pi_\xi\big) \ . \tag{1}$$

If $\gamma = \mathbf{n}$, it is known that (only) the KL divergence telescopes [Hut05, Sol78]:

$$\underset{h,\mathbf{n}}{\mathrm{KL}}\big(P^\pi_\nu \| P^\pi_\xi\big) \equiv \sum_{h' \in \mathcal{H}^{n-\ell(h)}} P^\pi_\nu(h'|h) \log \frac{P^\pi_\nu(h'|h)}{P^\pi_\xi(h'|h)} \ . \tag{2}$$

Information gain value. Let h be a history and $h' \in \mathcal{H}$ one further interaction, then the instantaneous information gain of observing h' after having observed h can be quantified in terms of how much the posterior $w_\nu(h)$ changes:

$$\underset{h,1}{\mathrm{IG}}(h') := \sum_{\nu \in \mathcal{M}} d(w_\nu(hh'), w_\nu(h)) \ , \qquad \underset{h_{<n},\gamma}{\mathrm{IG}}(h_{1:\infty}) := \sum_{t=n}^{\infty} \gamma_t \underset{h_{<t},1}{\mathrm{IG}}(h_t) \ . \tag{3}$$

The right expression is the natural ∞-step generalisation where instantaneous information gains are discounted by discount vector γ. Again, the default distance for information gain is KL. Ideally, the knowledge-seeking agent should maximise some version of the μ-expected information gain where μ is the true environment, but since the latter is unknown the natural choice is to maximise the Bayes expected information gain. If d is an f-divergence, this can be written

$$\begin{aligned}
\mathbb{E}^\pi_\xi\left[\underset{h,1}{\mathrm{IG}}(h')\right] &\overset{(a)}{=} \sum_{hh' \in \mathcal{H}^t} P^\pi_\xi(hh') \sum_{\nu \in \mathcal{M}} w_\nu(hh')\; f\!\left(\frac{w_\nu(h)}{w_\nu(hh')}\right) \\
&\overset{(b)}{=} \sum_{hh' \in \mathcal{H}^t} \sum_{\nu \in \mathcal{M}} w_\nu P^\pi_\nu(hh')\; f\!\left(\frac{w_\nu P^\pi_\nu(h)}{P^\pi_\xi(h)} \frac{P^\pi_\xi(hh')}{w_\nu P^\pi_\nu(hh')}\right) \\
&\overset{(c)}{=} \sum_{h \in \mathcal{H}^{t-1}} \sum_{\nu \in \mathcal{M}} w_\nu P^\pi_\nu(h) \sum_{h' \in \mathcal{H}} P^\pi_\nu(h'|h)\; f\!\left(\frac{P^\pi_\xi(h'|h)}{P^\pi_\nu(h'|h)}\right) \\
&\overset{(d)}{=} \sum_{\nu \in \mathcal{M}} w_\nu \sum_{h \in \mathcal{H}^{t-1}} P^\pi_\nu(h)\, \underset{h,1}{\mathrm{D}}\big(P^\pi_\nu \| P^\pi_\xi\big)
\end{aligned} \tag{4}$$

where (a) is the definition of the information gain and expectation, (b) by substituting the definition of the posterior, (c) by expanding the probabilities via the chain rule, and (d) by rearranging and substituting the definition of D. If we sum both sides of (4) over $\sum_{t=1}^{\infty}\gamma_t$ and use definitions (1) and (3) we get

$$\mathbb{E}^{\pi}_{\xi}\left[\operatorname*{IG}_{\epsilon,\gamma}(h_{1:\infty})\right] = \sum_{\nu\in\mathcal{M}} w_\nu \operatorname*{D}_{\epsilon,\gamma}\big(P^{\pi}_{\nu}\|P^{\pi}_{\xi}\big) \ .$$

Essentially the same derivation but with all quantities conditioned on h gives

$$\mathbb{E}^{\pi}_{\xi}\left[\operatorname*{IG}_{h,\gamma}(h_{1:\infty})\Big|h\right] = \sum_{\nu\in\mathcal{M}} w_\nu(h) \operatorname*{D}_{h,\gamma}\big(P^{\pi}_{\nu}\|P^{\pi}_{\xi}\big) \ .$$

This leads to a natural definition of the value of a policy π:

Definition 1. *The value of policy π having observed history h with respect to discount function γ is defined to be the ξ-expected discounted information gain. We also define the optimal policy π^* to be the policy maximising the value function and V^*_γ to be the value of the optimal policy.*

$$\boxed{\begin{aligned} V^{\pi}_{\gamma}(h) &:= \sum_{\nu\in\mathcal{M}} w_\nu(h) \operatorname*{D}_{h,\gamma}\big(P^{\pi}_{\nu}\|P^{\pi}_{\xi}\big) & V^{\pi}_{\gamma} &:= V^{\pi}_{\gamma}(\epsilon) \\ \pi^* &:= \arg\max_{\pi} V^{\pi}_{\gamma} & V^*_{\gamma} &:= \sup_{\pi} V^{\pi}_{\gamma} \end{aligned}} \tag{5}$$

Existence of values and policies. There are a variety of conditions required for the existence of the optimal value and policy respectively.

Theorem 2. *If γ is summable and $D \le KL$, then $V^*_\gamma < \infty$ and π^* exists.*

Proof. Let h be an arbitrary history of length n, then the value function can be written

$$V^{\pi}_{\gamma}(h) \overset{(a)}{\le} \sum_{\nu\in\mathcal{M}} w_\nu(h) \operatorname*{KL}_{h,\gamma}\big(P^{\pi}_{\nu}\|P^{\pi}_{\xi}\big) \overset{(b)}{=} \sum_{t=n}^{\infty}\gamma_t \sum_{\nu\in\mathcal{M}} w_\nu(h) \sum_{h'\in\mathcal{H}^t} P^{\pi}_{\nu}(h'|h) \operatorname*{KL}_{hh',1}\big(P^{\pi}_{\nu}\|P^{\pi}_{\xi}\big)$$

$$\overset{(c)}{=} \sum_{t=n}^{\infty}\gamma_t \sum_{h'\in\mathcal{H}^t} P^{\pi}_{\xi}(h'|h) \sum_{\nu\in\mathcal{M}} w_\nu(hh') \operatorname*{KL}_{hh',1}\big(P^{\pi}_{\nu}\|P^{\pi}_{\xi}\big)$$

$$\overset{(d)}{\le} \sum_{t=n}^{\infty}\gamma_t \sum_{h'\in\mathcal{H}^t} P^{\pi}_{\xi}(h'|h) \log|\mathcal{H}| \overset{(e)}{\le} \log|\mathcal{H}| \sum_{t=n}^{\infty}\gamma_t$$

where (a) by definition of the value function and D $\le$ KL assumption, (b) is the definition of the discounted KL divergence, (c) from $w_\nu(h)P^{\pi}_{\nu}(h'|h) = w_\nu P^{\pi}_{\nu}(hh')/P^{\pi}_{\xi}(h) = w_\nu(hh')P^{\pi}_{\xi}(h'|h)$ by inserting the definition of $w_\nu(\cdot)$, (d) by Lemma 14 in the Appendix, and (e) since ξ is a measure. Therefore

$$\lim_{n\to\infty}\sup_{\pi\in\Pi}\sum_{h\in\mathcal{H}^n} P^{\pi}_{\xi}(h)V^{\pi}_{\gamma}(h) \le \lim_{n\to\infty}\log|\mathcal{H}|\sum_{t=n}^{\infty}\gamma_t = 0 \ ,$$

which is sufficient to guarantee the existence of the optimal policy [LH11b]. ■

Theorem 3. *For all policies π and discount vectors γ and $D \leq KL$, we obtain $V_\gamma^\pi \leq \mathrm{Ent}(w)$.*

Proof. The result follows from dominance $P_\xi^\pi(h) \geq w_\nu P_\nu^\pi(h)$ for all h and ν:

$$\begin{aligned} V_\gamma^\pi &\overset{\text{(a)}}{\leq} V_{\boldsymbol{\infty}}^\pi \overset{\text{(b)}}{=} \lim_{n\to\infty} V_{\mathbf{n}}^\pi \overset{\text{(c)}}{\leq} \lim_{n\to\infty} \sum_{\nu\in\mathcal{M}} w_\nu \operatorname*{KL}_{\epsilon,\mathbf{n}}\big(P_\nu^\pi \| P_\xi^\pi\big) \\ &\overset{\text{(d)}}{=} \lim_{n\to\infty} \sum_{\nu\in\mathcal{M}} w_\nu \sum_{h\in\mathcal{H}^n} P_\nu^\pi(h) \log \frac{P_\nu^\pi(h)}{P_\xi^\pi(h)} \\ &\overset{\text{(e)}}{\leq} \lim_{n\to\infty} \sum_{\nu\in\mathcal{M}} w_\nu \sum_{h\in\mathcal{H}^n} P_\nu^\pi(h) \log \frac{1}{w_\nu} \overset{\text{(f)}}{=} \lim_{n\to\infty} \sum_{\nu\in\mathcal{M}} w_\nu \log \frac{1}{w_\nu} \overset{\text{(g)}}{=} \mathrm{Ent}(w) \end{aligned}$$

where (a) is by the positivity of the KL divergence and because $\gamma_k \leq 1$ for all k, (b) is by the definitions of $\boldsymbol{\infty}$, $\mathbf{n}$ and the monotone convergence theorem, (c) by the definition of the value and assumption D $\leq$ KL, (d) by the definition of the value function and the telescoping property (2), (e) by the dominance $P_\xi^\pi(h) \geq w_\nu P_\nu^\pi(h)$ for all h and $\nu \in \mathcal{M}$, and (f) and (g) by the definitions of expectation and entropy respectively. ■

We have seen that a summable discount vector ensures the existence of both V_γ^* and π^*. This solution may not be entirely satisfying as it encourages the agent to sacrifice long-term information for short-term (but maybe less) information. If the entropy of the prior is finite, then the optimal value is guaranteed to be finite, but the optimal policy may still not exist as demonstrated in Section 5. In this case it is possible to construct a δ-optimal policy.

Definition 4. *The δ-optimal policy is given by*

$$\pi^{*,\delta} \in \left\{\pi : V_\gamma^\pi \geq V_\gamma^* - \delta\right\} \ . \tag{6}$$

where the choice within the set on the right-hand-side is made arbitrarily.

Note that if at some history h it holds that $V_\gamma^*(h) < \delta$, then the δ-optimal policy may cease exploring. Table 1 summarises the consequences on the existence of optimal values/policies based on the discount vector and entropy of the prior. For D = KL we name KL-KSA the agent defined by the optimal policy π^* and KL-KSA$_\delta$ the agent defined by policy $\pi^{*,\delta}$.

Table 1. Parameter choices for D = KL.

Discount γ	Entropy	$V_\gamma^* < \infty$	π^* exists	$\pi^{*,\delta}$exists	Myopic	Stops Exploring
$\sum_{t=1}^\infty \gamma_t < \infty$	$\mathrm{Ent}(w) < \infty$	yes	yes	yes	yes	no
	$\mathrm{Ent}(w) = \infty$	yes	yes	yes	yes	no
$\sum_{t=1}^\infty \gamma_t = \infty$	$\mathrm{Ent}(w) < \infty$	yes	no?	yes	no	yes?
	$\mathrm{Ent}(w) = \infty$	no	no	no	no	?

Learning. Before presenting the new theorem showing that π^* learns to predict off-policy, we present an easier on-policy result that holds for all policies.

Theorem 5 (On-policy prediction). *Let $\mu \in \mathcal{M}$ and π be a policy and γ a discount vector (possibly non-summable), then*

$$\lim_{n\to\infty} \Gamma_n^{-1} \mathbb{E}^\pi_\mu \underset{h_{1:n},\gamma}{\mathrm{KL}}\big(P^\pi_\mu \| P^\pi_\xi\big) = 0 \ .$$

The proof requires a small lemma. Note the normalising factor Γ_n^{-1} is used to prove a non-vacuous result for summable discount vectors.

Lemma 6. *The KL divergence satisfies a chain rule:*

$$\underset{\epsilon,\boldsymbol{\infty}}{\mathrm{KL}}\big(P^\pi_\nu \| P^\pi_\xi\big) = \underset{\epsilon,\mathbf{n}}{\mathrm{KL}}\big(P^\pi_\nu \| P^\pi_\xi\big) + \sum_{h\in\mathcal{H}^n} P^\pi_\nu(h) \underset{h,\boldsymbol{\infty}}{\mathrm{KL}}\big(P^\pi_\nu \| P^\pi_\xi\big) \ .$$

The proof is well-known and follows from definitions of expectation and properties of the logarithm.

Proof of Theorem 5.

$$\begin{aligned}
\lim_{n\to\infty} \Gamma_n^{-1} \mathbb{E}^\pi_\mu \underset{h_{1:n},\gamma}{\mathrm{KL}}\big(P^\pi_\mu \| P^\pi_\xi\big) &\overset{\text{(a)}}{\leq} \lim_{n\to\infty} \frac{1}{\Gamma_n w_\mu} \sum_{\nu\in\mathcal{M}} w_\nu \mathbb{E}^\pi_\nu \underset{h_{1:n},\gamma}{\mathrm{KL}}\big(P^\pi_\nu \| P^\pi_\xi\big) \\
&\overset{\text{(b)}}{\leq} \frac{1}{w_\mu} \lim_{n\to\infty} \sum_{\nu\in\mathcal{M}} w_\nu \mathbb{E}^\pi_\nu \underset{h_{1:n},\boldsymbol{\infty}}{\mathrm{KL}}\big(P^\pi_\nu \| P^\pi_\xi\big) \qquad (\star) \\
&\overset{\text{(c)}}{=} \frac{1}{w_\mu} \lim_{n\to\infty} \sum_{\nu\in\mathcal{M}} w_\nu \left(\underset{\epsilon,\boldsymbol{\infty}}{\mathrm{KL}}\big(P^\pi_\nu \| P^\pi_\xi\big) - \underset{\epsilon,\mathbf{n}}{\mathrm{KL}}\big(P^\pi_\nu \| P^\pi_\xi\big) \right) \overset{\text{(d)}}{\underset{n\to\infty}{\longrightarrow}} 0
\end{aligned}$$

where (a) follows by the positivity of the KL divergence and by introducing the sum, (b) since $\gamma_k/\Gamma_n \leq 1$ for all $k \geq n$, (c) by rearranging terms in Lemma 6, and (d) from well known $\mathrm{KL}_{\epsilon,\boldsymbol{\infty}}\Big(P^\pi_\nu \| P^\pi_\xi\Big) \leq \log \frac{1}{w_\nu} < \infty$. ■

Theorem 5 shows that $P^\pi_\xi(\cdot|h_{<t})$ converges in expectation to $P^\pi_\mu(\cdot|h_{<t})$ where the difference between the two measures is taken with respect to the expected cumulative discounted KL-divergence. This implies that $P^\pi_\xi(\cdot|h_{<t})$ is in expectation a good estimate for the unknown $P^\pi_\mu(\cdot|h_{<t})$.

The following result is perhaps the most important theoretical justification for the definition of π^*. We show that if $h_{1:\infty}$ is generated by following π^*, then $P^\pi_\xi(\cdot|h_{<n})$ converges in expectation to $P^\pi_\mu(\cdot|h_{<n})$ for all π. More informally, this means that as a longer history is observed the agent learns to predict the counterfactuals *"what would happen if I follow another policy π instead"*. For example, if the observation also included a reward signal, then the agent would asymptotically be able to learn (but not follow) the policy maximising the expected discounted reward. In fact, the policy maximising the Bayes-expected reward would converge to optimal. This kind of off-policy prediction is not usually satisfied by arbitrary policies where the agent can typically only learn what will happen on-policy in the sense of Theorem 5, not what would happen if it chose to follow another policy.

Theorem 7 (On-policy learning, off-policy prediction). *Let $\mu \in \mathcal{M}$ and γ be a discount vector (possibly non-summable). If π^* based on $D \leq KL$ exists,*

then

$$\lim_{n\to\infty} \Gamma_n^{-1} \mathbb{E}_\mu^{\pi^*} \sup_{\pi\in\Pi(h_{1:n})} \mathrm{D}_{h_{1:n},\gamma}\left(P_\mu^\pi \| P_\xi^\pi\right) = 0$$

where the expectation is taken over $h_{1:n}$.

Proof. We use the properties of π^* and the proof of Theorem 5:

$$\begin{aligned}
\Delta_n &:= \Gamma_n^{-1} \mathbb{E}_\mu^{\pi^*} \sup_{\pi\in\Pi(h_{1:n})} \mathrm{D}_{h_{1:n},\gamma}\left(P_\nu^\pi \| P_\xi^\pi\right) \\
&\overset{\text{(a)}}{\leq} \Gamma_n^{-1} \mathbb{E}_\mu^{\pi^*} \frac{1}{w_\mu(h_{1:n})} \sup_{\pi\in\Pi(h_{1:n})} \sum_{\nu\in\mathcal{M}} w_\nu(h_{1:n}) \mathrm{D}_{h_{1:n},\gamma}\left(P_\nu^\pi \| P_\xi^\pi\right) \\
&\overset{\text{(b)}}{=} \Gamma_n^{-1} \mathbb{E}_\mu^{\pi^*} \frac{1}{w_\mu(h_{1:n})} \sum_{\nu\in\mathcal{M}} w_\nu(h_{1:n}) \mathrm{D}_{h_{1:n},\gamma}\left(P_\nu^{\pi^*} \| P_\xi^{\pi^*}\right) \\
&\overset{\text{(c)}}{\leq} \frac{1}{\Gamma_n w_\mu} \mathbb{E}_\xi^{\pi^*} \sum_{\nu\in\mathcal{M}} w_\nu(h_{1:n}) \mathrm{D}_{h_{1:n},\gamma}\left(P_\nu^{\pi^*} \| P_\xi^{\pi^*}\right) \\
&\overset{\text{(d)}}{=} \frac{1}{\Gamma_n w_\mu} \sum_{\nu\in\mathcal{M}} w_\nu \mathbb{E}_\nu^{\pi^*} \mathrm{D}_{h_{1:n},\gamma}\left(P_\nu^{\pi^*} \| P_\xi^{\pi^*}\right) \\
&\overset{\text{(e)}}{\leq} \frac{1}{\Gamma_n w_\mu} \sum_{\nu\in\mathcal{M}} w_\nu \mathbb{E}_\nu^{\pi^*} \mathrm{KL}_{h_{1:n},\gamma}\left(P_\nu^{\pi^*} \| P_\xi^{\pi^*}\right)
\end{aligned}$$

where (a) follows from the positivity of the KL divergence, (b) because π^* is chosen to maximise the quantity inside the supremum for $n = 0$ and due to time consistency [LH11b] also for $n > 0$, (c) by the definition of $w_\mu(h_{1:n})$ and the definition of expectation, (d) by exchanging the sum and expectation and then using the definition of $w_\nu(h_{1:n})$ and the definition of expectation, and (e) by assumption $\mathrm{D} \leq \mathrm{KL}$. Combining the above with $(\star)$ for $\pi = \pi^*$ leads to

$$0 \leq \lim_{n\to\infty} \Delta_n \leq \lim_{n\to\infty} \frac{1}{\Gamma_n w_\mu} \sum_{\nu\in\mathcal{M}} w_\nu \mathbb{E}_\nu^{\pi^*} \mathrm{KL}_{h_{1:n},\gamma}\left(P_\mu^{\pi^*} \| P_\xi^{\pi^*}\right) \overset{(\star)}{=} 0$$

as required. ■

Deterministic case. Although KL-KSA is a new algorithm, it shares some similarities with Shannon-KSA [Ors11]. In particular, if $\mathcal{M}$ contains only deterministic environments, then up to technical details KL-KSA reduces to Shannon-KSA when the horizon m_t in [Ors11] is set to infinity in that paper:

Proposition 8. *When* $\mathcal{M}$ *contains only deterministic environments and D=KL, then*

$$V_\infty^* = \sup_{\pi\in\Pi} \lim_{n\to\infty} \sum_{h\in\mathcal{H}^n} P_\xi^\pi(h) \log \frac{1}{P_\xi^\pi(h)} \ . \tag{7}$$

The proof, omitted due to lack of space, follows from definitions and the fact that for fixed policy a deterministic environment concentrates on a single history.

Noise insensitivity. Let h be some finite history. A policy π is said to be uninformative if the conditional measure $P_{\nu_1}^\pi(\cdot|h) = P_{\nu_2}^\pi(\cdot|h)$ for all $\nu_1, \nu_2 \in$

$\mathcal{M}(h)$, which implies that $\mathrm{KL}_{h,\boldsymbol{\infty}}\left(P^{\pi}_{\nu_1} \| P^{\pi}_{\nu_2}\right) = 0$; that is, if the measure induced by π and $\nu \in \mathcal{M}(h)$ is independent of the choice of ν. A policy is informative if it is not uninformative. The following result is immediate from the definitions and shows that unlike Shannon-KSA and Square-KSA, KL-KSA always prefers informative policies over uninformative ones as demonstrated in the experiments in Section 5.

Proposition 9. *Suppose $\gamma_k > 0$ for all k. Then $V^{\pi}_{\gamma}(h) > 0$ if and only if π is informative.*

Avoiding traps. Theorem 7 implies that the agent tends to learn everything it can learn about its environment. Although this is a strong result, it cannot alone define scientific behaviour. In particular, the agent could jump knowingly into an inescapable trap (provided there is one) where the observations of the agent are no longer informative. Since it would have no possibility to acquire any more information about its environment, it would have converged to optimal behaviour in the sense of Theorem 7. After some history h, the agent is said to be in a trap if *all* policies after h are uninformative: It cannot gain any information, and cannot escape this situation. The following proposition is immediate from the definitions, and shows that π^* will not take actions leading surely to a trap unless there is no alternative:

Proposition 10. *$V^{*}_{\gamma}(h) = 0$ if the agent is in a trap after h.*

A deterministic trap is a trap where observations are deterministic depending on the history. Since for deterministic environments Shannon-KSA and KL-KSA are identical, Shannon-KSA avoids jumping into deterministic traps (see experiments in Section 5) but, unlike KL-KSA, it may not avoid stochastic ones, *i.e.* traps with noise. Note that KL-KSA may still end up in a trap, *e.g.* if it has low probability or if it is unavoidable.

4 Choosing $\mathcal{M}$ and w

Until now we have ignored the question of choosing the environment class $\mathcal{M}$ and prior w. Since our aim is to construct an agent that is as explorative as possible we should choose $\mathcal{M}$ as large as possible. By the (strong) Church-Turing thesis we assume that the universe is computable and so the most natural choice for $\mathcal{M}$ is the set of all (semi-)computable environments $\mathcal{M}_U$ exactly as used by [Hut05], but with rewards ignored. To choose the prior we follow [Hut05] and combine Epicurus principle of multiple explanations and Occam's razor, to define $w_\nu := 2^{-K(\nu)}$ where $K(\nu)$ is the prefix Kolmogorov complexity of ν. This prior has several important properties. First, except for a constant multiplicative factor it assigns more weight to every environment than any other semi-computable prior [LV08]. Secondly, it satisfies the maximum entropy principle as demonstrated by the following theorem.

Proposition 11. *If $\mathcal{M} = \mathcal{M}_U$, then $\sum_{\nu\in\mathcal{M}} 2^{-K(\nu)} K(\nu) = \infty$.*

The proof follows from a straight-forward adaptation of [LV08, Ex. 4.3.4]. Unfortunately, this result can also be used to show that $V^*_{\boldsymbol{\infty}} = \infty$.

Proposition 12. *If $D = KL$, $\mathcal{M}$ contains all computable deterministic environments and $w_\nu = 2^{-K(\nu)}$, then $V^*_{\boldsymbol{\infty}} = \infty$.*

Proof. Assume without loss of generality that $|\mathcal{A}| = 1$, $\mathcal{O} = \{0, 1\}$ and $\pi = \pi^*$ is the only possible policy. Then we drop the dependence on actions and view history sequences as sequences of observations. Let $k \in \mathbb{N}$ and define environment ν_k to deterministically generate observation 0 until time-step k followed by observation 1 for all subsequent time-steps. It is straightforward to check that there exists a $c_1 \in \mathbb{R}$ such that $K(\nu_k) < K(k) + c_1$ for all $k \in \mathbb{N}$. By simple properties of the Kolmogorov complexity and [LV08, Ex.4.5.2] we have that there exist constants $c_i \in \mathbb{R}$ such that

$$\begin{aligned} -\log P^\pi_\xi(0^k 1^\infty) \geq -\log P^\pi_\xi(0^k 1) &> K(0^k 1) - 2\log K(0^k 1) + c_2 \\ &> K(k) - 2\log K(k) + c_3 > \tfrac{1}{2}K(k) - c_4 \ . \end{aligned}$$

Then

$$\begin{aligned} V^*_{\boldsymbol{\infty}} &\overset{\text{(a)}}{=} \sum_{\nu\in\mathcal{M}} w_\nu \underset{\epsilon,\boldsymbol{\infty}}{\mathrm{KL}}\big(P^\pi_\nu \| P^\pi_\xi\big) \overset{\text{(b)}}{\geq} \sum_{k\in\mathbb{N}} w_{\nu_k} \underset{\epsilon,\boldsymbol{\infty}}{\mathrm{KL}}\big(P^\pi_{\nu_k} \| P^\pi_\xi\big) \overset{\text{(c)}}{=} \sum_{k\in\mathbb{N}} 2^{-K(\nu_k)} \log \frac{1}{P^\pi_\xi(0^k 1^\infty)} \\ &\overset{\text{(d)}}{\geq} 2^{-c_1-1} \sum_{k\in\mathbb{N}} 2^{-K(k)} K(k) - 2^{-c_1} c_4 \sum_{k\in\mathbb{N}} 2^{-K(k)} \overset{\text{(e)}}{=} \infty - O(1) \end{aligned}$$

where (a) is the definition of the value function, (b) follows by dropping all environments except ν_k for $k \in \mathbb{N}$, (c) by substituting the definitions of the KL divergence and the prior and noting that ν_k is deterministic, (d) by the bounds in the previous display, and (e) by the well known fact that $\sum_{k\in\mathbb{N}} 2^{-K(k)} K(k) = \infty$ analogous to Proposition 11. ■

To avoid this problem the prior may be biased further towards simplicity by defining $w_\nu := 2^{-(1+\varepsilon)K(\nu)}$ where $\varepsilon \ll 1$ is chosen very small.

Proposition 13. *For all $\varepsilon > 0$, $\sum_{\nu\in\mathcal{M}} 2^{-(1+\varepsilon)K(\nu)}(1+\varepsilon)K(\nu) < \infty$.*

Proof. For each $k \in \mathbb{N}$, define $\mathcal{M}_k := \{\nu \in \mathcal{M} : K(\nu) = k\}$. The number of programs is bounded by $|\mathcal{M}_k| \leq 2^k$, thus we have

$$\begin{aligned} \sum_{\nu\in\mathcal{M}} 2^{-(1+\varepsilon)K(\nu)}(1+\varepsilon)K(\nu) &= \sum_{k=1}^{\infty} \sum_{\nu\in\mathcal{M}_k} 2^{-(1+\varepsilon)K(\nu)}(1+\varepsilon)K(\nu) \\ &\leq \sum_{k=1}^{\infty} 2^k 2^{-(1+\varepsilon)k}(1+\varepsilon)k = \sum_{k=1}^{\infty} 2^{-\varepsilon k}(1+\varepsilon)k < \infty \end{aligned}$$

as required. ■

Therefore, if we choose $w_\nu := 2^{-(1+\varepsilon)K(\nu)}$, then $\mathrm{Ent}(w) < \infty$ and so $V^*_{\boldsymbol{\infty}}(h) < \infty$ by Theorem 3. Unfortunately, this approach introduces an arbitrary parameter ε for which there seems to be no well-motivated single choice. Worse, the finiteness of $V^*_{\boldsymbol{\infty}}$ is by itself insufficient to ensure the existence of π^* for $\gamma = \boldsymbol{\infty}$. The issue

is circumvented by using a δ-optimal policy for some arbitrarily small δ, which introduces another parameter.

5 Experiments and Examples

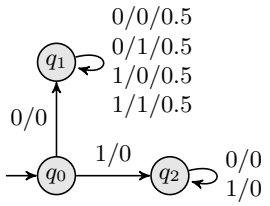
To give the intuition that the agent KL-KSA behaves according to expectation, we present a variety of toy experiments in particular situations. For each experiment we choose $\mathcal{M}$ to be a finite set of (possibly stochastic) environments, which will typically be representable as partially observable MDPs, but may occasionally be non-Markovian. Although the definitions of the environments are (mostly) finite state automata, they are not MDPs, as the agent receives only the observations and not the current state. The action and observation sets are $\mathcal{A} = \mathcal{O} = \{0, 1\}$. Unless otherwise stated, for each environment ν we set $w'_\nu := 1$, before normalisation to give the prior $w_\nu := w'_\nu / \sum_\nu w'_\nu$.

The horizon of the agents is the length n of the action sequences under consideration, *i.e.* the agents must maximise information gain in n steps. For KL-KSA, we thus use $V_{\mathbf{n}}^{\pi^*}(\epsilon)$, the agent Shannon-KSA is defined similarly by Equation (7), and the agent Square-KSA is defined by the policy [Ors11]

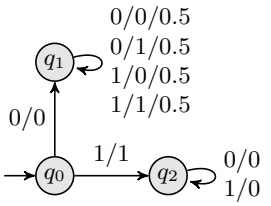
$$\pi_{\text{Square-KSA}} := \arg\max_\pi \sum_{h \in \mathcal{H}^n} -P_\xi^\pi(h)^2 \ .$$

Noise insensitivity. The first experiment shows that unlike Square-KSA and Shannon-KSA, KL-KSA is resistant to noise. Consider the two environments in Figure 1. The only difference between the two of them is that when the agent takes action 1 in state q_0, it receives observation either 0 or 1. The values of different actions sequences for Square-KSA, Shannon-KSA and KL-KSA are summarised in the table.

Fig. 1. Noisy environments μ_1 and μ_2: Edge labels are written action/observation/probability. The probability is omitted if it is 1. Here, only action 1 in q_0 is actually informative. The table contains values of various action sequences for various agents in $\mathcal{M} = \{\mu_1, \mu_2\}$.



(a) μ_1



(b) μ_2

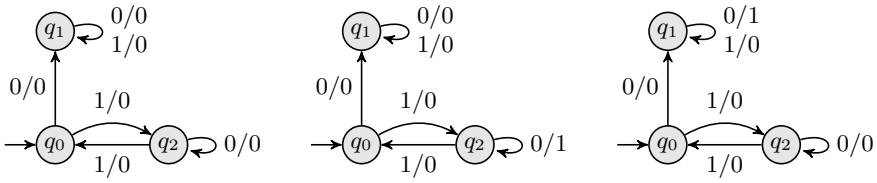
	Value V_γ^π of agent		
Actions	Square-KSA	Sh.-KSA	KL-KSA
00	- 0.5	1	0
000	- 0.25	2	0
0000	**- 0.125**	**3**	0
11	- 0.5	1	1
111	- 0.5	1	1
1111	- 0.5	1	**1**

We see that for Square-KSA and Shannon-KSA, each time a stochastic observation is received, the value of the action sequence increases. In particular, Shannon-KSA (wrongly) estimates a gain of 1 bit of information each time it observes a coin toss. Thus they both tend to follow actions that lead to stochastic observations.

On the other hand, KL-KSA always prefers to go to q_2, in order to gain information about which environment is the true one, and considers that it can only gain one bit of information, whatever the length of the action sequence of 1s. This shows its noise insensitivity, which makes it not interested in observing coin tosses. Note that KL-KSA's value does not depend here on the length of the horizon, and thus behaves likewise with an infinite horizon.

Trap avoidance. We now show a situation where KL-KSA avoids jumping into a trap if it can gain more information before doing so. Note that Square-KSA and Shannon-KSA behave similarly in these experiments, which rely only on deterministic environments. The environments are described in Figure 2 and the results are summarised in the left half of Table 2. We consider actions sequences of length 5. Increasing this number, even to infinity, does not change the results.

Fig. 2. Environments μ_3, μ_4, μ_5. The trap is in q_1, where the agent eventually cannot separate μ_3 and μ_4.



(a) μ_3 (b) μ_4 (c) μ_5

Table 2. Values of various action sequences for various KSA agents in the trap environments.

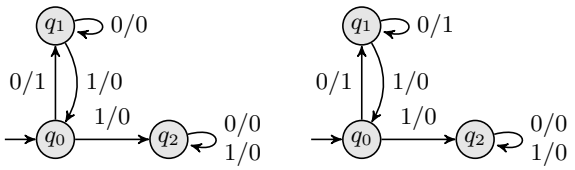
	$\mathcal{M} = \{\mu_3, \mu_4, \mu_5\}$			$\mathcal{M} = \{\mu_3..\mu_7\}$		
Actions	Square	Shannon	KL	Square	Shannon	KL
11111	- 1	0	0	- 1	0	0
01111	- 1	0	0	- 0.987	0.057	0.057
10000	- 0.556	0.918	0.918	- 0.557	0.916	0.916
00000	- 0.556	0.918	0.918	- 0.548	0.976	0.976
10100	**- 0.333**	**1.585**	**1.585**	**- 0.333**	**1.585**	**1.585**

Remarks:

- We see that Shannon-KSA and KL-KSA have the same values in classes of deterministic environments, as per Theorem 8.
- All agents prefer action 10100, which has the highest value among all the 2^5 possible action sequences of length 5, and allows the agents to identify the true environment with certainty.
- The trap in q_1 is initially avoided, in order to first gain information about the rest of the environment.
- All agents still go into q_1 in the end, because this allows them to separate μ_3 and μ_4 from μ_5, which is why action 00000 still has a relatively high value.
- Action sequence 11111 brings no information at all, since all environments would output the same observations, and would thus not be separated.

Getting caught in a trap. In addition to the environments of the last subsection, let us consider two environments μ_6 and μ_7 shown in Figure 3 of low weight $w'_\nu := 0.01$ (before normalisation). These two environments are thus very improbable compared to the other 3 environments. This low weight reflects either some preference (low prior, *e.g.* based on the complexity of the environments), or the fact that these environments have been made less probable (low posterior) after some hypothetical interaction history. The results are summarised in the right half of Table 2.

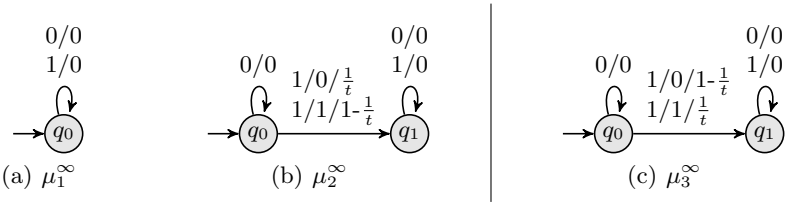
Fig. 3. Environments μ_6 and μ_7 with trap in q_2, and only differ in q_1.



Among the 5 environments, if μ_6 is actually the true environment, by doing action 1 as is suggested by the action sequence of optimal value, the agent immediately gets caught in a trap, and will never be able to separate the three environment μ_6, μ_7 and μ_3. Since if the agent chose to start with action 0 instead to not be caught in μ_6 and μ_7's trap it would get caught in the trap of the 3 other environments, it has to make a choice, based on the current weights of the environments. In contrast, if we take $\mathcal{M} = \{\mu_3, \mu_6, \mu_7\}$, one of the optimal action sequences of length 2 is 00 (of value 0.352), which first action first discards either μ_3 or both μ_6 and μ_7, and in case of the latter, the second action discards one of the two remaining environments.

Non-existence of π^* for $\gamma = \infty$. Consider the environments in Figure 4. When $\mathcal{M} = \{\mu_1^\infty, \mu_2^\infty\}$, the longer the agent stays in q_0 by taking action 0, the higher the probability that taking action 1 will lead to a gain of information. Taking the limit of this policy makes the agent stay in q_0 for ever, and actually never gain information. This means that the optimal policy π^* for $\gamma = \infty$ does

Fig. 4. Environments μ_1^∞, μ_2^∞ and μ_3^∞. The transition probability may depend on the time step number t. If $\mathcal{M} = \{\mu_1^\infty, \mu_2^\infty\}$, the optimal non-discounted policy is to remain in q_0 for ever, in order to increase the probability of gaining information when eventually choosing action 1.



not exist for $\mathcal{M} = \{\mu_1^\infty, \mu_2^\infty\}$, and here we must either use a summable discount vector or KL-KSA$_\delta$ with a prior of finite entropy. Note that this example alone is however not sufficient to prove the non-existence of the optimal policy for the

case where $\mathcal{M} = \mathcal{M}_U$ contains all computable (semi-)measures, since $\mathcal{M}$ must then also necessarily contain μ_3^∞, of complexity roughly equal to that of μ_2^∞. For these three environments, the optimal policy is now actually to start with action 1 instead of postponing it, because at least 2 environments cannot be separated, but action 1 separates μ_1^∞ and μ_3^∞ with certainty.

6 Conclusion

We extended previous work on knowledge-seeking agents [Ors11] by generalising from deterministic classes to the full stochastic case. As far as we are aware this is the first definition of a universal knowledge-seeking agent for this very general setting.

We gave a convergence result by showing that KL-KSA learns the true environment in the sense that it learns to predict the consequences of any future actions, even the counterfactual actions it ultimately chooses not to take. Furthermore, this new agent has been shown to be resistant to non-informative noise and, where reasonable, avoid traps from which it cannot escape.

One important concern lies in the choice of parameters/discount vector. If discounting is not used and the prior has infinite entropy, then the value function may be infinite and even approximately optimal policies do not exist. If the prior has finite entropy, then the value function is uniformly bounded and approximately optimal policies exist. For universal environment classes this precludes the use of the universal prior as it has infinite entropy.

An alternative is to use a summable discount vector. In this case optimal policies exist, but the knowledge seeking agent may be somewhat myopic. We are not currently convinced which option is best: the approximately optimal undiscounted agent that may eventually cease exploring, or the optimal discounted agent that is myopic.

References

[BO13] A. Baranes and P.-Y. Oudeyer. Active Learning of Inverse Models with Intrinsically Motivated Goal Exploration in Robots. *Robotics and Autonomous Systems*, 61(1):69–73, 2013.

[Hut05] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, 2005.

[LH11a] T. Lattimore and M. Hutter. Asymptotically optimal agents. In *Algorithmic Learning Theory (ALT)*, volume 6925 of *LNAI*, pages 368–382, Espoo, Finland, 2011. Springer.

[LH11b] T. Lattimore and M. Hutter. Time Consistent Discounting. In *Algorithmic Learning Theory*, volume 6925 of *LNAI*, pages 383–397. Springer, Berlin, 2011.

[LV08] M. Li and P.M.B. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer-Verlag, New York, 3rd edition, 2008.

[Ors11] L. Orseau. Universal Knowledge-Seeking Agents. In *Algorithmic Learning Theory (ALT)*, volume 6925 of *LNAI*, pages 353–367, Espoo, Finland, 2011. Springer.

[Ors13] L. Orseau. Asymptotic non-learnability of universal agents with computable horizon functions. *Theoretical Computer Science*, 473:149 – 156, 2013.

[RH11] S. Rathmanner and M. Hutter. A philosophical treatise of universal induction. *Entropy*, 13(6):1076–1136, 2011.

[SB98] R. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

[Sch06] J. Schmidhuber. Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts. *Connection Science*, 18(2):173–188, 2006.

[SGS11] Y. Sun, F. Gomez, and J. Schmidhuber. Planning to Be Surprised: Optimal Bayesian Exploration in Dynamic Environments. In *Artificial General Intelligence*, volume 6830 of *Lecture Notes in Computer Science*, pages 41–51. Springer Berlin Heidelberg, 2011.

[SHS95] J. Storck, S. Hochreiter, and J. Schmidhuber. Reinforcement driven information acquisition in non-deterministic environments. In *Proceedings of the International Conference on Artificial Neural Networks, Paris*, volume 2, pages 159–164. EC2 & Cie, 1995.

[Sol78] R. Solomonoff. Complexity-based induction systems: comparisons and convergence theorems. *Information Theory, IEEE Transactions on*, 24(4):422–432, 1978.

A Technical Results

Lemma 14. *Let $\mathcal{M}$ be a countable set of distributions on finite space X and $w : \mathcal{M} \to [0,1]$ be a distribution on $\mathcal{M}$. If $\xi(x) := \sum_{\rho\in\mathcal{M}} w_\rho \rho(x)$, then*

$$\sum_{\rho\in\mathcal{M}} w_\rho \sum_{x\in X} \rho(x) \log\left(\rho(x)/\xi(x)\right) \leq \log|X| \ .$$

Proof. We use properties of the KL divergence. Define distribution $R(x) := 1/|X|$. Then

$$\sum_{\rho\in\mathcal{M}} w_\rho \sum_{x\in X} \rho(x) \log \frac{\rho(x)}{\xi(x)} \overset{\text{(a)}}{\leq} \sum_{x\in X} \sum_{\rho\in\mathcal{M}} w_\rho \rho(x) \log \frac{1}{\xi(x)} \overset{\text{(b)}}{=} \sum_{x\in X} \xi(x) \log \frac{1}{\xi(x)} \overset{\text{(c)}}{\leq} \log|X|$$

where (a) follows from monotonicity of log and $\rho(x) \leq 1$. (b) by definition of ξ and (c) by Gibb's inequality. ■