Using Localization and Factorization to Reduce the Complexity of Reinforcement Learning

Peter Sunehag[1,2] and Marcus Hutter[1]
Sunehag@google.com, Marcus.Hutter@anu.edu.au

[1] Research School of Computer Science, Australian National University
Canberra Australia

[2] Google - DeepMind
London, UK

Abstract. General reinforcement learning is a powerful framework for artificial intelligence that has seen much theoretical progress since introduced fifteen years ago. We have previously provided guarantees for cases with finitely many possible environments. Though the results are the best possible in general, a linear dependence on the size of the hypothesis class renders them impractical. However, we dramatically improved on these by introducing the concept of environments generated by combining laws. The bounds are then linear in the number of laws needed to generate the environment class. This number is identified as a natural complexity measure for classes of environments. The individual law might only predict some feature (factorization) and only in some contexts (localization). We here extend previous deterministic results to the important stochastic setting.

Keywords: reinforcement learning, laws, optimism, bounds

1 Introduction

General reinforcement learning [2, 3, 12] is a theoretical foundation for artificial intelligence that has now been developed over the last fifteen years. A recent line of work starting with [8, 9] has studied finite classes of completely general environments and primarily optimistic agents that can be proven to eventually achieve optimality regardless of which environment turns out to be true. [8] presented finite-error bounds for the deterministic case and asymptotic guarantees for stochastic environments while [5] proved near-optimal sample complexity bounds for the latter stochastic case.

The bounds given in [8] have a linear dependence on the number of environments in the class. While this rate is easily seen to be the best one can have in general [5], it is exponentially worse than what we are used to from Markov Decision Processes (MDPs) [4] where the linear (up to logarithms) dependence is on the size of the state space instead. In [10] we introduced the concept of deterministic laws that predict some but not all features (factorization) and only in some contexts (localization), and environments generated by sets of such laws. We presented bounds that are linear in the number of laws instead of the number of environments. All deterministic environment classes are trivially generated by sets of laws that equal the environments but some can also be generated by exponentially fewer laws than there are environments.

We here expand the formal analysis of optimistic agents with hypothesis classes based on laws, from the deterministic to the stochastic case and we further consider fruitful combinations of those two basic cases.

Outline. Section 2 provides background on general reinforcement learning agents. Section 3 introduces the concept of environments generated by laws and extends previous concepts and results from the determinstic to the stochastic case as well as to the mixed setting. Section 4 concludes.

2 Background

We begin by introducing general reinforcement learning as well as the agent framework.

2.1 General reinforcement learning

We will consider an agent [6, 2] 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 $V^\pi_\nu(h_{t-1}) = \sum_{i=t}^\infty \gamma^{i-t} r_i$ 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 $V^\pi_\nu(h_{t-1}) := \mathbb{E}_{\nu(\cdot|\pi,h_{t-1})} \sum_{i=t}^\infty \gamma^{i-t} r_i := \lim_{j\to\infty} \mathbb{E}_{\nu(\cdot|\pi,h_{t-1})} \sum_{i=t}^j \gamma^{i-t} r_i$ and $V^*_\nu(h_{t-1}) := \max_\pi V^\pi_\nu(h_{t-1})$.

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 [2]. The above agent, and only agents of that form, satisfies the strict rationality axioms presented first in [7] while the slightly looser version we presented in [9] enables optimism. The optimist chooses its next action based on

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

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 [11]. We will rely on an agent framework presented in [11].

2.2 Agents based on decision functions and hypothesis generating 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. We are here primarily using optimistic decision functions.

Definition 1 (Optimistic decision function). *We call a decision function f optimistic if $f(\mathcal{M}) = \pi(\epsilon)$ for an optimistic policy π, i.e. for*

$$\pi \in \arg\max_{\tilde{\pi}} \max_{\nu \in \mathcal{M}} V_\nu^{\tilde{\pi}}. \tag{3}$$

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.

3 Environments defined by laws

We consider observations of the form of a feature vector $o = \boldsymbol{x} = (x_j)_{j=1}^m \in \mathcal{O} = \times_{j=1}^m \mathcal{O}_j$ including the reward as one coefficient where x_j is an element of some finite alphabet $\mathcal{O}_j$. 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.

Definition 2 (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. We below expand these notions, defined in [10, 11], from deterministic laws to stochastic laws.

Definition 3 (Stochastic law). *A stochastic law is a function $\tau : \mathcal{H} \times \mathcal{A} \times \mathcal{O}_\perp \to [0,1]$ such that*

$$\forall h \forall a \sum_{o \in \mathcal{O}_\perp} \tau(h,a,o) = 1$$

and

$$\forall h \forall a \forall j \in \{1,...,m\} \sum_{o \in \mathcal{O}_\perp : o_j = \perp} \tau(h,a,o) \in \{0,1\},$$

i.e. the marginal probability of the "no prediction" symbol $\perp$ always equals zero or one. We will use the notation $\tau(o|h,a) := \tau(h,a,o)$.

Definition 4 (Stochastic laws making predictions or not). *If τ is a law and*

$$\sum_{o \in \mathcal{O}_\perp : o_j = \perp} \tau(h,a,o) = 0$$

we say that τ does not make a prediction for j given h, a and write $\tau(h,a)_j = \perp$. Otherwise, i.e. when

$$\sum_{o \in \mathcal{O}_\perp : o_j = \perp} \tau(h,a,o) = 1,$$

we say that τ does make a prediction for j given h, a and write $\tau(h,a)_j \neq \perp$.

As in the deterministic case we need to define what it means for a set of stochastic laws to be complete and then we can define an environment from such a set. The definition is an extension of the deterministic counter-part. That we only demand completeness and not coherence in the stochastic case is because we are going to study the stochastic case with a domination assumption instead of excluding laws. The result is that the generated class is infinite even when the set of laws is finite.

Definition 5 (Complete set of stochastic laws). *A set $\mathcal{T}$ of stochastic laws is complete if*

$$\forall h, a \ \exists \tau_i \in \mathcal{T} \ \exists J_i \subset \{1,...,m\} = \dot\cup_i J_i : \tau_i(h,a)_j \neq \perp \iff j \in J_i.$$

Let $\hat{\mathcal{C}}(\mathcal{T})$ denote the set of complete subsets of $\mathcal{T}$.

Definition 6 (Environments from stochastic laws). *Given a complete class of stochastic laws $\mathcal{T}$, we define the class of environments $\Xi(\mathcal{T})$ generated by $\mathcal{T}$ as consisting of all ν for which there are τ_i and J_i as in Definition 5 such that*

$$\nu(\boldsymbol{x}|h,a) = \Pi_i \tau_i|_{J_i}(h,a)(\boldsymbol{x}|_{J_i}).$$

Error analysis. We first consider deterministic environments and deterministic laws and the optimistic agent from [8]. 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 [8] to hold but with $|\mathcal{M}|$ replaced by $|\mathcal{T}|$ (see Theorem 1 below) which can be exponentially smaller. We have presented this result previously [10, 11] but here we extend from the deterministic to stochastic settings.

Theorem 1 (Finite error bound when using laws). *Suppose that $\mathcal{T}$ is a finite class of deterministic laws and let $\mathcal{G}(h) = \{\nu(\cdot|h) \mid \nu \in \mathcal{M}(\{\tau|\ \tau \in \mathcal{T}$ consistent with $h\})\}$. We define $\bar{\pi}$ by combining $\mathcal{G}$ with the optimistic decision function (Definition 1). 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*

$$V_\mu^{\bar{\pi}}(h_t) \geq \max_\pi V_\mu^\pi(h_t) - \varepsilon \tag{4}$$

for all but at most $|\mathcal{T}|\frac{-\log\varepsilon(1-\gamma)}{1-\gamma}$ time steps t.

We now introduce optimistic agents with classes of stochastic dominant laws. To define what dominant means for a law we first introduce the notion of a restriction. We will say that a law τ is a restriction of a stochastic environment ν if it assigns the same probabilities to what τ predicts. We then also say that ν is an extension of τ. Similarly a law can be a restriction or an extension of another law. If τ is a restriction of some environment ν that μ is absolutely continuous w.r.t. (for every policy), then we say that μ is absolutely continuous (for every policy) with respect to τ. We here make use of the slightly more restrictive notion of dominance. We say that ν dominates μ if there is $c > 0$ such that $\nu(\cdot) \geq c\mu(\cdot)$. We extend this concept to laws.

Example 1 (Stochastic laws based on estimators). Consider again a binary vector of length m where each coefficient is an i.i.d. Bernoulli process, i.e. there is a fixed probability with which the coefficient equals 1. Consider laws that are such that there is one for each coefficient and they predict a 1 with probability $\frac{a+1/2}{a+b+1}$ where a is the number of 1s that have occurred before for that coefficient and b is the number of 0s. Then we have a complete set of stochastic laws that are based on the so called Krichevsky-Trofimov (KT) estimator. Also, they satisfy the absolute continuity property. These laws can e.g. be combined with laws based on the Laplace estimator which assigns probability $\frac{a+1}{a+b+2}$ instead.

Example 2 (Dominant laws, AIXI-CTW). Consider the AIXI agent defined by (1) with ξ being the mixture of all context tree environments up to a certain depth as defined in [13]. A context is defined by a condition on what the last few cycles of the history is. The context tree contains contexts of variable length upto the maximum depth. The Context Tree Weighting (CTW) algorithm relied on by [13], which is originally from [14], defines a prediction for each context using a Krichevsky-Trofimov estimator. ξ is a mixture of all of those predictions. Given a context, we can define a law as the restriction of ξ to the histories for which we are in the given context. All of these laws will be absolutely continuous for any context tree environment, hence so are all of these laws. If we consider the same restrictions for other dominant mixtures than ξ, e.g. by using the CTW construction on other/all possible binarizations of the environment, we have defined a large set of laws.

Theorem 2 (Convergence for stochastic laws). *Suppose that $\mathcal{T}$ is a finite class of stochastic laws as in Definition 6 and that they all are absolutely continuous w.r.t. the true environment μ and that for every h, there is an environment $\nu_h \in \Xi(\mathcal{T})$ such that $V^*_{\nu_h}(h) \geq V^*_\mu(h)$. Let $\mathcal{G}(h) = \{\nu(\cdot|h) \mid \nu \in \Xi(\mathcal{T})\}$. We define $\tilde{\pi}$ by combining $\mathcal{G}$ with an optimistic decision function. Then almost surely $V^{\tilde{\pi}}_\mu(h_t) \to V^*_\mu(h_t)$ as $t \to \infty$.*

Proof. Any $\nu \in \Xi(\mathcal{T})$ is such that $\nu(\cdot) \geq c\mu(\cdot)$ where c is the smallest constant such that all the laws in $\mathcal{T}$ are dominant with that constant. For each law $\tau \in \mathcal{T}$ pick an environment $\nu \in \Xi(\mathcal{T})$ such that τ is a restriction of ν, i.e. ν predicts according to τ whenever τ predicts something. We use the notation ν_τ for the environment chosen for τ. The Blackwell-Dubins Theorem says that ν_τ merges with μ almost surely under the policy followed (but not necessarily off that policy) and therefore τ merges with μ, i.e. with the restriction of μ to what τ makes predictions for, under the followed policy. Given $\varepsilon > 0$, let T be such that

$$\forall t \geq T : \max_{\tau \in \mathcal{T}} d(\nu_\tau(\cdot|h_t, \tilde{\pi}), \mu(\cdot|\tilde{\pi})) < \varepsilon$$

which implies that

$$\forall t \geq T : \max_{\nu \in \Xi(\mathcal{T})} d(\nu(\cdot|h_t, \tilde{\pi}), \mu(\cdot|\tilde{\pi})) < \varepsilon$$

and applying this to ν_{h_t} proves that $|V^{\tilde{\pi}}_\mu(h_t) - V^*_\mu(h_t)| < \varepsilon \; \forall t \geq T$ by Lemma 1 in [9]. Since there is, almost surely, such a T for every $\varepsilon > 0$ the claim is proved. ■

Excluding stochastic laws and sample complexity. To prove sample complexity bounds one typically needs to assume that the truth belongs to the class which is stronger than assuming domination. This agent would need to exclude implausible environments from the class. In the deterministic case that can be done with certainty after one contradiction, while [1] shows that in the stochastic case this can be done after a finite number m of sufficiently large contradiction. m depends on the confidence required, $m = O(\frac{1}{\varepsilon^2} \log \frac{k}{\delta}$ where ε is the accuracy, δ the confidence and k the number of hypothesis, and after m disagreements the environment that aligned worse with observations is excluded. The analysis closely follows the structure learning case in [1] where it relies on a more general theorem for predictions based on k possible algorithms. The main difference is that that they could do this per feature which we cannot since we are in a much more general setting where a law sometimes makes a prediction for a feature and sometimes not. One can have at most mk^2 disagreements (actually slightly fewer) where k is the number of laws. It is possible that this square dependence can be improved to linear, but it is already an exponential improvement for many cases compared to a linear dependence on the number of environments. There can only be errors when there is sufficient disagreement. The above argument works under a coherence assumption and for $\gamma = 0$ while for $\gamma > 0$ there are horizon effects that adds extra technical difficulty to proving optimal bounds avoiding losing a factor $1/(1-\gamma)$. [5] shows how such complications can be dealt with.

Having a background environment. The earlier deterministic results demanded that the set of laws in the class is rich enough to combine into complete environments and in particular to the true one. This might require such a large class of laws that the linear dependence on the number of laws in the error bound, though much better than depending on the number of environments, still is large. The problem is simplified if the agent has access to a background environment, which is here something that given previous history and the next features predicted by laws, assigns probabilities for the rest of the feature vector. A further purpose for this section is to prepare for classes with a mix of deterministic laws and stochastic laws. In this case the stochastic laws learn what we in this section call a background environment. Computer games provide a simple example where it is typically clear that we have a background and then objects. If the agent has already learnt a model of the background, then what remains is only the subproblem of finding laws related to how objects behave and affect the environment. As an alternative, we might not be able to deterministically predict the objects but we can learn a cruder probabilistic model for them and this is background that completes the deterministic world model the agent learns for the rest.

Example 3 (Semi-deterministic environment). Consider a binary vector of length m where some elements are fixed and some fluctuate randomly with probability 1/2. Consider the background environment where all coefficients are Bernoulli processes with probability 1/2 and consider the $2m$ laws that each always makes a deterministic prediction for one coefficient and it is fixed. The laws that make a prediction for a fluctuating coefficient will quickly get excluded and then the agent will have learnt the environment.

Definition 7 (Predicted and not predicted features). *Given a set of deterministic laws $\mathcal{T}$, let*

$$q_1(h,a,\mathcal{T}) := \{j \in \{1,...,m\} \mid \nu(h,a)_j = \bot \ \forall \nu \in \Xi(\mathcal{T})\}$$

be the features $\mathcal{T}$ cannot predict and $q_2(h,a,\mathcal{T}) := \{1,...,m\} \setminus q_1(h,a,\mathcal{T})$ the predicted features.

Since we are now working with sets of laws that are not complete, subsets can also not be complete, but they can be maximal in the sense that they predict all that any law in the full set predicts.

Definition 8 (Coherent and maximal sets of laws). *Given a set of deterministic laws, the set of maximal subsets of laws $\bar{\mathcal{C}}(\mathcal{T})$ consists of sets $\tilde{\mathcal{T}} \subset \mathcal{T}$ with the property*

$$\forall h, a \forall j \in q_2(h, a, \mathcal{T}) \exists \tau \in \tilde{\mathcal{T}} : \tau(h,a)_j \neq \bot.$$

If

$$\forall h, a \forall j \in q_2(h, a, \mathcal{T}) \forall \tau, \tilde{\tau} \in \tilde{\mathcal{T}}\ \tilde{\tau}(h,a)_j \in \{\bot, \tau(h,a)_j\}$$

we say that $\tilde{\mathcal{T}}$ is coherent.

A semi-deterministic environment is defined by combining the predictions of a number of laws with background probabilities for what the laws do not predict. We abuse notation by letting $\nu(h,a) = (o,r)$ mean that ν assigns probability 1 to the next observation and reward being (o,r). We then also let $\nu(h,a)$ represent the event predicted. As before, we use x_k to denote individual features.

Definition 9 (Semi-deterministic environment). *Given a coherent set of laws $\tilde{\mathcal{T}}$ and background probabilities $P(\boldsymbol{x}|x_{k_1},...,x_{k_n},h)$ where $\boldsymbol{x} = (x_1,...,x_m)$ for any subset $\{k_1,...,k_n\} \subset \{1,...,m\}$ of the features and previous history h, we let $\nu(P,\tilde{\mathcal{T}})$ be the environment ν which is such that*

$$\forall h, a \forall j \in q_2(h, a, \mathcal{T}) \exists \tau \in \tilde{\mathcal{T}} : \nu(h,a)_j = \tau(h,a)_j$$

and

$$\nu\big(\boldsymbol{x} \mid h, a,\ \boldsymbol{x}|_{q_2(h,a,\mathcal{T})} = \nu(h,a)|_{q_2(h,a,\mathcal{T})}\big) = P\big(\boldsymbol{x} \mid \boldsymbol{x}|_{q_2(h,a,\mathcal{T})} = \nu(h,a)_{q_2(h,a,\mathcal{T})}\big).$$

The last expression above says that the features not predicted by laws (denoted by q_1) are predicted by P where we condition on the predicted features (denoted by q_2).

Definition 10 (Semi-deterministic environments from laws and background). *Given a set of deterministic laws $\mathcal{T}$ and background probabilities $P(\boldsymbol{x}|x_{k_1},...,x_{k_n},h,a)$,*
we let

$$\bar{\mathcal{M}}(P,\mathcal{T}) := \{\nu(P,\tilde{\mathcal{T}}) \mid \tilde{\mathcal{T}} \in \bar{\mathcal{C}}(\mathcal{T})\}.$$

The resulting error bound theorem has almost identical formulation as the previous case (Theorem 1) and is true for exactly the same reasons. However, the class $\bar{\mathcal{M}}$ contains stochasticity but of the predefined form.

Theorem 3 (Finite error bound when using laws and background). *Suppose that $\mathcal{T}$ is a finite class of deterministic laws and P is background. Let $\mathcal{G}(h) = \{\nu(\cdot|h) \mid \nu \in \bar{\mathcal{M}}(P, \{\tau \in \mathcal{T}$ consistent with $h\})\}$. We define $\bar{\pi}$ by* *combining $\mathcal{G}$ with the optimistic decision function (Definition 1). Following $\bar{\pi}$ with a finite class of deterministic laws $\mathcal{T}$ in an environment $\mu \in \bar{\mathcal{M}}(P, \mathcal{T})$, for $0 < \varepsilon < \frac{1}{1-\gamma}$ we have that*

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

for all but at most $|\mathcal{T}|\frac{-\log \varepsilon(1-\gamma)}{1-\gamma}$ time steps t.

Mixing deterministic and stochastic laws. When we introduced the concept of background environment we mentioned that it prepared for studying sets of laws that mix deterministic laws with absolutely continuous stochastic laws. Given an $\tilde{\varepsilon} > 0$, the environment formed by combining the stochastic laws with a coherent and maximal set of true deterministic laws eventually have a value function that for the followed policy is within $\tilde{\varepsilon}$ of the true one. Combining the remaining deterministic laws with the dominant stochastic laws into semi-deterministic environments exactly as with the background probabilities, then yields the results as before but with the accuracy only being $\varepsilon + \tilde{\varepsilon}$ instead of ε and where we only count errors happening after sufficient merging has taken place.

Example 4 (Mixing deterministic laws and stochastic laws). Consider a binary vector of length m where some elements are fixed and some fluctuate randomly with a probability unknown to an agent. Consider the laws based on KT-estimators from Example 1 and consider the $2m$ laws that each always makes a fixed prediction for one coefficient. The laws that make a deterministic prediction for a fluctuating coefficient will quickly get excluded and then the agent will have to fall back on the KT-estimate for this coefficient.

Example 5 (AIXI-CTW as background). Consider the AIXI-CTW environment ξ described in Example 2. Also, consider two deterministic law for each context in the context tree, one always predicts 1 and the other 0. Combining those two, we will have an agent that uses deterministic laws to predict until all laws for a certain feature in a certain context (including its subcontexts) are contradicted. Then it falls back on ξ for that situation. Predicting as much as possible with deterministic laws is very helpful for planning.

4 Conclusions

We have further developed the theory of optimistic agents with hypothesis classes defined by combining laws. Previous results were restricted to the deterministic setting while stochastic environments are necessary for any hope of real application. We here remedied this by introducing and studying stochastic laws and environments generated by such.

Acknowledgement. This work was supported by ARC grant DP120100950.

References

1. Carlos Diuk, Lihong Li, and Bethany R. Leffler. The adaptive k-meteorologists problem and its application to structure learning and feature selection in reinforcement learning. In Andrea Pohoreckyj Danyluk, Lon Bottou, and Michael L. Littman, editors, *ICML*, volume 382 of *ACM International Conference Proceeding Series*, 2009.
2. M. Hutter. *Universal Articial Intelligence: Sequential Decisions based on Algorithmic Probability.* Springer, Berlin, 2005.
3. T. Lattimore. *Theory of General Reinforcement Learning.* PhD thesis, Australian National University, 2014.
4. 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.
5. 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, 2013.
6. S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach.* Prentice Hall, Englewood Cliffs, NJ, 3^{nd} edition, 2010.
7. 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.
8. 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, 2012. Springer.
9. 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, 2012.
10. P. Sunehag and M. Hutter. Learning agents with evolving hypothesis classes. In *AGI*, pages 150–159, 2013.
11. P. Sunehag and M. Hutter. A dual process theory of optimistic cognition. In *Annual conference of the cognitive science society (CogSci'2014)*, 2014.
12. P. Sunehag and M. Hutter. Rationality, Optimism and Guarantees in General Reinforcement Learning. *Journal of Machine Learning Reserch, to appear*, 2015.
13. J. Veness, K. S. Ng, M. Hutter, W. Uther, and D. Silver. A Monte-Carlo AIXI approximation. *Journal of Artifiicial Intelligence Research*, 40(1):95–142, 2011.
14. F. Willems, Y. Shtarkov, and T. Tjalkens. The context tree weighting method: Basic properties. *IEEE Transactions on Information Theory*, 41:653–664, 1995.