A Dual Process Theory of Optimistic Cognition

Peter Sunehag (Peter.Sunehag@anu.edu.au) and Marcus Hutter (Marcus.Hutter@anu.edu.au)
Research School of Computer Science
The Australian National University, Canberra Australia

Abstract

Optimism is a prevalent bias in human cognition including variations like self-serving beliefs, illusions of control and overly positive views of one's own future. Further, optimism has been linked with both success and happiness. In fact, it has been described as a part of human mental well-being which has otherwise been assumed to be about being connected to reality. In reality, only people suffering from depression are realistic. Here we study a formalization of optimism within a dual process framework and study its usefulness beyond human needs in a way that also applies to artificial reinforcement learning agents. Optimism enables systematic exploration which is essential in an (partially) unknown world. The key property of an optimistic hypothesis is that if it is not contradicted when one acts greedily with respect to it, then one is well rewarded even if it is wrong.

Keywords: Rationality, Optimism, Optimality, Reinforcement Learning

Introduction

The optimistic bias is (together with the simplicity bias (Chater and Vitanyi, 2003)) perhaps the most fundamental and prevalent among the human cognitive biases (Taylor and Brown, 1988; Sharot et al., 2007; Carver et al., 2010; Kahneman, 2011) It has been found to be correlated with high achievement, happiness and resilience when challenged (Carver et al., 2010) but also with dangerous risk-seeking in e.g. traffic or severely underestimating costs in ruinous public building projects (Kahneman, 2011). Some of the drawbacks can be viewed as falling for implausible (cockeyed instead of cautious (Wallston, 1994)) optimism where one has not learned from other examples through the base rate for the sort of situation one is in. Besides resilience in the face of challenges and the effect on others, the optimism bias has uses beyond the strictly human. In reinforcement learning, artificial agents are often equipped with an optimism bias to enable systematically explorative behavior in an unknown world (Szita and Lörincz, 2008). Here we mathematically study plausibly optimistic general reinforcement learning agents as an alternative (both normative and descriptive) paradigm to Bayesian models of cognition (Griffiths et al., 2008).

The general reinforcement learning problem in an unknown environment is an extremely challenging problem (Hutter, 2005). If one has access to a suitable (i.e. probability mass placed mainly on the actually plausible) a priori environment, e.g. a Bayesian mixture over all environments in a certain hypothesis class, and the computational resources to compute the policy that maximizes the desired quality measure, this is the natural choice and optimal by definition. The quality measure can for example be expected accumulated reward during the life of the agent, or the maximum accumulated reward that is guaranteed with a certain given probability (e.g. 0.95). The two immediate problems are that one needs such an a priori environment, e.g. through a prior on a hypothesis class, and that performing the mentioned computation is too hard, even approximately. We argue that the second of these motivates optimism as a rational strategy, since when optimizing over a too short horizon, realism leads to insufficiently explorative behavior. A property that makes optimism particularly appealing is that as long as the outcome is as predicted, high rewards are received. This is regardless of the correctness of the hypothesis. Humans are often trying to avoid contradiction of their hypothesis and often avoid contradicting each other's hypothesis (Taylor and Brown, 1988). Managing to enforce ones hypothesis in a group is primarily useful if it is optimistic, i.e. self-serving.

We are here going to discuss agents with limited resources within a dual process agent framework where a limited number of hypotheses are generated by one system and the other system is making a choice by excluding implausible hypotheses and choosing optimistically/greedily among the rest. Within such a framework we use normative principles such as rationality considerations as in the foundations of decision theory together with the kind of performance guarantees studied in reinforcement learning as a subfield of artificial intelligence. Dual Process theories with an implicit and an explicit part (often called system 1 and system 2) have a strong position in cognitive science with ample empirical support (Evans, 2003), though still considered an approximation of a more complex reality with overlapping functionality.

In our framework, an agent consists 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. We define agents within this framework by combining optimistic decision functions with hypothesis generating functions defined by enumerating a countable class and introducing new environments from this list (which could be sampled incrementally with a simplicity bias formalized by Kolmogorov complexity (Sunehag and Hutter, 2013)) when we are within a given error budget. We present results for generic countable classes by extending the agents introduced in Sunehag and Hutter (2012a) from the finite case.

The best bounds for fully general reinforcement learning have a linear dependence on the number of environments in the class. Though this is easily seen to be the best one can do in general (Lattimore et al., 2013), it is bad (exponentially worse) compared to 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. We introduce environment classes that

are much more general than MDPs, while they are finitely generated (by laws) in a way enabling a good bound.

Outline. We begin by introducing background and notation for general reinforcement learning and then we introduce our agent framework and provide examples of agents that fit within it. Within this framework we then provide bounds on the number of errors that an optimistic agent makes in the case of an infinite countable class of possible environments which the agent includes incrementally in its environment class. Finally we extend (and improve) the results to the case of environments generated by laws and then conclude.

General Reinforcement Learning. 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 O and rewards r_t from a finite set $\mathcal{R} \subset [0,1]$ resulting in a history $h_t := o_1 r_1 a_1, ..., o_t r_t$. These sets can be allowed to depend on time or context but we do not write this out explicitly. Let $\mathcal{H} := \cup_n (O \times R \times \mathcal{A})^n \times (O \times \mathcal{R})$ be the set of histories and let ε be the empty history. A function $\nu : \mathcal{H} \times \mathcal{A} \to O \times \mathcal{R}$ 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 (discount factor $0 \leq \gamma < 1$) 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 $\mathcal{H} \times \mathcal{A} \to O \times \mathcal{R}$ we can equivalently write it as a function $\nu : \mathcal{H} \times \mathcal{A} \times O \times \mathcal{R} \to \{0,1\}$ where we also write $\nu(o,r|h,a)$ for the function value $\nu(h,a,o,r)$ (which is not the probability of the four-tuple). 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 $\nu : \mathcal{H} \times \mathcal{A} \times O \times \mathcal{R} \to [0,1]$ such that $\sum_{o,r} \nu(o,r|h,a) = 1 \ \forall h,a$. Furthermore, we define $\nu(h_t|\pi) := \Pi_{i=1}^t \nu(o_i r_i | a_{i-1}, h_{i-1})$ where $a_i = \pi(h_i)$. $\nu(\cdot|\pi)$ is a probability measure over strings and we define $\nu(\cdot|\pi, h_{t-1})$ by conditioning $\nu(\cdot|\pi)$ on h_{t-1}. $V_\nu^\pi(h_{t-1}) := \mathbb{E}_{\nu(\cdot|\pi,h_{t-1})} \sum_{i=t}^\infty \gamma^{i-t} r_i$ and $V_\nu^*(h_{t-1}) := \max_\pi V_\nu^\pi(h_{t-1})$.

Examples of agents: AIXI and Optimist. 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 (in its general form) is defined by following the policy $\pi^* := \arg\max_\pi V_\xi^\pi(\emptyset)$. The above agent, and only agents of that form, satisfy the strict rationality axioms presented in Sunehag and Hutter (2011) while the slightly looser version from Sunehag and Hutter (2012b) allows for optimism. The optimist takes the decision

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

for a finite set of beliefs (environments) Ξ.

An agent framework with growing classes

In this section, we introduce an agent framework that we can fit some existing successful agents into by a choice of what we will call a decision function and a hypothesis generating function. Within this framework, we will extend the agents and analysis from the previous section to arbitrary infinitely countable classes.

Decision Functions

The primary component of our agent framework is a decision function $f : \mathbb{M} \to \mathcal{A}$ ($\mathbb{M}$ is the set of finite sets of environments) only depending on a class of environments $\mathcal{M}$. The decision function is independent of the history, however, the class $\mathcal{M}$ fed to the decision function introduce an indirect dependence. For example, the environments at time $t+1$ can be the environments at time t, conditioned on the new observation. In this setting we will often write the value function without an argument due to this independence. $V_{\nu_t}^{\tilde{\pi}} = V_{\nu_0}^\pi(h_t)$ 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 so it starts at a later stage, meaning $\tilde{\pi}(h) = \pi(h_t h)$ where $h_t h$ is a concatenation.

Definition 1 (Rational Decision Function)**.** *Given alphabets $\mathcal{A}$, 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 and finite history produces an action $a \in \mathcal{A}$. We say that f is strictly rational for the class $\mathcal{M}$ if there are $\omega_\nu \geq 0$, $\nu \in \mathcal{M}, \sum_{\nu \in \mathcal{M}} w_\nu = 1$ such that $a = \pi(\varepsilon)$ for a policy*

$$\pi \in \arg\max_\pi \sum_{\nu \in \mathcal{M}} \omega_\nu V_\nu^\pi. \tag{1}$$

Agents who are as in Definition 1 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 (1) then becomes

$$\pi \in \arg\max_\pi V_\nu^\pi$$

where ν is the environment in $\mathcal{M}$. The more general case connects to this by letting $\tilde{\nu}(\cdot) := \sum_{\nu \in \mathcal{M}} w_\nu \nu(\cdot)$. The next definition defines optimistic decision functions. They only coincide with strictly rational ones (as defined above, see Sunehag and Hutter (2011) for details) for the case $|\mathcal{M}| = 1$. However, agents based on such decision functions satisfy the looser axioms that define (a weaker form of) rationality in Sunehag and Hutter (2012b).

Definition 2 (Optimistic Decision Function)**.** *We call a decision function f optimistic if $f(\mathcal{M}) = a$ implies that $a = \pi(\varepsilon)$ for an optimistic policy π, i.e. for*

$$\pi \in \arg\max_\pi \max_{\nu \in \mathcal{M}} V_\nu^\pi. \tag{2}$$

Agents Based on Decision Functions

Given a decision function, what remains to create a complete agent is a hypothesis generating function $g(h) = \mathcal{M}$ that for any history $h \in \mathcal{H}$ produces a set of environments $\mathcal{M}$. A

special case of a hypothesis generating function is defined by combining the initial $g(\varepsilon) = \mathcal{M}_0$ with an update function $\psi(\mathcal{M}_{t-1}, h_t) = \mathcal{M}_t$. An agent, i.e. a function from histories to actions, is defined from a hypothesis generating function g and a decision function f by choosing action $a = f(g(h))$ after seeing history h. We discuss a number of examples below to elucidate the framework and as a basis for the results we present later.

Example 3. *Suppose that ν is a stochastic environment and $g(h) = \{\nu(\cdot|h)\}$ for all ν and let f be a strictly rational decision function. This agent is a rational agent in the stricter sense . Also, if $g(h) = \{\nu(\cdot|h) \mid \nu \in \mathcal{M}\}$ 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 $a = \pi(\varepsilon)$ for a policy*

$$\pi \in \arg\max_\pi \sum_{\nu\in g(h)} \omega_\nu V_\nu^\pi, \tag{3}$$

then we say that we have a Bayesian agent, which can be represented more simply in the first way by $g(h) = \{\sum w_\nu \nu(\cdot|h)\}$.

Example 4. *Suppose that $\mathcal{M}$ is a finite class of deterministic environments and let $g(h) = \{\nu(\cdot|h) \mid \nu \in \mathcal{M}$ consistent with $h\}$. If we combine g with the optimistic decision function we have defined the optimistic agents for classes of deterministic environments from Sunehag and Hutter (2012a). We here extend the analysis to infinite classes by letting $g(h_t)$ contain new environments that were not in $g(h_{t-1})$.*

Example 5. *Suppose that $\mathcal{M}$ is a finite class of stochastic environments and that $g(h) = \{\nu(\cdot|h) \mid \nu \in \mathcal{M}\}$. If we combine g with the optimistic decision function we have defined the optimistic AIXI agent from Sunehag and Hutter (2012b). If instead $g(h) = \{\nu(\cdot|h) \mid \nu \in \mathcal{M} : \frac{\nu(h)}{\max_{\bar\nu} \bar\nu(h)} \geq z\}$ for some $z > 0$ we have defined the optimistic agent with stochastic environments from Sunehag and Hutter (2012a).*

Example 6. *The Model Based Interval Estimation (MBIE) (Strehl et al., 2009) method for Markov Decision Processes (MDPs) defines $g(h)$ 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.*

Example 7. *Agents that switch explicitly between exploration and exploitation are typically not satisfying our (weak) rationality demand. An example is Lattimore et al. (2013) 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 it is proven that MERL satisfies near-optimal sample complexity for general reinforcement learning with finite classes. Another example of this kind of agent is BayesExp (Lattimore, 2013).*

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. In the choice of hypothesis generating functions we are going to focus on what kind of performance can be guaranteed in terms of how many suboptimal decisions will be taken. First, however, we want to restrain ourselves 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 make n hypothesis unlikely as in Example 5. We will consider the hypothesis generating function to satisfy Epicurus if the update function is such that it might add new environments in any way while removing environments if a hypothesis is implausible (likely to be false) in light of the observations made.

Aside from satisfying Epicurus' principle, we will design hypothesis generating functions based mainly on wanting few mistakes to be made. For this purpose we first define the term ε-error. We are going to formulate the rest of the definitions and results in this section for $\gamma = 0$ for simplicity and brevity.

Definition 8 (ε-error)**.** *Given $0 \leq \varepsilon < 1$, we define the number of ε-errors in history h to be*

$$m(h, \varepsilon) = |\{i \leq \ell(h) \mid V_\mu^{a_i}(h_i) < V_\mu^*(h_i) - \varepsilon\}|$$

where μ is the true environment, $\ell(h)$ is the length of h, a_i is the i:th action and $V_\mu^*(h) = \arg\max_a V_\mu^a(h)$. Each such timepoint is called an ε-error.

Since we consider a setting where 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 $g(h_t)$, 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 that the distance between any pair of environments in the class is small, then one can guarantee that an ε-error is not made (Sunehag and Hutter, 2012a). Since we do not know if this extra assumption holds for $g(h_t)$, we will use the terms ε-confident and ε-inconfident.

If the value functions in the class $g(h_t)$ 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 $g(h_t)$. We call such points ε-inconfidence points.

Definition 9 (ε-(in)confidence)**.** *Given $0 < \varepsilon < 1$, we define the number of ε-inconfidence point in the history h to be*

$$n(h, \varepsilon) := |\{i \leq l(h) \mid \max_{\nu_1, \nu_2 \in g(h_i)} |V_{\nu_1}^{\pi^*} - V_{\nu_2}^{\pi^*}| > \varepsilon\}|$$

where $\pi^* := \arg\max_\pi \max_{\nu\in g(h_t)} V_\nu^\pi$. In the $\gamma = 0$ case studied here, we can equivalently use $a^* := \arg\max_a \max_{\nu\in g(h_t)} V_\nu^a$ instead of π^*. The individual time-points 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 count-

able enumerated class $\mathcal{M}$ based on a *budget function* for ε-inconfidence. The idea is simply that when the number of ε-inconfidence points is below budget we introduce the next environment in the class. The intuition is that if the current hypotheses are frequently contradictory, then one should resolve this 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.

Definition 10 (Hypothesis generation with budget and exclusion function)**.**
Suppose we have a chosen accuracy $\varepsilon > 0$, an enumerated countable class of environments $\mathcal{M}$, a finite initial class $\mathcal{M}^0 \subset \mathcal{M}$ a non-decreasing budget function $N : \mathbb{N} \to \mathbb{N}$ such that $N(t) \to \infty$ as $t \to \infty$, an exclusion function (criterion) $\phi(\tilde{\mathcal{M}}, h) = \hat{\mathcal{M}}$ for $\tilde{\mathcal{M}} \subset \mathcal{M}$ and $h \in \mathcal{H}$ such that $\hat{\mathcal{M}} \subset \tilde{\mathcal{M}}$. Then the hypothesis generating function g with class $\mathcal{M}$, initial class $\mathcal{M}^0$, accuracy $\varepsilon > 0$, budget N and exclusion criterion ϕ is defined recursively as follows:

Let $g(\varepsilon) = \mathcal{M}^0$. If $n(h_t, \varepsilon) \geq N(t)$, then

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

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 $\cup_{i=1}^{t-1}\{\nu \in \mathcal{M} \mid \nu(\cdot|h_i) \in g(h_i)\}$ (i.e. the next environment to introduce) and let $g(h_t) =$

$$\{\nu(\cdot|h_t) \mid \nu \in \{\tilde{\nu} \cup \phi(\{\nu \in \mathcal{M} \mid \nu(\cdot|h_{t-1}) \in g(h_{t-1})\}, h_t)\}\}.$$

Error Analysis

We will now extend the agents described in Example 4 and Example 5 by removing the demand for the class $\mathcal{M}$ being finite and analyze the effect on the number of ε-errors made. We will still use the optimistic decision function and apply it to finite classes but we will keep adding environments from the full class to the finite working class of environments. The resulting agents differ from agents such as the one in Example 7 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 which may cause new explorative behavior if it becomes the optimistic hypothesis and it deviates significantly from the other environments. A nice point about our results is that one chooses the asymptotic rate, however one gets a worse constant the better rate one chooses. This is due to the fact that if one includes new environments at a slower rate it takes longer until the right environment is introduced while the error rate afterwards is better. If one knew that one had included the right one, then one would stop introducing more.

We extend the agent for finite classes of deterministic environments in Example 4 to the countable case and we leave the extension of the stochastic case to a longer report. In the finite case with a fixed class, the proof of the finite error-bound in Sunehag and Hutter (2012a) 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 by only being able to have at most $|\mathcal{M}-1|$ contradictions. In the case where environments are being added one can have errors 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 having encountered a contradiction, 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 we can have errors.

Theorem 11. *Suppose we have a countable class of deterministic environments $\mathcal{M}$ (with a chosen enumeration and containing the true one). Also suppose we have a hypothesis generating function g with a finite initial class $g(\varepsilon) := \mathcal{M}^0 \subset \mathcal{M}$, budget function $N : \mathbb{N} \to \mathbb{N}$, accuracy $\varepsilon = 0$ and suppose that g excludes contradicted environments. π° is defined by combining g with an optimistic decision function. The number of 0-errors $m(h_t, 0)$ is at most $n(h_t, 0) + C$ for some constant $C > 0$ (which is the number of steps before the true environment is introduced and depends on the choice of budget function N but not on t). Furthermore, $\forall i \in \mathbb{N}$ there is $t_i \in \mathbb{N}$ such that $t_i < t_{i+1}$ and $n(h_{t_i}, 0) < N(t_i)$.*

The last claim is the most important saying 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. If we wanted the errors to always be within $N(t)$ (except for a constant) we could forbid the agent from introducing more environments than $N(t)$ before time t. This is because the number of excluded hypotheses cannot exceed the number of introduced hypotheses including the ones in $g(\varepsilon)$ and we only have an error when an environment is excluded (and not always then) or when $g(h_t)$ is empty which it is not again after the true environment is introduced.

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

$$V_\mu^{\pi^\circ}(h_t) = \max_a V_\mu^a(h_t) \tag{4}$$

since all the environments in $g(h_t)$ 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 it 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 we are guaranteed to have a contradiction, $n(0, h_t)$ 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. □

Remark 12 (Extensions: $\gamma > 0$, Separable classes)**.** *As in the deterministic case, what makes the $\gamma = 0$ case simpler than the $0 < \gamma < 1$ case is that ε-errors can for $\gamma > 0$ also occur within $\frac{-\log\varepsilon(1-\gamma)}{1-\gamma}$ time-steps before a new environment is introduced. Further, one can extend our algorithms for countable classes to separable classes since they can be covered by countable many balls of arbitrarily small radius.*

Environments Defined by Laws

In this section we will investigate classes of environments of a special but generic form combining a number of laws that partially determine what happens next. Classes of such form have the property that one can exclude (or merge) laws and thereby exclude (or merge) whole classes of environments like when one learns about a state transition when working with MDPs. Our setting is, however, far more general than MDPs and shares characteristics with what has been studied in the Learning Classifier Systems literature (Holland, 1986; Drugowitsch, 2007).

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

Definition 13. *A law is a function $\tau : \mathcal{H} \times \mathcal{A} \to \tilde{O}$ where $\tilde{O}$ consists of the feature vectors from O but where some elements are replaced by a special letter $\perp$ meaning that there is no prediction for this feature, i.e. $\tilde{O} = \times_{j=1}^m (O_j \cup \{\perp\})$.*

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.

We first consider deterministic laws. 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 define a setting where the set of laws is such that in every situation and for every feature there is at least one law that makes a prediction of this feature in the given situation. We can then directly define a set of environments by combining such laws.

Definition 14. *[Environments from deterministic laws] Given a finite set of laws $\mathcal{T}$ (maps from $\mathcal{H} \times \mathcal{A}$ to $\tilde{O}$) such that $O = \times_j O_j$ have m features labelled $1, ..., m$. We define the class of environments*

$$\mathcal{M}(\mathcal{T}) := \{\nu \mid \exists \tilde{\mathcal{T}} \in \mathcal{C}(\mathcal{T}) : \nu = \nu(\tilde{\mathcal{T}})\}$$

where the set of coherent and complete sets of deterministic laws $\mathcal{C}(\mathcal{T})$ is defined by

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

$$\nu(h,a)(j) \neq \perp \wedge\ \tilde{\tau}(h,a)(j) = \perp \forall \tilde{\tau} \in \tilde{\mathcal{T}} \setminus \{\tau\}\}$$

and for $\tilde{\mathcal{T}} \in \mathcal{C}(\mathcal{T})$, $\nu(\tilde{T})$ is the environment ν which is such that

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

Example 15. *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.*

We analyze the optimistic agent from Example 4 in this new setting. Every contradiction of an environment is a contradiction of at least one law and there are finitely many laws and this is what is needed for the finite error result from before to hold but with $|\mathcal{M}|$ replaced by $|\mathcal{T}|$ (see Theorem 16 below) which can be exponentially smaller. Furthermore, the extension to countable $\mathcal{T}$ works the same as in Theorem 11.

Theorem 16 (Finite error-bound when using laws)**.** *Suppose that $\mathcal{T}$ is a finite class of deterministic laws and let $g(h) = \{\nu(\cdot|h) \mid \nu \in \mathcal{M}(\{\tau | \tau \in \mathcal{T} \text{ consistent with } h\})\}$. We define $\bar{\pi}$ by combining g with the optimistic decision function. Following $\bar{\pi}$ with a finite class of deterministic laws $\mathcal{T}$ in an environment $\mu \in \mathcal{M}(\mathcal{T})$,*

$$|V_\mu^{\bar{\pi}}(h_t) - \max_\pi V_\mu^\pi(h_t)| < \varepsilon \tag{5}$$

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 is true for the same reason as the finite-error bound in Sunehag and Hutter (2012a) since there are at most $|\mathcal{T} - l|$ time-steps with a contradiction and errors occur only at times that are at most $\frac{-\log\varepsilon(1-\gamma)}{1-\gamma}$ steps before a contradiction. This is due to time-consistency of geometric discounting.

□

Background Environments. Further improvement in the rate of errors can be achieved if we have a background environment and the laws are only replacing the prediction of the background environment for the part they say something about. Again if one formalizes this notion correctly, one can prove bounds linear in the number of laws which can be much fewer in this situation.

Computing the optimistic decision as one planning problem. Finding the optimistic decision with a collection of laws results in a computation as in an auction-based multi-agent planning system (as in e.g. Allard and Shekh (2012)), though many of the laws are making incorrect predictions. The combined choice of laws and action to use forms a larger action space as in Asmuth et al. (2009), though for a much more general situation. Monte-Carlo Tree Search methods (Silver and Veness, 2010), which are also planning methods, could be applied despite that some of the laws (which are also selected instead of only choosing actions in the search) are incorrect and contradict each other. Value predictions resulting from function approximation can be very useful for guiding the tree search (Silver et al., 2012). Human cognition likely involves both estimation of models as well as direct value estimation (Shteingart and Loewenstein, 2014).

Conclusions

We introduced a dual process framework based on a hypothesis function and a decision function. An optimistic decision function was found to be useful for achieving optimality guarantees while a simplicity bias can be useful for hypothesis generation (Sunehag and Hutter, 2013). A key point is that optimism encourages exploration, which is important if one cannot optimize a strategy for one's whole life. Further when acting according to an optimistic hypothesis it is only important that it is not contradicted while it does not have to be correct for other circumstances. One will still be highly rewarded.

References

Allard, T. and Shekh, S. (2012). Hierarchical multi-agent distribution planning. In *AI 2012: Advances in Artificial Intelligence*, volume 7691 of *Lecture Notes in Computer Science*, pages 755–766.

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

Carver, C. S., Scheier, M. F., and Segerstrom, S. C. (2010). Optimism. *Clinical Psychology Review*, 30(7):879–889.

Chater, N. and Vitanyi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, 7:19–22.

Drugowitsch, J. (2007). *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.

Evans, J. S. (2003). In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences*, 7(10):454 – 459.

Griffiths, T. L., Kemp, C., and Tenenbaum, J. B. (2008). Bayesian models of cognition. In Sun, R., editor, *Cambridge handbook of computational cognitive modeling*, pages 59–100. Cambridge University Press, Cambridge.

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

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

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux, New York.

Lattimore, T. (2013). *Theory of General Reinforcement Learning (submitted)*. PhD thesis, Australian National University.

Lattimore, T. and Hutter, M. (2012). PAC Bounds for Discounted MDPs. In Bshouty, N. H., Stoltz, G., Vayatis, N., and Zeugmann, T., editors, *ALT*, volume 7568 of *Lecture Notes in Computer Science*, pages 320–334. Springer.

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

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

Sharot, T., Riccardi, A. M., Raio, C. M., and Phelps, E. A. (2007). Neural mechanisms mediating optimism bias. *Nature*, pages 1–5.

Shteingart, H. and Loewenstein, Y. (2014). Reinforcement learning and human behavior. *Current Opinion in Neurobiology*, 25(0):93 – 98.

Silver, D., Sutton, R., and Müller, M. (2012). Temporal-difference search in computer Go. *Machine Learning*, 87(2):183–219.

Silver, D. and Veness, J. (2010). Monte-Carlo Planning in Large POMDPs. In *Advances in Neural Information Processing Systems 23: 2010.*, pages 2164–2172.

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

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

Sunehag, P. and Hutter, M. (2012a). 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. Springer.

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

Sunehag, P. and Hutter, M. (2013). Learning agents with evolving hypothesis classes. In *Proceedings of the 6th International Conference on AGI*, volume 7999 of *Lecture Notes in Computer Science*, pages 150–159. Springer.

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

Taylor, S. E. and Brown, J. D. (1988). Illusion and well-being: a social psychological perspective on mental health. *Psychological bulletin*, 103(2):193.

Wallston, K. A. (1994). Cautious optimism vs. cockeyed optimism. *Psychology & Health*, 9(3):201–203.