Journal of Artificial Intelligence Research 38 (2010) 475-511 Submitted 04/10; published 08/10

A Minimum Relative Entropy Principle for Learning and Acting

Pedro A. Ortega PEORTEGA@DCC.UCHILE.CL
Daniel A. Braun DAB54@CAM.AC.UK
Department of Engineering
University of Cambridge
Cambridge CB2 1PZ, UK

Abstract

This paper proposes a method to construct an adaptive agent that is universal with respect to a given class of experts, where each expert is designed specifically for a particular environment. This adaptive control problem is formalized as the problem of minimizing the relative entropy of the adaptive agent from the expert that is most suitable for the unknown environment. If the agent is a passive observer, then the optimal solution is the well-known Bayesian predictor. However, if the agent is active, then its past actions need to be treated as causal interventions on the I/O stream rather than normal probability conditions. Here it is shown that the solution to this new variational problem is given by a stochastic controller called the Bayesian control rule, which implements adaptive behavior as a mixture of experts. Furthermore, it is shown that under mild assumptions, the Bayesian control rule converges to the control law of the most suitable expert.

1. Introduction

When the behavior of an environment under any control signal is fully known, then the designer can choose an agent that produces the desired dynamics. Instances of this problem include hitting a target with a cannon under known weather conditions, solving a maze having its map and controlling a robotic arm in a manufacturing plant. However, when the environment is unknown, then the designer faces the problem of *adaptive control*. For example, shooting the cannon lacking the appropriate measurement equipment, finding the way out of an unknown maze and designing an autonomous robot for Martian exploration. Adaptive control turns out to be far more difficult than its non-adaptive counterpart. This is because any good policy has to carefully trade off explorative versus exploitative actions, i.e. actions for the identification of the environment's dynamics versus actions to control it in a desired way. Even when the environment's dynamics are known to belong to a particular class for which optimal agents are available, constructing the corresponding optimal adaptive agent is in general computationally intractable even for simple toy problems (Duff, 2002). Thus, finding tractable approximations has been a major focus of research.

Recently, it has been proposed to reformulate the problem statement for some classes of control problems based on the minimization of a relative entropy criterion. For example, a large class of optimal control problems can be solved very efficiently if the problem statement is reformulated as the minimization of the deviation of the dynamics of a controlled system from the uncontrolled system (Todorov, 2006, 2009; Kappen, Gomez, & Opper, 2010). In this work, a similar approach is introduced for adaptive control. If a class of agents is given, where each agent is tailored to a different environment, then adaptive controllers can be derived from a minimum relative entropy principle. In particular, one can construct an adaptive agent that is universal with respect to this class by minimizing the average relative entropy from the environment-specific agent.

However, this extension is not straightforward. There is a syntactical difference between actions and observations that has to be taken into account when formulating the variational problem. More specifically, actions have to be treated as interventions obeying the rules of causality (Pearl, 2000; Spirtes, Glymour, & Scheines, 2000; Dawid, 2010). If this distinction is made, the variational problem has a unique solution given by a stochastic control rule called the Bayesian control rule. This control rule is particularly interesting because it translates the adaptive control problem into an on-line inference problem that can be applied forward in time. Furthermore, this work shows that under mild assumptions, the adaptive agent converges to the environment-specific agent.

The paper is organized as follows. Section 2 introduces notation and sets up the adaptive control problem. Section 3 formulates adaptive control as a minimum relative entropy problem. After an initial, naïve approach, the need for causal considerations is motivated. Then, the Bayesian control rule is derived from a revised relative entropy criterion. In Section 4, the conditions for convergence are examined and a proof is given. Section 5 illustrates the usage of the Bayesian control rule for the multi-armed bandit problem and undiscounted Markov decision processes. Section 6 discusses properties of the Bayesian control rule and relates it to previous work in the literature. Section 7 concludes.

2. Preliminaries

In the following both agent and environment are formalized as causal models over I/O sequences. Agent and environment are coupled to exchange symbols following a standard interaction protocol having discrete time, observation and control signals. The treatment of the dynamics are fully probabilistic, and in particular, *both* actions *and* observations are random variables, which is in contrast to the typical decision-theoretic agent formulation treating only observations as random variables (Russell & Norvig, 2010). All proofs are provided in the appendix.

Notation. A set is denoted by a calligraphic letter like $\mathcal{A}$. The words *set* & *alphabet* and *element* & *symbol* are used to mean the same thing respectively. *Strings* are finite concatenations of symbols and *sequences* are infinite concatenations. $\mathcal{A}^n$ denotes the set of strings of length n based on $\mathcal{A}$, and $\mathcal{A}^* := \bigcup_{n\geq 0} \mathcal{A}^n$ is the set of finite strings. Furthermore, $\mathcal{A}^\infty := \{a_1a_2\ldots | a_i \in \mathcal{A} \text{ for all } i = 1,2,\ldots\}$ is defined as the set of one-way infinite sequences based on the alphabet $\mathcal{A}$. Tuples are written with parentheses (a_1, a_2, a_3) or as strings $a_1a_2a_3$. The notation $a_{\leq i} := a_1a_2\ldots a_i$ is a shorthand for a string starting from the first index. Also, symbols are underlined to glue them together like $\underline{ao}$ in $\underline{ao}_{\leq i} := a_1o_1a_2o_2\ldots a_io_i$. The function $\log(x)$ is meant to be taken w.r.t. base 2, unless indicated otherwise.

Interactions. The possible I/O symbols are drawn from two finite sets. Let $\mathcal{O}$ denote the set of *inputs* (observations) and let $\mathcal{A}$ denote the set of *outputs* (actions). The set $\mathcal{Z} := \mathcal{A}\times\mathcal{O}$ is the *interaction set*. A string $\underline{ao}_{\leq t}$ or $\underline{ao}_{<t}a_t$ is an *interaction string* (optionally ending in

a_t or o_t) where $a_k \in \mathcal{A}$ and $o_k \in \mathcal{O}$. Similarly, a one-sided infinite sequence $a_1 o_1 a_2 o_2 \ldots$ is an *interaction sequence*. The set of interaction strings of length t is denoted by $\mathcal{Z}^t$. The sets of (finite) interaction strings and sequences are denoted as $\mathcal{Z}^*$ and $\mathcal{Z}^\infty$ respectively. The interaction string of length 0 is denoted by ϵ.

I/O System. Agents and environments are formalized as I/O systems. An *I/O system* is a probability distribution $\mathbf{Pr}$ over interaction sequences $\mathcal{Z}^\infty$. $\mathbf{Pr}$ is uniquely determined by the conditional probabilities

$$\mathbf{Pr}(a_t | \underline{ao}_{<t}), \quad \mathbf{Pr}(o_t | \underline{ao}_{<t} a_t) \tag{1}$$

for each $\underline{ao}_{<t} \in \mathcal{Z}^*$. These conditional probabilities can either represent a *generative law* ("propensity") in case of issuing a symbol or an *evidential probability* ("plausibility") in the case of observing a symbol. Which of the two interpretations applies in a particular case becomes apparent once the I/O system is coupled to another I/O system.

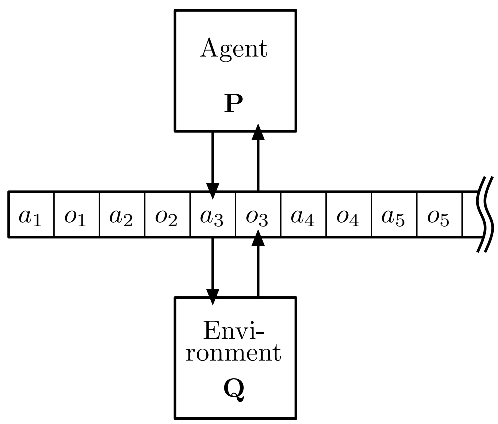

Figure 1: The model of interactions. The agent $\mathbf{P}$ and the environment $\mathbf{Q}$ define a probability distribution over interaction sequences.

Interaction System. Let $\mathbf{P}$, $\mathbf{Q}$ be two I/O systems. An *interaction system* $(\mathbf{P}, \mathbf{Q})$ is a coupling of the two systems giving rise to the *generative distribution* $\mathbf{G}$ that describes the probabilities that actually govern the I/O stream once the two systems are coupled. $\mathbf{G}$ is specified by the equations

$$\mathbf{G}(a_t | \underline{ao}_{<t}) := \mathbf{P}(a_t | \underline{ao}_{<t})$$
$$\mathbf{G}(o_t | \underline{ao}_{<t} a_t) := \mathbf{Q}(o_t | \underline{ao}_{<t} a_t)$$

valid for all $\underline{ao}_t \in \mathcal{Z}^*$. Here, $\mathbf{G}$ models the *true* probability distribution over interaction sequences that arises by coupling two systems through their I/O streams. More specifically, for the system $\mathbf{P}$, $\mathbf{P}(a_t | \underline{ao}_{<t})$ is the probability of producing action $a_t \in \mathcal{A}$ given history $\underline{ao}_{<t}$ and $\mathbf{P}(o_t | \underline{ao}_{<t} a_t)$ is the predicted probability of the observation $o_t \in \mathcal{O}$ given history

$\underline{ao}_{<t}a_t$. Hence, for $\mathbf{P}$, the sequence $o_1o_2\ldots$ is its input stream and the sequence $a_1a_2\ldots$ is its output stream. In contrast, the roles of actions and observations are reversed in the case of the system $\mathbf{Q}$. Thus, the sequence $o_1o_2\ldots$ is its output stream and the sequence $a_1a_2\ldots$ is its input stream. The previous model of interaction is fairly general, and many other interaction protocols can be translated into this scheme. As a convention, given an interaction system $(\mathbf{P},\mathbf{Q})$, $\mathbf{P}$ is an agent to be constructed by the designer, and $\mathbf{Q}$ is an environment to be controlled by the agent. Figure 1 illustrates this setup.

Control Problem. An environment $\mathbf{Q}$ is said to be *known* iff the agent $\mathbf{P}$ has the property that for any $\underline{ao}_{\leq t} \in \mathcal{Z}^*$,

$$\mathbf{P}(o_t|\underline{ao}_{<t}a_t) = \mathbf{Q}(o_t|\underline{ao}_{<t}a_t).$$

Intuitively, this means that the agent "knows" the statistics of the environment's future behavior under any past, and in particular, it "knows" the effects of given controls. If the environment is known, then the designer of the agent can build a custom-made policy into $\mathbf{P}$ such that the resulting generative distribution $\mathbf{G}$ produces interaction sequences that are desirable. This can be done in multiple ways. For instance, the controls can be chosen such that the resulting policy maximizes a given utility criterion; or such that the resulting trajectory of the interaction system stays close enough to a prescribed trajectory. Formally, if $\mathbf{Q}$ is known, and if the conditional probabilities $\mathbf{P}(a_t|\underline{ao}_{<t})$ for all $\underline{ao}_{\leq t} \in \mathcal{Z}^*$ have been chosen such that the resulting generative distribution $\mathbf{G}$ over interaction sequences given by

$$\begin{aligned}\mathbf{G}(a_t|\underline{ao}_{<t}) &= \mathbf{P}(a_t|\underline{ao}_{<t})\\ \mathbf{G}(o_t|\underline{ao}_{<t}a_t) &= \mathbf{Q}(o_t|\underline{ao}_{<t}a_t) = \mathbf{P}(o_t|\underline{ao}_{<t}a_t)\end{aligned}$$

is *desirable*, then $\mathbf{P}$ is said to be *tailored* to $\mathbf{Q}$.

Adaptive Control Problem. If the environment $\mathbf{Q}$ is *unknown*, then the task of designing an appropriate agent $\mathbf{P}$ constitutes an *adaptive control problem*. Specifically, this work deals with the case when the designer already has a class of agents that are tailored to the class of possible environments. Formally, it is assumed that $\mathbf{Q}$ is going to be drawn with probability $P(m)$ from a set $\mathcal{Q} := \{\mathbf{Q}_m\}_{m\in\mathcal{M}}$ of possible systems before the interaction starts, where $\mathcal{M}$ is a countable set. Furthermore, one has a set $\mathcal{P} := \{\mathbf{P}_m\}_{m\in\mathcal{M}}$ of systems such that for each $m \in \mathcal{M}$, $\mathbf{P}_m$ is tailored to $\mathbf{Q}_m$ and the interaction system $(\mathbf{P}_m, \mathbf{Q}_m)$ has a generative distribution $\mathbf{G}_m$ that produces desirable interaction sequences. How can the designer construct a system $\mathbf{P}$ such that its behavior is as close as possible to the custom-made system $\mathbf{P}_m$ under any realization of $\mathbf{Q}_m \in \mathcal{Q}$?

3. Adaptive Systems

The main goal of this paper is to show that the problem of adaptive control outlined in the previous section can be reformulated as a universal compression problem. This can be informally motivated as follows. Suppose the agent $\mathbf{P}$ is implemented as a machine that is interfaced with the environment $\mathbf{Q}$. Whenever the agent interacts with the environment, the agent's state changes as a *necessary consequence* of the interaction. This "change in state" can take place in many possible ways: by updating the internal memory; consulting a random number generator; changing the physical location and orientation; and so forth. Naturally, the design of the agent facilitates some interactions while it complicates others. For instance, if the agent has been designed to explore a natural environment, then it might incur into a very low memory footprint when recording natural images, while being very memory-inefficient when recording artificially created images. If one abstracts away from the inner workings of the machine and decides to encode the state transitions as binary strings, then the minimal amount of resources in bits that are required to implement these state changes can be derived directly from the associated probability distribution $\mathbf{P}$. In the context of adaptive control, an agent can be constructed such that it minimizes the expected amount of changes necessary to implement the state transitions, or equivalently, such that it maximally compresses the experience. Thereby, compression can be taken as a *stand-alone principle to design adaptive agents*.

3.1 Universal Compression and Naïve Construction of Adaptive Agents

In coding theory, the problem of compressing a sequence of observations from an unknown source is known as the adaptive coding problem. This is solved by constructing universal compressors, i.e. codes that adapt on-the-fly to any source within a predefined class (MacKay, 2003). Such codes are obtained by minimizing the average deviation of a predictor from the true source, and then by constructing codewords using the predictor. In this subsection, this procedure will be used to derive an adaptive agent (Ortega & Braun, 2010).

Formally, the deviation of a predictor $\mathbf{P}$ from the true distribution $\mathbf{P}_m$ is measured by the *relative entropy*[1]. A first approach would be to construct an agent $\mathbf{B}$ so as to minimize the total expected relative entropy to $\mathbf{P}_m$. This is constructed as follows. Define the history-dependent relative entropies over the action a_t and observation o_t as

$$D_m^{a_t}(\underline{ao}_{<t}) := \sum_{a_t} \mathbf{P}_m(a_t|\underline{ao}_{<t}) \log \frac{\mathbf{P}_m(a_t|\underline{ao}_{<t})}{\mathbf{Pr}(a_t|\underline{ao}_{<t})}$$

$$D_m^{o_t}(\underline{ao}_{<t}a_t) := \sum_{o_t} \mathbf{P}_m(o_t|\underline{ao}_{<t}a_t) \log \frac{\mathbf{P}_m(o_t|\underline{ao}_{<t}a_t)}{\mathbf{Pr}(o_t|\underline{ao}_{<t}a_t)},$$

where $\mathbf{P}_m(o_t|\underline{ao}_{<t}a_t) = \mathbf{Q}_m(o_t|\underline{ao}_{<t}a_t)$ because the $\mathbf{Q}_m$ are known and where $\mathbf{Pr}$ will be the argument of the variational problem. Then, one removes the dependency on the past by averaging over all possible histories:

$$D_m^{a_t} := \sum_{\underline{ao}_{<t}} \mathbf{P}_m(\underline{ao}_{<t}) D_m^{a_t}(\underline{ao}_{<t})$$

$$D_m^{o_t} := \sum_{\underline{ao}_{<t}a_t} \mathbf{P}_m(\underline{ao}_{<t}a_t) D_m^{o_t}(\underline{ao}_{<t}a_t).$$

Finally, the total expected relative entropy of $\mathbf{Pr}$ from $\mathbf{P}_m$ is obtained by summing up all time steps and then by averaging over all choices of the true environment:

$$D := \limsup_{t\to\infty} \sum_m P(m) \sum_{\tau=1}^{t} \big(D_m^{a_\tau} + D_m^{o_\tau}\big). \tag{2}$$

1. The *relative entropy* is also known as the *KL-divergence* and it measures the average amount of extra bits that are necessary to encode symbols due to the usage of the (wrong) predictor.

Using (2), one can define a variational problem with respect to $\mathbf{Pr}$. The agent $\mathbf{B}$ that one is looking for is the system $\mathbf{Pr}$ that minimizes the total expected relative entropy in (2), i.e.

$$\mathbf{B} := \arg\min_{\mathbf{Pr}} D(\mathbf{Pr}). \tag{3}$$

The solution to Equation 3 is the system $\mathbf{B}$ defined by the set of equations

$$\begin{aligned} \mathbf{B}(a_t|\underline{ao}_{<t}) &= \sum_m \mathbf{P}_m(a_t|\underline{ao}_{<t}) w_m(\underline{ao}_{<t}) \\ \mathbf{B}(o_t|\underline{ao}_{<t}a_t) &= \sum_m \mathbf{P}_m(o_t|\underline{ao}_{<t}a_t) w_m(\underline{ao}_{<t}a_t) \end{aligned} \tag{4}$$

valid for all $\underline{ao}_{\leq t} \in \mathcal{Z}^*$, where the mixture weights are

$$\begin{aligned} w_m(\underline{ao}_{<t}) &:= \frac{P(m)\mathbf{P}_m(\underline{ao}_{<t})}{\sum_{m'} P(m')\mathbf{P}_{m'}(\underline{ao}_{<t})} \\ w_m(\underline{ao}_{<t}a_t) &:= \frac{P(m)\mathbf{P}_m(\underline{ao}_{<t}a_t)}{\sum_{m'} P(m')\mathbf{P}_{m'}(\underline{ao}_{<t}a_t)}. \end{aligned} \tag{5}$$

For reference, see the work of Haussler and Opper (1997) and Opper (1998). It is clear that $\mathbf{B}$ is just the Bayesian mixture over the agents $\mathbf{P}_m$. If one defines the conditional probabilities

$$\begin{aligned} P(a_t|m, \underline{ao}_{<t}) &:= \mathbf{P}_m(a_t|\underline{ao}_{<t}) \\ P(o_t|m, \underline{ao}_{<t}a_t) &:= \mathbf{P}_m(a_t|\underline{ao}_{<t}a_t) \end{aligned} \tag{6}$$

for all $\underline{ao}_{\leq t} \in \mathcal{Z}^*$, then Equation 4 can be rewritten as

$$\begin{aligned} \mathbf{B}(a_t|\underline{ao}_{<t}) &= \sum_m P(a_t|m, \underline{ao}_{<t}) P(m|\underline{ao}_{<t}) = P(a_t|\underline{ao}_{<t}) \\ \mathbf{B}(o_t|\underline{ao}_{<t}a_t) &= \sum_m P(o_t|m, \underline{ao}_{<t}a_t) P(m|\underline{ao}_{<t}a_t) = P(o_t|\underline{ao}_{<t}a_t) \end{aligned} \tag{7}$$

where the $P(m|\underline{ao}_{<t}) = w_m(\underline{ao}_{<t})$ and $P(m|\underline{ao}_{<t}a_t) = w_m(\underline{ao}_{<t}a_t)$ are just the posterior probabilities over the elements in $\mathcal{M}$ given the past interactions. Hence, the conditional probabilities in (4) that minimize the total expected divergence are just the predictive distributions $P(a_t|\underline{ao}_{<t})$ and $P(o_t|\underline{ao}_{<t}a_t)$ that one obtains by standard probability theory, and in particular, Bayes' rule. This is interesting, as it provides a teleological interpretation for Bayes' rule.

The behavior of $\mathbf{B}$ can be described as follows. At any given time t, $\mathbf{B}$ maintains a mixture over systems $\mathbf{P}_m$. The weighting over them is given by the mixture coefficients w_m. Whenever a new action a_t *or* a new observation o_t is produced (by the agent or the environment respectively), the weights w_m are updated according to Bayes' rule. In addition, $\mathbf{B}$ issues an action a_t suggested by a system $\mathbf{P}_m$ drawn randomly according to the weights w_t.

However, there is an important problem with $\mathbf{B}$ that arises due to the fact that it is not only a system that is passively observing symbols, but also *actively generating* them. In the subjective interpretation of probability theory, conditionals play the role of observations made by the agent that have been generated by an external source. This interpretation suits the symbols $o_1, o_2, o_3, \ldots$ because they have been issued by the environment. However, symbols that are generated by the system itself require a fundamentally different belief update. Intuitively, the difference can be explained as follows. Observations provide information that allows the agent inferring properties about the environment. In contrast, actions do not carry information about the environment, and thus have to be incorporated differently into the belief of the agent. In the following section we illustrate this problem with a simple statistical example.

3.2 Causality

Causality is the study of the *functional dependencies* of events. This stands in contrast to statistics, which, on an abstract level, can be said to study the *equivalence dependencies* (i.e. co-occurrence or correlation) amongst events. Causal statements differ fundamentally from statistical statements. Examples that highlight the differences are many, such as "do smokers *get* lung cancer?" as opposed to "do smokers *have* lung cancer?"; "*assign* $y \leftarrow f(x)$" as opposed to "*compare* $y = f(x)$" in programming languages; and "$a \leftarrow F/m$" as opposed to "$F = m\,a$" in Newtonian physics. The study of causality has recently enjoyed considerable attention from researchers in the fields of statistics and machine learning. Especially over the last decade, significant progress has been made towards the formal understanding of causation (Shafer, 1996; Pearl, 2000; Spirtes et al., 2000; Dawid, 2010). In this subsection, the aim is to provide the essential tools required to understand causal interventions. For a more in-depth exposition of causality, the reader is referred to the specialized literature.

To illustrate the need for causal considerations in the case of generated symbols, consider the following thought experiment. Suppose a statistician is asked to design a model for a simple time series $X_1, X_2, X_3, \ldots$ and she decides to use a Bayesian method. Assume she collects a first observation $X_1 = x_1$. She computes the posterior probability density function (pdf) over the parameters θ of the model given the data using Bayes' rule:

$$p(\theta|X_1 = x_1) = \frac{p(X_1 = x_1|\theta)p(\theta)}{\int p(X_1 = x_1|\theta')p(\theta')\,d\theta'},$$

where $p(X_1 = x_1|\theta)$ is the likelihood of x_1 given θ and $p(\theta)$ is the prior pdf of θ. She can use the model to predict the next observation by drawing a sample x_2 from the predictive pdf

$$p(X_2 = x_2|X_1 = x_1) = \int p(X_2 = x_2|X_1 = x_1, \theta)\,p(\theta|X_1 = x_1)\,d\theta,$$

where $p(X_2 = x_2|X_1 = x_1, \theta)$ is the likelihood of x_2 given x_1 and θ. Note that x_2 is *not* drawn from $p(X_2 = x_2|X_1 = x_1, \theta)$. She understands that the nature of x_2 is very different from x_1: *while x_1 is informative and does change the belief state of the Bayesian model, x_2 is non-informative and thus is a reflection of the model's belief state.* Hence, she would never use x_2 to further condition the Bayesian model. Mathematically, she seems to imply that

$$p(\theta|X_1 = x_1, X_2 = x_2) = p(\theta|X_1 = x_1)$$

if x_2 has been generated from $p(X_2|X_1 = x_1)$ itself. But this simple independence assumption is not correct as the following elaboration of the example will show.

The statistician is now told that the source is waiting for the simulated data point x_2 in order to produce a next observation $X_3 = x_3$ which does depend on x_2. She hands in x_2 and obtains a new observation x_3. Using Bayes' rule, the posterior pdf over the parameters is now

$$\frac{p(X_3 = x_3|X_1 = x_1, X_2 = x_2, \theta)\, p(X_1 = x_1|\theta)\, p(\theta)}{\int p(X_3 = x_3|X_1 = x_1, X_2 = x_2, \theta')\, p(X_1 = x_1|\theta')\, p(\theta')\, d\theta'} \tag{8}$$

where $p(X_3 = x_3|X_1 = x_1, X_2 = x_2, \theta)$ is the likelihood of the new data x_3 given the old data x_1, the parameters θ *and the simulated data* x_2. Notice that this looks almost like the posterior pdf $p(\theta|X_1 = x_1, X_2 = x_2, X_3 = x_3)$ given by

$$\frac{p(X_3 = x_3|X_1 = x_1, X_2 = x_2, \theta)\, p(X_2 = x_2|X_1 = x_1, \theta)\, p(X_1 = x_1|\theta)\, p(\theta)}{\int p(X_3 = x_3|X_1 = x_1, X_2 = x_2, \theta')\, p(X_2 = x_2|X_1 = x_1, \theta')\, p(X_1 = x_1|\theta')\, p(\theta')\, d\theta'}$$

with the exception that in the latter case, the Bayesian update contains the likelihoods of the simulated data $p(X_2 = x_2|X_1 = x_1, \theta)$. This suggests that Equation 8 is a variant of the posterior pdf $p(\theta|X_1 = x_1, X_2 = x_2, X_3 = x_3)$ but where the simulated data x_2 is treated in a different way than the data x_1 and x_3.

Define the pdf p' such that the pdfs $p'(\theta)$, $p'(X_1|\theta)$, $p'(X_3|X_1, X_2, \theta)$ are identical to $p(\theta)$, $p(X_1|\theta)$ and $p(X_3|X_2, X_1, \theta)$ respectively, but differ in $p'(X_2|X_1, \theta)$:

$$p'(X_2|X_1, \theta) = \delta(X_2 - x_2).$$

where δ is the Dirac delta function. That is, p' is identical to p but it assumes that the value of X_2 is fixed to x_2 given X_1 and θ. For p', the simulated data x_2 is non-informative:

$$-\log_2 p'(X_2 = x_2|X_1, \theta) = 0.$$

If one computes the posterior pdf $p'(\theta|X_1 = x_1, X_2 = x_2, X_3 = x_3)$, one obtains the result of Equation 8:

$$\begin{aligned}\frac{p'(X_3 = x_3|X_1 = x_1, X_2 = x_2, \theta)\, p'(X_2 = x_2|X_1 = x_1, \theta)\, p'(X_1 = x_1|\theta)\, p'(\theta)}{\int p'(X_3 = x_3|X_1 = x_1, X_2 = x_2, \theta') p'(X_2 = x_2|X_1 = x_1, \theta')\, p'(X_1 = x_1|\theta')\, p'(\theta')\, d\theta'} \\ = \frac{p(X_3 = x_3|X_1 = x_1, X_2 = x_2, \theta)\, p(X_1 = x_1|\theta)\, p(\theta)}{\int p(X_3 = x_3|X_1 = x_1, X_2 = x_2, \theta')\, p(X_1 = x_1|\theta')\, p(\theta')\, d\theta'}.\end{aligned}$$

Thus, in order to explain Equation 8 as a posterior pdf given the observed data x_1 and x_3 and the generated data x_2, one has to *intervene* p in order to account for the fact that x_2 *is non-informative given* x_1 *and* θ. In other words, the statistician, by defining the value of X_2 herself[2], has changed the (natural) regime that brings about the series $X_1, X_2, X_3, \ldots$, which is mathematically expressed by redefining the pdf.

Two essential ingredients are needed to carry out interventions. First, one needs to know the functional dependencies amongst the random variables of the probabilistic model. This is provided by the *causal model*, i.e. the unique factorization of the joint probability

2. Note that this is conceptually broken down into two steps: first, she samples x_2 from $p(X_2|X_1 = x_1)$; and second, she imposes the value $X_2 = x_2$ by setting $p'(X_2|X_1, \theta) = \delta(X_2 - x_2)$.

distribution over the random variables encoding the causal dependencies. In the general case, this defines a partial order over the random variables. In the previous thought experiment, the causal model of the joint pdf $p(\theta, X_1, X_2, X_3)$ is given by the set of conditional pdfs

$$p(\theta), p(X_1|\theta), p(X_2|X_1, \theta), p(X_3|X_1, X_2, \theta).$$

Second, one defines the *intervention* that sets X to the value x, denoted as $X \leftarrow x$, as the operation on the causal model replacing the conditional probability of X by a Dirac delta function $\delta(X - x)$ or a Kronecker delta δ_x^X for a continuous or a discrete variable X respectively. In our thought experiment, it is easily seen that

$$p'(\theta, X_1 = x_1, X_2 = x_2, X_3 = x_3) = p(\theta, X_1 = x_1, X_2 \leftarrow x_2, X_3 = x_3)$$

and thereby,

$$p'(\theta|X_1 = x_1, X_2 = x_2, X_3 = x_3) = p(\theta|X_1 = x_1, X_2 \leftarrow x_2, X_3 = x_3).$$

Causal models contain additional information that is not available in the joint probability distribution alone. The appropriate model for a given situation depends on the story that is being told. Note that an intervention can lead to different results if the respective causal models differ. Thus, if the causal model had been

$$p(X_3), p(X_2|X_3), p(X_1|X_2, X_3), p(\theta|X_1, X_2, X_3)$$

then the intervention $X_2 \leftarrow x_2$ would differ from p', i.e.

$$p'(\theta, X_1 = x_1, X_2 = x_2, X_3 = x_3) \neq p(\theta, X_1 = x_1, X_2 \leftarrow x_2, X_3 = x_3),$$

even though both causal models represent the same joint probability distribution. In the following, this paper will use the shorthand notation $\hat{x} := X \leftarrow x$ when the random variable is obvious from the context.

3.3 Causal Construction of Adaptive Agents

Following the discussion in the previous section, an adaptive agent $\mathbf{P}$ is going to be constructed by minimizing the expected relative entropy to the expected $\mathbf{P}_m$, but this time treating actions as interventions. Based on the definition of the conditional probabilities in Equation 6, the total expected relative entropy to characterize $\mathbf{P}$ using interventions is going to be defined. Assuming the environment is chosen first, and that each symbol depends functionally on the environment and all the previously generated symbols, the causal model is given by

$$P(m), P(a_1|m), P(o_1|m, a_1), P(a_2|m, a_1, o_1), P(o_2|m, a_1, o_1, a_2), \ldots$$

Importantly, interventions index a set of intervened probability distributions derived from a base probability distribution. Hence, the set of fixed intervention sequences of the form $\hat{a}_1, \hat{a}_2, \ldots$ indexes probability distributions over observation sequences $o_1, o_2, \ldots$. Because of this, one defines a set of criteria indexed by the intervention sequences, but it will be clear that they all have the same solution. Define the history-dependent intervened relative entropies over the action a_t and observation o_t as

$$C_m^{a_t}(\underline{\hat{a}o}_{<t}) := \sum_{a_t} P(a_t|m, \underline{\hat{a}o}_{<t}) \log_2 \frac{P(a_t|m, \underline{\hat{a}o}_{<t})}{\mathbf{Pr}(a_t|\underline{ao}_{<t})}$$
$$C_m^{o_t}(\underline{\hat{a}o}_{<t}\hat{a}_t) := \sum_{o_t} P(o_t|m, \underline{\hat{a}o}_{<t}\hat{a}_t) \log_2 \frac{P(o_t|m, \underline{\hat{a}o}_{<t}\hat{a}_t)}{\mathbf{Pr}(o_t|\underline{ao}_{<t}a_t)},$$

where $\mathbf{Pr}$ is a given arbitrary agent. Note that past actions are treated as interventions. In particular, $P(a_t|m, \underline{\hat{a}o}_{<t})$ represents the knowledge state when the past actions have already been issued but the next action a_t is not known yet. Then, averaging the previous relative entropies over all pasts yields

$$C_m^{a_t} = \sum_{\underline{ao}_{<t}} P(\underline{\hat{a}o}_{<t}|m) C_m^{a_t}(\underline{\hat{a}o}_{<t})$$
$$C_m^{o_t} = \sum_{\underline{ao}_{<t}a_t} P(\underline{\hat{a}o}_{<t}\hat{a}_t|m) C_m^{o_t}(\underline{\hat{a}o}_{<t}\hat{a}_t).$$

Here again, because of the knowledge state in time represented by $C_m^{a_t}(\underline{\hat{a}o}_{<t})$ and $C_m^{o_t}(\underline{\hat{a}o}_{<t}\hat{a}_t)$, the averages are taken treating past actions as interventions. Finally, define the total expected relative entropy of $\mathbf{Pr}$ from $\mathbf{P}_m$ as the sum of $(C_m^{a_t} + C_m^{o_t})$ over time, averaged over the possible draws of the environment:

$$C := \limsup_{t\to\infty} \sum_m P(m) \sum_{\tau=1}^{t} \big(C_m^{a_\tau} + C_m^{o_\tau}\big). \tag{9}$$

The variational problem consists in choosing the agent $\mathbf{P}$ as the system $\mathbf{Pr}$ minimizing $C = C(\mathbf{Pr})$, i.e.

$$\mathbf{P} := \arg\min_{\mathbf{Pr}} C(\mathbf{Pr}). \tag{10}$$

The following theorem shows that this variational problem has a unique solution, which will be the central theme of this paper.

Theorem 1. *The solution to Equation 10 is the system* $\mathbf{P}$ *defined by the set of equations*

$$\begin{aligned}\mathbf{P}(a_t|\underline{ao}_{<t}) = P(a_t|\underline{\hat{a}o}_{<t}) &= \sum_m P(a_t|m, \underline{ao}_{<t}) v_m(\underline{ao}_{<t})\\ \mathbf{P}(o_t|\underline{ao}_{<t}a_t) = P(o_t|\underline{\hat{a}o}_{<t}\hat{a}_t) &= \sum_m P(o_t|m, \underline{ao}_{<t}a_t) v_m(\underline{ao}_{<t}a_t)\end{aligned} \tag{11}$$

valid for all $\underline{ao}_{\leq t} \in \mathcal{Z}^*$, *where the mixture weights are*

$$v_m(\underline{ao}_{<t}a_t) = v_m(\underline{ao}_{<t}) := \frac{P(m)\prod_{\tau=1}^{t-1} P(o_\tau|m, \underline{ao}_{<\tau}a_\tau)}{\sum_{m'} P(m')\prod_{\tau=1}^{t-1} P(o_\tau|m', \underline{ao}_{<\tau}a_\tau)}. \tag{12}$$

Bayesian Control Rule: Given a set of operation modes $\{P(\cdot|m,\cdot)\}_{m\in\mathcal{M}}$ over interaction sequences in $\mathcal{Z}^\infty$ and a prior distribution $P(m)$ over the parameters $\mathcal{M}$, the probability of the action a_{t+1} is given by

$$P(a_{t+1}|\hat{\underline{ao}}_{\leq t}) = \sum_m P(a_{t+1}|m, \underline{ao}_{\leq t})P(m|\hat{\underline{ao}}_{\leq t}), \tag{13}$$

where the posterior probability over operation modes is given by the recursion

$$P(m|\hat{\underline{ao}}_{\leq t}) = \frac{P(o_t|m, \underline{ao}_{<t})P(m|\hat{\underline{ao}}_{<t})}{\sum_{m'} P(o_t|m', \underline{ao}_{<t})P(m'|\hat{\underline{ao}}_{<t})}.$$

Table 1: Summary of the Bayesian control rule.

The theorem says that the optimal solution to the variational problem in (10) is precisely the predictive distribution over actions and observations treating actions as interventions and observations as conditionals, i.e. it is the solution that one would obtain by applying *only standard probability and causal calculus*. This provides a teleological interpretation for the agent **P** akin to the naïve agent **B** constructed in Section 3.1. The behavior of **P** differs in an important aspect from **B**. At any given time t, **P** maintains a mixture over systems $\mathbf{P}_m$. The weighting over these systems is given by the mixture coefficients v_m. In contrast to **B**, **P** updates the weights v_m *only* whenever a new observation o_t is produced by the environment. The update follows Bayes' rule but treats past actions as interventions by dropping the evidence they provide. In addition, **P** issues an action a_t suggested by an system m drawn randomly according to the weights v_m.

3.4 Summary

Adaptive control is formalized as the problem of designing an agent for an unknown environment chosen from a class of possible environments. If the environment-specific agents are known, then the Bayesian control rule allows constructing an adaptive agent by combining these agents. The resulting adaptive agent is universal with respect to the environment class. In this context, the constituent agents are called the *operation modes* of the adaptive agent. They are represented by causal models over the interaction sequences, i.e. conditional probabilities $P(a_t|m, \underline{ao}_{<t})$ and $P(o_t|m, \underline{ao}_{<t})$ for all $\underline{ao}_{\leq t} \in \mathcal{Z}^*$, and where $m \in \mathcal{M}$ is the index or parameter characterizing the operation mode. The probability distribution over the input stream (output stream) is called the *hypothesis (policy)* of the operation mode. Table 1 collects the essential equations of the Bayesian control rule. In particular, there the rule is stated using a recursive belief update.

4. Convergence

The aim of this section is to develop a set of sufficient conditions of convergence and then to provide a proof of convergence. To simplify the exposition, the analysis has been limited to the case of controllers having a finite number of input-output models.

4.1 Policy Diagrams

In the following we use "policy diagrams" as a useful informal tool to analyze the effect of policies on environments. Figure 2 illustrates an example.

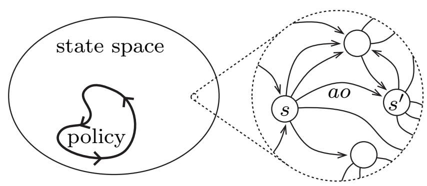

Figure 2: A policy diagram. One can imagine an environment as a collection of states connected by transitions labeled by I/O symbols. The zoom highlights a state s where taking action $a \in \mathcal{A}$ and collecting observation $o \in \mathcal{O}$ leads to state s'. Sets of states and transitions are represented as enclosed areas similar to a Venn diagram. Choosing a particular policy in an environment amounts to partially controlling the transitions taken in the state space, thereby choosing a probability distribution over state transitions (e.g. a Markov chain given by the environmental dynamics). If the probability mass concentrates in certain areas of the state space, choosing a policy can be thought of as choosing a *subset* of the environment's dynamics. In the following, a policy is represented by a subset in state space (enclosed by a directed curve) as illustrated above.

Policy diagrams are especially useful to analyze the effect of policies on different hypotheses about the environment's dynamics. An agent that is endowed with a set of operation modes $\mathcal{M}$ can be seen as having *hypotheses* about the environment's underlying dynamics, given by the observation models $P(o_t|m, \underline{ao}_{<t}a_t)$, and associated *policies*, given by the action models $P(a_t|m, \underline{ao}_{<t})$, for all $m \in \mathcal{M}$. For the sake of simplifying the interpretation of policy diagrams, we will assume the existence of a state space $\mathcal{T} : (\mathcal{A} \times \mathcal{O})^* \to \mathcal{S}$ mapping I/O histories into states. Note however that no such assumptions are made to obtain the results of this section.

4.2 Divergence Processes

The central question in this section is to investigate whether the Bayesian control rule converges to the correct control law or not. That is, whether $P(a_t|\hat{\underline{ao}}_t) \to P(a_t|m^*, \underline{ao}_{<t})$ as $t \to \infty$ when m^* is the true operation mode, i.e. the operation mode such that $P(o_t|m^*, \underline{ao}_{<t}a_t) = Q(o_t|\underline{ao}_{<t}a_t)$. As will be obvious from the discussion in the rest of this section, this is in general not true.

As it is easily seen from Equation 13, showing convergence amounts to show that the posterior distribution $P(m|\hat{\underline{ao}}_{<t})$ concentrates its probability mass on a subset of operation

modes $\mathcal{M}^*$ having essentially the same output stream as m^*,

$$\sum_{m \in \mathcal{M}} P(a_t|m, \underline{ao}_{<t})P(m|\underline{\hat{a}o}_{<t}) \approx \sum_{m \in \mathcal{M}^*} P(a_t|m^*, \underline{ao}_{<t})P(m|\underline{\hat{a}o}_{<t}) \approx P(a_t|m^*, \underline{ao}_{<t}).$$

Hence, understanding the asymptotic behavior of the posterior probabilities

$$P(m|\underline{\hat{a}o}_{\leq t})$$

is crucial here. In particular, we need to understand under what conditions these quantities converge to zero. The posterior can be rewritten as

$$P(m|\underline{\hat{a}o}_{\leq t}) = \frac{P(\underline{\hat{a}o}_{\leq t}|m)P(m)}{\sum_{m' \in \mathcal{M}} P(\underline{\hat{a}o}_{\leq t}|m')P(m')} = \frac{P(m)\prod_{\tau=1}^{t} P(o_\tau|m, \underline{ao}_{<\tau}a_\tau)}{\sum_{m' \in \mathcal{M}} P(m')\prod_{\tau=1}^{t} P(o_\tau|m', \underline{ao}_{<\tau}a_\tau)}.$$

If all the summands but the one with index m^* are dropped from the denominator, one obtains the bound

$$P(m|\underline{\hat{a}o}_{\leq t}) \leq \frac{P(m)}{P(m^*)} \prod_{\tau=1}^{t} \frac{P(o_\tau|m, \underline{ao}_{<\tau}a_\tau)}{P(o_\tau|m^*, \underline{ao}_{<\tau}a_\tau)},$$

which is valid for all $m^* \in \mathcal{M}$. From this inequality, it is seen that it is convenient to analyze the behavior of the stochastic process

$$d_t(m^*\|m) := \sum_{\tau=1}^{t} \ln \frac{P(o_\tau|m^*, \underline{ao}_{<\tau}a_\tau)}{P(o_\tau|m, \underline{ao}_{<\tau}a_\tau)}$$

which is the *divergence process* of m from the reference m^*. Indeed, if $d_t(m^*\|m) \to \infty$ as $t \to \infty$, then

$$\lim_{t \to \infty} \frac{P(m)}{P(m^*)} \prod_{\tau=1}^{t} \frac{P(o_\tau|m, \underline{ao}_{<\tau}a_\tau)}{P(o_\tau|m^*, \underline{ao}_{<\tau}a_\tau)} = \lim_{t \to \infty} \frac{P(m)}{P(m^*)} \cdot e^{-d_t(m^*\|m)} = 0,$$

and thus clearly $P(m|\underline{\hat{a}o}_{\leq t}) \to 0$. Figure 3 illustrates simultaneous realizations of the divergence processes of a controller. Intuitively speaking, these processes provide lower bounds on accumulators of surprise value measured in information units.

A divergence process is a random walk whose value at time t depends on the whole history up to time $t-1$. What makes these divergence processes cumbersome to characterize is the fact that their statistical properties depend on the particular policy that is applied; hence, a given divergence process can have different growth rates depending on the policy (Figure 4). Indeed, the behavior of a divergence process might depend critically on the distribution over actions that is used. For example, it can happen that a divergence process stays stable under one policy, but diverges under another. In the context of the Bayesian control rule this problem is further aggravated, because in each time step, the policy that is applied is determined stochastically. More specifically, if m^* is the true operation mode, then $d_t(m^*\|m)$ is a random variable that depends on the realization $\underline{ao}_{\leq t}$ which is drawn from

$$\prod_{\tau=1}^{t} P(a_\tau|m_\tau, \underline{ao}_{<\tau})P(o_\tau|m^*, \underline{ao}_{<\tau}a_\tau),$$

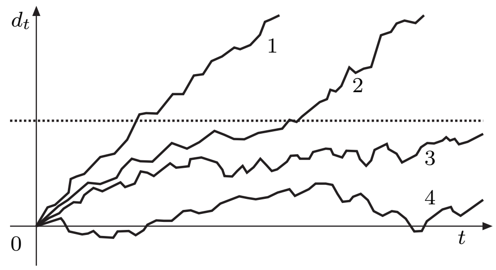

Figure 3: Realization of the divergence processes 1 to 4 associated to a controller with operation modes m_1 to m_4. The divergence processes 1 and 2 diverge, whereas 3 and 4 stay below the dotted bound. Hence, the posterior probabilities of m_1 and m_2 vanish.

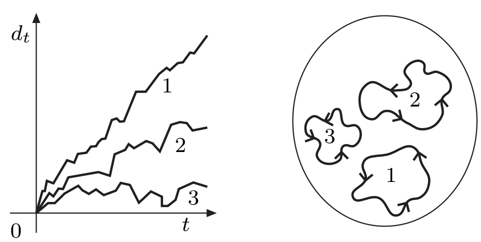

Figure 4: The application of different policies lead to different statistical properties of the same divergence process.

where the $m_1, m_2, \dots, m_t$ are drawn themselves from $P(m_1), P(m_2|\underline{\hat{a}o}_1), \dots, P(m_t|\underline{\hat{a}o}_{<t})$.

To deal with the heterogeneous nature of divergence processes, one can introduce a temporal decomposition that demultiplexes the original process into many sub-processes belonging to unique policies. Let $\mathcal{N}_t := \{1, 2, \dots, t\}$ be the set of time steps up to time t. Let $\mathcal{T} \subset \mathcal{N}_t$, and let $m, m' \in \mathcal{M}$. Define a *sub-divergence* of $d_t(m^*\|m)$ as a random variable

$$g_{m'}(m; \mathcal{T}) := \sum_{\tau \in \mathcal{T}} \ln \frac{P(o_\tau|m^*, \underline{ao}_{<\tau}a_\tau)}{P(o_\tau|m, \underline{ao}_{<\tau}a_\tau)}$$

drawn from

$$P_{m'}(\{\underline{ao}_\tau\}_{\tau\in\mathcal{T}}|\{\underline{ao}_\tau\}_{\tau\in\mathcal{T}^\complement}) := \Big(\prod_{\tau\in\mathcal{T}} P(a_\tau|m', \underline{ao}_{<\tau})\Big)\Big(\prod_{\tau\in\mathcal{T}} P(o_\tau|m^*, \underline{ao}_{<\tau}a_\tau)\Big),$$

where $\mathcal{T}^\complement := \mathcal{N}_t \setminus \mathcal{T}$ and where $\{\underline{ao}_\tau\}_{\tau\in\mathcal{T}^\complement}$ are given conditions that are kept constant. In this definition, m' plays the role of the policy that is used to sample the actions in the time steps $\mathcal{T}$. Clearly, any realization of the divergence process $d_t(m^*\|m)$ can be decomposed into a sum of sub-divergences, i.e.

$$d_t(m^*\|m) = \sum_{m'} g_{m'}(m; \mathcal{T}_{m'}), \tag{14}$$

where $\{\mathcal{T}_m\}_{m\in\mathcal{M}}$ forms a partition of $\mathcal{N}_t$. Figure 5 shows an example decomposition.

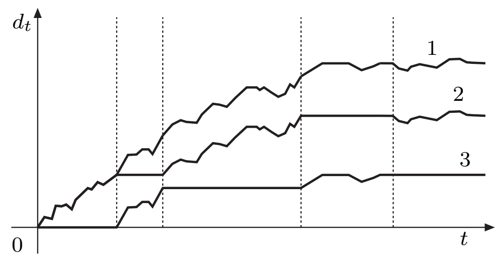

Figure 5: Decomposition of a divergence process (1) into sub-divergences (2 & 3).

The averages of sub-divergences will play an important role in the analysis. Define the average over all realizations of $g_{m'}(m; \mathcal{T})$ as

$$G_{m'}(m; \mathcal{T}) := \sum_{(\underline{ao}_\tau)_{\tau\in\mathcal{T}}} P_{m'}(\{\underline{ao}_\tau\}_{\tau\in\mathcal{T}}|\{\underline{ao}_\tau\}_{\tau\in\mathcal{T}^\complement}) g_{m'}(m; \mathcal{T}).$$

Notice that for any $\tau \in \mathcal{N}_t$,

$$G_{m'}(m; \{\tau\}) = \sum_{\underline{ao}_\tau} P(a_\tau|m', \underline{ao}_{<\tau}) P(o_\tau|m^*, \underline{ao}_{<\tau}a_\tau) \ln \frac{P(o_\tau|m^*, \underline{ao}_{<\tau}a_\tau)}{P(o_\tau|m, \underline{ao}_{<\tau}a_\tau)} \geq 0,$$

because of Gibbs' inequality. In particular,

$$G_{m'}(m^*; \{\tau\}) = 0.$$

Clearly, this holds as well for any $\mathcal{T} \subset \mathcal{N}_t$:

$$\begin{aligned} \forall m \quad G_{m'}(m; \mathcal{T}) &\geq 0, \\ G_{m'}(m^*; \mathcal{T}) &= 0. \end{aligned} \tag{15}$$

4.3 Boundedness

In general, a divergence process is very complex: virtually all the classes of distributions that are of interest in control go well beyond the assumptions of i.i.d. and stationarity. This increased complexity can jeopardize the analytic tractability of the divergence process, such that no predictions about its asymptotic behavior can be made anymore. More specifically, if the growth rates of the divergence processes vary too much from realization to realization, then the posterior distribution over operation modes can vary qualitatively between realizations. Hence, one needs to impose a stability requirement akin to ergodicity to limit the class of possible divergence-processes to a class that is analytically tractable. For this purpose the following property is introduced.

A divergence process $d_t(m^*\|m)$ is said to have *bounded variation* in $\mathcal{M}$ iff for any $\delta > 0$, there is a $C \geq 0$, such that for all $m' \in \mathcal{M}$, all t and all $\mathcal{T} \subset \mathcal{N}_t$

$$\left| g_{m'}(m;\mathcal{T}) - G_{m'}(m;\mathcal{T}) \right| \leq C$$

with probability $\geq 1 - \delta$.

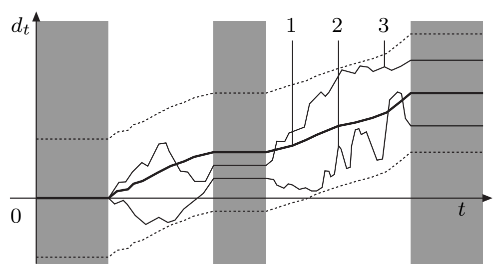

Figure 6: If a divergence process has bounded variation, then the realizations (curves 2 & 3) of a sub-divergence stay within a band around the mean (curve 1).

Figure 6 illustrates this property. Boundedness is the key property that is going to be used to construct the results of this section. The first important result is that the posterior probability of the true input-output model is bounded from below.

Theorem 2. *Let the set of operation modes of a controller be such that for all $m \in \mathcal{M}$ the divergence process $d_t(m^*\|m)$ has bounded variation. Then, for any $\delta > 0$, there is a $\lambda > 0$, such that for all $t \in \mathbb{N}$,*

$$P(m^*|\underline{\hat{a}o}_{\leq t}) \geq \frac{\lambda}{|\mathcal{M}|}$$

with probability $\geq 1 - \delta$.

4.4 Core

If one wants to identify the operation modes whose posterior probabilities vanish, then it is not enough to characterize them as those modes whose hypothesis does not match the true hypothesis. Figure 7 illustrates this problem. Here, three hypotheses along with their associated policies are shown. H_1 and H_2 share the prediction made for region A but differ in region B. Hypothesis H_3 differs everywhere from the others. Assume H_1 is true. As long as we apply policy P_2, hypothesis H_3 will make wrong predictions and thus its divergence process will diverge as expected. However, no evidence against H_2 will be accumulated. It is only when one applies policy P_1 *for long enough time* that the controller will eventually enter region B and hence accumulate counter-evidence for H_2.

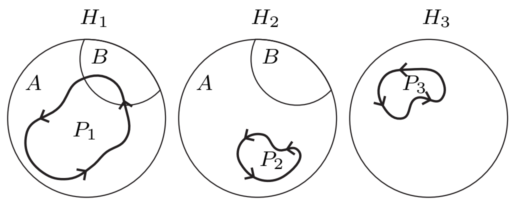

Figure 7: If hypothesis H_1 is true and agrees with H_2 on region A, then policy P_2 cannot disambiguate the three hypotheses.

But what does "long enough" mean? If P_1 is executed only for a short period, then the controller risks not visiting the disambiguating region. But unfortunately, neither the right policy nor the right length of the period to run it are known beforehand. Hence, an agent needs a clever time-allocating strategy to test all policies for all finite time intervals. This motivates the following definition.

The *core* of an operation mode m^*, denoted as $[m^*]$, is the subset of $\mathcal{M}$ containing operation modes behaving like m^* under its policy. More formally, an operation mode $m \notin [m^*]$ (i.e. is *not* in the core) iff for any $C \geq 0$, $\delta > 0$, there is a $\xi > 0$ and a $t_0 \in \mathbb{N}$, such that for all $t \geq t_0$,

$$G_{m^*}(m; \mathcal{T}) \geq C$$

with probability $\geq 1 - \delta$, where $G_{m^*}(m; \mathcal{T})$ is a sub-divergence of $d_t(m^* \| m)$, and $\mathbf{Pr}\{\tau \in \mathcal{T}\} \geq \xi$ for all $\tau \in \mathcal{N}_t$.

In other words, if the agent was to apply m^*'s policy in each time step with probability at least ξ, and under this strategy the expected sub-divergence $G_{m^*}(m; \mathcal{T})$ of $d_t(m^* \| m)$ grows unboundedly, then m is not in the core of m^*. Note that demanding a strictly positive probability of execution in each time step guarantees that the agent will run m^* for all possible finite time-intervals. As the following theorem shows, the posterior probabilities of the operation modes that are not in the core vanish almost surely.

Theorem 3. *Let the set of operation modes of an agent be such that for all $m \in \mathcal{M}$ the divergence process $d_t(m^* \| m)$ has bounded variation. If $m \notin [m^*]$, then $P(m|\underline{\hat{a}o}_{\leq t}) \to 0$ as $t \to \infty$ almost surely.*

4.5 Consistency

Even if an operation mode m is in the core of m^*, i.e. given that m is essentially indistinguishable from m^* under m^*'s control, it can still happen that m^* and m have different policies. Figure 8 shows an example of this. The hypotheses H_1 and H_2 share region A but differ in region B. In addition, both operation modes have their policies P_1 and P_2 respectively confined to region A. Note that both operation modes are in the core of each other. However, their policies are different. This means that it is unclear whether multiplexing the policies in time will ever disambiguate the two hypotheses. This is undesirable, as it could impede the convergence to the right control law.

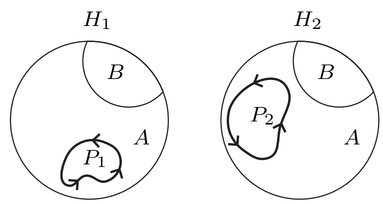

Figure 8: An example of inconsistent policies. Both operation modes are in the core of each other, but have different policies.

Thus, it is clear that one needs to impose further restrictions on the mapping of hypotheses into policies. With respect to Figure 8, one can make the following observations:

1. Both operation modes have policies that select subsets of region A. Therefore, the dynamics in A are preferred over the dynamics in B.

2. Knowing that the dynamics in A are preferred over the dynamics in B allows us to drop region B from the analysis when choosing a policy.

3. Since both hypotheses agree in region A, *they have to choose the same policy in order to be consistent in their selection criterion.*

This motivates the following definition. An operation mode m is said to be *consistent* with m^* iff $m \in [m^*]$ implies that for all $\varepsilon < 0$, there is a t_0, such that for all $t \geq t_0$ and all $\underline{ao}_{<t}a_t$,

$$\left| P(a_t|m, \underline{ao}_{\leq t}) - P(a_t|m^*, \underline{ao}_{\leq t}) \right| < \varepsilon.$$

In other words, if m is in the core of m^*, then m's policy has to converge to m^*'s policy. The following theorem shows that consistency is a sufficient condition for convergence to the right control law.

Theorem 4. *Let the set of operation modes of an agent be such that: for all $m \in \mathcal{M}$ the divergence process $d_t(m^* \| m)$ has bounded variation; and for all $m, m^* \in \mathcal{M}$, m is consistent with m^*. Then,*

$$P(a_t|\hat{\underline{ao}}_{<t}) \to P(a_t|m^*, \underline{ao}_{<t})$$

almost surely as $t \to \infty$.

4.6 Summary

In this section, a proof of convergence of the Bayesian control rule to the true operation mode has been provided for a finite set of operation modes. For this convergence result to hold, two necessary conditions are assumed: boundedness and consistency. The first one, *boundedness*, imposes the stability of divergence processes under the partial influence of the policies contained within the set of operation modes. This condition can be regarded as an ergodicity assumption. The second one, *consistency*, requires that if a hypothesis makes the same predictions as another hypothesis within its most relevant subset of dynamics, then both hypotheses share the same policy. This relevance is formalized as the *core* of an operation mode. The concepts and proof strategies strengthen the intuition about potential pitfalls that arise in the context of controller design. In particular we could show that the asymptotic analysis can be recast as the study of concurrent *divergence processes* that determine the evolution of the posterior probabilities over operation modes, thus abstracting away from the details of the classes of I/O distributions. The extension of these results to infinite sets of operation modes is left for future work. For example, one could think of partitioning a continuous space of operation modes into "essentially different" regions where representative operation modes subsume their neighborhoods (Grünwald, 2007).

5. Examples

In this section we illustrate the usage of the Bayesian control rule on two examples that are very common in the reinforcement learning literature: multi-armed bandits and Markov decision processes.

5.1 Bandit Problems

Consider the *multi-armed bandit problem* (Robbins, 1952). The problem is stated as follows. Suppose there is an N-armed bandit, i.e. a slot-machine with N levers. When pulled, lever i provides a reward drawn from a Bernoulli distribution with a bias h_i specific to that lever. That is, a reward $r = 1$ is obtained with probability h_i and a reward $r = 0$ with probability $1-h_i$. The objective of the game is to maximize the time-averaged reward through iterative pulls. There is a continuum range of stationary strategies, each one parameterized by N probabilities $\{s_i\}_{i=1}^N$ indicating the probabilities of pulling each lever. The difficulty arising in the bandit problem is to balance reward maximization based on the knowledge already acquired with attempting new actions to further improve knowledge. This dilemma is known as the exploration versus exploitation tradeoff (Sutton & Barto, 1998).

This is an ideal task for the Bayesian control rule, because each possible bandit has a known optimal agent. Indeed, a bandit can be represented by an N-dimensional bias vector $m = [m_1, \ldots, m_N] \in \mathcal{M} = [0; 1]^N$. Given such a bandit, the optimal policy consists in pulling the lever with the highest bias. That is, an operation mode is given by:

$$h_i = P(o_t = 1|m, a_t = i) = m_i \qquad s_i = P(a_t = i|m) = \begin{cases} 1 & \text{if } i = \max_j\{m_j\}, \\ 0 & \text{else.} \end{cases}$$

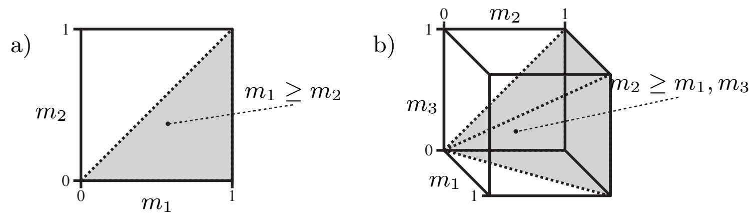

Figure 9: The space of bandit configurations can be partitioned into N regions according to the optimal lever. Panel a and b show the 2-armed and 3-armed bandit cases respectively.

To apply the Bayesian control rule, it is necessary to fix a prior distribution over the bandit configurations. Assuming a uniform distribution, the Bayesian control rule is

$$P(a_{t+1} = i|\underline{\hat{a}o}_{\leq t}) = \int_{\mathcal{M}} P(a_{t+1} = i|m)P(m|\underline{\hat{a}o}_{\leq t}) \tag{16}$$

with the update rule given by

$$P(m|\underline{\hat{a}o}_{\leq t}) = \frac{P(m)\prod_{\tau=1}^{t} P(o_\tau|m, a_\tau)}{\int_{\mathcal{M}} P(m')\prod_{\tau=1}^{t} P(o_\tau|m', a_\tau)\,dm'} = \prod_{j=1}^{N} \frac{m_j^{r_j}(1-m_j)^{f_j}}{\mathrm{B}(r_j+1, f_j+1)} \tag{17}$$

where r_j and f_j are the counts of the number of times a reward has been obtained from pulling lever j and the number of times no reward was obtained respectively. Observe that here the summation over discrete operation modes has been replaced by an integral over the continuous space of configurations. In the last expression we see that the posterior distribution over the lever biases is given by a product of N Beta distributions. Thus, sampling an action amounts to first sample an operation mode m by obtaining each bias m_i from a Beta distribution with parameters $r_i + 1$ and $f_i + 1$, and then choosing the action corresponding to the highest bias $a = \arg\max_i m_i$. The pseudo-code can be seen in Algorithm 1.

Simulation: The Bayesian control rule described above has been compared against two other agents: an ε-greedy strategy with decay (on-line) and Gittins indices (off-line). The test bed consisted of bandits with $N = 10$ levers whose biases were drawn uniformly at the beginning of each run. Every agent had to play 1000 runs for 1000 time steps each. Then, the performance curves of the individual runs were averaged. The ε-greedy strategy selects a random action with a small probability given by $\varepsilon\alpha^{-t}$ and otherwise plays the lever with highest expected reward. The parameters have been determined empirically to the values $\varepsilon = 0.1$, and $\alpha = 0.99$ after several test runs. They have been adjusted in a way to maximize the average performance in the last trials of our simulations. For the Gittins method, all the indices were computed up to horizon 1300 using a geometric discounting of $\alpha = 0.999$, i.e. close to one to approximate the time-averaged reward. The results are shown in Figure 10.

Algorithm 1 BCR bandit.

```
for all i = 1, ..., N do
  Initialize r_i and f_i to zero.
end for

for t = 1, 2, 3, ... do
  Sample m using (17).

  { Interaction }
  Set a ← arg max_i m_i and issue a.
  Obtain o from environment.

  {Update belief}
  if o = 1 then
    r_a = r_a + 1
  else
    f_a = f_a + 1
  end if
end for
```

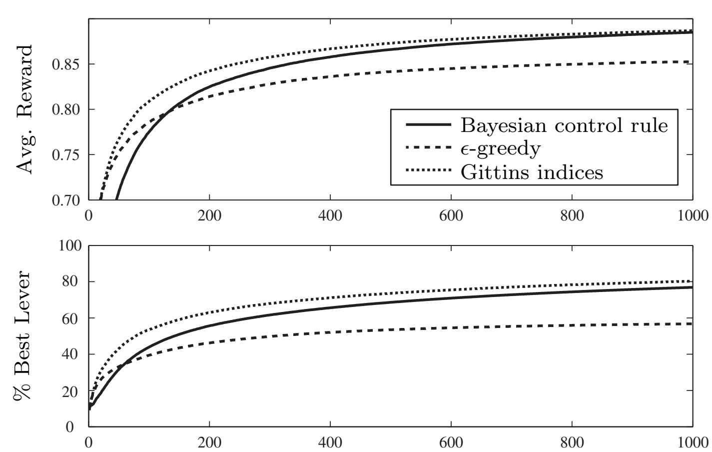

Figure 10: Comparison in the N-armed bandit problem of the Bayesian control rule (solid line), an ε-greedy agent (dashed line) and using Gittins indices (dotted line). 1,000 runs have been averaged. The top panel shows the evolution of the average reward. The bottom panel shows the evolution of the percentage of times the best lever was pulled.

It is seen that ε-greedy strategy quickly reaches an acceptable level of performance, but then seems to stall at a significantly suboptimal level, pulling the optimal lever only 60% of the time. In contrast, both the Gittins strategy and the Bayesian control rule show essentially the same asymptotic performance, but differ in the initial transient phase where the Gittins strategy significantly outperforms the Bayesian control rule. There are at least three observations that are worth making here. First, Gittins indices have to be pre-computed off-line. The time complexity scales quadratically with the horizon, and the computations for the horizon of 1300 steps took several hours on our machines. In contrast, the Bayesian control rule could be applied without pre-computation. Second, even though the Gittins method actively issues the optimal information gathering actions while the Bayesian control rule passively samples the actions from the posterior distribution over operation modes, in the end both methods rely on the convergence of the underlying Bayesian estimator. This implies that both methods have the same information bottleneck, since the Bayesian estimator requires the same amount of information to converge. Thus, active information gathering actions only affect the utility of the transient phase, not the permanent state. Other efficient algorithms for bandit problems can be found in the literature (Auer, CesaBianchi, & Fischer, 2002).

5.2 Markov Decision Processes

A Markov Decision Process (*MDP*) is defined as a tuple $(\mathcal{X}, \mathcal{A}, T, r)$: $\mathcal{X}$ is the state space; $\mathcal{A}$ is the action space; $T_a(x; x') = \mathbf{Pr}(x'|a, x)$ is the probability that an action $a \in \mathcal{A}$ taken in state $x \in \mathcal{X}$ will lead to state $x' \in \mathcal{X}$; and $r(x, a) \in \mathcal{R} := \mathbb{R}$ is the immediate reward obtained in state $x \in \mathcal{X}$ and action $a \in \mathcal{A}$. The interaction proceeds in time steps $t = 1, 2, \ldots$ where at time t, action $a_t \in \mathcal{A}$ is issued in state $x_{t-1} \in \mathcal{X}$, leading to a reward $r_t = r(x_{t-1}, a_t)$ and a new state x_t that starts the next time step $t+1$. A stationary closed-loop control policy $\pi : \mathcal{X} \to \mathcal{A}$ assigns an action to each state. For MDPs there always exists an optimal stationary deterministic policy and thus one only needs to consider such policies. In undiscounted MDPs the average reward per time step for a fixed policy π with initial state x is defined as $\rho^\pi(x) = \lim_{t\to\infty} \mathbf{E}^\pi[\frac{1}{t}\sum_{\tau=0}^{t} r_\tau]$. It can be shown (Bertsekas, 1987) that $\rho^\pi(x) = \rho^\pi(x')$ for all $x, x' \in \mathcal{X}$ under the assumption that the Markov chain for policy π is ergodic. Here, we assume that the MDPs are ergodic for all stationary policies.

In order to keep the intervention model particularly simple[3], we follow the Q-notation of Watkins (1989). The optimal policy π^* can then be characterized in terms of the optimal average reward ρ and the optimal relative Q-values $Q(x, a)$ for each state-action pair (x, a) that are solutions to the following system of non-linear equations (Singh, 1994): for any

3. The "brute-force" adaptive agent for this problem would roughly look as follows. First, the agent starts with a prior distribution over all MDPs, e.g. product of Dirichlet distributions over the transition probabilities. Then, in each cycle, the agent samples a full transition matrix from the distribution and solves it using dynamic programming. Once it has computed the optimal policy, it uses it to issue the next action, and then discards the policy. Subsequently, it updates the distribution over MDPs using the next observed state. However, in the main text we follow a different approach that avoids solving an MDP in every time step.

state $x \in \mathcal{X}$ and action $a \in \mathcal{A}$,

$$\begin{aligned} Q(x,a) + \rho &= r(x,a) + \sum_{x' \in \mathcal{X}} \mathbf{Pr}(x'|x,a) \Big[\max_{a'} Q(x',a') \Big] \\ &= r(x,a) + \mathbf{E}_{x'} \Big[\max_{a'} Q(x',a') \Big| x, a \Big]. \end{aligned} \tag{18}$$

The optimal policy can then be defined as $\pi^*(x) := \arg\max_a Q(x,a)$ for any state $x \in \mathcal{X}$.

Again this setup allows for a straightforward solution with the Bayesian control rule, because each learnable MDP (characterized by the Q-values and the average reward) has a known solution π^*. Accordingly, an operation mode m is given by $m = [Q, \rho] \in \mathcal{M} = \mathbb{R}^{|\mathcal{A}| \times |\mathcal{O}| + 1}$. To obtain a likelihood model for inference over m, we realize that Equation 18 can be rewritten such that it predicts the instantaneous reward $r(x,a)$ as the sum of a mean instantaneous reward ξ_m plus a noise term ν given the Q-values and the average reward ρ for the MDP labeled by m

$$r(x,a) = \underbrace{Q(x,a) + \rho - \max_{a'} Q(x',a')}_{\text{mean instantaneous reward } \xi_m(x,a,x')} \; \underbrace{+ \max_{a'} Q(x',a') - \mathbf{E}[\max_{a'} Q(x',a') | x, a]}_{\text{noise } \nu}$$

Assuming that ν can be reasonably approximated by a normal distribution $\mathrm{N}(0, 1/p)$ with precision p, we can write down a likelihood model for the immediate reward r using the Q-values and the average reward, i.e.

$$P(r|m,x,a,x') = \sqrt{\frac{p}{2\pi}} \exp\Big\{ -\frac{p}{2} (r - \xi_m(x,a,x'))^2 \Big\}. \tag{19}$$

In order to determine the intervention model for each operation mode, we can simply exploit the above properties of the Q-values, which gives

$$P(a|m,x) = \begin{cases} 1 & \text{if } a = \arg\max_{a'} Q(x,a') \\ 0 & \text{else.} \end{cases} \tag{20}$$

To apply the Bayesian control rule, the posterior distribution $P(m|\hat{a}_{\leq t}, x_{\leq t})$ needs to be computed. Fortunately, due to the simplicity of the likelihood model, one can easily devise a conjugate prior distribution and apply standard inference methods (see Appendix A.5). Actions are again determined by sampling operation modes from this posterior and executing the action suggested by the corresponding intervention models. The resulting algorithm is very similar to Bayesian Q-learning (Dearden, Friedman, & Russell, 1998; Dearden, Friedman, & Andre, 1999), but differs in the way actions are selected. The pseudo-code is listed in Algorithm 2.

Simulation: We have tested our MDP-agent in a grid-world example. To give an intuition of the achieved performance, the results are contrasted with those achieved by R-learning. We have used the R-learning variant presented in the work of Singh (1994, Algorithm 3) together with the uncertainty exploration strategy (Mahadevan, 1996). The corresponding update equations are

$$\begin{aligned} Q(x,a) &\leftarrow (1-\alpha) Q(x,a) + \alpha\big(r - \rho + \max_{a'} Q(x',a')\big) \\ \rho &\leftarrow (1-\beta)\rho + \beta\big(r + \max_{a'} Q(x',a') - Q(x,a)\big), \end{aligned} \tag{21}$$

Algorithm 2 BCR-MDP Gibbs sampler.

Initialize entries of λ and μ to zero.
Set initial state to $x \leftarrow x_0$.
for $t = 1, 2, 3, \ldots$ **do**
  {Gibbs sweep}
  Sample ρ using (30).
  for all $Q(y, b)$ of visited states **do**
    Sample $Q(y, b)$ using (31).
  end for

  { Interaction }
  Set $a \leftarrow \arg\max_{a'} Q(x, a')$ and issue a.
  Obtain $o = (r, x')$ from environment.

  {Update hyperparameters}
  $\mu(x, a, x') \leftarrow \frac{\lambda(x,a,x')\mu(x,a,x') + p\,r}{\lambda(x,a,x') + p}$
  $\lambda(x, a, x') \leftarrow \lambda(x, a, x') + p$

  Set $x \leftarrow x'$.
end for

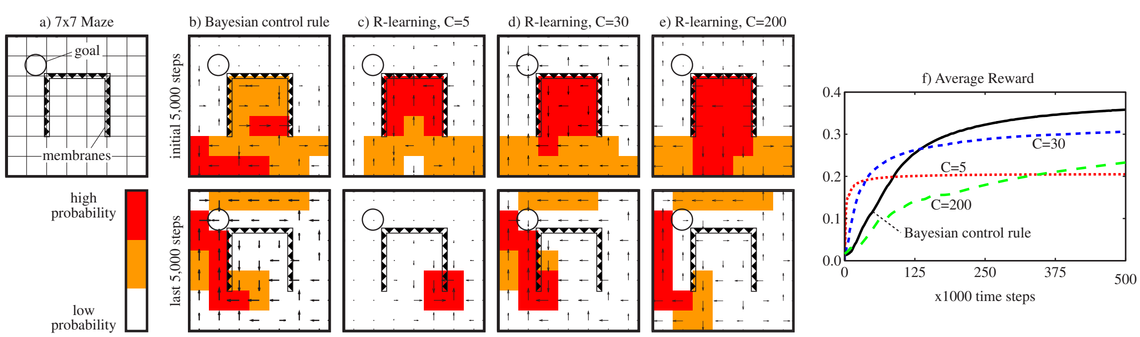

Figure 11: Results for the 7×7 grid-world domain. Panel (a) illustrates the setup. Columns (b)-(e) illustrate the behavioral statistics of the algorithms. The upper and lower row have been calculated over the first and last 5,000 time steps of randomly chosen runs. The probability of being in a state is color-encoded, and the arrows represent the most frequent actions taken by the agents. Panel (f) presents the curves obtained by averaging ten runs.

	Average Reward
BCR	0.3582 ± 0.0038
R-learning, $C = 200$	0.2314 ± 0.0024
R-learning, $C = 30$	0.3056 ± 0.0063
R-learning, $C = 5$	0.2049 ± 0.0012

Table 2: Average reward attained by the different algorithms at the end of the run. The mean and the standard deviation has been calculated based on 10 runs.

where $\alpha, \beta > 0$ are learning rates. The exploration strategy chooses with fixed probability $p_{\mathrm{exp}} > 0$ the action a that maximizes $Q(x, a) + \frac{C}{F(x,a)}$, where C is a constant, and $F(x, a)$ represents the number of times that action a has been tried in state x. Thus, higher values of C enforce increased exploration.

In a study (Mahadevan, 1996), a grid-world is described that is especially useful as a test bed for the analysis of RL algorithms. For our purposes, it is of particular interest because it is easy to design experiments containing *suboptimal limit-cycles*. Figure 11, panel (a), illustrates the 7×7 grid-world. A controller has to learn a policy that leads it from any initial location to the goal state. At each step, the agent can move to any adjacent space (up, down, left or right). If the agent reaches the goal state then its next position is randomly set to any square of the grid (with uniform probability) to start another trial. There are also "one-way membranes" that allow the agent to move into one direction but not into the other. In these experiments, these membranes form "inverted cups" that the agent can enter from any side but can only leave through the bottom, playing the role of a local maximum. Transitions are stochastic: the agent moves to the correct square with probability $p = \frac{9}{10}$ and to any of the free adjacent spaces (uniform distribution) with probability $1 - p = \frac{1}{10}$. Rewards are assigned as follows. The default reward is $r = 0$. If the agent traverses a membrane it obtains a reward of $r = 1$. Reaching the goal state assigns $r = 2.5$. The parameters chosen for this simulation were the following. For our MDP-agent, we have chosen hyperparameters $\mu_0 = 1$ and $\lambda_0 = 1$ and precision $p = 1$. For R-learning, we have chosen learning rates $\alpha = 0.5$ and $\beta = 0.001$, and the exploration constant has been set to $C = 5$, $C = 30$ and to $C = 200$. A total of 10 runs were carried out for each algorithm. The results are presented in Figure 11 and Table 2. R-learning only learns the optimal policy given sufficient exploration (panels d & e, bottom row), whereas the Bayesian control rule learns the policy successfully. In Figure 11f, the learning curve of R-learning for $C = 5$ and $C = 30$ is initially steeper than the Bayesian controller. However, the latter attains a higher average reward around time step 125,000 onwards. We attribute this shallow initial transient to the phase where the distribution over the operation modes is flat, which is also reflected by the initially random exploratory behavior.

6. Discussion

The key idea of this work is to extend the minimum relative entropy principle, i.e. the variational principle underlying Bayesian estimation, to the problem of adaptive control.

From a coding point of view, this work extends the idea of maximal compression of the observation stream to the whole experience of the agent containing both the agent's actions and observations. This not only minimizes the amount of bits to write when *saving/encoding* the I/O stream, but it also minimizes the amount of bits required to *produce/decode* an action (MacKay, 2003, Ch. 6).

This extension is non-trivial, because there is an important caveat for coding I/O sequences: unlike observations, actions do not carry any information that could be used for inference in adaptive coding because actions are issued by the decoder itself. The problem is that doing inference on ones own actions is logically inconsistent and leads to paradoxes (Nozick, 1969). This seemingly innocuous issue has turned out to be very intricate and has been investigated intensely in the recent past by researchers focusing on the issue of causality (Pearl, 2000; Spirtes et al., 2000; Dawid, 2010). Our work contributes to this body of research by providing further evidence that actions cannot be treated using probability calculus alone.

If the causal dependencies are carefully taken into account, then minimizing the relative entropy leads to a rule for adaptive control which we called the Bayesian control rule. This rule allows combining a class of task-specific agents into an agent that is universal with respect to this class. The resulting control law is a simple stochastic control rule that is completely general and parameter-free. As the analysis in this paper shows, this control rule converges to the true control law under mild assumptions.

6.1 Critical Issues

- *Causality.* Virtually every adaptive control method in the literature successfully treats actions as conditionals over observation streams and never worries about causality. Thus, why bother about interventions? In a decision-theoretic setup, the decision maker chooses a policy $\pi^* \in \Pi$ maximizing the expected utility U over the outcomes $\omega \in \Omega$, i.e. $\pi^* := \arg\max_\pi \mathbf{E}[U|\pi] = \sum_\omega \mathbf{Pr}(\omega|\pi)U(\omega)$. "Choosing π^*" is formally equivalent to choosing the Kronecker delta function $\delta^\pi_{\pi^*}$ as the probability distribution over policies. In this case, the conditional probabilities $\mathbf{Pr}(\omega|\pi)$ and $\mathbf{Pr}(\omega|\hat{\pi})$ coincide, since
$$\mathbf{Pr}(\omega, \pi) = \mathbf{Pr}(\pi)\mathbf{Pr}(\omega|\pi) = \delta^\pi_{\pi^*}\mathbf{Pr}(\omega|\pi) = \mathbf{Pr}(\omega, \hat{\pi}).$$
In this sense, the choice of the policy causally precedes the interactions. As we have discussed in Section 3 however, when there is uncertainty about the policy (i.e. $\mathbf{Pr}(\pi) \neq \delta^\pi_{\pi^*}$), then causal belief updates are crucial. Essentially, this problem arises because the uncertainty over the policy is resolved during the interactions. Hence, treating actions as interventions seamlessly extends them to the status of random variables.

- *Where do prior probabilities/likelihood models/policies come from?* The predictor in the Bayesian control rule is essentially a Bayesian predictor and thereby entails (almost) the same modeling paradigm. The designer has to define a class of hypotheses over the environments, construct appropriate likelihood models, and choose a suitable prior probability distribution to capture the model's uncertainty. Similarly, under sufficient domain knowledge, an analogous procedure can be applied to construct suitable operation modes. However, there are many situations where this is a difficult or even intractable problem in itself. For example, one can design a class of operation modes by pre-computing the optimal policies for a given class of environments. Formally, let Θ be a class of hypotheses modeling environments and let Π be class of policies. Given a utility criterion U, define the set of operation modes $\mathcal{M} := \{m_\theta\}_{\theta\in\Theta}$ by constructing each operation mode as $m_\theta := (\theta, \pi^*)$, $\pi^* \in \pi$, where $\pi^* := \arg\max_\pi \mathbf{E}[U|\theta, \pi]$. However, computing the optimal policy π^* is in many cases intractable. In some cases, this can be remedied by characterizing the operation modes through optimality equations which are solved by probabilistic inference as in the example of the MDP agent in Section 5.2. Recently, we have applied a similar approach to adaptive control problems with linear quadratic regulators (Braun & Ortega, 2010).

- *Problems of Bayesian methods.* The Bayesian control rule treats an adaptive control problem as a Bayesian inference problem. Hence, all the problems typically associated with Bayesian methods carry over to agents constructed with the Bayesian control rule. These problems are of both analytical and computational nature. For example, there are many probabilistic models where the posterior distribution does not have a closed-form solution. Also, exact probabilistic inference is in general computationally very intensive. Even though there is a large literature in efficient/approximate inference algorithms for particular problem classes (Bishop, 2006), not many of them are suitable for on-line probabilistic inference in more realistic environment classes.

- *Bayesian control rule versus Bayes-optimal control.* Directly maximizing the (subjective) expected utility for a given environment class is not the same as minimizing the expected relative entropy for a given class of operation modes. *The two methods are based on different assumptions and optimality principles.* As such, the Bayesian control rule is not a Bayes-optimal controller. Indeed, it is easy to design experiments where the Bayesian control rule converges exponentially slower (or does not converge at all) than a Bayes-optimal controller to the maximum utility. Consider the following simple example: Environment 1 is a k-state MDP in which only k consecutive actions A reach a state with reward $+1$. Any interception with a B-action leads back to the initial state. Consider a second environment which is like the first but actions A and B are interchanged. A Bayes-optimal controller figures out the true environment in k actions (either k consecutive A's or B's). Consider now the Bayesian control rule: The optimal action in Environment 1 is A, in Environment 2 is B. A uniform $(\frac{1}{2}, \frac{1}{2})$ prior over the operation modes stays a uniform posterior as long as no reward has been observed. Hence the Bayesian control rule chooses at each time-step A and B with equal probability. With this policy it takes about 2^k actions to accidentally choose a row of A's (or B's) of length k. From then on the Bayesian control rule is optimal too. So a Bayes-optimal controller converges in time k, while the Bayesian control rule needs exponentially longer. One way to remedy this problem might be to allow the Bayesian control rule to sample actions from the same operation mode for several time steps in a row rather than randomizing controllers in every cycle. However, if one considers non-stationary environments this strategy can also break down. Consider, for example, an increasing MDP with $k = \lceil 10\sqrt{t}\,\rceil$, in which a Bayes-optimal controller converges in 100 steps, while the Bayesian control rule does not converge at all in most realizations, because the boundedness assumption is violated.

6.2 Relation to Existing Approaches

Some of the ideas underlying this work are not unique to the Bayesian control rule. The following is a selection of previously published work in the recent Bayesian reinforcement learning literature where related ideas can be found.

- *Compression principles.* In the literature, there is an important amount of work relating compression to intelligence (MacKay, 2003; Hutter, 2004b). In particular, it has been even proposed that compression ratio is an objective quantitative measure of intelligence (Mahoney, 1999). Compression has also been used as a basis for a theory of curiosity, creativity and beauty (Schmidhuber, 2009).

- *Mixture of experts.* Passive sequence prediction by mixing experts has been studied extensively in the literature (Cesa-Bianchi & Lugosi, 2006). In a study on online-predictors (Hutter, 2004a), Bayes-optimal predictors are mixed. Bayes-mixtures can also be used for universal prediction (Hutter, 2003). For the control case, the idea of using mixtures of expert-controllers has been previously evoked in models like the MOSAIC-architecture (Haruno, Wolpert, & Kawato, 2001). Universal learning with Bayes mixtures of experts in reactive environments has been studied in the work of Poland and Hutter (2005) and Hutter (2002).

- *Stochastic action selection.* The idea of using actions as random variables, and the problems that this entails, has been expressed in the work of Hutter (2004b, Problem 5.1). The study in Section 3 can be regarded as a thorough investigation of this open problem. Other stochastic action selection approaches are found in the thesis of Wyatt (1997) who examines exploration strategies for (PO)MDPs, in learning automata (Narendra & Thathachar, 1974) and in probability matching (Duda, Hart, & Stork, 2001) amongst others. In particular, the thesis discusses theoretical properties of an extension to *probability matching* in the context of multi-armed bandit problems. There, it is proposed to choose a lever according to how likely it is to be optimal and it is shown that this strategy converges, thus providing a simple method for guiding exploration.

- *Relative entropy criterion.* The usage of a minimum relative entropy criterion to derive control laws underlies the KL-control methods developed in the work of Todorov (2006, 2009) and Kappen et al. (2010). There, it has been shown that a large class of optimal control problems can be solved very efficiently if the problem statement is reformulated as the minimization of the deviation of the dynamics of a controlled system from the uncontrolled system. A related idea is to conceptualize planning as an inference problem (Toussaint, Harmeling, & Storkey, 2006). This approach is based on an equivalence between maximization of the expected future return and likelihood maximization which is both applicable to MDPs and POMDPs. Algorithms based on this duality have become an active field of current research. See for example the work of Rasmussen and Deisenroth (2008), where very fast model-based RL techniques are used for control in continuous state and action spaces.

7. Conclusions

This work introduces the Bayesian control rule, a Bayesian rule for adaptive control. The key feature of this rule is the special treatment of actions based on causal calculus and the decomposition of an adaptive agent into a mixture of operation modes, i.e. environment-specific agents. The rule is derived by minimizing the expected relative entropy from the true operation mode and by carefully distinguishing between actions and observations. Furthermore, the Bayesian control rule turns out to be exactly the predictive distribution over the next action given the past interactions that one would obtain by using only probability and causal calculus. Furthermore, it is shown that agents constructed with the Bayesian control rule converge to the true operation mode under mild assumptions: boundedness, which is related to ergodicity; and consistency, demanding that two indistinguishable hypotheses share the same policy.

We have presented the Bayesian control rule as a way to solve adaptive control problems based on a minimum relative entropy principle. Thus, the Bayesian control rule can either be regarded as a new principled approach to adaptive control under a novel optimality criterion or as a heuristic approximation to traditional Bayes-optimal control. Since it takes on a similar form to Bayes' rule, the adaptive control problem could then be translated into an on-line inference problem where actions are sampled stochastically from a posterior distribution. It is important to note, however, that the problem statement as formulated here and the usual Bayes-optimal approach in adaptive control are *not* the same. In the future the relationship between these two problem statements deserves further investigation.

Acknowledgments

We thank Marcus Hutter, David Wingate, Zoubin Ghahramani, José Aliste, José Donoso, Humberto Maturana and the anonymous reviewers for comments on earlier versions of this manuscript and/or inspiring discussions. We thank the Ministerio de Planificación de Chile (MIDEPLAN) and the Böhringer-Ingelheim-Fonds (BIF) for funding.

Appendix A. Proofs

A.1 Proof of Theorem 1

Proof. The proof follows the same line of argument as the solution to Equation 3 with the crucial difference that actions are treated as interventions. Consider without loss of generality the summand $\sum_m P(m)C_m^{a_t}$ in Equation 9. Note that the relative entropy can be written as a difference of two logarithms, where only one term depends on $\mathbf{Pr}$ to be varied. Therefore, one can pull out the other term and write it as a constant c. This yields

$$c - \sum_m P(m) \sum_{\underline{ao}_{<t}} P(\hat{\underline{ao}}_{<t}|m) \sum_{a_t} P(a_t|m, \hat{\underline{ao}}_{<t}) \ln \mathbf{Pr}(a_t|\underline{ao}_{<t}).$$

Substituting $P(\underline{\hat{a}o}_{<t}|m)$ by $P(m|\underline{\hat{a}o}_{<t})P(\underline{\hat{a}o}_{<t})/P(m)$ using Bayes' rule and further rearrangement of the terms leads to

$$= c - \sum_{m} \sum_{\underline{ao}_{<t}} P(m|\underline{\hat{a}o}_{<t}) P(\underline{\hat{a}o_{<t}}) \sum_{a_t} P(a_t|m, \underline{\hat{a}o}_{<t}) \ln \mathbf{Pr}(a_t|\underline{ao}_{<t})$$
$$= c - \sum_{\underline{ao}_{<t}} P(\underline{\hat{a}o}_{<t}) \sum_{a_t} P(a_t|\underline{\hat{a}o}_{<t}) \ln \mathbf{Pr}(a_t|\underline{ao}_{<t}).$$

The inner sum has the form $-\sum_x p(x) \ln q(x)$, i.e. the cross-entropy between $q(x)$ and $p(x)$, which is minimized when $q(x) = p(x)$ for all x. Let $\mathbf{P}$ denote the optimum distribution for $\mathbf{Pr}$. By choosing this optimum one obtains $\mathbf{P}(a_t|\underline{ao}_{<t}) = P(a_t|\underline{\hat{a}o}_{<t})$ for all a_t. Note that the solution to this variational problem is independent of the weighting $P(\underline{\hat{a}o}_{<t})$. Since the same argument applies to any summand $\sum_m P(m)C_m^{a_\tau}$ and $\sum_m P(m)C_m^{o_\tau}$ in Equation 9, their variational problems are mutually independent. Hence,

$$\mathbf{P}(a_t|\underline{ao}_{<t}) = P(a_t|\underline{\hat{a}o}_{<t}) \qquad \mathbf{P}(o_t|\underline{ao}_{<t}) = P(o_t|\underline{\hat{a}o}_{<t}\hat{a}_t)$$

for all $\underline{ao}_{\leq t} \in \mathcal{Z}^*$. For $P(a_t|\underline{\hat{a}o}_{<t})$, introduce the variable m via a marginalization and then apply the chain rule:

$$P(a_t|\underline{\hat{a}o}_{<t}) = \sum_m P(a_{t+1}|m, \underline{\hat{a}o}_{<t}) P(m|\underline{\hat{a}o}_{<t}).$$

The term $P(m|\underline{\hat{a}o}_{\leq t})$ can be further developed as

$$\begin{aligned} P(m|\underline{\hat{a}o}_{<t}) &= \frac{P(\underline{\hat{a}o}_{<t}|m)P(m)}{\sum_{m'} P(\underline{\hat{a}o}_{<t}|m')P(m')} \\ &= \frac{P(m)\prod_{\tau=1}^{t-1} P(\hat{a}_\tau|m, \underline{\hat{a}o}_{<\tau})P(o_\tau|m, \underline{\hat{a}o}_{<\tau}\hat{a}_\tau)}{\sum_{m'} P(m')\prod_{\tau=1}^{t-1} P(\hat{a}_\tau|m', \underline{\hat{a}o}_{<\tau})P(o_\tau|m', \underline{\hat{a}o}_{<\tau}\hat{a}_\tau)} \\ &= \frac{P(m)\prod_{\tau=1}^{t-1} P(o_\tau|m, \underline{ao}_{<\tau}a_\tau)}{\sum_{m'} P(m')\prod_{\tau=1}^{t-1} P(o_\tau|m', \underline{ao}_{<\tau}a_\tau)}. \end{aligned}$$

The first equality is obtained by applying Bayes' rule and the second by using the chain rule for probabilities. To get the last equality, one applies the interventions to the causal factorization. Thus, $P(\hat{a}_\tau|m, \underline{\hat{a}o}_{<\tau}) = 1$ and $P(o_\tau|m, \underline{\hat{a}o}_{<\tau}\hat{a}_\tau) = P(o_\tau|m, \underline{ao}_{<\tau}a_\tau)$. The equations characterizing $P(o_t|\underline{\hat{a}o}_{<t}\hat{a}_t)$ are obtained similarly. □

A.2 Proof of Theorem 2

Proof. As has been pointed out in (14), a particular realization of the divergence process $d_t(m^*\|m)$ can be decomposed as

$$d_t(m^*\|m) = \sum_{m'} g_m(m'; \mathcal{T}_{m'}),$$

where the $g_m(m'; \mathcal{T}_{m'})$ are sub-divergences of $d_t(m^*\|m)$ and the $\mathcal{T}_{m'}$ form a partition of $\mathcal{N}_t$. However, since $d_t(m^*\|m)$ has bounded variation for all $m \in \mathcal{M}$, one has for all $\delta' > 0$, there is a $C(m) \geq 0$, such that for all $m' \in \mathcal{M}$, all $t \in \mathcal{N}_t$ and all $\mathcal{T} \subset \mathcal{N}_t$, the inequality

$$\Big| g_m(m'; \mathcal{T}_{m'}) - G_m(m'; \mathcal{T}_{m'}) \Big| \leq C(m)$$

holds with probability $\geq 1-\delta'$. However, due to (15),

$$G_m(m'; \mathcal{T}_{m'}) \geq 0$$

for all $m' \in \mathcal{M}$. Thus,

$$g_m(m'; \mathcal{T}_{m'}) \geq -C(m).$$

If all the previous inequalities hold simultaneously then the divergence process can be bounded as well. That is, the inequality

$$d_t(m^* \| m) \geq -MC(m) \tag{22}$$

holds with probability $\geq (1-\delta')^M$ where $M := |\mathcal{M}|$. Choose

$$\beta(m) := \max\{0, \ln \tfrac{P(m)}{P(m^*)}\}.$$

Since $0 \geq \ln \frac{P(m)}{P(m^*)} - \beta(m)$, it can be added to the right hand side of (22). Using the definition of $d_t(m^* \| m)$, taking the exponential and rearranging the terms one obtains

$$P(m^*) \prod_{\tau=1}^{t} P(o_\tau | m^*, \underline{ao}_{<\tau} a_\tau) \geq e^{-\alpha(m)} P(m) \prod_{\tau=1}^{t} P(o_\tau | m, \underline{ao}_{<\tau} a_\tau)$$

where $\alpha(m) := MC(m) + \beta(m) \geq 0$. Identifying the posterior probabilities of m^* and m by dividing both sides by the normalizing constant yields the inequality

$$P(m^* | \underline{\hat{a}o}_{\leq t}) \geq e^{-\alpha(m)} P(m | \underline{\hat{a}o}_{\leq t}).$$

This inequality holds simultaneously for all $m \in \mathcal{M}$ with probability $\geq (1-\delta')^{M^2}$ and in particular for $\lambda := \min_m\{e^{-\alpha(m)}\}$, that is,

$$P(m^* | \underline{\hat{a}o}_{\leq t}) \geq \lambda P(m | \underline{\hat{a}o}_{\leq t}).$$

But since this is valid for any $m \in \mathcal{M}$, and because $\max_m\{P(m|\underline{\hat{a}o}_{\leq t})\} \geq \frac{1}{M}$, one gets

$$P(m^* | \underline{\hat{a}o}_{\leq t}) \geq \frac{\lambda}{M},$$

with probability $\geq 1-\delta$ for arbitrary $\delta > 0$ related to δ' through the equation $\delta' := 1 - \sqrt[M^2]{1-\delta}$. □

A.3 Proof of Theorem 3

Proof. The divergence process $d_t(m^* \| m)$ can be decomposed into a sum of sub-divergences (see Equation 14)

$$d_t(m^* \| m) = \sum_{m'} g_{m'}(m; \mathcal{T}_{m'}). \tag{23}$$

Furthermore, for every $m' \in \mathcal{M}$, one has that for all $\delta > 0$, there is a $C \geq 0$, such that for all $t \in \mathbb{N}$ and for all $\mathcal{T} \subset \mathcal{N}_t$

$$\Big| g_{m'}(m; \mathcal{T}) - G_{m'}(m; \mathcal{T}) \Big| \leq C(m)$$

with probability $\geq 1 - \delta'$. Applying this bound to the summands in (23) yields the lower bound

$$\sum_{m'} g_{m'}(m; \mathcal{T}_{m'}) \geq \sum_{m'} \bigl(G_{m'}(m; \mathcal{T}_{m'}) - C(m)\bigr)$$

which holds with probability $\geq (1-\delta')^M$, where $M := |\mathcal{M}|$. Due to Inequality 15, one has that for all $m' \neq m^*$, $G_{m'}(m; \mathcal{T}_{m'}) \geq 0$. Hence,

$$\sum_{m'} \bigl(G_{m'}(m; \mathcal{T}_{m'}) - C(m)\bigr) \geq G_{m^*}(m; \mathcal{T}_{m^*}) - MC$$

where $C := \max_m \{C(m)\}$. The members of the set $\mathcal{T}_{m^*}$ are determined stochastically; more specifically, the i^{th} member is included into $\mathcal{T}_{m^*}$ with probability $P(m^*|\underline{\hat{a}o}_{\leq i}) \geq \lambda/M$ for some $\lambda > 0$ by Theorem 2. But since $m \notin [m^*]$, one has that $G_{m^*}(m; \mathcal{T}_{m^*}) \to \infty$ as $t \to \infty$ with probability $\geq 1 - \delta'$ for arbitrarily chosen $\delta' > 0$. This implies that

$$\lim_{t\to\infty} d_t(m^* \| m) \geq \lim_{t\to\infty} G_{m^*}(m; \mathcal{T}_{m^*}) - MC \nearrow \infty$$

with probability $\geq 1 - \delta$, where $\delta > 0$ is arbitrary and related to δ' as $\delta = 1 - (1-\delta')^{M+1}$. Using this result in the upper bound for posterior probabilities yields the final result

$$0 \leq \lim_{t\to\infty} P(m|\underline{\hat{a}o}_{\leq t}) \leq \lim_{t\to\infty} \frac{P(m)}{P(m^*)} e^{-d_t(m^*\|m)} = 0.$$

$\square$

A.4 Proof of Theorem 4

Proof. We will use the abbreviations $p_m(t) := P(a_t|m, \underline{\hat{a}o}_{<t})$ and $w_m(t) := P(m|\underline{\hat{a}o}_{<t})$. Decompose $P(a_t|\underline{\hat{a}o}_{<t})$ as

$$P(a_t|\underline{\hat{a}o}_{<t}) = \sum_{m \notin [m^*]} p_m(t) w_m(t) + \sum_{m \in [m^*]} p_m(t) w_m(t). \tag{24}$$

The first sum on the right-hand side is lower-bounded by zero and upper-bounded by

$$\sum_{m \notin [m^*]} p_m(t) w_m(t) \leq \sum_{m \notin [m^*]} w_m(t)$$

because $p_m(t) \leq 1$. Due to Theorem 3, $w_m(t) \to 0$ as $t \to \infty$ almost surely. Given $\varepsilon' > 0$ and $\delta' > 0$, let $t_0(m)$ be the time such that for all $t \geq t_0(m)$, $w_m(t) < \varepsilon'$. Choosing $t_0 := \max_m \{t_0(m)\}$, the previous inequality holds for all m and $t \geq t_0$ simultaneously with probability $\geq (1-\delta')^M$. Hence,

$$\sum_{m \notin [m^*]} p_m(t) w_m(t) \leq \sum_{m \notin [m^*]} w_m(t) < M\varepsilon'. \tag{25}$$

To bound the second sum in (24) one proceeds as follows. For every member $m \in [m^*]$, one has that $p_m(t) \to p_{m^*}(t)$ as $t \to \infty$. Hence, following a similar construction as above, one can choose t'_0 such that for all $t \geq t'_0$ and $m \in [m^*]$, the inequalities

$$\Bigl| p_m(t) - p_{m^*}(t) \Bigr| < \varepsilon'$$

hold simultaneously for the precision $\varepsilon' > 0$. Applying this to the second sum in Equation 24 yields the bounds

$$\sum_{m \in [m^*]} \big(p_{m^*}(t) - \varepsilon'\big) w_m(t) \leq \sum_{m \in [m^*]} p_m(t) w_m(t) \leq \sum_{m \in [m^*]} \big(p_{m^*}(t) + \varepsilon'\big) w_m(t).$$

Here $\big(p_{m^*}(t) \pm \varepsilon'\big)$ are multiplicative constants that can be placed in front of the sum. Note that

$$1 \geq \sum_{m \in [m^*]} w_m(t) = 1 - \sum_{m \notin [m^*]} w_m(t) > 1 - \varepsilon.$$

Use of the above inequalities allows simplifying the lower and upper bounds respectively:

$$\begin{aligned} \big(p_{m^*}(t) - \varepsilon'\big) \sum_{m \in [m^*]} w_m(t) &> p_{m^*}(t)(1 - \varepsilon') - \varepsilon' \geq p_{m^*}(t) - 2\varepsilon', \\ \big(p_{m^*}(t) + \varepsilon'\big) \sum_{m \in [m^*]} w_m(t) &\leq p_{m^*}(t) + \varepsilon' < p_{m^*}(t) + 2\varepsilon'. \end{aligned} \tag{26}$$

Combining the inequalities (25) and (26) in (24) yields the final result:

$$\Big| P(a_t | \underline{\hat{a}o}_{<t}) - p_{m^*}(t) \Big| < (2 + M)\varepsilon' = \varepsilon,$$

which holds with probability $\geq 1 - \delta$ for arbitrary $\delta > 0$ related to δ' as $\delta' = 1 - \sqrt[M]{1 - \delta}$ and arbitrary precision ε. □

A.5 Gibbs Sampling Implementation for MDP Agent

Inserting the likelihood given in Equation 19 into Equation 13 of the Bayesian control rule, one obtains the following expression for the posterior

$$\begin{aligned} P(m|\hat{a}_{\leq t}, o_{\leq t}) &= \frac{P(x'|m, x, a) P(r|m, x, a, x') P(m|\hat{a}_{<t}, o_{<t})}{\int_{\mathcal{M}} P(x'|m', x, a) P(r|m', x, a, x') P(m'|\hat{a}_{<t}, o_{<t})\, dm'} \\ &= \frac{P(r|m, x, a, x') P(m|\hat{a}_{<t}, o_{<t})}{\int_{\mathcal{M}} P(r|m', x, a, x') P(m'|\hat{a}_{<t}, o_{<t})\, dm'}, \end{aligned} \tag{27}$$

where we have replaced the sum by an integration over m', the finite-dimensional real space containing only the average reward and the Q-values of the observed states, and where we have simplified the term $P(x'|m, x, a)$ because it is constant for all $m' \in \mathcal{M}$.

The likelihood model $P(r|m', x, a, x')$ in Equation 27 encodes a set of independent normal distributions over the immediate reward with means $\xi_m(x, a, x')$ indexed by triples $(x, a, x') \in \mathcal{X} \times \mathcal{A} \times \mathcal{X}$. In other words, given (x, a, x'), the rewards are drawn from a normal distribution with unknown mean $\xi_m(x, a, x')$ and known variance σ^2. The sufficient statistics are given by $n(x, a, x')$, the number of times that the transition $x \to x'$ under action a, and $\bar{r}(x, a, x')$, the mean of the rewards obtained in the same transition. The conjugate prior distribution is well known and given by a normal distribution with hyperparameters μ_0 and λ_0:

$$P(\xi_m(x, a, x')) = \mathrm{N}(\mu_0, 1/\lambda_0) = \sqrt{\frac{\lambda_0}{2\pi}} \exp\Big\{ -\tfrac{\lambda_0}{2} \big(\xi_m(x, a, x') - \mu_0\big)^2 \Big\}. \tag{28}$$

The posterior distribution is given by

$$P(\xi_m(x,a,x')|\hat{a}_{\le t}, o_{\le t}) = \mathrm{N}(\mu(x,a,x'), 1/\lambda(x,a,x'))$$

where the posterior hyperparameters are computed as

$$\begin{aligned} \mu(x,a,x') &= \frac{\lambda_0\,\mu_0 + p\,n(x,a,x')\,\bar{r}(x,a,x')}{\lambda_0 + p\,n(x,a,x')} \\ \lambda(x,a,x') &= \lambda_0 + p\,n(x,a,x'). \end{aligned} \tag{29}$$

By introducing the shorthand $V(x) := \max_a Q(x,a)$, we can write the posterior distribution over ρ as

$$P(\rho|\hat{a}_{\le t}, o_{\le t}) = \mathrm{N}(\bar{\rho}, 1/S) \tag{30}$$

where

$$\bar{\rho} = \frac{1}{S} \sum_{x,a,x'} \lambda(x,a,x')(\mu(x,a,x') - Q(x,a) + V(x')),$$
$$S = \sum_{x,a,x'} \lambda(x,a,x').$$

The posterior distribution over the Q-values is more difficult to obtain, because each $Q(x,a)$ enters the posterior distribution both linearly and non-linearly through μ. However, if we fix $Q(x,a)$ within the max operations, which amounts to treating each $V(x)$ as a constant within a single Gibbs step, then the conditional distribution can be approximated by

$$P(Q(x,a)|\hat{a}_{\le t}, o_{\le t}) \approx \mathrm{N}\Big(\bar{Q}(x,a), 1/S(x,a)\Big) \tag{31}$$

where

$$\bar{Q}(x,a) = \frac{1}{S(x,a)} \sum_{x'} \lambda(x,a,x')(\mu(x,a,x') - \rho + V(x')),$$
$$S(x,a) = \sum_{x'} \lambda(x,a,x').$$

We expect this approximation to hold because the resulting update rule constitutes a contraction operation that forms the basis of most stochastic approximation algorithms (Mahadevan, 1996). As a result, the Gibbs sampler draws all the values from normal distributions. In each cycle of the adaptive controller, one can carry out several Gibbs sweeps to obtain a sample of m to improve the mixing of the Markov chain. However, our experimental results have shown that a *single Gibbs sweep per state transition* performs reasonably well. Once a new parameter vector m is drawn, the Bayesian control rule proceeds by taking the optimal action given by Equation 20. Note that only the μ and λ entries of the transitions that have occurred need to be represented explicitly; similarly, only the Q-values of visited states need to be represented explicitly.

References

Auer, P., CesaBianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, *47*, 235–256.

Bertsekas, D. (1987). *Dynamic Programming: Deterministic and Stochastic Models*. Prentice-Hall, Upper Saddle River, NJ.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Braun, D. A., & Ortega, P. A. (2010). A minimum relative entropy principle for adaptive control in linear quadratic regulators. In *The 7th conference on informatics in control, automation and robotics*, Vol. 3, pp. 103–108.

Cesa-Bianchi, N., & Lugosi, G. (2006). *Prediction, Learning and Games*. Cambridge University Press.

Dawid, A. P. (2010). Beware of the DAG!. *Journal of Machine Learning Research*, *(to appear)*.

Dearden, R., Friedman, N., & Andre, D. (1999). Model based bayesian exploration. In *In Proceedings of Fifteenth Conference on Uncertainty in Artificial Intelligence*, pp. 150–159.

Dearden, R., Friedman, N., & Russell, S. (1998). Bayesian q-learning. In *AAAI '98/IAAI '98: Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence*, pp. 761–768. American Association for Artificial Intelligence.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern Classification* (Second edition). Wiley & Sons, Inc.

Duff, M. O. (2002). *Optimal learning: computational procedures for bayes-adaptive markov decision processes*. Ph.D. thesis. Director-Andrew Barto.

Grünwald, P. (2007). *The Minimum Description Length Principle*. The MIT Press.

Haruno, M., Wolpert, D., & Kawato, M. (2001). Mosaic model for sensorimotor learning and control. *Neural Computation*, *13*, 2201–2220.

Haussler, D., & Opper, M. (1997). Mutual information, metric entropy and cumulative relative entropy risk. *The Annals of Statistics*, *25*, 2451–2492.

Hutter, M. (2002). Self-optimizing and pareto-optimal policies in general environments based on bayes-mixtures. In *COLT*.

Hutter, M. (2003). Optimality of universal Bayesian prediction for general loss and alphabet. *Journal of Machine Learning Research*, *4*, 971–997.

Hutter, M. (2004a). Online prediction – bayes versus experts. Tech. rep.. Presented at the EU PASCAL Workshop on Learning Theoretic and Bayesian Inductive Principles (LTBIP-2004).

Hutter, M. (2004b). *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin.

Kappen, B., Gomez, V., & Opper, M. (2010). Optimal control as a graphical model inference problem. *JMLR (to appear)*.

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

Mahadevan, S. (1996). Average reward reinforcement learning: Foundations, algorithms, and empirical results. *Machine Learning*, *22*(1-3), 159–195.

Mahoney, M. V. (1999). Text compression as a test for artificial intelligence. In *AAAI/IAAI*, pp. 486–502.

Narendra, K., & Thathachar, M. A. L. (1974). Learning automata - a survey. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-4*(4), 323–334.

Nozick, R. (1969). Newcomb's problem and two principles of choice. In Rescher, N. (Ed.), *Essays in Honor of Carl G. Hempel*, pp. 114–146. Reidel.

Opper, M. (1998). A bayesian approach to online learning. *Online Learning in Neural Networks*, 363–378.

Ortega, P. A., & Braun, D. A. (2010). A bayesian rule for adaptive control based on causal interventions. In *The third conference on artificial general intelligence*, pp. 121–126.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, UK.

Poland, J., & Hutter, M. (2005). Defensive universal learning with experts. In *ALT*.

Rasmussen, C. E., & Deisenroth, M. P. (2008). *Recent Advances in Reinforcement Learning*, Vol. 5323 of *Lecture Notes on Computer Science, LNAI*, chap. Probabilistic Inference for Fast Learning in Control, pp. 229–242. Springer-Verlag.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin American Mathematical Socierty*, *58*, 527–535.

Russell, S., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach* (3rd edition). Prentice-Hall.

Schmidhuber, J. (2009). Simple algorithmic theory of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. *Journal of SICE*, *48*(1), 21–32.

Shafer, G. (1996). *The art of causal conjecture*. The MIT Press.

Singh, S. P. (1994). Reinforcement learning algorithms for average-payoff markovian decision processes. In *National Conference on Artificial Intelligence*, pp. 700–705.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, Prediction and Search* (2nd edition). Springer-Verlag, New York.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA.

Todorov, E. (2006). Linearly solvable markov decision problems. In *Advances in Neural Information Processing Systems*, Vol. 19, pp. 1369–1376.

Todorov, E. (2009). Efficient computation of optimal actions. *Proceedings of the National Academy of Sciences U.S.A.*, *106*, 11478–11483.

Toussaint, M., Harmeling, S., & Storkey, A. (2006). Probabilistic inference for solving (po)mdps. Tech. rep. EDI-INF-RR-0934, University of Edinburgh.

Watkins, C. (1989). *Learning from Delayed Rewards*. Ph.D. thesis, University of Cambridge, Cambridge, England.

Wyatt, J. (1997). *Exploration and Inference in Learning from Reinforcement*. Ph.D. thesis, Department of Artificial Intelligence, University of Edinburgh.