A Bayesian Rule for Adaptive Control based on Causal Interventions

Pedro A. Ortega
Department of Engineering
University of Cambridge
Cambridge CB2 1PZ, UK
peortega@dcc.uchile.cl

Daniel A. Braun
Department of Engineering
University of Cambridge
Cambridge CB2 1PZ, UK
dab54@cam.ac.uk

Abstract

Explaining adaptive behavior is a central problem in artificial intelligence research. Here we formalize adaptive agents as mixture distributions over sequences of inputs and outputs (I/O). Each distribution of the mixture constitutes a 'possible world', but the agent does not know which of the possible worlds it is actually facing. The problem is to adapt the I/O stream in a way that is compatible with the true world. A natural measure of adaptation can be obtained by the Kullback-Leibler (KL) divergence between the I/O distribution of the true world and the I/O distribution expected by the agent that is uncertain about possible worlds. In the case of pure input streams, the Bayesian mixture provides a well-known solution for this problem. We show, however, that in the case of I/O streams this solution breaks down, because outputs are issued by the agent itself and require a different probabilistic syntax as provided by intervention calculus. Based on this calculus, we obtain a Bayesian control rule that allows modeling adaptive behavior with mixture distributions over I/O streams. This rule might allow for a novel approach to adaptive control based on a minimum KL-principle.

Keywords: Adaptive behavior, Intervention calculus, Bayesian control, Kullback-Leibler-divergence

Introduction

The ability to adapt to unknown environments is often considered a hallmark of intelligence [Beer, 1990, Hutter, 2004]. Agent and environment can be conceptualized as two systems that exchange symbols in every time step [Hutter, 2004]: the symbol issued by the agent is an action, whereas the symbol issued by the environment is an observation. Thus, both agent and environment can be conceptualized as probability distributions over sequences of actions and observations (I/O streams).

If the environment is perfectly known then the I/O probability distribution of the agent can be tailored to suit this particular environment. However, if the environment is unknown, but known to belong to a set of possible environments, then the agent faces an adaptation problem. Consider, for example, a robot that has been endowed with a set of behavioral primitives and now faces the problem of how to act while being ignorant as to which is the correct primitive. Since we want to model both agent and environment as probability distributions over I/O sequences, a natural way to measure the degree of adaptation would be to measure the 'distance' in probability space between the I/O distribution represented by the agent and the I/O distribution conditioned on the true environment. A suitable measure (in terms of its information-theoretic interpretation) is readily provided by the KL-divergence [MacKay, 2003]. In the case of passive prediction, the adaptation problem has a well-known solution. The distribution that minimizes the KL-divergence is a Bayesian mixture distribution over all possible environments [Haussler and Opper, 1997, Opper, 1998]. The aim of this paper is to extend this result for distributions over both inputs and outputs. The main result of this paper is that this extension is only possible if we consider the special syntax of actions in probability theory as it has been suggested by proponents of causal calculus [Pearl, 2000].

Preliminaries

We restrict the exposition to the case of discrete time with discrete stochastic observations and control signals. Let $\mathcal{O}$ and $\mathcal{A}$ be two finite sets, the first being the *set of observations* and the second being the *set of actions*. We use $a_{\leq t} \equiv a_1 a_2 \ldots a_t$, $\underline{ao}_{\leq t} \equiv a_1 o_1 \ldots a_t o_t$ etc. to simplify the notation of strings. Using $\mathcal{A}$ and $\mathcal{O}$, a set of interaction sequences is constructed. Define the *set of interactions* as $\mathcal{Z} \equiv \mathcal{A} \times \mathcal{O}$. A pair $(a, o) \in \mathcal{Z}$ is called an *interaction*. The set of interaction strings of length $t \geq 0$ is denoted by $\mathcal{Z}^t$. Similarly, the set of (finite) interaction strings is $\mathcal{Z}^* \equiv \bigcup_{t \geq 0} \mathcal{Z}^t$ and the set of (infinite) interaction sequences is $\mathcal{Z}^\infty \equiv \{w : w = a_1 o_1 a_2 o_2 \ldots\}$, where each $(a_t, o_t) \in \mathcal{Z}$. The interaction string of length 0 is denoted by ϵ.

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 $\underline{ao}_{\leq t} \in \mathcal{Z}^*$. However, the semantics of the probability distribution $\mathbf{Pr}$ are only fully defined once it is coupled to another 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

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

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. This model of interaction is very general in that it can accommodate many specific regimes of interaction. Note that an agent $\mathbf{P}$ can perfectly predict its environment $\mathbf{Q}$ iff for all $\underline{ao}_{\leq t} \in \mathcal{Z}^*$,

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

In this case we say that $\mathbf{P}$ is *tailored* to $\mathbf{Q}$.

Adaptive Systems: Naïve Construction

Throughout this paper, we use the convention that $\mathbf{P}$ is an *agent* to be constructed by a designer, which is then going to be interfaced with a preexisting but unknown *environment* $\mathbf{Q}$. The designer assumes that $\mathbf{Q}$ is going to be drawn with probability $P(m)$ from a set $\mathcal{Q} \equiv \{\mathbf{Q}_m\}_{m\in\mathcal{M}}$ of possible systems before the interaction starts, where $\mathcal{M}$ is a countable set.

Consider the case when the designer knows beforehand which environment $\mathbf{Q} \in \mathcal{Q}$ is going to be drawn. Then, not only can $\mathbf{P}$ be tailored to $\mathbf{Q}$, but also a custom-made policy for $\mathbf{Q}$ can be designed. That is, the output stream $\mathbf{P}(a_t|\underline{ao}_{<t})$ is such that the true probability $\mathbf{G}$ of the resulting interaction system $(\mathbf{P}, \mathbf{Q})$ gives rise to interaction sequences that the designer considers *desirable.*

Consider now the case when the designer does not know which environment $\mathbf{Q}_m \in \mathcal{Q}$ is going to be drawn, and assume he has a set $\mathcal{P} \equiv \{\mathbf{P}_m\}_{m\in\mathcal{M}}$ 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 $\mathbf{Q}_m \in \mathcal{Q}$?

A convenient measure of how much $\mathbf{P}$ deviates from $\mathbf{P}_m$ is given by the KL-divergence. A first approach would be to construct an agent $\tilde{\mathbf{P}}$ so as to minimize the total expected KL-divergence to $\mathbf{P}_m$. This is constructed as follows. Define the history-dependent KL-divergences over the action a_t and observation o_t as

$$\begin{aligned}D^{a_t}_m(\underline{ao}_{<t}) &\equiv \sum_{a_t} \mathbf{P}_m(a_t|\underline{ao}_{<t}) \log_2 \frac{\mathbf{P}_m(a_t|\underline{ao}_{<t})}{\mathbf{Pr}(a_t|\underline{ao}_{<t})}\\ D^{o_t}_m(\underline{ao}_{<t}a_t) &\equiv \sum_{o_t} \mathbf{P}_m(o_t|\underline{ao}_{<t}a_t) \log_2 \frac{\mathbf{P}_m(o_t|\underline{ao}_{<t}a_t)}{\mathbf{Pr}(o_t|\underline{ao}_{<t}a_t)},\end{aligned}$$

where $\mathbf{Pr}$ is a given arbitrary agent. Then, define the average KL-divergences over a_t and o_t as

$$\begin{aligned}D^{a_t}_m &= \sum_{\underline{ao}_{<t}} \mathbf{P}_m(\underline{ao}_{<t}) D^{a_t}_m(\underline{ao}_{<t})\\ D^{o_t}_m &= \sum_{\underline{ao}_{<t}a_t} \mathbf{P}_m(\underline{ao}_{<t}a_t) D^{o_t}_m(\underline{ao}_{<t}a_t).\end{aligned}$$

Finally, we define the total expected KL-divergence of $\mathbf{Pr}$ to $\mathbf{P}_m$ as

$$D \equiv \limsup_{t\to\infty} \sum_m P(m) \sum_{\tau=1}^{t} \bigl(D^{a_\tau}_m + D^{o_\tau}_m\bigr).$$

We construct the agent $\tilde{\mathbf{P}}$ as the system that minimizes $D = D(\mathbf{Pr})$:

$$\tilde{\mathbf{P}} \equiv \arg\min_{\mathbf{Pr}} D(\mathbf{Pr}). \tag{2}$$

The solution to Equation 2 is the system $\tilde{\mathbf{P}}$ defined by the set of equations

$$\begin{aligned}\tilde{\mathbf{P}}(a_t|\underline{ao}_{<t}) &= \sum_m \mathbf{P}_m(a_t|\underline{ao}_{<t}) w_m(\underline{ao}_{<t})\\ \tilde{\mathbf{P}}(o_t|\underline{ao}_{<t}a_t) &= \sum_m \mathbf{P}_m(o_t|\underline{ao}_{<t}a_t) w_m(\underline{ao}_{<t}a_t)\end{aligned} \tag{3}$$

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

$$\begin{aligned}w_m(\underline{ao}_{<t}) &\equiv \frac{P(m)\mathbf{P}_m(\underline{ao}_{<t})}{\sum_{m'} P(m')\mathbf{P}_{m'}(\underline{ao}_{<t})}\\ w_m(\underline{ao}_{<t}a_t) &\equiv \frac{P(m)\mathbf{P}_m(\underline{ao}_{<t}a_t)}{\sum_{m'} P(m')\mathbf{P}_{m'}(\underline{ao}_{<t}a_t)}.\end{aligned} \tag{4}$$

For reference, see Haussler and Opper [1997], Opper [1998]. It is clear that $\tilde{\mathbf{P}}$ is just the Bayesian mixture over the agents $\mathbf{P}_m$. If we define the conditional probabilities

$$\begin{aligned}P(a_t|m, \underline{ao}_{<t}) &\equiv \mathbf{P}_m(a_t|\underline{ao}_{<t})\\ P(o_t|m, \underline{ao}_{<t}a_t) &\equiv \mathbf{P}_m(a_t|\underline{ao}_{<t}a_t)\end{aligned} \tag{5}$$

for all $\underline{ao}_{\leq t} \in \mathcal{Z}^*$, then Equation 3 can be rewritten as

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

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 Equation 5, together with the prior probabilities $P(m)$, define a Bayesian model over interaction sequences with hypotheses $m \in \mathcal{M}$.

The behavior of $\tilde{\mathbf{P}}$ can be described as follows. At any given time t, $\tilde{\mathbf{P}}$ 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 is produced (by the agent or the environment respectively), the weights w_m are updated according to Bayes' rule. In addition, $\tilde{\mathbf{P}}$ 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 $\tilde{\mathbf{P}}$ that arises due to the fact that it is not only a system that is passively observing symbols, but also *actively generating* them. Therefore, an action that is generated by the agent should not provide the same information than an observation that is issued by its environment. Intuitively, it does not make any sense to use one's own actions to do inference. In the following section we illustrate this problem with a simple statistical example.

The Problem of Causal Intervention

Suppose a statistician is asked to design a model for a given data set $\mathcal{D}$ and she decides to use a Bayesian method. She computes the posterior probability density function (pdf) over the parameters θ of the model given the data using Bayes' rule:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta')p(\theta')\,d\theta'},$$

where $p(\mathcal{D}|\theta)$ is the likelihood of $\mathcal{D}$ given θ and $p(\theta)$ is the prior pdf of θ. She can simulate the source by drawing a sample data set $\mathcal{S}$ from the predictive pdf

$$p(\mathcal{S}|\mathcal{D}) = \int p(\mathcal{S}|\mathcal{D},\theta)p(\theta|\mathcal{D})\,d\theta,$$

where $p(\mathcal{S}|\mathcal{D},\theta)$ is the likelihood of $\mathcal{S}$ given $\mathcal{D}$ and θ. She decides to do so, obtaining a sample set $\mathcal{S}'$. She understands that the nature of $\mathcal{S}'$ is very different from $\mathcal{D}$: *while $\mathcal{D}$ is informative and does change the belief state of the Bayesian model, $\mathcal{S}'$ is non-informative and thus is a reflection of the model's belief state.* Hence, she would never use $\mathcal{S}'$ to further condition the Bayesian model. Mathematically, she seems to imply that

$$p(\theta|\mathcal{D},\mathcal{S}') = p(\theta|\mathcal{D})$$

if $\mathcal{S}'$ has been generated from $p(\mathcal{S}|\mathcal{D})$ 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 simulation results $\mathcal{S}'$ in order to produce a next data set $\mathcal{D}'$ which does depend on $\mathcal{S}'$. She hands in $\mathcal{S}'$ and obtains a new data set $\mathcal{D}'$. Using Bayes' rule, the posterior pdf over the parameters is now

$$\frac{p(\mathcal{D}'|\mathcal{D},\mathcal{S}',\theta)p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}'|\mathcal{D},\mathcal{S}',\theta')p(\mathcal{D}|\theta')p(\theta')\,d\theta'} \tag{7}$$

where $p(\mathcal{D}'|\mathcal{D},\mathcal{S}',\theta)$ is the likelihood of the new data $\mathcal{D}'$ given the old data $\mathcal{D}$, the parameters θ *and the simulated data $\mathcal{S}'$*. Notice that this looks almost like the posterior pdf $p(\theta|\mathcal{D},\mathcal{S}',\mathcal{D}')$ given by

$$\frac{p(\mathcal{D}'|\mathcal{D},\mathcal{S}',\theta)p(\mathcal{S}'|\mathcal{D},\theta)p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}'|\mathcal{D},\mathcal{S}',\theta')p(\mathcal{S}'|\mathcal{D},\theta')p(\mathcal{D}|\theta')p(\theta')\,d\theta'}$$

with the exception that now the Bayesian update contains the likelihoods of the simulated data $p(\mathcal{S}'|\mathcal{D},\theta)$. This suggests that Equation 7 is a variant of the posterior pdf $p(\theta|\mathcal{D},\mathcal{S}',\mathcal{D}')$ but where the simulated data $\mathcal{S}'$ is treated in a different way than the data $\mathcal{D}$ and $\mathcal{D}'$.

Define the pdf p' such that the pdfs $p'(\theta)$, $p'(\mathcal{D}|\theta)$, $p'(\mathcal{D}'|\mathcal{D},\mathcal{S}',\theta)$ are identical to $p(\theta)$, $p(\mathcal{D}|\theta)$ and $p(\mathcal{D}'|\mathcal{D},\mathcal{S}',\theta)$ respectively, but differ in $p'(\mathcal{S}|\mathcal{D},\theta)$:

$$p'(\mathcal{S}|\mathcal{D},\theta) = \begin{cases} 1 & \text{if } \mathcal{S}' = \mathcal{S}, \\ 0 & \text{else.} \end{cases}$$

That is, p' is identical to p but it assumes that the value of $\mathcal{S}$ is fixed to $\mathcal{S}'$ given $\mathcal{D}$ and θ. For p', the simulated data $\mathcal{S}'$ is non-informative:

$$-\log_2 p(\mathcal{S}'|\mathcal{D},\theta) = 0.$$

If one computes the posterior pdf $p'(\theta|\mathcal{D},\mathcal{S}',\mathcal{D}')$, one obtains the result of Equation 7:

$$\frac{p'(\mathcal{D}'|\mathcal{D},\mathcal{S}',\theta)p'(\mathcal{S}'|\mathcal{D},\theta)p'(\mathcal{D}|\theta)p'(\theta)}{\int p'(\mathcal{D}'|\mathcal{D},\mathcal{S}',\theta')p'(\mathcal{S}'|\mathcal{D},\theta')p'(\mathcal{D}|\theta')p'(\theta')\,d\theta'} = \frac{p(\mathcal{D}'|\mathcal{D},\mathcal{S}',\theta)p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}'|\mathcal{D},\mathcal{S}',\theta')p(\mathcal{D}|\theta')p(\theta')\,d\theta'}.$$

Thus, in order to explain Equation 7 as a posterior pdf given the data sets $\mathcal{D}$, $\mathcal{D}'$ and the simulated data $\mathcal{S}'$, one has to *intervene* p in order to account for the fact that *$\mathcal{S}'$ is non-informative given $\mathcal{D}$ and θ.*

In statistics, there is a rich literature on causal intervention. In particular, we will use the formalism developed by Pearl [2000], because it suits the needs to formalize interactions in systems and has a convenient notation—compare Figures 1a & b. Given a *causal model*[1] variables that are intervened are denoted by a hat as in $\hat{\mathcal{S}}$. In the previous example, the causal model of the joint pdf $p(\theta,\mathcal{D},\mathcal{S},\mathcal{D}')$ is given by the set of conditional pdfs

$$\mathcal{C}_p = \big\{p(\theta), p(\mathcal{D}|\theta), p(\mathcal{S}|\mathcal{D},\theta), p(\mathcal{D}'|\mathcal{D},\mathcal{S},\theta)\big\}.$$

If $\mathcal{D}$ and $\mathcal{D}'$ are observed from the source and $\mathcal{S}$ is intervened to take on the value $\mathcal{S}'$, then the posterior pdf over the parameters θ is given by $p(\theta|\mathcal{D},\hat{\mathcal{S}}',\mathcal{D}')$ which is just

$$\frac{p(\mathcal{D}'|\mathcal{D},\hat{\mathcal{S}}',\theta)p(\hat{\mathcal{S}}'|\mathcal{D},\theta)p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}'|\mathcal{D},\hat{\mathcal{S}}',\theta')p(\hat{\mathcal{S}}'|\mathcal{D},\theta')p(\mathcal{D}|\theta')p(\theta')\,d\theta'} = \frac{p(\mathcal{D}'|\mathcal{D},\mathcal{S}',\theta)p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}'|\mathcal{D},\mathcal{S}',\theta')p(\mathcal{D}|\theta')p(\theta')\,d\theta'}.$$

[1] For our needs, it is enough to think about a causal model as a complete factorization of a probability distribution into conditional probability distributions representing the causal structure.

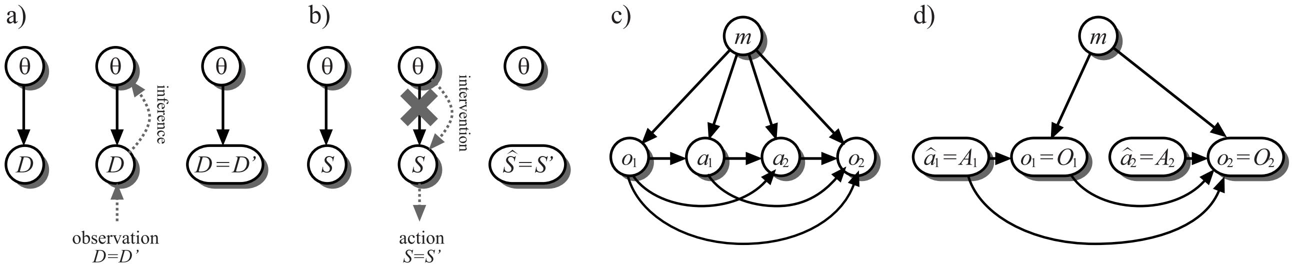

Figure 1: (a-b) Two causal networks, and the result of conditioning on $D = D'$ and intervening on $S = S'$. Unlike the condition, the intervention is set endogenously, thus removing the link to the parent θ. (c-d) A causal network representation of an I/O system with four variables $a_1o_1a_2o_2$ and latent variable m. (c) The initial, un-intervened network. (d) The intervened network after experiencing $\hat{a}_1o_1\hat{a}_2o_2$.

because $p(\mathcal{D}'|\mathcal{D},\hat{\mathcal{S}}',\theta) = p(\mathcal{D}'|\mathcal{D},\mathcal{S}',\theta)$, which corresponds to applying rule 2 in Pearl's intervention calculus, and because $p(\hat{\mathcal{S}}'|\mathcal{D},\theta') = p'(\mathcal{S}'|\mathcal{D},\theta') = 1$.

Adaptive Systems: Causal Construction

Following the discussion in the previous section, we want to construct an adaptive agent $\mathbf{P}$ by minimizing the KL-divergence to the $\mathbf{P}_m$, but this time treating actions as interventions. Based on the definition of the conditional probabilities in Equation 5, we construct now the KL-divergence criterion to characterize $\mathbf{P}$ using intervention calculus. Importantly, interventions index a set of intervened probability distribution derived from an initial 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 $o_1o_2\ldots$. Because of this, we are going to construct a set of criteria indexed by the intervention sequences, but we will see that they all have the same solution. Define the history-dependent intervened KL-divergences over the action a_t and observation o_t as

$$C^{a_t}_m(\underline{\hat{a}o}_{<t}) \equiv \sum_{a_t} P(a_t|m,\underline{\hat{a}o}_{<t}) \log_2 \frac{P(a_t|m,\underline{\hat{a}o}_{<t})}{\mathbf{Pr}(a_t|\underline{ao}_{<t})}$$

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

where $\mathbf{Pr}$ is a given arbitrary agent. Note that past actions are treated as interventions. Then, define the average KL-divergences over a_t and o_t as

$$\begin{aligned} C^{a_t}_m &= \sum_{\underline{ao}_{<t}} P(\underline{\hat{a}o}_{<t}|m) C^{a_t}_m(\underline{\hat{a}o}_{<t}) \\ C^{o_t}_m &= \sum_{\underline{ao}_{<t}a_t} P(\underline{\hat{a}o}_{<t}a_t|m) C^{o_t}_m(\underline{\hat{a}o}_{<t}\hat{a}_t). \end{aligned}$$

Finally, we define the total expected KL-divergence of $\mathbf{P}$ to $\mathbf{P}_m$ as

$$C \equiv \limsup_{t\to\infty} \sum_m P(m) \sum_{\tau=1}^{t} \bigl(C^{a_\tau}_m + C^{o_\tau}_m\bigr). \qquad (8)$$

We construct the agent $\mathbf{P}$ as the system that minimizes $C = C(\mathbf{Pr})$:

$$\mathbf{P} \equiv \arg\min_{\mathbf{Pr}} C(\mathbf{Pr}). \qquad (9)$$

The solution to Equation 9 is the system $\mathbf{P}$ defined by the set of equations

$$\begin{aligned} \mathbf{P}(a_t|\underline{ao}_{<t}) &= P(a_t|\underline{\hat{a}o}_{<t}) \\ &= \sum_m P(a_t|m,\underline{\hat{a}o}_{<t}) v_m(\underline{\hat{a}o}_{<t}) \\ \mathbf{P}(o_t|\underline{ao}_{<t}a_t) &= P(o_t|\underline{\hat{a}o}_{<t}\hat{a}_t) \\ &= \sum_m P(o_t|m,\underline{\hat{a}o}_{<t}\hat{a}_t) v_m(\underline{\hat{a}o}_{<t}\hat{a}_t) \end{aligned} \qquad (10)$$

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

$$\begin{aligned} v_m(\underline{ao}_{<t}a_t) = v_m(\underline{ao}_{<t}) &\equiv \frac{P(m)P(\underline{\hat{a}o}_{<t}|m)}{\sum_{m'} P(m')P(\underline{\hat{a}o}_{<t}|m)} \\ &= \frac{P(m)\prod_{\tau=1}^{t-1} P(o_\tau|m,\underline{\hat{a}o}_{<\tau}\hat{a}_\tau)}{\sum_{m'} P(m')\prod_{\tau=1}^{t-1} P(o_\tau|m',\underline{\hat{a}o}_{<\tau}\hat{a}_\tau)}. \end{aligned} \qquad (11)$$

The proof follows the same line of argument as the solution to Equation 2 with the crucial difference that actions are treated as interventions. Consider without loss of generality the summand $\sum_m P(m)C^{a_t}_m$ in Equation 8. Note that the KL-divergence can be written as a difference of two logarithms, where only one term depends on $\mathbf{Pr}$ that we want to vary. Therefore, we can integrate out the other term and write it as a constant c. Then we get

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

Substituting $P(\underline{\hat{a}o}_{<t}|m)$ by $P(m|\underline{\hat{a}o}_{<t})P(\underline{\hat{a}o}_{<t})/P(m)$ and identifying $\mathbf{P}$ characterized by Equations 10 and 11 we obtain

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

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. By choosing this optimum one obtains $\mathbf{Pr}(a_t|\hat{\underline{ao}}_{<t}) = \mathbf{P}(a_t|\hat{\underline{ao}}_{<t})$ for all a_t. Note that the solution to this variational problem is independent of the weighting $P(\hat{\underline{ao}}_{<t})$. 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 8, their variational problems are mutually independent.

The behavior of $\mathbf{P}$ differs in an important aspect from $\tilde{\mathbf{P}}$. At any given time t, $\mathbf{P}$ maintains a mixture over systems $\mathbf{P}_m$. The weighting over these systems is given by the mixture coefficients v_m. In contrast to $\tilde{\mathbf{P}}$, $\mathbf{P}$ updates the weights v_m *only* whenever a new observation o_t is produced by the environment respectively. The update follows Bayes' rule but treating past actions as interventions, i.e. dropping the evidence they provide. In addition, $\mathbf{P}$ issues an action a_t suggested by an system m drawn randomly according to the weights v_m—see Figures 1c & d.

If we use the following equalities connecting the weights and the intervened posterior distributions

$$v_m(\underline{ao}_{<t}) = P(m|\hat{\underline{ao}}_{<t}) = P(m|\hat{\underline{ao}}_{<t}\hat{a}_t) = v_m(\underline{ao}_{<t}a_t)$$

and substitute interventions by observations in the conditionals

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

which corresponds to rule 2 of Pearl's intervention calculus, we can rewrite Equations 10 and 11 as

$$\begin{aligned} \mathbf{P}(a_t|\underline{ao}_{<t}) &= P(a_t|\hat{\underline{ao}}_{<t}) \\ &= \sum_m P(a_t|m, \underline{ao}_{<t})P(m|\hat{\underline{ao}}_{<t}) \qquad (12) \\ \mathbf{P}(o_t|\underline{ao}_{<t}a_t) &= P(o_t|\hat{\underline{ao}}_{<t}\hat{a}_t) \\ &= \sum_m P(o_t|m, \underline{ao}_{<t}a_t)P(m|\hat{\underline{ao}}_{<t}) \qquad (13) \end{aligned}$$

where the intervened posterior probabilities are

$$P(m|\hat{\underline{ao}}_{<t}) = \frac{P(m)\prod_{\tau=1}^{t-1} P(o_\tau|m, \underline{ao}_{<\tau}a_\tau)}{\sum_{m'} P(m')\prod_{\tau=1}^{t-1} P(o_\tau|m', \underline{ao}_{<\tau}a_\tau)}. \qquad (14)$$

Equations 12, 13 and 14 are important because they describe the behavior of $\mathbf{P}$ only in terms of known probabilities, i.e. probabilities that are computable from the causal model associated to P given by

$$C_P = \big\{P(m), P(a_t|m, \underline{ao}_{<t}), P(o_t|m, \underline{ao}_{<t}a_t) : t \geq 1\big\}.$$

Importantly, Equation 12 describes a stochastic method to produce desirable actions that differs fundamentally from an agent that is constructed by choosing an optimal policy with respect to a given utility criterion. We call this action selection rule the *Bayesian control rule.*

Experimental Results

Here we design a very simple toy experiment to illustrate the behavior of an agent $\tilde{\mathbf{P}}$ based on a Bayesian mixture compared to an agent $\mathbf{P}$ based on the Bayesian control rule.

Let $\mathbf{Q}_0$, $\mathbf{Q}_1$, $\mathbf{P}_0$ and $\mathbf{P}_1$ be four agents with binary I/O sets $\mathcal{A} = \mathcal{O} = \{0, 1\}$ defined as follows. $\mathbf{P}_1$ is such that $\mathbf{P}_1(a_t|\underline{ao}_{<t}) = \mathbf{P}_1(a_t)$ and $\mathbf{P}_1(o_t|\underline{ao}_{<t}a_t) = \mathbf{P}_1(o_t)$ for all $\underline{ao}_{\leq t} \in \mathcal{Z}^*$, where

$$\mathbf{P}_1(a_t) = \begin{cases} 0.1 & \text{if } a_t = 0 \\ 0.9 & \text{if } a_t = 1 \end{cases}, \quad \mathbf{P}_1(o_t) = \begin{cases} 0.4 & \text{if } a_t = 0 \\ 0.6 & \text{if } a_t = 1 \end{cases}.$$

Let $\mathbf{P}_0$ be such that

$$\begin{aligned} \mathbf{P}_0(a_t|\underline{ao}_{<t}) &= 1 - \mathbf{P}_1(a_t|\underline{ao}_{<t}) \\ \mathbf{P}_0(o_t|\underline{ao}_{<t}a_t) &= 1 - \mathbf{P}_0(o_t|\underline{ao}_{<t}a_t) \end{aligned}$$

for all $\underline{ao}_{<t} \in \mathcal{Z}^*$. Thus, $\mathbf{P}_0$ and $\mathbf{P}_1$ are agents that are biased towards observing and acting 0's and 1's respectively. Furthermore, $\mathbf{Q}_0 = \mathbf{P}_0$ and $\mathbf{Q}_1 = \mathbf{P}_1$. Assume a uniform distribution over $\mathcal{Q} = \{\mathbf{Q}_0, \mathbf{Q}_1\}$, i.e. $P(m = 0) = P(m = 1) = \frac{1}{2}$.

Assume $\mathbf{Q}_0 \in \mathcal{Q}$ is drawn. In this case, one wants the agents $\tilde{\mathbf{P}}$ and $\mathbf{P}$ to minimize the deviation from $\mathbf{P}_0$. Consider the following instantaneous measure

$$\begin{aligned} d(t) \equiv &\sum_{a'_t} \mathbf{P}_0(a'_t)\log_2 \frac{\mathbf{P}_0(a'_t)}{\mathbf{Pr}(a'_t|\underline{ao}_{<t})} \\ &+ \sum_{o'_t} \mathbf{P}_0(o'_t)\log_2 \frac{\mathbf{P}_0(o'_t)}{\mathbf{Pr}(o'_t|\underline{ao}_{<t}a_t)} \end{aligned}$$

where $a_1o_1a_2o_2\ldots$ is a realization of the interaction system $(\mathbf{Pr}, \mathbf{Q}_0)$. $d(t)$ measures how much $\mathbf{Pr}$'s action and observation probabilities deviate from $\mathbf{P}_0$ at time t.

Recall that both $\tilde{\mathbf{P}}$ and $\mathbf{P}$ maintain a mixture over $\mathbf{P}_0$ and $\mathbf{P}_1$. The instantaneous I/O probabilities of such a system can always be written as

$$\begin{aligned} &w\mathbf{P}_0(a_t) + (1 - w)\mathbf{P}_1(a_t) \\ &w\mathbf{P}_0(o_t) + (1 - w)\mathbf{P}_1(o_t). \end{aligned}$$

where $w \in [0, 1]$. Thus, it is easy to see that the instantaneous I/O deviation takes on the minimum value when $w = 1$ and the maximum value when $w = 0$: In the case $w = 1$, $d(t) = 0$ bits; In the case $w = 0$, $d(t) \approx 2.653$.

We have simulated realizations of the instantaneous I/O deviation using the agents $\tilde{\mathbf{P}}$ and $\mathbf{P}$. The results are summarized in Figure 2. For $\tilde{\mathbf{P}}$, $d(t)$ happens to be non-ergodic: it either converges to $d(t) \to 0$ or to $d(t) \to\approx 2.654$, implying that either $\tilde{\mathbf{P}} \to \mathbf{P}_0$ or $\tilde{\mathbf{P}} \to \mathbf{P}_1$ respectively. In contrast, $d(t) \to 0$ always for $\mathbf{P}$, implying that $\mathbf{P} \to \mathbf{P}_0$.

Analogous results are obtained when $\mathbf{Q}_1 \in \mathcal{Q}$ is drawn instead: For $\tilde{\mathbf{P}}$, $d(t)$ converges either to 0 or to ≈ 2.654, whereas for $\mathbf{P}$, $d(t) \to\approx 2.654$ always implying that $\mathbf{P} \to \mathbf{P}_1$. Hence, $\mathbf{P}$ shows the correct adaptive behavior while $\tilde{\mathbf{P}}$ does not.

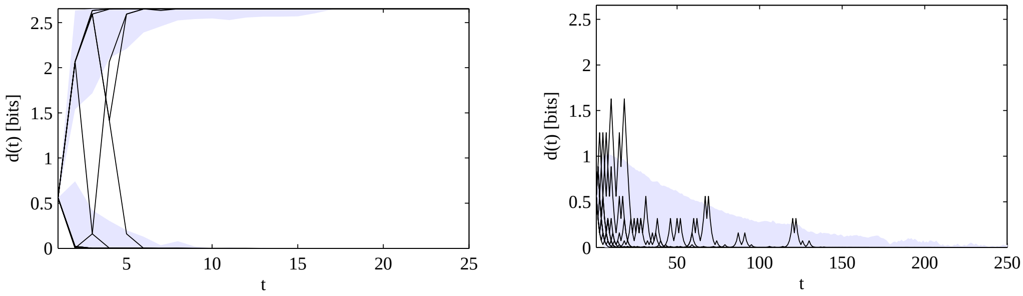

Figure 2: 10 realizations of the instantaneous deviation $d(t)$ for the agents $\tilde{\mathbf{P}}$ (left panel) and $\mathbf{P}$ (right panel). The shaded region represents the standard deviation barriers computed over 1000 realizations. Since $d(t)$ is non-ergodic for $\tilde{\mathbf{P}}$, we have separated the realizations converging to 0 from the realizations converging to ≈ 2.654 to compute the barriers. Note that the time scales differ in one order of magnitude.

Conclusions

We propose 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 agents into Bayesian mixture of I/O distributions. The question of how to integrate information generated by an agent's probabilistic model into the agent's information state lies at the very heart of adaptive agent design. We show that the naïve application of Bayes' rule to I/O distributions leads to inconsistencies, because outputs don't provide the same type of information as genuine observations. Crucially, these inconsistencies vanish if intervention calculus is applied [Pearl, 2000].

Some of the presented key ideas are not unique to the Bayesian control rule. The idea of representing agents and environments as I/O streams has been proposed by a number of other approaches, such as predictive state representation (PSR) [Littman et al., 2002] and the universal AI approach by Hutter [2004]. The idea of breaking down a control problem into a superposition of controllers has been previously evoked in the context of "mixture of experts"-models like the MOSAIC-architecture Haruno et al. [2001]. Other stochastic action selection approaches are found in exploration strategies for (PO)MDPs [Wyatt, 1997], learning automata [Narendra and Thathachar, 1974] and in probability matching [R.O. Duda, 2001] amongst others. The usage of compression principles to select actions has been proposed by AI researchers, for example Schmidhuber [2009]. The main contribution of this paper is the derivation of a stochastic action selection and inference rule by minimizing KL-divergences of intervened I/O distributions.

An important potential application of the Bayesian control rule would naturally be the realm of adaptive control problems. 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.

References

R. Beer. *Intelligence as Adaptive Behavior.* Academic Press, Inc., 1990.

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

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

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

M. Littman, R. Sutton, and S. Singh. Predictive representations of state. In *Neural Information Processing Systems (NIPS)*, number 14, pages 1555–1561, 2002.

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

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

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

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

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

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

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