Adaptive Coding of Actions and Observations

Pedro A. Ortega & Daniel A. Braun
Max Planck Institute for Intelligent Systems
Max Planck Institute for Biological Cybernetics
{pedro.ortega|daniel.braun}@tuebingen.mpg.de

Abstract

The application of expected utility theory to construct adaptive agents is both computationally intractable and statistically questionable. To overcome these difficulties, agents need the ability to delay the choice of the optimal policy to a later stage when they have learned more about the environment. How should agents do this optimally? An information-theoretic answer to this question is given by the Bayesian control rule—the solution to the adaptive coding problem when there are not only observations but also actions. This paper reviews the central ideas behind the Bayesian control rule.

Keywords: Adaptive control, Bayesian control, causal interventions, adaptive coding

1 Introduction

The design of adaptive interactive systems is the quintessential problem of artificial intelligence. *In theory*, it is solved by choosing the policy that maximizes the expected utility of the interaction sequence generated by the agent and the environment. The only difference to the non-adaptive case is that expected utilities are taken with respect to a Bayesian model over possible environments (instead of the distribution belonging to any particular environment) [2]. *In practice* however, finding the optimal policy is intractable. The obvious reason why this is so is because the policy space grows exponentially with the planning horizon. However, there is a second—more subtle—reason. *Even if planning was tractable*, it is statistically wasteful having to calculate the optimal policy before having interacted with the environment even once because the predictions of the model are not supported by any data [4]. Both aforementioned problems have a solution that is straightforward: namely, to delay the choice of the optimal policy to a later stage when it is justified by the data.

The central question addressed in this paper is: *how do we choose the optimal policy dynamically?* As we will see, this question requires solving the following subproblems:

1. How is uncertainty over the policy represented?
2. How are actions issued when the policy is uncertain?
3. How is this uncertainty reduced?
4. How is computation modeled?

We argue that every adaptive control or reinforcement learning (RL) algorithm has to deal with each one of these problems, either explicitly or implicitly. For instance, popular RL algorithms like Q-Learning, which keep track of a point estimate of the optimal policy (and hence do not explicitly represent the uncertainty over the optimal policy), turn out to implicitly represent policy uncertainty when interpreted as a stochastic approximation method [9].

An optimal information-theoretic answer to our central question was given in [5]. There, a rule for adaptive control—called the *Bayesian control rule*—was derived as the solution to the adaptive

coding problem when there are not only observations but also actions. This paper recapitulates the central ideas behind the Bayesian control rule.

2 Preliminaries

Notation. 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.

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

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

for each $\underline{ao}_{\leq t} \in \mathcal{Z}^*$. Graphically, an I/O system can be best thought of as a tree where nodes denote interaction pasts (i.e. the probability condition) and edges represent transitions (i.e. the probability argument). Depending on how the I/O system is interfaced with another I/O system, these conditional probabilities will describe how symbols are either *generated* or *predicted* by the system.

Let $\mathbf{P}$, $\mathbf{Q}$ be two I/O systems. Throughout this paper, we use the convention that $\mathbf{P}$ is an *agent* to be constructed, which is then going to be interfaced with a preexisting (but possibily unknown) *environment* $\mathbf{Q}$. An *interaction system* is a pair $(\mathbf{P}, \mathbf{Q})$ giving rise to the *generative measure* $\mathbf{G}$ that describes the probability law governing the interaction sequences once the two systems are coupled. $\mathbf{G}$ is defined as

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

for all $\underline{ao}_t \in \mathcal{Z}^*$. Intuitively, these equations say that actions a_t are generated by the agent and observations o_t are generated by the environment, and that these interactions depend on the interaction history. This models a fairly general interaction protocol that can accommodate many others.

Policy and Predictor. For a given agent $\mathbf{P}$, we call the action probabilities $\mathbf{P}(a_t | \underline{ao}_{<t})$ the *policy*, and the observation probabilities $\mathbf{P}(o_t | \underline{ao}_{<t})$ the *predictor*. The predictor captures the assumptions the agent makes about the statistics of the environment. We assume that we have access to a *set of policies* $\{\mathbf{P}(a_t | \pi, \underline{ao}_{<t}) : \pi \in \Pi\}$ and a *set of predictors* $\{\mathbf{P}(o_t | \theta, \underline{ao}_{<t} a_t) : \theta \in \Theta\}$ from which we can pick the policy and the predictor respectively. These sets can contain multi-armed bandits, MDPs, POMDPs, or any other controllable stochastic processes. With a slight abuse of language, we will say "the policy π" and "the predictor θ" when we really mean the corresponding probabilistic models.

At a first glance, it seems sufficient to specify just the policy in order to characterize an agent, since practically all it needs to know is how to generate actions given the history. However, a complete information-theoretic characterization requires both the policy and the predictor, because together they specify the number of bits needed to *both generate and record experience*.

Preferences, Utility, and the Relation between Policy and Predictor. If the environment $\mathbf{Q}$ is known, then we pick the matching predictor $\theta \in \Theta$ such that $\mathbf{P}(o_t | \theta, \underline{ao}_{<t} a_t) = \mathbf{Q}(o_t | \underline{ao}_{<t} a_t)$ and the corresponding *preferred/optimal policy*, which—according to expected utility theory—is the one that maximizes the expected utility of the interaction sequence. Since interaction sequences are infinitely long, expected utililities are calculated as limit processes. Formally, let $\{\mathbf{U}\}_{t \in \mathbb{N}}$ be a collection of real-valued functions over finite interaction strings in $\mathcal{Z}^*$, such that for each infinite interaction sequence $z = a_1 o_1 a_2 o_2 \ldots \in \mathcal{Z}^\infty$, the limit

$$\lim_{t \to \infty} \mathbf{U}_t(\underline{ao}_{\leq t}) =: \mathbf{U}(z)$$

exists. The limit function $\mathbf{U} : \mathcal{Z}^\infty \to \mathbb{R}$ is the *utility function*. Hence, the optimal policy $\pi \in \Pi$ is the one that maximizes

$$\lim_{t\to\infty} \sum_{\underline{ao}_{\leq t}} \mathbf{P}_{\pi,\theta}(\underline{ao}_{\leq t}) \mathbf{U}(\underline{ao}_{\leq t})$$

where $\mathbf{P}_{\pi,\theta}$ is the agent having policy $\pi \in \Pi$ and predictor $\theta \in \Theta$. In practice, utility functions are constructed as cumulative functions of instantaneous reward or cost functions.

The point is that, regardless of the underlying utility function, the optimal policy is a function of the predictor, i.e. every predictor $\theta \in \Theta$ has an associated optimal policy $\pi(\theta) \in \Pi$. From an information-theoretic point of view, the parameter of the predictor contains all the information needed to uniquely identify the optimal policy: knowing θ implies knowing π, and conversely, not knowing π implies not knowing θ. Loosely speaking, expected utility theory provides us with a "conceptual glue" between predictor and policy that allows us to shift our attention to full dynamics. Consequently, in what follows we use a single index $\theta \in \Theta$ to specify a complete stochastic process consisting of its predictor and its associated optimal policy.

3 Adaptive Agents

Agents under Policy Uncertainty. We now consider the case when the environment $\mathbf{Q}$ is unknown. This means that we don't know which predictor—and hence neither which policy—to choose. We express our uncertainty by placing probabilities $\mathbf{P}(\theta)$ over $\theta \in \Theta$. Thus, the question is: accepting that we do not know the optimal policy, how do we design an agent that learns the optimal dynamics in the most efficient way? To make this question precise, we rephrase it in terms of adaptive coding as: how do we maximally compress the interaction sequence when the environment is uncertain? This amounts to finding an agent $\mathbf{Pr}$ that minimizes the collection of functionals

$$\sum_\theta \mathbf{P}(\theta) \left\{ \sum_{\underline{ao}_{\leq t}} \mathbf{P}(\underline{ao}_{\leq t}|\theta) \log \frac{\mathbf{P}(\underline{ao}_{\leq t}|\theta)}{\mathbf{Pr}(\underline{ao}_{\leq t})} \right\} \qquad \text{for all } t \in \mathbb{N}, \tag{2}$$

that is, the average relative entropy to the target dynamics for any planning horizon. In [5], it was proven that the solution is given by the Bayesian mixture

$$\mathbf{Pr}(\underline{ao}_{\leq t}) = \mathbf{P}(\underline{ao}_{\leq t}) = \sum_\theta \mathbf{P}(\theta)\mathbf{P}(\underline{ao}_{\leq t}|\theta) \qquad \text{for all } t \in \mathbb{N}. \tag{3}$$

This agent is well defined because the solutions for different planning horizons $t \in \mathbb{N}$ are all consistent with each other. Inspecting (3), we see that the resulting agent is a weighted superposition of all the possible agents with weights given by their prior plausibilities. Note that inserting (3) into (2) yields the mutual information $\mathbf{I}(\theta; \underline{ao}_{\leq t})$ between the environment and the interaction sequence.

Acting & Observing. One of the insights of [5] and ultimately of statistical causality [6, 7] is that actions can only reduce uncertainty via its effects, but not by themselves. Therefore, actions have to be treated as causal interventions—unlike observations, which are treated as normal conditions. Formally, this means that actions have to be drawn from

$$a_t \sim \mathbf{P}(a_t|\underline{\hat{a}o}_{<t}),$$

where the "hat"-notation $\hat{a}_t$ denotes causal intervention rather than Bayesian conditioning. Informally, an intervention is a mechanism to inform the agent that the information content of its own action is zero after it has been issued. Using causal calculus one can show that

$$\mathbf{P}(a_t|\underline{\hat{a}o}_{<t}) = \sum_\theta \mathbf{P}(a_t|\theta, \underline{ao}_{<t})\mathbf{P}(\theta|\underline{\hat{a}o}_{<t}). \tag{4}$$

This can be read as follows: action a_t is generated by first sampling a belief $\bar{\theta} \sim \mathbf{P}(\theta|\underline{\hat{a}o}_{<t})$ from the posterior, and then by sampling the action $a_t \sim \mathbf{P}(a_t|\bar{\theta}, \underline{\hat{a}o}_{<t})$ from policy $\bar{\theta}$ *as if* it was the optimal policy. This mechanism of "randomly instantiating beliefs" is also the central idea behind the *random beliefs* and *Thompson sampling* schemes [1, 3, 8].

For the posterior, the following recursive expression is illuminating [5]:

$$\mathbf{P}(\theta|\underline{\hat{a}o}_{\leq t}\hat{a}_{t+1}) = \mathbf{P}(\theta|\underline{\hat{a}o}_{\leq t}) = \frac{\mathbf{P}(o_t|\theta, \underline{ao}_{<t}a_t)\mathbf{P}(\theta|\underline{\hat{a}o}_{<t}\hat{a}_t)}{\sum_{\theta'}\mathbf{P}(o_t|\theta', \underline{ao}_{<t}a_t)\mathbf{P}(\theta'|\underline{\hat{a}o}_{<t}\hat{a}_t)}. \tag{5}$$

The first equality means that actions do not change the posterior. The second equality is just Bayes' rule. Hence, the posterior is only updated after observations. It is clear that this agent converges to the optimal policy when the predictor converges. The conditions of convergence are beyond the scope of this paper.

Modeling Computation. The previous section described a method to design adaptive agents that dynamically discover their optimal policies in a fully observation-driven way. How do we fit agents that precalculate their optimal policies into this scheme?

To understand this connection, it is helpful to think about the uncertainties that are involved during the calculation of optimal policies. The reason why this calculation is done in the first place is precisely because we hope to resolve our uncertainty over the optimal policy. Hence, we can think of individual calculation steps as interaction cycles that reduce the uncertainty over the optimal policy in the same sense interactions with the world would do. This is because, by definition, *any device that transforms an agent's beliefs over the dynamics is part of the environment.*

4 Conclusions

The application of expected utility theory to construct adaptive agents is both computationally intractable and statistically wasteful. To overcome these difficulties, agents need the ability to delay the choice of the optimal policy to a later stage when they have collected more data about the environment. We have argued that the Bayesian control rule presented in [5] is the information-theoretic optimal solution to this adaptive control problem. A key feature is that it distinguishes between the nature of actions and observations—actions are treated as causal interventions and observations as standard Bayesian conditions.

References

[1] J. Friedman and C. Mezzetti. Random belief equilibrium in normal form games. *Games and Economic Behavior*, 51(2):296–323, 2005.

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

[3] B.C. May and D.S. Leslie. Simulation studies in optimistic Bayesian sampling in contextual-bandit problems. Technical Report 11:02, Statistics Group, Department of Mathematics, University of Bristol, 2011.

[4] P.A. Ortega. *A Unified Framework for Resource-Bounded Autonomous Agents Interacting with Unknown Environments*. PhD thesis, Dept. of Engineering, University of Cambridge, 2011.

[5] P.A. Ortega and D.A. Braun. A minimum relative entropy principle for learning and acting. *Journal of Artificial Intelligence Research*, 38:475–511, 2010.

[6] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, UK, 2nd edition, 2009.

[7] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, 2nd edition, 2000.

[8] W.R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3–4):285–294, 1933.

[9] J.N. Tsitsiklis. Asynchronous stochastic approximation and q-learning. *Machine Learning*, 16(3):185–202, 1993.