Ortega and Braun *Complex Adaptive Systems Modeling* 2014, 2:2
http://www.casmodeling.com/content/2/1/2

Complex Adaptive Systems Modeling
a SpringerOpen Journal

RESEARCH Open Access

Generalized Thompson sampling for sequential decision-making and causal inference

Pedro A Ortega[1*] and Daniel A Braun[2*]

*Correspondence:
ope@seas.upenn.edu;
daniel.braun@tuebingen.mpg.de
[1] GRASP Laboratory, Electrical and Systems Engineering Department, University of Pennsylvania, Philadelphia, PA 19104, USA
[2] Max Planck Institute for Biological Cybernetics and Max Planck Institute for Intelligent Systems, Speemanstrasse 38, Tübingen 72076, Germany

Abstract

Purpose: Sampling an action according to the probability that the action is believed to be the optimal one is sometimes called Thompson sampling.

Methods: Although mostly applied to bandit problems, Thompson sampling can also be used to solve sequential adaptive control problems, when the optimal policy is known for each possible environment. The predictive distribution over actions can then be constructed by a Bayesian superposition of the policies weighted by their posterior probability of being optimal.

Results: Here we discuss two important features of this approach. First, we show in how far such generalized Thompson sampling can be regarded as an optimal strategy under limited information processing capabilities that constrain the sampling complexity of the decision-making process. Second, we show how such Thompson sampling can be extended to solve causal inference problems when interacting with an environment in a sequential fashion.

Conclusion: In summary, our results suggest that Thompson sampling might not merely be a useful heuristic, but a principled method to address problems of adaptive sequential decision-making and causal inference.

Keywords: Thompson sampling; Adaptive control; Bounded rationality; Decision-making; Causal inference

Background

In a research paper from 1933, Thompson studied the problem of finding out which one of two drugs was better when testing them on a patient population under the constraint that as few people as possible should be subjected to the inferior drug during the course of testing (Thompson 1933). Given a current (Bayesian) probability estimate P of one treatment being better than the other, he suggested that it might be a good idea to adjust the proportions of future test subjects that take the two drugs to the respective probabilities P and $1 - P$. This way one would not run into the danger of *permanently* cutting off all future test subjects from a potentially superior treatment that so far seems inferior due to statistical fluctuations, while only *temporarily* risking exposure to a potentially inferior drug for a smaller proportion of the population. Randomizing actions based on the probability that this action is believed to be optimal when faced with an unknown environment is now sometimes called *Thompson sampling*.

Today, Thompson's problem is generally thought of as a *bandit problem* that consists in determining which lever to pull at which point in time when facing a set of one-armed slot machines, each one having an unknown distribution over a reward variable (Russell and Norvig 1995; Sutton and Barto 1998). In the case of known prior probabilities and geometrically discounted future rewards, Gittins (1979) provides an optimal policy for the bandit problem that maximizes the expected future cumulative discounted reward. In contrast, Thompson sampling is usually considered as a heuristic approach to solve bandit problems (Wyatt 1997; Granmo 2008, 2010; Asmuth et al. 2009; Graepel et al. 2010; Scott 2010; May and Leslie 2011; Chapelle and Li 2011; Agrawal and Goyal 2011; Granmo and Glimsdal 2013; May et al. 2012; Kaufmann et al. 2012; Russo and Roy 2013; Korda et al. 2013; Bubeck and Liu 2013). However, the basic idea of Thompson sampling—that is, sampling actions from a mixture distribution of policies according to their probability of being optimal–can also be applied to solve more general problems in sequential adaptive control (Dearden et al. 1998; Strens 2000; Ortega and Braun 2010a, 2010b, 2012a; Braun and Ortega 2010; Osband and Russo 2013; Cao and Ray 2012; Tziortziotis et al. 2013a, 2013; Dimitrakakis 2013; Dimitrakakis and Tziortziotis 2013; Mellor and Shapiro 2013).

In this paper, our aim is to discuss some of the basic properties of Thompson sampling as a modeling approach. In particular, we want to argue that

1. Thompson sampling can be considered as an application of Bayes' rule for acting where actions are treated as causally intervened random variables within the framework of statistical causality.
2. Thompson sampling can be considered as a form of optimal adaptive control under bounded rationality where limited information processing capabilities are modeled by entropic search costs.
3. Thompson sampling provides a natural strategy for causal induction when interacting with an environment with unknown causal structure.

Although the third section contains an algorithmic extension to previous work (Ortega and Braun 2010a; 2010b), it should be emphasized that the main contribution of the paper is not so much to present a novel algorithm, but to discuss basic properties of Thompson sampling, in particular how it relates to the information-theoretic bounded rationality model in (Ortega and Braun 2013), how this boundedness can be interpreted in terms of sampling complexity, and how this method can be applied to solve problems of causal inference.

The paper is structured as follows. In Section "Problem statement" we clarify the problem statement and recapitulate the main result of (Ortega and Braun 2010b). In Section "Decision-making with limited resources" we analyze the decision-making problem faced by agents that are unable to compute the single best policy. In Section "Causal induction" we investigate how this approach can be applied to adaptive agents that need to discover the causal structure of their environment. Finally, we discuss the significance of these results in Section "Discussion".

Methods

Problem statement

In an adaptive control problem a decision-maker faces an environment Q_θ drawn from a set of potential environments $\mathcal{Q} = \{Q_\theta | \theta \in \Theta\}$. In general θ could be a continuous variable, but we restrict our exposition to the discrete case. Each environment Q_θ can be characterized by a set of conditional distributions $Q(o_t|\theta, a_{\leq t}, o_{<t})$ that indicate the probability of observing o_t given past observations $o_{<t} = o_1 \dots o_{t-1}$ and past actions $a_{\leq t} = a_1 \dots a_t$. This class of environments is very general, and it encompasses multi-armed bandits, (partially observable) Markov decision processes and others –compare Chapter 3 (Legg 2008). To allow for self-optimizing agents, the environment is typically assumed to be ergodic, so agents can recover from their mistakes –compare Section 3.5 (Legg 2008). The decision-maker has perfectly fitting prediction models $P(o_t|\theta, a_{\leq t}, o_{<t}) = Q(o_t|\theta, a_{\leq t}, o_{<t})$, but is uncertain about θ. The uncertainty about θ can be represented by a prior distribution $P(\theta)$. The interaction proceeds as follows. First an environment θ is sampled from $P(\theta)$. The agent picks an action a_0 and receives an observation o_0, to which the agent responds with a_1 and receives observation o_1 etc. The agent's policy can be described by a set of conditional distributions $P(a_t|o_{<t}, a_{<t})$.

Problem statement: decision-theory

In order to solve the problem within the framework of maximum expected utility theory, one requires

- a prior $P(\theta)$ over possible environments Q_θ
- a class of prediction models $P(o_t|\theta, a_{\leq t}, o_{<t})$
- a utility function $U(o_{\leq T}, a_{\leq T})$.

Then one can reduce the problem of the unknown environment to a problem with known environment. Such a "known" environment can be created from the set of possible environments by marginalizing over the parameter of the possible environments, thus, obtaining the Bayesian mixture distribution, where

$$P(o_t|o_{<t}, a_{\leq t}) = \sum_\theta P(\theta|o_{<t}, a_{\leq t}) P(o_t|\theta, o_{<t}, a_{\leq t}). \tag{1}$$

The adaptive control problem is then solved by finding the optimal policy $p(a_t|a_{<t}, o_{<t}) = \delta(a_t - a_t^*)$, where

$$\begin{aligned} a_t^* = \arg\max_{a_t} \sum_{o_t} P(o_t|o_{<t}, a_{\leq t}) \max_{a_{t+1}} \sum_{o_{t+1}} P(o_{t+1}|o_{\leq t}, a_{\leq t+1}) \cdots \\ \cdots \max_{a_T} \sum_{o_T} P(o_T|o_{<T}, a_{\leq T}) U(o_{\leq T}, a_{\leq T}) \end{aligned} \tag{2}$$

maximizes the expected utility under the mixture distribution (Hutter 2004). Equations (1) and (2) define a Bayesian adaptive control problem (Martin 1967; Duff 2002). This problem formulation becomes quickly intractable, as the number of reachable information states grows exponentially in the time horizon (Duff 2002).

Problem statement: probability theory & statistical causality

Ignoring the notion of utility for a moment and treating actions purely as (causally intervened) random variables (Pearl 2000), one could think of another kind of adaptive control problem that is defined entirely in probabilistic and causal terms. This requires the following ingredients

- a prior $P(\theta)$ over possible environments Q_θ
- a class of prediction models $P(o_t|\theta, a_{\leq t}, o_{<t})$
- a class of policy models $P(a_t|\theta, a_{<t}, o_{<t})$

such that for every possible environment indexed by θ there is a perfectly fitting predictor $P(o_t|\theta, a_{\leq t}, o_{<t}) = Q_\theta(o_t|a_{\leq t}, o_{<t})$ and a desirable custom-built[a] policy $P(a_t|\theta, a_{<t}, o_{<t})$. The problem statement is: What is the next action a_t given the uncertainty over θ and given the history of previous actions $\hat{a}_{<t}$ and previous observations $o_{<t}$? As for any random variable, answering this question for the random variable a_t simply requires computing the predictive distribution conditioned on the past, that is

$$P(a_t|\hat{a}_{<t}, o_{<t}) = \sum_\theta P(a_t|\theta, \hat{a}_{<t}, o_{<t}) P(\theta|\hat{a}_{<t}, o_{<t}). \tag{3}$$

As there can be only one action at any one time, single actions can be obtained as samples from $P(a_t|\hat{a}_{<t}, o_{<t})$. Importantly, sampling from $P(a_t|\hat{a}_{<t}, o_{<t})$ is equivalent to first sampling a random belief θ from the posterior $P(\theta|\hat{a}_{<t}, o_{<t})$ and then sampling an action from $P(a_t|\theta, \hat{a}_{<t}, o_{<t})$. This componentwise sampling from a mixture distribution is known as *hierarchical sampling* and corresponds here to a generalized Thompson sampling procedure, where first a random belief is sampled and then the associated policy with respect to this belief is executed. If we assume now that each of the custom-built policies is optimal in their respective environments, we effectively select an action according to the probability that it is the optimal action, because we first sample the environment θ according to its posterior probability of being the true environment and then we perform the policy that is optimal in that environment. The question is how this problem formulation can be reconciled with a decision-theoretic problem statement that involves utilities. This is the topic of Section "Decision-making with limited resources".

Statistical causality

While both actions and observations are treated as random variables in (3), there is an important difference between actions and observations. Observations are produced by the environment and can be used to update the agent's state of knowledge about the environment. In contrast, actions are set by the agent itself and hence they do not provide information about the environment. This distinction becomes crucial when conditioning on the history of actions and observations. The theory that deals with the distinction between exogenous and endogenous information is *statistical causality* (Pearl 2000; Glymour et al. 2000).

What is a causal intervention?

A typical example for causal intervention is the manipulation of a barometer (Pearl 2000). In a barometer the atmospheric pressure changes the height of the mercury: if it rises, we expect good weather; and if it drops rapidly, we expect rain. A simple Bayesian model captures this relation:

$$P(w|b) = \frac{P(b|w)P(w)}{P(b)},$$

where w and b are the weather and the Barometer variables respectively, $P(w)$ is the prior probability of the weather (e.g. good or bad) and $P(b|w)$ is the likelihood of the barometer change given the weather. The posterior $P(w|b)$ allows us to infer the weather from the barometer reading.

Now, imagine you decide to change the level of the mercury yourself, say (using a bit of imagination) by means of a pressurizing device. Now, *you* set the value of the random variable—and intuition tells us that we cannot predict the weather anymore from the barometer reading. Apparently, our previous Bayesian model is useless now. This shouldn't come as a surprise, as our intervention effectively changed the relation between the barometer and the weather.

Mathematically, we can model the causal relationship between different random variables by a particular factorization of the joint probability distribution into conditional probabilities reflecting generative mechanisms, *where causes are in the conditional and effects in the argument*. In our example the weather causes the barometer to rise and fall and not the other way around, that is in causal terms we have $P(b,w) = P(b|w)P(w)$ and not $P(b,w) = P(w|b)P(b)$, even though in purely probabilistic terms the two factorizations are of course equivalent. The causal factorization becomes important when modifying the causal relationships by intervention, as in our example. Manipulating the barometer setting directly severs the causal between weather and barometer. Consequently, we have to modify our joint probability distribution from $P(b,w)$ to

$$P(\hat{b},w) = \delta(b)P(w),$$

that is, where $P(b|w)$, viewed as a generative mechanism, has been replaced by $\delta(b)$, thereby rendering the two random variables independent. The hat-notation (due to (Pearl 2000)) is just a shorthand referring to this particular transformation of the probability distribution. When we evaluate the posterior under the intervention $\hat{b}$, we get

$$P(w|\hat{b}) = \frac{P(w,\hat{b})}{\sum_w P(w,\hat{b})} = \frac{1 \cdot P(w)}{1} = P(w).$$

In other words, we don't gain knowledge about the weather—as expected. Notice that intervening the alternative factorization, $P(\hat{b},w) = \delta(b)P(w|b)$, would give a different result that is inconsistent with our causal story: we have assumed that the mercury level of the barometer depends functionally on the weather, and not the other way around. The reason for this special treatment of actions is that when we set the value of a random variable ourselves, we change Nature's probability law.

Causal interventions in Thompson sampling

To calculate the effect of an intervention, the causal model has to be known, that is the unique factorization of the joint distribution into conditional probabilities matching the causal dependencies over the random variables. In our case, the causal dependencies are straightforward: first, the environment secretly chooses a true parameter $\theta^* \in \Theta$, and then the interactions $a_1, o_1, a_2, o_2, \ldots$ follow chronologically. This causal model allows us to study the effect of interventions in the problem statement given in (3). We start by re-expressing the posterior $P(\theta|\hat{a}_{<t}, o_{<t})$ as

$$\begin{aligned}
&P(\theta|\hat{a}_{<t}, o_{<t}) \\
&\overset{(1.)}{=} \frac{P(\theta, \hat{a}_{<t}, o_{<t})}{\sum_{\theta'} P(\theta', \hat{a}_{<t}, o_{<t})} \\
&\overset{(2.)}{=} \frac{P(\theta) \prod_{k=1}^{t} P(\hat{a}_k|\theta, \hat{a}_{<k}, o_{<k}) P(o_k|\theta, \hat{a}_{\leq k}, o_{<k})}{\sum_{\theta'} P(\theta') \prod_{k=1}^{t} P(\hat{a}_k|\theta', \hat{a}_{<k}, o_{<k}) P(o_k|\theta', \hat{a}_{\leq k}, o_{<k})} \\
&\overset{(3.)}{=} \frac{P(\theta) \prod_{k=1}^{t} P(\hat{a}_k|\theta, a_{<k}, o_{<k}) P(o_k|\theta, a_{\leq k}, o_{<k})}{\sum_{\theta'} P(\theta') \prod_{k=1}^{t} P(\hat{a}_k|\theta', a_{<k}, o_{<k}) P(o_k|\theta', a_{\leq k}, o_{<k})} \\
&\overset{(4.)}{=} \frac{P(\theta) \prod_{k=1}^{t} P(o_k|\theta, a_{\leq k}, o_{<k})}{\sum_{\theta'} P(\theta') \prod_{k=1}^{t} P(o_k|\theta', a_{\leq k}, o_{<k})},
\end{aligned} \tag{4}$$

where we first expand the probabilities in terms of the joint distribution, second rewrite the joint distribution as the causal factorization, third remove the intervention tags from the intervened random variables that are in the probability conditions (Pearl's second rule of do-calculus (Pearl 2000)), and fourth replace each conditional probability having an intervened variable in the argument by a delta function over its chosen value—compare Chapter 4.2 in (Pearl 2000).

These equations show that beliefs are updated only using past observations, and that past actions provide no further evidence. Intuitively, the reason for this is that the agent can be surprised about his past observations and learn from them, but he cannot be surprised about his own actions chosen by himself in the past. Also, due to Pearl's second law of do-calculus (Pearl 2000) we have

$$P(a_t|\theta, \hat{a}_{<t}, o_{<t}) = P(a_t|\theta, a_{<t}, o_{<t}). \tag{5}$$

Using (3), (4) and (5), we arrive at the desired result, that is the predictive distribution for $a_t \in \mathcal{A}$ given by

$$P(a_t|\hat{a}_{<t}, o_{<t}) = \sum_{\theta} P(a_t|\theta, a_{<t}, o_{<t}) P(\theta|\hat{a}_{<t}, o_{<t}), \tag{6}$$

obtained only by applying probability theory and causal calculus.

Results

Decision-making with limited resources

The difference between the random belief model supposed by Thompson sampling and the standard maximum expected utility model for decision-making is highlighted by contrasting two simple decision scenarios depicted in Figure 1. The goal is to predict the outcome when throwing one of two possible biased coins. A rational decision maker places bets (shown inside speech bubbles) such that his subjective expected utility is maximized. These subjective beliefs are delimited within dotted boxes. A Thompson sampling agent first samples a random belief and then chooses the best prediction with respect to this belief.

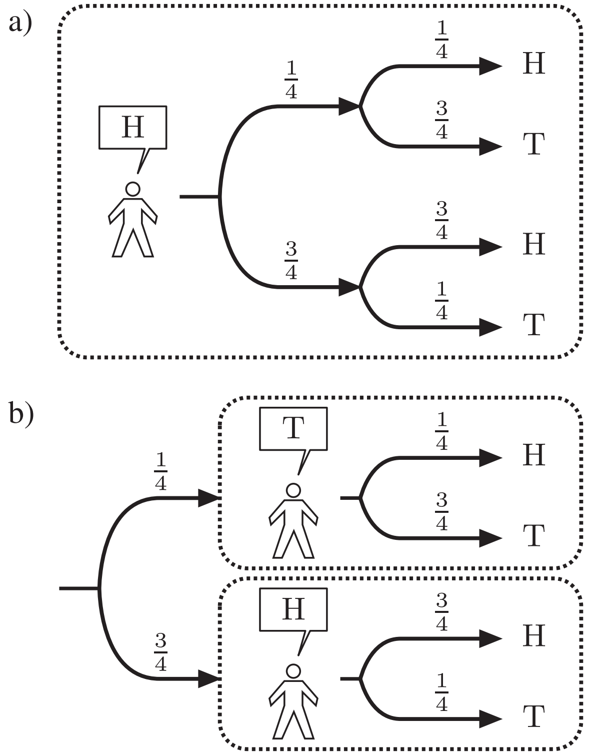

Figure 1 Comparison between (a) a fully rational decision maker and (b) a decision maker with random beliefs. There are two biased coins. The probability of throwing the first coin is $\frac{1}{4}$, the probability of throwing the second coin is $\frac{3}{4}$. The first coin has a probability of $\frac{1}{4}$ for the outcome Head and the second coin has a probability of $\frac{3}{4}$ for Head. The decision-maker has to predict the most likely outcome. A correct guess is rewarded with \$1, an incorrect guess yields \$0.

The difference between the two becomes clear by inspecting the expected utility in each case: they are

$$\max_{P'} \sum_{\theta} P(\theta) \left\{ \sum_{o} P'(a|\theta) P(o|\theta, a) U(o) \right\}, \qquad \text{(a)}$$

$$\text{and} \quad \sum_{\theta} P(\theta) \max_{P'} \left\{ \sum_{o} P'(a|\theta) P(o|\theta, a) U(o) \right\} \qquad \text{(b)}$$

respectively, where the labels (a) and (b) correspond to the labels in Figure 1. Here it is clearly seen that the difference between the two lies in the order in which we apply the expectation (over the environment parameter) and the maximization operator. It should also be noted that the expected utility of (a) is an upper bound on the expected utility of (b). Yet, both cases can constitute optimal decisions depending on constraints. In (a), the decision-maker picks his action taking into account the uncertainty over the bias, while in (b), the decision-maker picks his action only after his beliefs over the coin bias are instantiated—that is, *he is optimal w.r.t. his random beliefs.* Here we consider how this optimality w.r.t. random beliefs can be considered as a form of optimal decision-making under information processing constraints.

Modeling bounded rational decision-making

Here we consider a particular information-theoretic model of bounded rational decision-making that formalizes limited information processing resources by a variational principle that trades off expected utility gains (or losses) and entropic information costs (Ortega 2011a; Ortega and Braun 2011, 2012a, 2013). Information processing costs are usually ignored in the study of perfectly rational decision-makers. Given a choice set $\mathcal{X}$ with choices $x \in \mathcal{X}$ and utilities $U(x)$, a perfectly rational decision-maker would always choose the best option $x^* = \arg\max_x U(x)$—presupposing there is a unique maximum. In general, a bounded rational decision-maker is unable to pick out the best option with certainty, and his choice can be described by a probability distribution $P(x)$ reflecting uncertainty. Improving the choice strategy $P(x)$ can be understood as a costly search process.

Let us assume the initial strategy of the decision-maker can be described by a probability distribution $P_0(x)$. The search process for the optimum transforms this initial choice into a final choice $P(x)$. In case of the perfectly rational decision-maker the final choice is $P(x) = \delta_{x,x^*}$. In the general case of the bounded rational decision-maker the search is costly and he will not be able to afford such a stark reduction in uncertainty. Assuming that search costs are real-valued, additive and higher for rare events (Ortega and Braun 2010c), it can be shown that the cost of the search is determined by the information distance D_{KL} between P_0 and P, that is $D_{KL} = \sum_x P(x) \log \frac{P(x)}{P_0(x)}$. Both Bayesian search (Jaynes 1985) and Koopman's random search (Stone 1998) are compatible with these assumptions, as well as energetic costs that would have to be paid by a Maxwellian demon for reducing uncertainty in statistical physical systems (Ortega and Braun 2013). How this information-theoretic model of search costs relates to computational resources such as space and time complexity is still an open problem (Vitanyi 2005).

Simple decisions

The decision process is modeled as a transformation of a prior choice probability P_0 into a posterior choice probability P by taking into account the utility gains (or losses) and the transformation costs arising from information processing, such that

$$P = \arg\max_{\tilde{P}} \left\{ \sum_x \tilde{P}(x) U(x) - \frac{1}{\alpha} \sum_x \tilde{P}(x) \log \frac{\tilde{P}(x)}{P_0(x)} \right\}, \tag{7}$$

where the $x \in \mathcal{X}$ are the possible outcomes, $P_0(x)$ are their prior probablities, $U(x)$ are their utilities. The *inverse temperature* $\alpha \geq 0$ can be regarded as a rationality parameter that translates the cost of information processing measured in units of information into units of utility. If the limits in information processing capabilities are given as a constraint $D_{KL}(P||P_0) \leq K$ with some positive constant K, then α is determined as a Lagrange multiplier. The maximizing distribution $\tilde{P} = P$ is the *equilibrium distribution*

$$P(x) = \frac{1}{Z_\alpha} P_0(x) e^{\alpha U(x)}, \qquad \text{where} \qquad Z_\alpha = \sum_x P_0(x) e^{\alpha U(x)}, \tag{8}$$

and represents the choice probabilities after deliberation—see Theorem 1.1.3 (Keller 1998) and (Ortega and Braun 2013) for a proof. The *value* V of the choice set $\mathcal{X}$ under choice probabilities P can be determined from the same variational principle

$$V[P] = \max_{\tilde{P}} \left\{ \sum_x \tilde{P}(x)U(x) - \frac{1}{\alpha} \sum_x \tilde{P}(x) \log \frac{\tilde{P}(x)}{P_0(x)} \right\}$$
$$= \frac{1}{\alpha} \log \left(\sum_x P_0(x) e^{\alpha U(x)} \right) = \frac{1}{\alpha} \log Z_\alpha. \qquad (9)$$

For the two different limits of α, the value and the equilibrium distribution take the asymptotic forms

$$\begin{array}{llll} \alpha \to +\infty & \frac{1}{\alpha}\log Z_\alpha = \max_x U(x) & P(x) = \mathcal{U}_{\max}(x) & \text{(perfectly rational)} \\ \alpha \to 0 & \frac{1}{\alpha}\log Z_\alpha = \sum_x P_0(x)U(x) & P(x) = P_0(x) & \text{(irrational)} \end{array}$$

where $\mathcal{U}_{\max}$ is the uniform distribution over the maximizing subset $\mathcal{X}_{\max} := \{x \in \mathcal{X} : U(x) = \max_{x'} U(x')\}$. It can be seen that a perfectly rational agent with $\alpha \to \infty$ is able to pick out the optimal action—which is a deterministic policy in the case of a single optimum—, whereas finitely rational agents have stochastic policies with non-zero probability of picking a sub-optimal action.

The model of bounded rational decision-making also lends itself to an interpretation in terms of sampling complexity. If we use a rejection sampling scheme to obtain samples from $p(x)$ by first sampling from $p_0(x)$, we can ask how many samples we will need on average from p_0 to obtain one sample from p. In this scheme, we produce a sample $x \sim p_0(x)$ and then decide whether to accept or reject the sample based on the criterion

$$u \leq \frac{e^{\alpha U(x)}}{e^{\alpha T}}, \qquad (10)$$

where u is drawn from the uniform $\mathcal{U}[0;1]$ and T is the acceptance target value with $T \geq \max_x U(x)$. The equality holds for the most efficient sampler, but requires knowledge of the maximum. With this sampling scheme, the accepted samples will be distributed according to Equation (8). The average number of samples needed from p_0 to produce one sample of p is then

$$\overline{\sharp Samples} = \frac{1}{\sum_x p_0(x) \frac{e^{\alpha U(x)}}{e^{\alpha T}}} = \frac{e^{\alpha T}}{Z_\alpha}. \qquad (11)$$

The important point about Equation (11) is that the average number of samples increases with increasing rationality parameter α. In fact, the average number of samples will grow exponentially for large α when $T > \max_x U(x)$, as

$$\frac{e^{\alpha T}}{Z_\alpha} \xrightarrow{\alpha \to \infty} \frac{e^{\alpha(T - U(x^*))}}{P_0(x^*)},$$

where $x^* = \arg\max U(x)$. It can also be straightforwardly seen that

$$\overline{\sharp Samples} \geq e^{D_{KL}(p||p_0)} = \frac{e^{\alpha \sum_x p(x)U(x)}}{Z_\alpha} \qquad (12)$$

because $\sum_x p(x)U(x) \leq T$, that is the exponential of the Kullback-Leibler divergence provides a lower bound on the average number of samples.

Decisions in the presence of latent variables

To model a Thompson sampling agent, we need at least a two-step decision with a variable x that has to be chosen by the agent, and a variable θ that is chosen by the environment. In the example described in Figure 1, the variable x is the agent's prediction for the outcome of a coin toss, the variable θ indicates nature's choice which one of the two coins is tossed. The agent's prediction can take on the values $x = H$ and $x = T$ corresponding to the outcomes Head and Tail. The variable θ takes on the two values $\theta = \frac{1}{4}$ and $\theta = \frac{3}{4}$ corresponding to the biases of the two coins. The prior probability over θ is $p_0\left(\theta = \frac{1}{4}\right) = \frac{1}{4}$ and $p_0\left(\theta = \frac{3}{4}\right) = \frac{3}{4}$. The expected rewards for all combinations of x and θ are then $U\left(x = H, \theta = \frac{1}{4}\right) = \frac{1}{4}$, $U\left(x = T, \theta = \frac{1}{4}\right) = \frac{3}{4}$, $U\left(x = H, \theta = \frac{3}{4}\right) = \frac{3}{4}$ and $U\left(x = T, \theta = \frac{3}{4}\right) = \frac{1}{4}$.

In the case of two-step decisions, the variational problem can in general be formulated as a nested expression (Ortega and Braun 2011, 2012a, 2013)

$$\arg\max_{\tilde{p}(x,\theta)} \sum_x \tilde{p}(x) \left[U(x) - \frac{1}{\alpha} \log \frac{\tilde{p}(x)}{p_0(x)} + \sum_\theta \tilde{p}(\theta|x) \left[U(x,\theta) - \frac{1}{\beta} \log \frac{\tilde{p}(\theta|x)}{p_0(\theta|x)} \right] \right]. \quad (13)$$

with the two different rationality parameters α and β for the two different variables x and θ. Limited information processing resources with respect to these variables can also be thought of as different degrees of control. For example, if α assumed a large value, the decision-maker could basically hand-pick a particular x, or if θ was determined by a coin toss that the agent cannot influence, we could model this by setting β to zero. The utility can in general depend on both action and observation variables. However, since the action by itself does not yield a reward in our case, we have $U(x) \equiv 0$. Moreover, we see that in our case, nature's probability of flipping either coin does not actually depend on the agent's prediction, so we can replace the conditional probabilities $p(\theta|x)$ by $p(\theta)$. We have then an inner variational problem:

$$\arg\max_{\tilde{p}(\theta)} \sum_\theta \tilde{p}(\theta) \left[-\frac{1}{\beta} \log \frac{\tilde{p}(\theta)}{p_0(\theta)} + U(x,\theta) \right] \quad (14)$$

with the solution

$$p(\theta) = \frac{1}{Z_\beta(x)} p_0(\theta) \exp\left(\beta U(x,\theta)\right) \quad (15)$$

and the normalization constant $Z_\beta(x) = \sum_\theta p_0(\theta) \exp\left(\beta U(x,\theta)\right)$ and an outer variational problem

$$\arg\max_{\tilde{p}(x)} \sum_x \tilde{p}(x) \left[-\frac{1}{\alpha} \log \frac{\tilde{p}(x)}{p_0(x)} + \frac{1}{\beta} \log Z_\beta(x) \right] \quad (16)$$

with the solution

$$p(x) = \frac{1}{Z_{\alpha\beta}} p_0(x) \exp\left(\frac{\alpha}{\beta} \log Z_\beta(x) \right) \quad (17)$$

and the normalization constant $Z_{\alpha\beta} = \sum_x p_0(x) \exp\left(\frac{\alpha}{\beta} \log Z_\beta(x)\right)$. From Equation (17) we can derive both the perfectly rational decision-maker and the Thompson sampling agent. To simplify, we assume in the following that the agent has no prior preference for x, that is $p_0(x) = \mathcal{U}(x)$.

The perfectly rational decision-maker is obtained in the limit $\alpha \to \infty$ and $\beta \to 0$. If we first take the limit $\lim_{\beta\to 0} \frac{1}{\beta} \log Z_\beta(x) = \sum_\theta p_0(\theta) U(x,\theta)$, a decision-maker with rationality α chooses x with probability

$$p(x) = \frac{e^{\alpha \sum_\theta p_0(\theta) U(x,\theta)}}{Z_{\alpha\beta}}. \tag{18}$$

The perfectly rational expected utility maximizer as depicted in Figure 1a is then obtained from Equation (18) by taking the limit $\alpha \to \infty$.

In contrast, the Thompson sampling agent is obtained when $\beta = \alpha$. In this case, the choice probability for x is given by

$$p(x) = \sum_\theta p_{0(\theta)} \frac{e^{\alpha U(x,\theta)}}{Z_\alpha(x)}. \tag{19}$$

The resulting agent is a probabilistic superposition of agents that act optimally for any given θ as depicted in Figure 1b. It can be seen that in Equation (19) and in Equation (18) the order of the expectation operation and the (soft-)maximization operation are reversed.

Again we can interpret this formalism in terms of sampling complexity. Here we should accept a sample $x \sim p_0(x)$ if it fulfils the criterion

$$u \leq \frac{e^{\alpha \frac{1}{\beta} \log Z_\beta(x)}}{e^{\alpha T}} = \left[\frac{Z_\beta(x)}{e^{\beta T}}\right]^{\frac{\alpha}{\beta}}, \tag{20}$$

where $u \sim \mathcal{U}[0;1]$ and $T \geq \max_x \max_\theta U(x,\theta)$. From Equation (11) we know that the ratio $Z_\beta(x)/e^{\beta T}$ is the acceptance probability of $\theta \sim p_0(\theta)$. In order to accept one sample from x, we thus need to accept $\frac{\alpha}{\beta}$ consecutive samples of θ, with acceptance criterion

$$u \leq \frac{e^{\beta U(x,\theta)}}{e^{\beta T}} \tag{21}$$

with $u \sim \mathcal{U}[0;1]$ and T as set above. Since $\alpha \gg \beta$ we can assume $\alpha \approx N\beta$ with $N \in \mathbb{N}$, and we can see easily that the perfectly rational agent will require infinitely many θ samples ($\alpha \to \infty$ and $\beta \to 0$) to obtain one sample of x, whereas the Thompson sampling agent will only require a single sample ($\alpha = \beta$). The Thompson sampling agent is therefore the agent that can solve the optimization problem of Equation (16) for a given α with the least amount of samples. This can also be seen from Equation (18), when doing the Monte Carlo approximation $\sum_\theta p_0(\theta) U(x,\theta) \approx \frac{1}{N} \sum_i U(x,\theta_i)$ by drawing N samples $\theta_i \sim p_0(\theta_i)$. For infinitely many samples, the average approximates the expectation, for a single sample we can rewrite Equation (18) into Equation (19). This sampling procedure also allows estimating the upper and lower bounds of the optimal utility (Tziortziotis et al. 2013). Of course, the Thompson sampling agent will not achieve the same expected utility as the perfectly rational agent. But both agents can be considered optimal under particular information processing constraints.

Causal induction

A generalized Thompson-sampling agent can be thought of as a probabilistic superposition of models θ, where each model θ is characterized by a likelihood model $P(o_t|\theta, a_{\leq t}, o_{<t})$ and a policy model $P(a_t|\theta, a_{<t}, o_{<t})$. In previous applications we assumed that all models θ have the same causal structure, i.e. considering multivariate random variables a_t and o_t, we assumed that the same variables a_t are intervened for all θ and the same causal model is used to predict the consequences of these interventions on the observational variables o_t. However, this need not be the case. In principle, different models θ could represent different causal structures and suggest intervention of different variables. Such a setup can be used for causal induction as illustrated in the following example.

> Imagine we are working on a medical treatment that involves two gene sites X and Y, each of which can be active or inactive. We encode the 'on' and 'off' states of X as $X = x$ and $X = \neg x$ and similarly $Y = y$ and $Y = \neg y$ to denote the 'on' and 'off' states of Y. Assume we are unsure about the causal mechanism between the two variables, that is we are unsure whether the activity of X influences the activity of Y or the other way around. Formally, we are interested in the explanatory power of two competing causal hypotheses: either 'X causes Y' ($\Theta = \theta$) or 'Y causes X' ($\Theta = \neg\theta$). Assume our goal is to have Y in an active state, but that it is much cheaper and easier to manipulate X instead of Y. This leaves us with the following policies. If X causes Y we prefer to manipulate X, because it is cheap and easy. If Y causes X we have no other choice, but to directly manipulate Y. When manipulating either gene, we can be 100% sure that the new state of the gene is set by us, but we only have a 50% chance that the state will be 'on'. Assume not manipulating either variable is not an option, because then both gene sites stay inactive. The question is how should we act if we do not know the causal dependency?

One of the main methods to deal with problems of causal inference is the framework of causal graphical models (Pearl 2000). Given a graph that represents a causal structure, we can intervene this graph and ask questions about the probabilities of the variables in the graph. However, in causal induction we would like to discover the causal structure itself, that is we would like to do inference over a multitude of graphs representing different causal structures (Heckerman et al. 1999). If one would like to represent the problem of causal discovery graphically, the main challenge is that the model Θ is a random variable that controls the causal structure itself. However, as argued in (Ortega 2011), this difficulty can be overcome by using a probability tree to model the causal structure over the random events. Probability trees can encode alternative causal realizations, and in particular alternative causal hypotheses (Shafer 1996). For instance, Figure 2a encodes the probabilities and functional dependencies among the random random variables of the previous problem.

In a probability tree, each (internal) node is a causal mechanism; hence a path from the root node to one of the leaves corresponds to a particular sequential real-

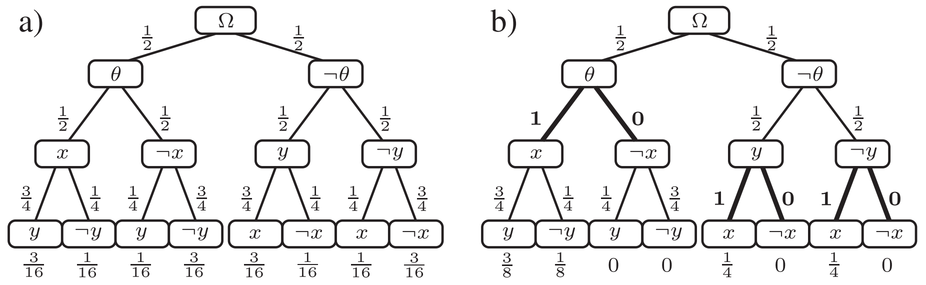

Figure 2 Panel a: An exemplary probability tree to represent the agent's prediction model about its environment. The probability of a realization (written under the leaves) is calculated by multiplying the probabilities starting from the root until a leave is reached. The immediate children of any note resolve the value of a single random variable. Note that the two hypotheses are statistically indistinguishable. Panel **b**: The probability tree resulting from **(a)** after intervening $\hat{x}$.

ization of causal mechanisms. The logic underlying the structure of this tree is as follows:

1. *Causal precedence:* A node causally precedes its descendants. For instance, the root node corresponding to the sure event Ω causally precedes all other nodes.
2. *Resolution of variables:* Each node resolves the value of a random variable. For instance, given the node corresponding to $\Theta = \theta$ and $X = \neg x$, either $Y = y$ will happen with probability $P(y|\theta, \neg x) = \frac{1}{4}$ or $Y = \neg y$ with probability $P(\neg y|\theta, \neg x) = \frac{3}{4}$.
3. *Heterogeneous order:* The resolution order of random variables can vary across different branches. For instance, X precedes Y under $\Theta = \theta$, but Y precedes X under $\Theta = \neg\theta$. This is precisely how we model competing causal hypotheses.

While the probability tree represents the agent's subjective model explaining the order in which the random values are resolved, it does not necessarily correspond to the temporal order in which the events are revealed to us. So for instance, under hypothesis $\Theta = \theta$, the value of the variable Y might be revealed before X, even though X causally precedes Y; and the causal hypothesis Θ, which precedes both X and Y, is never observed.

Consider a Thompson sampling agent that uses the beliefs outlined in Figure 2 that runs a single experiment. The agent does so by first manipulating X and observing Y:

1. *Manipulating X:* First, the agent instantiates his random beliefs by sampling the value of Θ from the prior:

$$P(\Theta = \theta) = P(\Theta = \neg\theta) = \frac{1}{2}.$$

Assume that the result is θ. Treating θ as if it was the true parameter, he proceeds to sample the action from $P(X|\theta)$ given by

$$P(X = x|\theta) = P(X = \neg x|\theta) = \frac{1}{2},$$

as indicated in the left branch of the probability tree. Assume that outcome is x, and this is the action that the agent executes. Because of this, the agent has to update its beliefs first by intervening the probability tree for $\hat{x}$ and second by conditioning on x. The intervention $\hat{x}$ is carried out by replacing all the nodes in the tree that resolve the value of X with new nodes assigning probability one to x and zero to $\neg x$. Figure 2b illustrates the result of this intervention. The posterior is then given by

$$P(\theta|\hat{x}) = \frac{P(\hat{x}|\theta)P(\theta)}{P(\hat{x}|\theta)P(\theta) + P(\hat{x}|\neg\theta)P(\neg\theta)} = \frac{1 \cdot \frac{1}{2}}{1 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2}} = \frac{1}{2}.$$

In other words, the agent has switched on X, and has so far learned nothing from this fact.

2. *Observing Y:* Now, the agent observes the activity of Y, and assume that it is active, i.e. $Y = y$. Then, the posterior beliefs of the agent are given as

$$\begin{aligned} P(\theta|\hat{x}, y) &= \frac{P(y|\theta,\hat{x})P(\hat{x}|\theta)P(\theta)}{P(y|\theta,\hat{x})P(\hat{x}|\theta)P(\theta) + P(\hat{x}|\neg\theta, y)P(y|\neg\theta)P(\neg\theta)} \\ &= \frac{\frac{3}{4} \cdot 1 \cdot \frac{1}{2}}{\frac{3}{4} \cdot 1 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} \cdot \frac{1}{2}} = \frac{3}{5}. \end{aligned}$$

Since $P(\theta) < P(\theta|\hat{x}, y)$, the agent has gathered evidence favoring the hypothesis "X causes Y". This was only possible because the intervention introduced a statistical asymmetry among the two hypotheses that did not exist in the beginning. In comparison, if the action is not treated as an intervention, then the posterior is

$$\begin{aligned} P(\theta|x, y) &= \frac{P(y|\theta,x)P(x|\theta)P(\theta)}{P(y|\theta,x)P(x|\theta)P(\theta) + P(x|\neg\theta, y)P(y|\neg\theta)P(\neg\theta)} \\ &= \frac{\frac{3}{4} \cdot \frac{1}{2} \cdot \frac{1}{2}}{\frac{3}{4} \cdot \frac{1}{2} \cdot \frac{1}{2} + \frac{3}{4} \cdot \frac{1}{2} \cdot \frac{1}{2}} = \frac{1}{2} = P(\theta), \end{aligned}$$

that is, the agent doesn't learn anything just from observing. This also highlights the importance of interventions (Box 1966).

Naturally, multiple interventions and observations can be executed in consecution. In this case Thompson sampling is used in each time step to decide which policy model to use, which implies the decision which variables to intervene. Then, after the intervention, all variables are revealed simultaneously at every time step of the inference process. The update of the observational probabilities is done the same way as in the one step case, taking into account which variables were intervened. A simulation of the repeated Thompson sampling process for causal induction of our example system is shown in Figure 3. This very simple example contains the principles of causal induction using Thompson sampling. Of course, more complex causal structures require richer model classes as is customary in Bayesian modeling. But importantly, the essence of causal induction is already contained in our simple illustration.

Discussion

The main contribution of the present paper is to show in how far generalized Thompson sampling can be regarded as an optimal solution method for adaptive decision-making in the presence of information-processing constraints and how this framework can be extended to solve problems of causal induction. We previously proposed Equation (3) as a *Bayesian rule for acting* in (Ortega and Braun 2010a, 2010b) that optimally solves the adaptive coding problem for actions and observations. In practice, it is implemented by

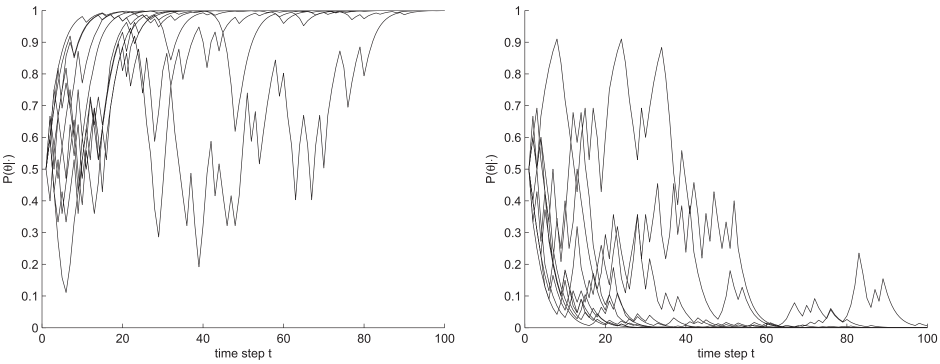

Figure 3 Thompson sampling for causal induction. (Left) Posterior distribution $P(\theta|\cdot)$ for 10 runs when the true system is given by $\Theta = \theta$. (Right) Posterior distribution $P(\theta|\cdot)$ for 10 runs when the true system is given by $\Theta = \neg\theta$. In both cases the agent is able to identify the causal structure of the environment with high confidence when $P(\theta|\cdots)$ is close to one or zero respectively.

sampling an environment parameter $\hat{\theta}_t$ for each time step from the posterior distribution $P(\theta|\hat{a}_{<t}, o_{<t})$, and then treating it as if it was the true parameter—that is, issuing the action a_t from $P(a_t|\hat{\theta}_t, a_{<t}, o_{<t})$. This action-sampling method where beliefs are randomly instantiated was first proposed as a heuristic in (Thompson 1933) and is now known as *Thompson sampling*. Importantly, this method can be generalized and applied to solve general sequential adaptive decision-making problems.

So far Thompson sampling has been mainly applied to multi-armed bandit problems. Multi-armed bandits can be represented by a parameter θ that summarizes the statistical properties of the reward obtained for each lever. Reward distributions range from Bernoulli to Gaussian (with unknown mean and variance), and they can also depend on the particular context or state (Graepel et al. 2010; May and Leslie 2011; Granmo 2010; Scott 2010). In particular, the work of (May and Leslie 2011) and the work of (Granmo 2010) prove asymptotic convergence of Thompson sampling. The performance of bandit algorithms has also been studied in terms of the rate of growth of the regret (Lai and Robbins 1995), and recent bandit algorithms have been shown to match this lower bound (Cappé et al. 2013), including Thompson sampling algorithms for Bernoulli bandits (Kaufmann et al. 2012). Also, the work of (Chapelle and Li 2011) presents empirical results that show Thompson sampling is highly competitive, matching or outperforming popular methods such as UCB (Lai and Robbins 1995; Auer et al. 2002).

Another class of problems, where Thompson sampling has been applied in the past, are Markov decision processes (MDPs). MDPs can be represented by parameterizing the dynamics and reward distribution (model-based) (Strens 2000) or by directly parameterizing the *Q-table* (model-free) (Dearden et al. 1998; Ortega and Braun 2010a). The first approach samples a full description of an MDP, solves it for the optimal policy, and then issues the optimal action. This is repeated in each time step. The second approach avoids the computational overhead of solving for the optimal policy in each time step by directly doing inference on the Q-tables. Actions are chosen by picking the one having the highest Q-value for the current state. The same ideas can also be applied to solve adaptive control problems with linear system equations, quadratic cost functions and Gaussian noise (Braun and Ortega 2010).

Optimality

While maximum expected utility is formally appealing as a principle for the construction of adaptive agents, its strict application is in practice often problematic. This is mainly due to two reasons:

1. *Computational complexity.* The computations required to find the optimal solution (for instance, the computational complexity of solving the Bellman optimality equations) are prohibitive in general and scale exponentially with the length of the horizon. The problem is tractable only in very special cases under assumptions that reduce the effective size of the problem.
2. *Causal precedence of policy choice.* The choice of the policy has to be made before the interaction with the environment starts. That is, an agent has to have a unique optimal policy before it has even interacted once with the environment. An optimal policy constructed by the maximum expected utility principle is therefore a very risky bet, as a lot of resources have to be spent before any evidence exists that the underlying model or prior is adequate.

Because of these two reasons, it is practically often impossible to apply the maximum expected utility principle. This has led to the development of theories of bounded rational decision-making that take the information processing limitations of decision-makers into account. The modern study of bounded rationality was famously broached by Simon (1956, 1972, 1984) and has since been extensively investigated in psychology (Gigerenzer and Selten 2001; Camerer 2003), cognitive science (Howes et al. 2009; Janssen et al. 2011; Lewis et al.), economics (Aumann 1997; Rubinstein 1998; Kahneman 2003), game theory (McKelvey and Palfrey 1995, 1998; Wolpert 2004), political science (Jones 2003), industrial organization (Spiegler 2011), computer science and artificial intelligence research (Lipman 1995; Russell 1995; Russell and Subramanian 1995). Different conceptions of bounded rationality are divided as to whether bounded rational behavior is thought to be fundamentally non-optimizing or whether it can be expressed as a (constrained) optimization problem and as to whether it involves any kind of meta-reasoning (Klein 2001). While the variational formulation in the free energy can also be thought of as a constrained optimization problem, this optimization is only implicit in an agent that runs an anytime algorithm to obtain samples that directly optimize the original (unconstrained) utility function. The average number of samples that can be afforded is determined by an inverse temperature parameter, such that the search for the optimum is aborted after some time, thereby generating some kind of *satisficing* solution. The free energy formulation of bounded rationality also allows reinterpreting a wider research program that has investigated relative entropy as a particular cost function for control (Kappen 2005; Todorov 2006, 2009; Theodorou et al. 2010; Peters et al. 2010; Braun and Ortega 2011; Kappen et al. 2012) and has inspired the formulation of optimal control problems as inference problems (Tishby and Polani 2011; Kappen et al. 2012; Rawlik et al. 2012). In Section "Decision-making with limited resources" we have argued that Thompson sampling can be regarded as an instantiation of free energy optimizing bounded rationality requiring the minimal amount of samples of the latent variable θ in the decision-making process determining the next action. An agent that follows such a Thompson sampling strategy randomly samples beliefs θ and acts optimally with respect to these random beliefs. In contrast, a perfectly rational agent optimizes his utility over the entire belief tree.

Policy Uncertainty. Given a problem specification in terms of the predictive model and the utility function, we can think about policy uncertainty in terms of policy search methods. The task of a policy search method is to calculate a policy that approximates the optimal policy. More specifically, let π be a parameter in a set Π indexing the set of candidate policies $P(a_t|\pi, a_{1:t-1}, o_{1:t-1})$ indexed by $\theta \in \Theta$. *Then, in the most general case, a policy search method returns a probability distribution* $P(\pi)$ *over* Π *representing the uncertainty over the optimal policy parameters.* If the algorithm solves the maximum expected utility problem, then the support of this distribution will exclusively cover the set of optimal policies $\Pi^* \subset \Pi$. Otherwise there remains uncertainty over the optimal policy parameters. However, many policy search methods do not explicitly deal with the uncertainty over the policy parameters. Some methods only return a point estimate $\hat{\pi} \in \Pi$. For instance, reinforcement learning algorithms (Sutton and Barto 1998) start from a randomly initialized point estimate $\hat{\pi}_0$ of the optimal policy and then generate refined point estimates $\hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_3, \ldots$ in each time step $t = 1, 2, 3, \ldots$ using the data provided by experience. In order to converge to the optimal policy, these algorithms have to deal with the exploration-exploitation trade-off. This means that the agents cannot just greedily act according to these point estimates; instead, they have to produce explorative actions as well, that is, actions that deviate from the current estimate of the optimal policy—for instance producing optimistic actions based on UCB (Lai and Robbins 1995; Auer et al. 2002).

Crucially, when sampling actions from the predictive distribution, the policy index π is identical to the index θ that identifies a particular environment with the likelihood model $P(o_t|a_{1:t-1}, o_{1:t-1})$. By turning the reinforcement learning problem thus into an inference problem, the exploration-exploitation trade-off becomes a *bias-variance trade-off* (Geman et al. 1992) in policy space. This highlights the essence of the exploration-exploitation trade-off: any action issued by the agent has to respect the uncertainty over the policy parameter—otherwise they are biased. In particular, if the agent acts deterministically and greedily (i.e. it treats the estimate $\hat{\pi}$ as if it were the true policy parameter) then it is overfitting the experience and introducing a bias; likewise, an agent that follows a stochastic policy introduces variance and will not produce the highest possible reward compared to the case when the optimal policy is known. An excessively stochastic agent therefore underfits its experience.

The operational distinction of having policy uncertainty has important algorithmic consequences. When there is policy uncertainty, the belief of the decision-maker is itself a random variable. This means that the very policy is undefined until the random variable is resolved. Hence, the computation of the optimal policy can be delayed and determined dynamically. It is precisely this fact that is (implicitly) exploited in popular reinforcement learning algorithms, and explicitly in the algorithms based on random beliefs. This is in stark contrast to the case when there is no policy uncertainty, where the policy is pre-computed and static. Another example where random beliefs play a crucial role is in *games with incomplete information* (Osborne and Rubinstein 1999). Here, having incomplete information about the other player leads to a infinite hierarchy of meta-reasoning about the other player's strategy. To avoid this difficulty, Harsanyi introduced *Bayesian games* (Harsanyi 1967). In a Bayesian game, incomplete knowledge is modeled by randomly instantiating the player's types, after which they choose their strategies optimally—thus eliminating the need for recurrent reasoning about the other players' strategy. Similarly, a Thompson sampling agent randomly instantiates his belief at every point in time and acts optimally with respect to this belief. An important consequence of this is that agents have uncertainty about their policy.

Adaptive Coding. The adaptive control problem can also be construed as an adaptive coding problem both for actions and observations (Ortega and Braun 2010b, 2012b). The question then is: How can we construct a system P defined by $P(o_t|\hat{a}_{\leq t}, o_{<t})$ and $P(a_t|\hat{a}_{<t}, o_{<t})$ such that its behavior is as close as possible to the custom-made system $P(o_t|\theta, \hat{a}_{\leq t}, o_{<t})$ and $P(a_t|\theta, \hat{a}_{<t}, o_{<t})$ under any realization of Q_θ? Using the Kullback-Leibler divergence as a distance measure, we can formulate a variational problem in **Pr**, where **Pr** defines an input-output system trough a distribution over interaction sequences $a_1 o_1 a_2 o_2 \ldots$, such that

$$P := \arg\min_{\mathbf{Pr}} \left\{ \limsup_{t \to \infty} \sum_\theta P(\theta) \sum_{\tau=1}^{t} \left(D_m^{a_\tau}(\mathbf{Pr}) + D_m^{o_\tau}(\mathbf{Pr}) \right) \right\}$$

with

$$D_\theta^{a_t}(\mathbf{Pr}) = \sum_{\hat{a}_{<t}, o_{<t}} P(\hat{a}_{<t}, o_{<t}|\theta) \sum_{a_t} P(a_t|\theta, \hat{a}_{<t}, o_{<t}) \log \frac{P(a_t|\theta, \hat{a}_{<t}, o_{<t})}{\mathbf{Pr}(a_t|\hat{a}_{<t}, o_{<t})}$$

$$D_\theta^{o_t}(\mathbf{Pr}) = \sum_{a_{\leq t}, o_{<t}} P(\hat{a}_{\leq t}, o_{<t}|\theta) \sum_{o_t} P(o_t|\theta, \hat{a}_{\leq t}, o_{<t}) \log \frac{P(o_t|\theta, \hat{a}_{\leq t}, o_{<t})}{\mathbf{Pr}(o_t|\hat{a}_{\leq t}, o_{<t})}.$$

In the case of observations, this is a well-known variational principle for Bayesian inference, as it describes a predictor that requires, on average, the least amount of extra bits to capture informational surprise stemming from the behavior of the environment. In the case of actions, the same principle can be harnessed to describe resourceful generation of actions in a way that requires random bits with minimum length on average, when trying to match the optimal policy most suitable for the unknown environment (MacKay 2003). When thinking about the adaptive control problem in this way, the aim of the adaptive agent is simply to avoid surprise. The fact that each custom-built policy $P(a_t|\theta, \hat{a}_{<t}, o_{<t})$ can be thought of as maximizing a utility in environment Q_θ is not crucial, as this policy could also be given by a teacher's demonstration in the absence of an explicitly stated utility function. The avoidance of surprise of adaptive systems has recently been discussed in the context of active inference and the free energy principle (Friston 2009, 2010).

Causality

In Section "Causal induction", we could demonstrate that generalized Thompson sampling can also be applied to the problem of causal induction, by designing policy and prediction models with different causal structures. This way generalized Thompson sampling can be used as a general method for causal induction that is Bayesian in nature. It is based on the idea of combining probability trees (Shafer 1996) with interventions (Pearl 2000) for predicting the behavior of a manipulated system with multiple causal hypotheses. Both the interventions and the constraints on the causal hypotheses introduce statistical asymmetries that permit the extraction of causal information. Unlike frameworks that aim to extract causal information from observational data alone (Shimizu et al. 2006; Griffiths and Tenenbaum 2009; Janzing and Schölkopf 2010), the proposed method is designed for agents that interact with their environment and use these interactions to discover causal relationships.

To construct the Bayes-causal solution (3), we needed to treat actions as interventions. This raises the question about why this distinction was not made for deriving classical expected utility solutions. Since,

$$P(a_t|a_{<t}, o_{<t}) = \sum_\theta P(a_t|\theta, a_{<t}, o_{<t}) P(\theta|\hat{a}_{<t}, o_{<t})$$
$$P(o_t|a_{\le t}, o_{<t}) = \sum_\theta P(o_t|\theta, a_{\le t}, o_{<t}) P(\theta|\hat{a}_{\le t}, o_{<t}),$$

determining the conditions boils down to analyzing when the equalities

$$\begin{aligned} P(\theta|a_{<t}, o_{<t}) &= P(\theta|\hat{a}_{<t}, o_{<t}) \\ P(\theta|\hat{a}_{\le t}, o_{<t}) &= P(\theta|a_{\le t}, o_{<t}) \end{aligned}$$

hold. Replacing both sides yields,

$$\frac{P(\theta)\prod_{k=1}^{t} P(a_k|\theta, a_{<k}, o_{<k}) P(o_k|\theta, a_{\le k}, o_{<k})}{\sum_{\theta'} P(\theta')\prod_{k=1}^{t} P(a_k|\theta', a_{<k}, o_{<k}) P(o_k|\theta', a_{\le k}, o_{<k})}$$
$$= \frac{P(\theta)\prod_{k=1}^{t} P(o_k|\theta, a_{\le k}, o_{<k})}{\sum_{\theta'} P(\theta')\prod_{k=1}^{t} P(o_k|\theta', a_{\le k}, o_{<k})}$$

and we conclude that

$$P(a_k|\theta, a_{<k}, o_{<k}) = \delta_{\bar{a}_k}(a_k),$$

i.e. the actions have to be issued deterministically (but possibly history-dependent) from a unique policy. Intuitively speaking, this is because the operations of intervening and conditioning coincide when the random variables are deterministic.

Convergence

There are important cases where random belief approaches can fail. Indeed, it is easy to devise experiments where having policy uncertainty converges exponentially slower (or does not converge at all) than the Bayes adaptive optimal policy. Consider, for example, two k-order Markov chains with only one observable state when applying k times the same action, but we do not know which action it is. For two possible actions and a uniform prior over the two possible environments the distribution over possible worlds stays uniform as long as no reward has been observed. Choosing actions randomly according to this distribution would require 2^k actions to accidentally choose a sequence of the same action of length k. Thus, the Bayes adaptive optimal policy converges in time k, while the agent with policy uncertainty needs exponentially longer. A simple way to remedy this problem is, of course, to sample random beliefs only every k time steps (Strens 2000). But this problem can be exacerbated in non-stationary environments. Take for instance, an increasing MDP with two actions and number of states $k = \lceil 10\sqrt{t} \rceil$, in which the optimal policy converges in 100 steps, while an agent with policy uncertainty would not converge at all in most realizations. Although (Ortega and Braun 2010b) prove asymptotic convergence for general environments fulfilling a restrictive form of ergodicity condition, this condition needs to be weakened for the convergence proof to be applicable to most real problems. But it is clear that a form of ergodicity is required for an agent with policy uncertainty to be able to learn to act optimally. Intuitively, this means that an agent can only learn if the environment has temporally stable statistical properties.

Conclusion

In this paper we have argued that Thompson sampling is a bounded rational strategy in decision-making that can be considered optimal under given information processing constraints. Thompson sampling agents have uncertainty over their policy, which is a natural phenomenon that arises whenever there are not enough computational resources to apply the maximum expected utility principle to single out a unique optimal policy. Having policy uncertainty effectively weakens the two assumptions of the maximum expected utility principle: the optimal policy can be chosen and refined during interactions, and the computational complexity is lower. We have shown that treating this uncertainty in a Bayesian way with actions as random variables that obey causal calculus naturally leads to Thompson sampling and its Bayesian generalization. This generalized Thompson sampling can be straightforwardly applied to the problem of causal induction. Maintaining and updating Bayesian probabilities is an optimally efficient way to deal with uncertainty—be it with respect to the policy or the environment (Ortega and Braun 2010a). As these random-belief approaches can be derived simply from probability theory and causal calculus we suggest that they cannot only be regarded as heuristic approximations to optimal decision-making, but as principled solution methods in their own right.

Endnote

[a] Each custom-built policy $P(a_t|\theta, a_{<t}, o_{<t})$ can be thought to maximize a utility function in its environment θ, but this is not essential—the policy could also just be given by a teacher's demonstration as in imitation learning (Schaal 1999).

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

DAB and PAO conceived the analysis and wrote the paper.

Acknowledgements
This study was supported by the DFG, Emmy Noether grant BR4164/1-1.

Received: 11 November 2013 Accepted: 17 February 2014
Published: 14 Mar 2014

References

Agrawal S, Goyal N: **Analysis of Thompson sampling for the multi-armed bandit problem.** In *JMLR: Workshop and Conference Proceedings vol 23 (2012) 39.1–39.26. 25th Annual Conference on Learning Theory*; 2011.

Asmuth J, Li L, Littman ML, Nouri A, Wingate D: **A Bayesian s+ in reinforcement learning.** In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. UAI '09, Arlington, Virginia, United States: AUAI Press; 2009:19–26.

Auer P, Cesa-Bianchi N, Fisher P: **Finite-time analysis of the multiarmed bandit problem.** *Machine Learning* 2002, **47:**235–256.

Aumann RJ: **Rationality and bounded rationality.** *Games and Econ Behavior* 1997, **21**(1-2):2–14.

Box G: **Use and abuse of regression.** *Technometrics* 1966, **8**(4):625–629.

Braun DA, Ortega PA: **A minimum relative entropy principle for adaptive control in linear quadratic regulators.** In *The 7th Conference on Informatics in Control, Automation and Robotics, Volume 3*; 2010:103–108.

Braun DA, Ortega PA, Theodorou E, Schaal S: **Path integral control and bounded rationality.** In *IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning*; 2011:202–209.

Bubeck S, Liu CY: **A note on the Bayesian regret of Thompson sampling with an arbitrary prior.** 2013. arXiv:1304.5758.

Camerer C: *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton: Princeton University Press; 2003.

Cao F, Ray S: **Bayesian hierarchical reinforcement learning.** In *Neural Information Processing Systems 25 (NIPS)*; 2012.

Cappé O, Garivier A, Maillard OA, Munos R, Stoltz G: **Kullback-Leibler upper confidence bounds for optimal sequential allocation.** *Ann Stat* 2013, **41**(3):1516–1541.

Chapelle O, Li L: **An empirical evaluation of Thompson sampling.** In *NIPS*; 2011:2249–2257.

Dearden R, Friedman N, Russell S: **Bayesian Q-learning.** In *AAAI '98/IAAI '98: Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence*. Menlo Park, CA, US: American Association for Artificial Intelligence; 1998:761–768.

Dimitrakakis C: **Monte-Carlo utility estimates for Bayesian reinforcement learning.** In *IEEE Conference on Decision and Control*; 2013.

Dimitrakakis C, Tziortziotis N: **ABC reinforcement learning.** In *Proceedings of The 30th International Conference on Machine Learning*; 2013:684–692.

Duff M: **Optimal learning: computational procedures for bayes-adaptive markov decision processes.** *PhD thesis* 2002. [Director-Andrew Barto].

Friston K: **The free-energy principle: a rough guide to the brain?** *Trends in Cognitive Science* 2009, **13:**293–301.

Friston K: **The free-energy principle: a unified brain theory?** *Nat Rev Neurosci* 2010, **11:**127–138.

Geman S, Bienenstock E, Doursat R: **Neural networks and the bias/variance dilemma.** *Neural Comput* 1992, **4:**1–58.

Gigerenzer G, Selten R: *Bounded Rationality: The Adaptive Toolbox*. Cambridge, MA: MIT Press; 2001.

Gittins J: **Bandit processes and dynamic allocation indices.** *J R Stat Soc Ser B , Methodological* 1979, **41:**148–177.

Glymour C, Spirtes P, Scheines R: *Causation, Prediction, and Search, 2nd edition*. Cambridge, Massachusetts, USA: MIT Press; 2000.

Graepel T, Quiñonero Candela J, Borchert T, Herbrich R: **Web-scale Bayesian click-through rate prediction for sponsored search advertising in Microsoft's Bing search engine.** In *Proceedings of the Twenty-Seventh International Conference on Machine Learning*; 2010:25–26.

Granmo OC: **A Bayesian learning automaton for solving two-armed bernoulli bandit problems.** In *Proceedings of the 2008 Seventh International Conference on Machine Learning and Applications*: ICMLA '08; 2008:23–30.

Granmo OC: **Solving two-armed Bernoulli bandit problems using a Bayesian learning automaton.** *Int J Intell Comput Cybernetics* 2010, **3**(2):207–234.

Granmo OC, Glimsdal S: **Accelerated Bayesian learning for decentralized two-armed bandit based decision making with applications to the Goore game.** *Applied intelligence* 2013, **38**(4):479–488.

Griffiths TL, Tenenbaum JB: **Theory-based causal induction.** *Psychological Rev* 2009, **116:**661–716.

Harsanyi J: **Games with incomplete information played by "Bayesian" players.** *Management Sci* 1967, **14**(3):159–182.

Heckerman D, Meek C, Cooper G: **A Bayesian approach to causal discovery.** *Computation, causation, and discovery* 1999, **19:**141–166.

Howes A, Lewis RL, Vera A: **Rational adaptation under task and processing constraints: implications for testing theories of cognition and action.** *Psychological Rev* 2009, **116**(4):717–751.

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

Janssen CP, Brumby DP, Dowell J, Chater N, Howes A: **Identifying optimum performance trade-offs using a cognitively bounded rational analysis model of discretionary task interleaving.** *Topics in Cognitive Sci* 2011, **3:**123–139.

Janzing D, Schölkopf B: **Causal inference using the algorithmic Markov condition.** *IEEE Trans Inf Theor* 2010, **56**(10):5168–5194.

Jaynes E: **Entropy and search theory.** In *Maximum entropy and Bayesian methods in inverse problems*. Heidelberg: Springer-Verlag; 1985.

Jones BD: **Bounded rationality and political science: lessons from public administration and public policy.** *J Public Administration Res Theory* 2003, **13**(4):395–412.

Kahneman D: **Maps of bounded rationality: psychology for behavioral economics.** *Am Econ Rev* 2003, **93**(5):1449–1475.

Kappen H: **A linear theory for control of non-linear stochastic systems.** *Phys Rev Lett* 2005, **95:**200201.

Kappen H, Gómez V, Opper M: **Optimal control as a graphical model inference problem.** *Machine Learn* 2012, **1:**1–11.

Kaufmann E, Korda N, Munos R: **Thompson sampling: an asymptotically optimal finite-time analysis.** In *ALT, Volume 7568 of, Lecture Notes in Computer Science*. Edited by Bshouty NH, Stoltz G, Vayatis N, Zeugmann T. Heidelberg, Germany: Springer; 2012:199–213.

Keller G: *Equilibrium States in Ergodic Theory*. London Mathematical Society Student Texts: Cambridge Univeristy Press; 1998.

Klein G: **The fiction of optimization.** In *Bounded rationality: The adaptive toolbox*. Edited by Gigerenzer G, Selten R. Cambridge, Massachusetts, USA: MIT Press; 2001.

Korda N, Kaufmann E, Munos R: **Thompson sampling for 1-dimensional exponential family bandits.** In *Advances in Neural Information Processing Systems*; 2013:1448–1456.

Lai T, Robbins H: **Asymptotically efficient adaptive allocation rules.** *Adv Appl Math* 1995, **6:**4–22.

Legg S: **Machine super intelligence.** *PhD thesis,* Department of Informatics, University of Lugano 2008.

Lewis R, Howes A, Singh S: **Computational rationality: linking mechanism and behavior through bounded utility maximization.** *Topics in Cognitive Science* 2014, (in press).

Lipman B: **Information processing and bounded rationality: a survey.** *Canadian J Econ* 1995, **28:**42–67.

MacKay D: *Information Theory, Inference, and Learning Algorithms*. Cambridge, UK: Cambridge University Press; 2003.

Martin J: *Bayesian Decision Problems and Markov Chains*: Publications in Operations Research, Wiley; 1967.

May B, Leslie D: **Simulation studies in optimistic Bayesian sampling in contextual-bandit problems.** In *Technical Report 11:02*. Statistics Group, Department of Mathematics, Bristol, UK: University of Bristol; 2011.

May BC, Korda N, Lee A, Leslie DS: **Optimistic Bayesian sampling in contextual-bandit problems.** *J Mach Learn Res* 2012, **98888:**2069–2106.

Mckelvey R, Palfrey TR: **Quantal response equilibria for extensive form games.** *Experimental Econ* 1998, **1:**9–41.

McKelvey RD, Palfrey TR: **Quantal response equilibria for normal form games.** *Games and Econ Behavior* 1995, **10:**6–38.

Mellor J, Shapiro J: **Thompson sampling in switching environments with Bayesian online change point detection.** 2013. arXiv:1302.3721.

Ortega PA: **A unified framework for resource-bounded autonomous agents interacting with unknown environments.** *PhD thesis,* Department of Engineering, University of Cambridge, UK 2011a.

Ortega PA: **Bayesian causal induction.** In *NIPS Workshop on Philosophy and Machine Learning,* Granada*,* 2011.

Ortega PA, Braun DA: **A Bayesian rule for adaptive control based on causal interventions.** In *Proceedings of the third conference on general artificial intelligence*. Paris, France: Atlantis Press; 2010a.

Ortega PA, Braun DA: **A minimum relative entropy principle for learning and acting.** *J Artif Intell Res* 2010b, **38:**475–511.

Ortega PA, Braun DA: **A conversion between utility and information.** In *Proceedings of the Third Conference on Artificial General Intelligence*. Paris, France: Atlantis Press; 2010c:115–120.

Ortega PA, Braun DA: **Information, utility and bounded rationality.** In *Lecture notes on artificial intelligence, Volume 6830*. Heidelberg, Germany: Springer-Verlag; 2011:269–274.

Ortega PA, Braun DA: **Free energy and the generalized optimality equations for sequential decision making.** In *European Workshop for Reinforcement Learning*. Edinburgh, UK; 2012a.

Ortega PA, Braun DA: **Adaptive coding of actions and observations.** *NIPS Workshop on Information in Perception and Action* 2012b.

Ortega PA, Braun DA: **Thermodynamics as a theory of decision-making with information-processing costs.** *Proc R Soc A: Mathematical, Physical and Engineering Science* 2013, **469:**2153.

Osband I, Russo D, Roy BV: **(More) efficient reinforcement learning via posterior sampling.** In *Advances in Neural Information Processing Systems*; 2013:3003–3011.

Osborne MJ, Rubinstein A: *A Course in Game Theory*. Cambridge, Massachusetts, USA: MIT Press; 1999.

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

Peters J, Mülling K, Altun Y: **Relative entropy policy search.** In *AAAI*; 2010.

Rawlik K, Toussaint M, Vijayakumar S: **On stochastic optimal control and reinforcement learning by approximate inference.** In *Proceedings of Robotics: Science and Systems*. Sydney, Australia; 2012.

Rubinstein A: *Modeling Bounded Rationality*. Cambridge, Massachusetts, USA: MIT Press; 1998.

Russell S: **Rationality and Intelligence.** In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*. Edited by Mellish C. Englewood Cliffs, New Jersey, USA: Prentice-Hall; 1995:950–957.

Russell S, Norvig P: *Artificial Intelligence: A Modern Approach, 1st edition*. Prentice-Hall: Englewood Cliffs, NJ; 1995.

Russell S, Subramanian D: **Provably bounded-optimal agents.** *J Artif Intell Res* 1995, **3:**575–609.

Russo D, Roy BV: **Learning to optimize via posterior sampling.** 2013. arXiv:abs/1301.2609.

Schaal S: **Is imitation learning the route to humanoid robots?** *Trends in cognitive sciences* 1999, **3**(6):233–242.

Scott S: **A modern Bayesian look at the multi-armed bandit.** *Applied Stochastic Models in Business and Industry* 2010, **26:**639–658.

Shafer G: *The Art of Causal Conjecture*. Cambridge, Massachusetts, USA: MIT Press; 1996.

Shimizu S, Hoyer PO, Hyvärinen A, Kerminen A: **A Linear Non-Gaussian Acyclic Model for Causal Discovery.** *J Mach Learn Res* 2006, **7:**2003–2030.

Simon HA: **Rational choice and the structure of the environment.** *Psychological Rev* 1956, **63**(2):129–138.

Simon HA: **Theories of bounded rationality.** In *Decision and Organization*. Edited by McGuire CB, Radner R. Amsterdam: North-Holland Publishing; 1972:161–176.

Simon H A: *Models of Bounded Rationality. Cambridge*. Cambridge, Massachusetts, USA: MIT Press; 1984.

Spiegler R: *Bounded Rationality and Industrial Organization*. Oxford: Oxford University Press; 2011.

Stone L: *Theory of Optimal Search*. New York: Academic Press; 1998.

Strens M: **A Bayesian framework for reinforcement learning.** In *Proceedings of the Seventeenth International Conference on Machine Learning*; 2000.

Sutton R, Barto A: *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press; 1998.

Theodorou E, Buchli J, Schaal S: **A generalized path integral approach to reinforcement learning.** *J Mach Learn Res* 2010, **11:**3137–3181.

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

Tishby N, Polani D: **Information theory of decisions and actions.** In *Perception-reason-action cycle: Models, algorithms and systems*. Edited by Vassilis T Hussain. Heidelberg: Springer-Verlag; 2011:601–636.

Todorov E: **Linearly solvable Markov decision problems.** In *Advances in Neural Information Processing Systems, Volume 19*; 2006:1369–1376.

Todorov E: **Efficient computation of optimal actions.** *Proceedings of the National Academy of Sciences USA* 2009, **106:**11478–11483.

Tziortziotis N, Dimitrakakis C, Blekas K: **Cover tree Bayesian reinforcement learning.** 2013a. arXiv: 1305.1809.

Tziortziotis N, Dimitrakakis C, Blekas K: **Linear Bayesian reinforcement learning.** In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*: AAAI Press; 2013:1721–1728.

Vitanyi P: **Time, space, and energy in reversible computing.** In *Proceedings of the 2nd ACM conference on Computing frontiers*; 2005:435–444.

Wolpert DH: **Information theory - the bridge connecting bounded rational game theory and statistical physics.** In *Complex Engineering Systems*. New York, USA: Perseus Books; 2004.

Wyatt J: **Exploration and inference in learning from reinforcement.** *PhD thesis,* Department of Artificial Intelligence, University of Edinburgh 1997.

10.1186/2194-3206-2-2

Cite this article as: Ortega and Braun: Generalized Thompson sampling for sequential decision-making and causal inference. *Complex Adaptive Systems Modeling* 2014, 2:2