VB free learner

*J. Daunizeau1,2*

*1* Brain and Spine Institute, Paris, France

*2* Wellcome Trust Centre for Neuroimaging, University College of London, United Kingdom

Address for Correspondence:

Jean Daunizeau

Motivation, Brain and Behaviour

Brain and Spine Institute

47, bvd de l'Hopital, Paris, France

Email: jean.daunizeau@gmail.com

**Abstract**

**Notation and acronyms**

We listed below acronyms and mathematical notations that may not be familiar to the reader.

BDT Bayesian Decision Theory

VB Variational Bayes

COTP Cue-Outcome Transition Probability

CRP Chinese Restaurant Process

 ‘is defined as’

or  the expectation of the random vector .

 the variance-covariance matrix of the random vector .

 ‘the random vector  follows a Gaussian distribution with mean  and covariance ’

1. **Introduction**
2. **Theory**
   1. **Description of the task**

In the following, we will assume the task can be framed as a variant the so-called multi-armed bandit problem. In brief, participants are asked to bet on the outcome of an idealized experiment (the ‘bandit’), which is influenced by the subject’s actions.

A bandit is modelled as a random number generator, with specific properties we detail below. First, we assume that bandits can deliver different outcomes , which may be associated with different subjective rewards. We assume the latter is captured by the utility of the outcome, i.e. , where is the so-called utility function.

Second, a set of cues (e.g., the bandit label) are accessible prior to the outcome, and we assume they are sampled from a set of  symbols. Third, the agent can choose in between  admissible actions  (c.f. ‘multi-armed bandit’). Last, we assume that the contingencies between the cues and the outcomes are dependent upon the chosen action at each trial. In other terms, the conditional distribution  fully determines the bandit behaviour. Critically, and though it may be experimentally controlled, this conditional distribution is unknown to the agent. This means that the agent has to learn the above cue-outcome transition probabilities (COTP) in order to act optimally, i.e. maximize its expected reward.

A key idea here is that the COTP *may vary* over time or trials, such that, for example, cues that were associated with winning now get associated with loosing. This is what we refer to as an ‘unstable’ environment. Thus the agent has to quickly revise what it may have learned at any point in time in order to be able to adapt to these changes. From a Bayesian perspective, this means that the agent must have priors about the structure of the environment (i.e., the bandit) that allow flexible learning and adaptive behaviour. We will detail such Bayesian agents below.

* 1. **Perceptual and response models**

The first step is to define the candidate response models that we wish to consider. In what follows, we will assume the agent summarizes the COTP by its first two moments, i.e. mean and variance. This is what we call the perceptual model, i.e. the prior assumptions, under which the observer assimilate new information. We will see that this yields a Bayesian belief update that is very close in form to the classical Rescorla-Wagner (delta) learning rule.

A complete BDT response model is fully specified by a perceptual model and an action emission law. We will see an example of this below.

* + 1. ***The perceptual model***

In the following, we will represent the cues and actions with binary vectors, whose dimension matches the number of possibilities, and with all-zero entries except for the relevant dimension, e.g.:  actually means "third type of cue". In contradistinction, the outcome  is continuous. We assume that that the observer summarizes the COTP using its two first moments. This allows us to represent the outcome’s likelihood as a mixture of Gaussian distributions, as follows:

 (1)

where  (resp. ) is a hidden state that captures the first- (resp. -log- second-) order moment of , given the cue and the chosen action.

Specifying priors  effectively shapes how subjects learn to predict outcomes from previous expositions to cue-outcome-actions triplets , and thus completes the definition of the perceptual model.

The key idea here is that the hidden states  (and hence the COTP) are constrained to vary (smoothly) across time or trials. Without loss of generality, we will assume  behaves as a random walk, with two variants:

* *D0*: Subjects assumed *a priori* that the hidden states  varies smoothly over time, according to a first-order Markov process. This is modelled as a random walk with a Gaussian transition density:

 (8)

Here,  are variance hyperparameter that represent the static roughness of changes in hidden states.

* *D1*: Similarly to the perceptual model above, subjects assumed *a priori* that the hidden states  varied smoothly over time, according to a first-order Markov process:

 (9)

Here, the roughness is now varying in time, and controlled by other hidden states , which we refer to as volatility states, and are assumed to also vary according to a similar first-order Markov process:

 (10)

Here, the variance hyperparameters  represent the roughness of changes in volatility, which we have assumed to be the same across cues and actions.

Once can see that it would be trivial to extend the hierarchy up to an arbitrary number of levels, where each level’s smoothness is controlled by the level above, but we do not consider these extensions in the remainder of this manuscript. This concludes the description of the *dynamic* perceptual models, whose complete generative models derives from the product of the likelihood  in Equation 7 and the Markov priors in Equation 8 or 9-10, depending on the model variant. Again, the parameter  that controls the prior belief will be estimated empirically, given the series of cues , outcomes  and actions  of each subject.

* + 1. ***Variational Bayesian recognition of the perceptual models***

Given the perceptual models described above, we can now specify the recognition process in terms of their variational Bayesian (VB) inversion. In brief, subjects update their belief on-line, using successive triplets  to optimise , the sufficient statistics of the posterior density on the trial . It suffices to say that the VB recognition process invariably derives from the minimization of the statistical surprise conveyed by outcomes at each trial.

* + - 1. Model *D0*.

At trial , the free energy of model *D0* writes:

 (19)

where the prior predictive density  derives from marginalizing out the transition density (Equation 8) with respect to their approximate posterior distribution.

We first characterize the variational energy of  and the ensuing VB update:

 (20)

From Equation 20, one can see that the density  is separable across  and  indices (no cross-talk between cue- and action-dependent estimates). Equation 20 is similar in form to a Rescorla-Wagner learning rule, where  is the prediction error and  is the learning rate (which is controlled by )

Let us now focus on the variational energy of  and its two first derivatives:



(21)

One then needs to derive the expectations under both approximate marginal posterior densities::

 (22)

This completes the VB belief update rules for model *D0*.

* + - 1. Model *D1*.

To be continued…

***Appendix 1: VB recognition: deriving the learning rules from the perceptual models***

Within a variational Bayesian framework, negative surprise is measured (or, more precisely, lower-bounded) via the so-called perceptual free-energy , where the expectation is taken under the approximate posterior density (representation)  and  denotes the Shannon entropy. Now variational Bayesian update equations derive from noting that:

 (A1)

where  is the joint pdf and the expectation is taken under all other approximate posterior densities . Whenever the generative model relies upon prior densities that are conjugate to the likelihood function, iteratively applying equation A1 above simply yields the update rules for the approximate posterior sufficient statistics. Otherwise, the so-called Laplace approximation simply performs a parametric approximation of this density, whereby the first-order moment is placed at the mode and the second-order moment is the inverse curvature of the variational energy  at the mode. We can even further simplify this by truncating a Taylor expansion of  to second order at the last representation yields the actual update rule:

 (A2)

where all derivatives are evaluated at the last expectation ,  and  are the first- and second- order moments of the representation of the approximate posterior density . Equation A2 is actually the first iteration of a full Laplace approximation (see Mathys et al., 2011). We will see that  plays the role of a learning rate, and the gradient , that of a prediction error.