# Reverse engineering neural networks to characterise their cost functions

Takuya Isomura[1], Karl Friston[2]

1 Laboratory for Neural Computation and Adaptation, RIKEN Center for Brain Science, Wako, Saitama 351-0198, Japan

2 Wellcome Centre for Human Neuroimaging, Institute of Neurology, University College London, 12 Queen Square, London, WC1N 3AR, UK

Corresponding author email: takuya.isomura@riken.jp

**Abstract**

This work considers a class of biologically plausible cost functions for neural networks, where the same cost function is minimised by both neural activity and plasticity. We show that such cost functions can be cast as a variational bound on model evidence under an implicit generative model. Using generative models based on Markov decision processes (MDP), we show, analytically, that neural activity and plasticity perform Bayesian inference and learning, respectively, by maximising model evidence. Using mathematical and numerical analyses, we then confirm that biologically plausible cost functions—used in neural networks—correspond to variational free energy under some prior beliefs about the prevalence of latent states that generate inputs. These prior beliefs are determined by particular constants (i.e., thresholds) that define the cost function. This means that the Bayes optimal encoding of latent or hidden states is achieved when, and only when, the network's implicit priors match the process that generates the inputs. Our results suggest that when a neural network minimises its cost function, it is implicitly minimising variational free energy under optimal or sub-optimal prior beliefs. This insight is potentially important because it suggests that any free parameter of a neural network's cost function can itself be optimised—by minimisation with respect to variational free energy.

Keywords: free-energy principle, variational Bayesian inference, learning algorithm, synaptic plasticity, Markov decision process, blind source separation

## 1. Introduction

Cost functions are ubiquitous in scientific fields that entail optimisation—including physics, chemistry, biology, engineering, and machine learning. Furthermore, any optimisation problem that can be specified using a cost function can be formulated as a gradient descent. In the neurosciences, this enables one to treat neuronal dynamics and plasticity as an optimisation process (Marr, 1969; Albus, 1971; Schultz et al., 1997; Sutton & Barto, 1998; Linsker, 1988; Brown et al., 2001). These examples highlight the importance of specifying a problem in terms of cost functions, from which neural and synaptic dynamics can be derived. In other words, cost functions provide a formal (i.e., normative) expression of the purpose of a neural network and prescribe the dynamics of that neural network. Crucially, once the cost function has been established and an initial condition has been selected, it is no longer necessary to solve the dynamics. Instead, one can characterise the neural network's behaviour in terms of fixed points, basin of attraction, and structural stability—based on and only on the cost function. In short, it is important to identify the cost function to understand the dynamics, plasticity, and function of a neural network.

A ubiquitous cost function in neurobiology, theoretical biology, and machine learning is model evidence, or equivalently, marginal likelihood or surprise; namely, the probability of some inputs or data under a model of how those inputs were generated by unknown or hidden causes. Generally, the evaluation of surprise is intractable. However, this evaluation can be converted into an optimisation problem by inducing a variational bound on surprise. In machine learning, this is known as an evidence lower bound (ELBO), while the same quantity is known as variational free energy in statistical physics and theoretical neurobiology.

Variational free energy minimisation is a candidate principle that governs neuronal activity and synaptic plasticity (Friston et al., 2006; Friston, 2010). Here, surprise reflects the improbability of sensory inputs given a model of how those inputs were caused. In turn, minimising variational free energy, as a proxy for surprise, corresponds to inferring the (unobservable) causes of (observable) consequences. To the extent that biological systems minimise variational free energy, it is possible to say that they infer and learn the hidden states and parameters that generate their sensory inputs (Helmholtz, 1925; Knill & Pouget, 2004; DiCarlo et al., 2012) and consequently predict those inputs (Rao & Ballard, 1999; Friston, 2005). This is generally referred to as perceptual inference based upon an internal generative model about the external world (Dayan et al., 1995; George & Hawkins, 2009; Bastos et al., 2012).

Variational free energy minimisation provides a unified mathematical formulation of these inference and learning processes in terms of self-organising neural networks that function as Bayes optimal encoders. Moreover, organisms can use the same cost function to control their surrounding environment by sampling predicted (i.e., preferred) inputs. This is known as active inference (Friston et al., 2011). The ensuing free-energy principle suggests that active inference and learning are mediated by changes in neural activity, synaptic strengths, and the behaviour of an organism to minimise variational free energy, as a proxy for surprise. Crucially, variational free energy and model evidence rest upon a generative model of continuous or discrete hidden states. A number of recent studies have used Markov decision process (MDP) generative models to elaborate schemes that minimise variational free energy (Friston, FitzGerald et al., 2016; Friston, FitzGerald et al., 2017; Friston, Parr et al., 2017). This minimisation reproduces various interesting dynamics and behaviours of real neuronal networks and biological organisms. However, it remains to be established whether variational free energy minimisation is an apt explanation for any given neural network, as opposed to the optimisation of alternative cost functions.

In principle, any neural network that produces an output or a decision can be cast as performing some form of inference, in terms of Bayesian decision theory. On this reading, the complete class theorem suggests that any neural network can be regarded as performing Bayesian inference under some prior beliefs; therefore, it can be regarded as minimising variational free energy. The complete class theorem (Wald, 1947; Brown, 1981) states that for any pair of decisions and cost functions, there are some prior beliefs (implicit in the generative model) that render the decisions Bayes optimal. This suggests that it should be theoretically possible to identify an implicit generative model within any neural network architecture, which renders its cost function a variational free energy or ELBO. In what follows, we show that such identification is possible for a fairly canonical form of a neural network and a generic form of a generative model.

In brief, we adopt a reverse engineering approach to identify a plausible cost function for neural networks—and show that the resulting cost function is formally equivalent to variational free energy. Here, we define a cost function as a function of sensory input, neural activity, and synaptic strengths and suppose that neural activity and synaptic plasticity follows a gradient descent on the cost function. For simplicity, we consider single-layer feed-forward neural networks comprising firing rate neuron models and focus on blind source separation (BSS); namely, the problem of separating sensory inputs into multiple hidden sources or causes (Belouchrani et al., 1997; Cichocki et al., 2009; Comon & Jutten, 2010), which provides the minimum setup for modelling causal inference. Previously, we observed BSS performed by *in vitro* neural networks (Isomura et al., 2015) and reproduced this self-supervised process using an MDP and variational free energy minimisation (Isomura & Friston, 2018). These works suggest that variational free energy minimisation offers a plausible account of the empirical behaviour of *in vitro* networks.

In this work, we ask whether variational free energy minimisation can account for the normative behaviour of a canonical neural network that minimises its cost function, by considering all possible cost functions, within a generic class. Using mathematical analysis, we identify a class of cost functions—from which update rules for both neural activity and synaptic plasticity can be derived—when a single-layer feed-forward neural network comprises firing rate neurons whose firing intensity is determined by the sigmoid activation function. The gradient descent on the ensuing cost function naturally leads to Hebbian plasticity with an activity-dependent homeostatic term. We show that these cost functions are formally homologous to variational free energy under an MDP. Finally, we discuss the implications of this result for explaining the empirical behaviour of neuronal networks, in terms of free energy minimisation under particular prior beliefs.

## 2. Methods

In this section, we first derive the form of a variational free energy cost function under a specific generative model; namely a Markov decision process[1]. We will go through the derivations carefully, with a focus on the form of the ensuing Bayesian belief updating. The

[1] Strictly speaking, the generative model used in this paper is a hidden Markov model (HMM) because we do not consider probabilistic transitions between hidden states that depend upon control variables. However, for consistency with the literature on variational treatments of discrete state space models, we retain the MDP formalism; noting that we are using a special case (with unstructured state transitions).

form of this update will re-emerge later, when reverse engineering the cost functions implicit in neural networks. This section starts with a description of Markov decision processes—as a general kind of generative model—and then considers the minimisation of variational free energy under these models.

***2.1 Generative models.*** Under an MDP model (Fig. 1A), a minimal BSS setup (in a discrete-space) reduces to the likelihood mapping from $N_s$ hidden sources or states $s_t \equiv \left(s_t^{(1)}, \dots, s_t^{(N_s)}\right)^T$ to $N_o$ observations $o_t \equiv \left(o_t^{(1)}, \dots, o_t^{(N_o)}\right)^T$. Each source and observation takes a value of one (ON state) or zero (OFF state) at each time step; i.e., $s_t^{(j)}, o_t^{(i)} \in \{1,0\}$. Throughout this paper, *j* denotes the *j*-th hidden state, while *i* denotes the *i*-th observation. The probability of $s_t^{(j)}$ follows a categorical distribution $P\left(s_t^{(j)}\right) = \mathrm{Cat}\left(D^{(j)}\right)$, where $D^{(j)} \equiv \left(D_1^{(j)}, D_0^{(j)}\right) \in \mathbb{R}^2$ with $D_1^{(j)} + D_0^{(j)} = 1$.

The probability of an outcome is determined by the likelihood mapping from all hidden states to each kind of observation in terms of a categorical distribution, $P\left(o_t^{(i)} | s_t, A^{(i)}\right) = \mathrm{Cat}\left(A^{(i)}\right)$. Here, each element of the tensor $A^{(i)} \in \mathbb{R}^{2 \times 2^{N_s}}$ parameterises the probability that $P\left(o_t^{(i)} = k | s_t = \vec{l}\right)$, where $k \in \{1,0\}$ are possible observations and $\vec{l} \in \{1,0\}^{N_s}$ encodes a particular combination of hidden states. The prior distribution of each column of $A^{(i)}$, denoted by $A_{\cdot \vec{l}}^{(i)}$, has a Dirichlet distribution $P\left(A_{\cdot \vec{l}}^{(i)}\right) = \mathrm{Dir}\left(a_{\cdot \vec{l}}^{(i)}\right)$ with concentration parameter $a_{\cdot \vec{l}}^{(i)} \in \mathbb{R}^2$. We use Dirichlet distributions, as they are tractable and widely used for random variables that take a continuous value between 0 and 1. Furthermore, learning the likelihood mapping leads to biologically plausible update rules, which have the form of associative or Hebbian plasticity: please see below and (Friston et al., 2016) for details.

We use $\tilde{o} \equiv (o_1, \dots, o_t)$ and $\tilde{s} \equiv (s_1, \dots, s_t)$ to denote sequences of observations and hidden states, respectively. With this notation in place, the generative model (i.e., the joint distribution over outcomes, hidden states, and the parameters of their likelihood mapping) can be expressed as

$$P(\tilde{o}, \tilde{s}, A) = P(A) \prod_{\tau=1}^{t} P(o_\tau | s_\tau, A) P(s_\tau)$$

$$= \prod_{i=1}^{N_o} P\left(A^{(i)}\right) \cdot \prod_{\tau=1}^{t} \left\{ \prod_{i=1}^{N_o} P\left(o_\tau^{(i)} | s_\tau, A^{(i)}\right) \prod_{j=1}^{N_s} P\left(s_\tau^{(j)}\right) \right\}. \quad (1)$$

Throughout this paper, $t$ denotes the current time, while $\tau$ denotes an arbitrary time from the past to the present, $1 \le \tau \le t$.

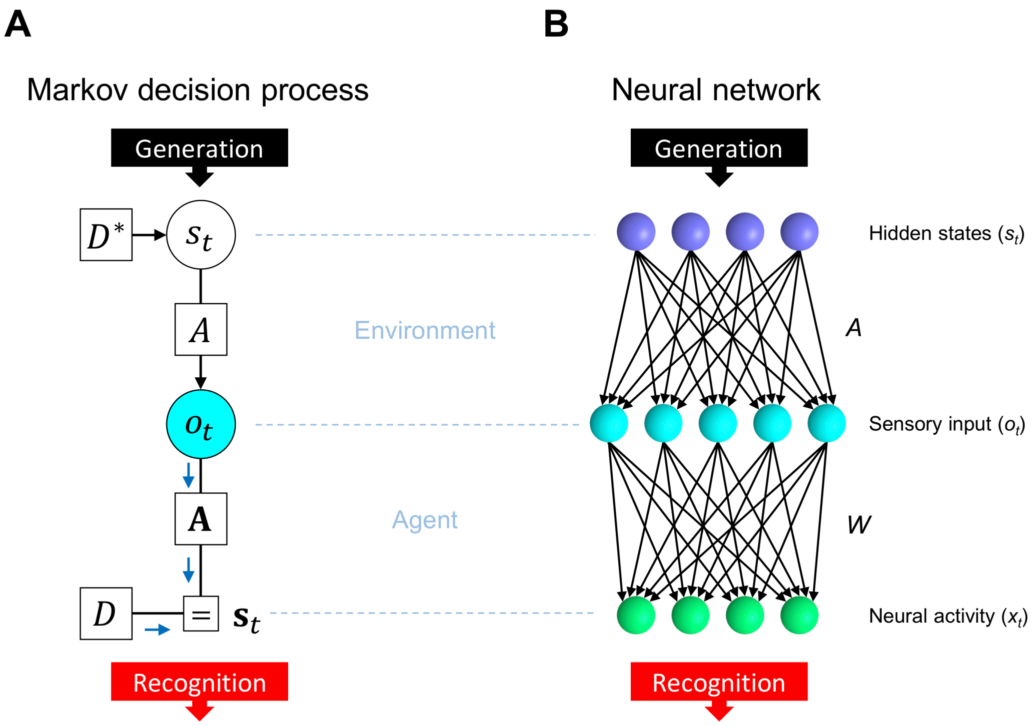


**Figure 1.** Comparison between an MDP scheme and a neural network. (**A**) MDP scheme expressed as a Forney factor graph (Forney, 2001; Dauwels, 2007) based upon the formulation in (Friston, Parr et al., 2017). In this BSS setup, the prior $D$ determines hidden states $s_t$, while $s_t$ determines observation $o_t$ through the likelihood mapping $A$. Inference corresponds to the inversion of this generative process. Here, $D^*$ indicates the true prior while $D$ indicates the prior under which the network operates. If $D = D^*$, the inference is optimal; otherwise, it is biased. (**B**) Neural network comprising a single layer feed-forward network with a sigmoid activation function. The network receives sensory inputs $o_t = \left(o_t^{(1)}, \ldots, o_t^{(N_o)}\right)^T$ that are generated from hidden states $s_t = \left(s_t^{(1)}, \ldots, s_t^{(N_s)}\right)^T$ and outputs neural activities $x_t = \left(x_{t1}, \ldots, x_{tN_x}\right)^T$. Here, $x_{tj}$ should encode the posterior expectation about a binary state $s_t^{(j)}$.

***2.2 Minimisation of variational free energy.*** In this MDP scheme, the aim is to minimise surprise by minimising variational free energy as a proxy; i.e., performing approximate or variational Bayesian inference. From the generative model, we can motivate a mean-field approximation to the posterior (i.e., recognition) density as follows:

$$Q(\tilde{s}, A) = Q(A)Q(\tilde{s}) = \prod_{i=1}^{N_o} Q\left(A^{(i)}\right) \cdot \prod_{\tau=1}^{t} \prod_{j=1}^{N_s} Q\left(s_\tau^{(j)}\right), \tag{2}$$

where $A^{(i)}$ is the likelihood mapping (i.e., tensor), and the marginal posterior distributions of $s_\tau^{(j)}$ and $A^{(i)}$ have a categorical $Q\left(s_\tau^{(j)}\right) = \mathrm{Cat}\left(\mathbf{s}_\tau^{(j)}\right)$ and Dirichlet distribution $Q\left(A^{(i)}\right) = \mathrm{Dir}\left(\mathbf{a}^{(i)}\right)$, respectively. For simplicity, we assume that $A^{(i)}$ factorises into the product of the likelihood mappings from the *j*-th hidden state to the *i*-th observation: $A_{k\cdot}^{(i)} \approx A_{k\cdot}^{(i,1)} \otimes \cdots \otimes A_{k\cdot}^{(i,N_s)}$ (where $\otimes$ denotes the outer product and $A^{(i,j)} \in \mathbb{R}^{2\times 2}$). This (mean field) approximation simplifies the computation of the state posteriors.

In what follows, a bold case variable indicates the posterior expectation of the corresponding variable in italics. For example, $s_\tau^{(j)}$ takes the value 0 or 1, while the posterior expectation $\mathbf{s}_\tau^{(j)} \in \mathbb{R}^2$ is the expected value of $s_\tau^{(j)}$ that lies between 0 and 1. Moreover, $\mathbf{a}^{(i,j)} \in \mathbb{R}^{2\times 2}$ denotes positive concentration parameters. Below, we use the posterior expectation of $\ln A^{(i,j)}$ to encode posterior beliefs about the likelihood, which are given by

$$\begin{aligned} \ln \mathbf{A}^{(i,j)} &\equiv \mathrm{E}_{Q\left(A^{(i,j)}\right)}\left[\ln A^{(i,j)}\right] = \psi\left(\mathbf{a}_{\cdot l}^{(i,j)}\right) - \psi\left(\mathbf{a}_{1l}^{(i,j)} + \mathbf{a}_{0l}^{(i,j)}\right) \\ &= \ln \mathbf{a}_{\cdot l}^{(i,j)} - \ln\left(\mathbf{a}_{1l}^{(i,j)} + \mathbf{a}_{0l}^{(i,j)}\right) + \mathcal{O}\left(\left(\mathbf{a}_{\cdot l}^{(i,j)}\right)^{-1}\right), \end{aligned} \tag{3}$$

where $l \in \{1,0\}$. Here, $\psi(\cdot) \equiv \Gamma'(\cdot)/\Gamma(\cdot)$ denotes the digamma function, which arises naturally from the definition of the Dirichlet distribution. Please see (Friston et al., 2016) for details. $\mathrm{E}_{Q\left(A^{(i,j)}\right)}[\cdot]$ denotes the expectation over the posterior of $A^{(i,j)}$.

The ensuing variational free energy of this generative model is then given by

$$\begin{aligned} &F\left(\tilde{o}, Q(\tilde{s}), Q(A)\right) \\ &\equiv \sum_{\tau=1}^{t} \left\{\mathrm{E}_{Q(s_\tau)Q(A)}\left[-\ln P(o_\tau | s_\tau, A)\right] + \mathcal{D}_{\mathrm{KL}}\left[Q(s_\tau) || P(s_\tau)\right]\right\} + \mathcal{D}_{\mathrm{KL}}\left[Q(A) || P(A)\right] \end{aligned}$$

$$= \underbrace{\sum_{j=1}^{N_s}\sum_{\tau=1}^{t} \mathbf{s}_{\tau}^{(j)} \cdot \left\{ -\sum_{i=1}^{N_o} \ln \mathbf{A}^{(i,j)} \cdot o_{\tau}^{(i)} + \ln \mathbf{s}_{\tau}^{(j)} - \ln D^{(j)} \right\}}_{\text{accuracy+state complexity}}$$

$$+ \underbrace{\sum_{i=1}^{N_o}\sum_{j=1}^{N_s} \left\{ \left(\mathbf{a}^{(i,j)} - a^{(i,j)}\right) \cdot \ln \mathbf{A}^{(i,j)} - \ln \mathcal{B}\left(\mathbf{a}^{(i,j)}\right) \right\}}_{\text{parameter complexity}}, \tag{4}$$

where $\ln \mathbf{A}^{(i,j)} \cdot o_{\tau}^{(i)}$ denotes the inner product of $\ln \mathbf{A}^{(i,j)}$ and a one-hot encoded vector of $o_{\tau}^{(i)}$, $\mathcal{D}_{\mathrm{KL}}[\cdot \,||\, \cdot]$ is the complexity as scored by the Kullback–Leibler divergence (Kullback & Leibler, 1951), and $\mathcal{B}\left(\mathbf{a}^{(i,j)}\right) \equiv \mathcal{B}\left(\mathbf{a}_{\cdot 1}^{(i,j)}\right) \mathcal{B}\left(\mathbf{a}_{\cdot 0}^{(i,j)}\right)$ with $\mathcal{B}\left(\mathbf{a}_{\cdot l}^{(i,j)}\right) \equiv \Gamma\left(\mathbf{a}_{1l}^{(i,j)}\right) \Gamma\left(\mathbf{a}_{0l}^{(i,j)}\right) / \Gamma\left(\mathbf{a}_{1l}^{(i,j)} + \mathbf{a}_{0l}^{(i,j)}\right)$ is the beta function. The first term in the final equality comprises the accuracy $\left(-\mathbf{s}_{\tau}^{(j)} \cdot \sum_{i=1}^{N_o} \ln \mathbf{A}^{(i,j)} \cdot o_{\tau}^{(i)}\right)$ and (state) complexity $\left(\mathbf{s}_{\tau}^{(j)} \cdot \left(\ln \mathbf{s}_{\tau}^{(j)} - \ln D^{(j)}\right)\right)$. The accuracy term is simply the expected log likelihood of an observation, while complexity scores the divergence between prior and posterior beliefs. In other words, complexity reflects the degree of belief updating or degrees of freedom required to provide an accurate account of observations. Both belief updates to states and parameters incur a complexity cost: the state complexity increases with time *t*, while parameter complexity increases in the order of $\ln t$—and is thus negligible when *t* is large (see Supplementary Methods S1 for details). This means that we can ignore parameter complexity, when the scheme has experienced a sufficient number of outcomes. We will drop the parameter complexity in subsequent sections. In the remainder of this section, we show how the minimisation of variational free energy transforms (i.e., updates) priors into posteriors, when the parameter complexity is evaluated explicitly.

Inference optimises posterior expectations about the hidden states by minimising variational free energy. The optimal posterior expectations are obtained by solving the variation of *F* to give

$$\mathbf{s}_{t}^{(j)} = \sigma\left(\sum_{i=1}^{N_o} \ln \mathbf{A}^{(i,j)} \cdot o_{t}^{(i)} + \ln D^{(j)}\right) = \sigma\left(\ln \mathbf{A}^{(\cdot,j)} \cdot o_t + \ln D^{(j)}\right), \tag{5}$$

where $\sigma(\cdot)$ is the softmax function. As $s_{t}^{(j)}$ is a binary value in this work, the posterior expectation of $s_{t}^{(j)}$ taking a value of one (ON state) can be expressed as

$$\mathbf{s}_{t1}^{(j)} = \frac{\exp\left(\ln \mathbf{A}_{\cdot 1}^{(\cdot,j)} \cdot o_t + \ln D_1^{(j)}\right)}{\exp\left(\ln \mathbf{A}_{\cdot 1}^{(\cdot,j)} \cdot o_t + \ln D_1^{(j)}\right) + \exp\left(\ln \mathbf{A}_{\cdot 0}^{(\cdot,j)} \cdot o_t + \ln D_0^{(j)}\right)}$$

$$= \mathrm{sig}\left(\ln \mathbf{A}_{\cdot 1}^{(\cdot,j)} \cdot o_t - \ln \mathbf{A}_{\cdot 0}^{(\cdot,j)} \cdot o_t + \ln D_1^{(j)} - \ln D_0^{(j)}\right) \tag{6}$$

using the sigmoid function $\mathrm{sig}(z) \equiv 1/(1+\exp(-z))$. Thus, the posterior expectation of $s_t^{(j)}$ taking a value 0 (OFF state) is $\mathbf{s}_{t0}^{(j)} = 1 - \mathbf{s}_{t1}^{(j)}$. Here, $D_1^{(j)}$ and $D_0^{(j)}$ are constants denoting the prior beliefs about hidden states. Bayes optimal encoding is obtained when, and only when, the prior beliefs match the genuine prior distribution; i.e., $D_1^{(j)} = D_0^{(j)} = 0.5$ in this BSS setup. This concludes our treatment of inference about hidden states under this minimal scheme. Note that the updates in Equation (5) have a biological plausibility in the sense that the posterior expectations can be associated with nonnegative sigmoid-shape firing rates (also known as neurometric functions (Tolhurst et al., 1983; Newsome et al., 1989)), while the arguments of the sigmoid (softmax) function can be associated with neuronal depolarisation; rendering the softmax function a voltage-firing rate activation function. Please see (Friston, FitzGerald et al., 2017) for a more comprehensive discussion—and simulations using this kind of variational message passing to reproduce empirical phenomena; such as place fields, mismatch negativity responses, phase-precession, pre-play activity, etc in systems neuroscience.

In terms of learning, by solving the variation of $F$ with respect to $\mathbf{a}^{(i,j)}$, the optimal posterior expectations about the parameters are given by

$$\mathbf{a}^{(i,j)} = a^{(i,j)} + \sum_{\tau=1}^{t} o_\tau^{(i)} \otimes \mathbf{s}_\tau^{(j)} = a^{(i,j)} + t\overline{o_t^{(i)} \otimes \mathbf{s}_t^{(j)}}, \tag{7}$$

where $a^{(i,j)}$ is the prior, $o_\tau^{(i)} \otimes \mathbf{s}_\tau^{(j)}$ expresses the outer product of a one-hot encoded vector of $o_\tau^{(i)}$ and $\mathbf{s}_\tau^{(j)}$, and $\overline{o_t^{(i)} \otimes \mathbf{s}_t^{(j)}} \equiv \frac{1}{t}\sum_{\tau=1}^{t} o_\tau^{(i)} \otimes \mathbf{s}_\tau^{(j)}$. Thus, the optimal posterior expectation of matrix $A$ is

$$\begin{cases} \mathbf{A}_{11}^{(i,j)} = \dfrac{\mathbf{a}_{11}^{(i,j)}}{\mathbf{a}_{11}^{(i,j)} + \mathbf{a}_{01}^{(i,j)}} = \dfrac{t\overline{o_t^{(i)} \mathbf{s}_{t1}^{(j)}} + a_{11}^{(i,j)}}{t\overline{\mathbf{s}_{t1}^{(j)}} + a_{11}^{(i,j)} + a_{01}^{(i,j)}} = \dfrac{\overline{o_t^{(i)} \mathbf{s}_{t1}^{(j)}}}{\overline{\mathbf{s}_{t1}^{(j)}}} + \mathcal{O}\left(\dfrac{1}{t}\right) \\ \mathbf{A}_{10}^{(i,j)} = \dfrac{\mathbf{a}_{10}^{(i,j)}}{\mathbf{a}_{10}^{(i,j)} + \mathbf{a}_{00}^{(i,j)}} = \dfrac{t\overline{o_t^{(i)} \mathbf{s}_{t0}^{(j)}} + a_{10}^{(i,j)}}{t\overline{\mathbf{s}_{t0}^{(j)}} + a_{10}^{(i,j)} + a_{00}^{(i,j)}} = \dfrac{\overline{o_t^{(i)} \mathbf{s}_{t0}^{(j)}}}{\overline{\mathbf{s}_{t0}^{(j)}}} + \mathcal{O}\left(\dfrac{1}{t}\right) \end{cases}, \tag{8}$$

where $\overline{o_t^{(i)}\mathbf{s}_{t1}^{(j)}} = \frac{1}{t}\sum_{\tau=1}^{t} o_\tau^{(i)}\mathbf{s}_{\tau1}^{(j)}$, $\overline{\mathbf{s}_{t1}^{(j)}} = \frac{1}{t}\sum_{\tau=1}^{t} \mathbf{s}_{\tau1}^{(j)}$, $\overline{o_t^{(i)}\mathbf{s}_{t0}^{(j)}} = \frac{1}{t}\sum_{\tau=1}^{t} o_\tau^{(i)}\mathbf{s}_{\tau0}^{(j)}$, and $\overline{\mathbf{s}_{t0}^{(j)}} = \frac{1}{t}\sum_{\tau=1}^{t} \mathbf{s}_{\tau0}^{(j)}$. Further, $\mathbf{A}_{01}^{(i,j)} = 1 - \mathbf{A}_{11}^{(i,j)}$ and $\mathbf{A}_{00}^{(i,j)} = 1 - \mathbf{A}_{10}^{(i,j)}$. The prior of parameters $a^{(i,j)}$ is in the order of 1 and is thus negligible when *t* is large. The matrix $\mathbf{A}^{(i,j)}$ express the optimal posterior expectations of $o_t^{(i)}$ taking the ON state when $s_t^{(j)}$ is ON ($\mathbf{A}_{11}^{(i,j)}$) or OFF ($\mathbf{A}_{10}^{(i,j)}$), or $o_t^{(i)}$ taking the OFF state when $s_t^{(j)}$ is ON ($\mathbf{A}_{01}^{(i,j)}$) or OFF ($\mathbf{A}_{00}^{(i,j)}$). Although this expression may seem complicated, it is fairly straightforward. The posterior expectations of the likelihood simply accumulate posterior expectations about the co-occurrence of states and their outcomes. These accumulated (Dirichlet) parameters are then normalised to give a likelihood or probability. Crucially, one can observe the associative or Hebbian aspect of this belief update, expressed here in terms of the outer products between outcomes and posteriors about states in Equation (7). We now turn to the equivalent update for neural activities and synaptic weights of a neural network.

***2.3 Neural activity and Hebbian plasticity models.*** Next, we consider the neural activity and synaptic plasticity in the neural network (Fig. 1B). We assume that the *j*-th neuron's activity $x_{tj}$ is given by

$$\dot{x}_{tj} \propto \underbrace{-f'(x_{tj})}_{\text{leakage}} + \underbrace{W_{j1}o_t - W_{j0}o_t}_{\text{synaptic input}} + \underbrace{h_{j1} - h_{j0}}_{\text{threshold}}. \tag{9}$$

We suppose that $W_{j1} \in \mathbb{R}^{N_o}$ and $W_{j0} \in \mathbb{R}^{N_o}$ comprise row vectors of synapses, and $h_{j1} \in \mathbb{R}$ and $h_{j0} \in \mathbb{R}$ are adaptive thresholds that depend on the values of $W_{j1}$ and $W_{j0}$, respectively. One may regard $W_{j1}$ and $W_{j0}$ as excitatory and inhibitory synapses, respectively. We further assume that the nonlinear leakage $f'(\cdot)$ (i.e., the leak current) is the inverse of the sigmoid function (i.e., the logit function), such that the fixed point of $x_{tj}$ is given by

$$\begin{aligned} x_{tj} &= \mathrm{sig}\left(W_{j1}o_t - W_{j0}o_t + h_{j1} - h_{j0}\right) \\ &= \frac{\exp\left(W_{j1}o_t + h_{j1}\right)}{\exp\left(W_{j1}o_t + h_{j1}\right) + \exp\left(W_{j0}o_t + h_{j0}\right)}. \end{aligned} \tag{10}$$

We further assume that synaptic strengths are updated following Hebbian plasticity with an activity-dependent homeostatic term as follows:

$$\begin{cases} \Delta W_{j1}(t) \equiv W_{j1}(t+1) - W_{j1}(t) \propto Hebb_1\left(x_{tj}, o_t, W_{j1}\right) + Home_1\left(x_{tj}, W_{j1}\right) \\ \Delta W_{j0}(t) \equiv W_{j0}(t+1) - W_{j0}(t) \propto Hebb_0\left(x_{tj}, o_t, W_{j0}\right) + Home_0\left(x_{tj}, W_{j0}\right) \end{cases}, \tag{11}$$

where $Hebb_1$ and $Hebb_0$ denote Hebbian plasticity as determined by the product of sensory inputs and neural outputs, and $Home_1$ and $Home_0$ denote homeostatic plasticity determined by output neural activity.

In the MDP scheme, posterior expectations about hidden states and parameters are usually associated with neural activity and synaptic strengths. Here, we can observe a formal similarity between the solutions for the state posterior (Equation (6)) and the activity in the neural network (Equation (10)). By this analogy, $x_{tj}$ can be regarded as encoding the posterior expectation of the ON state $\mathbf{s}_{t1}^{(j)}$. Moreover, $W_{j1}$ and $W_{j0}$ correspond to $\ln \mathbf{A}_{11}^{(\cdot,j)} - \ln\left(\vec{1} - \mathbf{A}_{11}^{(\cdot,j)}\right) = \mathrm{sig}^{-1}\left(\mathbf{A}_{11}^{(\cdot,j)}\right)$ and $\ln \mathbf{A}_{10}^{(\cdot,j)} - \ln\left(\vec{1} - \mathbf{A}_{10}^{(\cdot,j)}\right) = \mathrm{sig}^{-1}\left(\mathbf{A}_{10}^{(\cdot,j)}\right)$, respectively, in the sense that they express the amplitude of $o_t$ influencing $x_{tj}$ or $\mathbf{s}_{t1}^{(j)}$. Here, $\vec{1} = (1, \dots, 1) \in \mathbb{R}^{N_o}$ is a vector of ones. In particular, the optimal posterior of a hidden state taking a value of one (Equation (6)) is given by the ratio of the beliefs about ON and OFF states, expressed as a sigmoid function. Thus, to be a Bayes optimal encoder, the fixed point of neural activity needs to be a sigmoid function. This requirement is straightforwardly ensured when $f'(x_{tj})$ is the inverse of the sigmoid function (see Equation (13) below). Under this condition, the fixed point or solution for $x_{tk}$ (Equation (10)) compares inputs from ON and OFF pathways, and thus $x_{tj}$ straightforwardly encodes the posterior of the *j*-th hidden state being ON (i.e., $x_{tj} \to \mathbf{s}_{t1}^{(j)}$). In short, the above neural network is effectively inferring the hidden state.

If the activity of the neural network is performing inference, does the Hebbian plasticity correspond to Bayes optimal learning? In other words, does the synaptic update rule in Equation (11) ensure that the neural activity and synaptic strengths asymptotically encode Bayes optimal posterior beliefs about hidden states ($x_{tj} \to \mathbf{s}_{t1}^{(j)}$) and parameters ($W_{j1} \to \mathrm{sig}^{-1}\left(\mathbf{A}_{11}^{(\cdot,j)}\right)$), respectively? To this end, below we will identify a class of cost functions from which the neural activity and synaptic plasticity can be derived, and consider the conditions under which the cost function becomes consistent with variational free energy.

***2.4 Neural network cost functions.*** Here, we consider a class of functions that constitute a cost function for both neural activity and synaptic plasticity. We start by assuming that the update of the *j*-th neuron's activity (Equation (9)) is determined by the gradient of cost function $L_j$; i.e., $\dot{x}_{tj} \propto -\partial L_j / \partial x_{tj}$. By integrating the right-hand side of Equation (9), we obtain a class of cost functions as

$$L_j = \sum_{\tau=1}^{t} \left( f(x_{\tau j}) - x_{\tau j} W_{j1} o_\tau - (1 - x_{\tau j}) W_{j0} o_\tau - x_{\tau j} h_{j1} - (1 - x_{\tau j}) h_{j0} \right) + \mathcal{O}(1)$$

$$= \sum_{\tau=1}^{t} \left( f(x_{\tau j}) - \begin{pmatrix} x_{\tau j} \\ 1 - x_{\tau j} \end{pmatrix}^T \left( \begin{pmatrix} W_{j1} \\ W_{j0} \end{pmatrix} o_\tau + \begin{pmatrix} h_{j1} \\ h_{j0} \end{pmatrix} \right) \right) + \mathcal{O}(1), \quad (12)$$

where the $\mathcal{O}(1)$ term, which depends on $W_{j1}$ and $W_{j0}$, is of a lower order than the other terms (as they are $\mathcal{O}(t)$) and is thus negligible when $t$ is large. Please see Supplementary Methods S3 for the case where we explicitly evaluate the $\mathcal{O}(1)$ term, to demonstrate the formal correspondence between the initial values of synaptic strengths and the parameter prior $p(A)$. The cost function of the entire network is defined by $L \equiv \sum_{j=1}^{N_x} L_j$. When $f'(x_{\tau j})$ is the inverse of the sigmoid function, we have

$$f(x_{\tau j}) = x_{\tau j} \ln x_{\tau j} + (1 - x_{\tau j}) \ln(1 - x_{\tau j}) \quad (13)$$

up to a constant term. We further assume that the synaptic weight update rule is derived from the same cost function $L_j$. Thus, the synaptic plasticity is given by

$$\begin{cases} \dot{W}_{j1} \propto -\dfrac{1}{t} \dfrac{\partial L_j}{\partial W_{j1}} = \overline{x_{tj} o_t} + \overline{x_{tj}} h'_{j1} \\ \dot{W}_{j0} \propto -\dfrac{1}{t} \dfrac{\partial L_j}{\partial W_{j0}} = \overline{(1 - x_{tj}) o_t} + \overline{1 - x_{tj}} h'_{j0} \end{cases}, \quad (14)$$

where $\overline{x_{tj} o_t} \equiv \frac{1}{t} \sum_{\tau=1}^{t} x_{\tau j} o_\tau$, $\overline{x_{tj}} \equiv \frac{1}{t} \sum_{\tau=1}^{t} x_{\tau j}$, $\overline{(1 - x_{tj}) o_t} \equiv \frac{1}{t} \sum_{\tau=1}^{t} (1 - x_{\tau j}) o_\tau$, $\overline{1 - x_{tj}} \equiv \frac{1}{t} \sum_{\tau=1}^{t} (1 - x_{\tau j})$, $h'_{j1} \equiv \partial h_{j1} / \partial W_{j1}$, and $h'_{j0} \equiv \partial h_{j0} / \partial W_{j0}$. Note that the update of $W_{j1}$ is not directly influenced by $W_{j0}$, and *vice versa*, because they encode parameters in physically distinct pathways (i.e., the updates are local learning rules (Lee et al., 2000)). The update rule for $W_{j1}$ can be viewed as Hebbian plasticity mediated by an additional activity-dependent term expressing homeostatic plasticity. Moreover, the update of $W_{j0}$ can be viewed as anti-Hebbian plasticity with a homeostatic term, in the sense that $W_{j0}$ is reduced when input ($o_t$) and output ($x_{tj}$) fire together. The fixed points of $W_{j1}$ and $W_{j0}$ are given by

$$\begin{cases} W_{j1} = h_1'^{-1} \left( -\dfrac{\overline{x_{tj} o_t}}{\overline{x_{tj}}} \right) \\ W_{j0} = h_0'^{-1} \left( -\dfrac{\overline{(1 - x_{tj}) o_t}}{\overline{1 - x_{tj}}} \right) \end{cases}. \quad (15)$$

Crucially, these synaptic strength updates are a subclass of the general synaptic plasticity rule in Equation (11); see also Supplementary Methods S2 for the mathematical explanation.

Therefore, if the synaptic update rule is derived from the cost function underlying neural activity, the synaptic update rule has a biologically plausible form comprising Hebbian plasticity and activity-dependent homeostatic plasticity.

***2.5 Comparison with variational free energy.*** Here, we establish a formal relationship between the cost function $L$ and variational free energy. We define $\widehat{W}_{j1} \equiv \operatorname{sig}(W_{j1})$ and $\widehat{W}_{j0} \equiv \operatorname{sig}(W_{j0})$ as the sigmoid functions of synaptic strengths. We consider the case in which neural activity is expressed as a sigmoid function and thus Equation (13) holds. As $W_{j1} = \ln \widehat{W}_{j1} - \ln(\vec{1} - \widehat{W}_{j1})$, Equation (12) becomes

$$L = \sum_{j=1}^{N_x} \sum_{\tau=1}^{t} \begin{pmatrix} x_{\tau j} \\ 1 - x_{\tau j} \end{pmatrix}^T \left\{ \begin{pmatrix} \ln x_{\tau j} \\ \ln(1 - x_{\tau j}) \end{pmatrix} - \begin{pmatrix} \ln \widehat{W}_{j1} & \ln(\vec{1} - \widehat{W}_{j1}) \\ \ln \widehat{W}_{j0} & \ln(\vec{1} - \widehat{W}_{j0}) \end{pmatrix} \begin{pmatrix} o_\tau \\ \vec{1} - o_\tau \end{pmatrix} - \begin{pmatrix} h_{j1} \\ h_{j0} \end{pmatrix} \right.$$
$$\left. + \begin{pmatrix} \ln(\vec{1} - \widehat{W}_{j1}) \\ \ln(\vec{1} - \widehat{W}_{j0}) \end{pmatrix} \vec{1} \right\} + \mathcal{O}(1), \tag{16}$$

where $\vec{1} = (1, \dots, 1) \in \mathbb{R}^{N_o}$. One can immediately see a formal correspondence between this cost function and variational free energy (Equation (4)). That is, when we assume that $x_{tj} = \mathbf{s}_{t1}^{(j)}$, $\widehat{W}_{j1} = \mathbf{A}_{11}^{(\cdot,j)}$, and $\widehat{W}_{j0} = \mathbf{A}_{10}^{(\cdot,j)}$, Equation (16) has exactly the same form as the sum of the accuracy and state complexity, which is the leading order term of variational free energy (see the first term in the last equality of Equation (4)).

Specifically, when the thresholds satisfy $h_{j1} = \ln(\vec{1} - \widehat{W}_{j1}) \cdot \vec{1} + \ln D_1^{(j)}$ and $h_{j0} = \ln(\vec{1} - \widehat{W}_{j0}) \cdot \vec{1} + \ln D_0^{(j)}$, Equation (16) becomes equivalent to Equation (4) up to the $\ln t$ order term (that disappears when $t$ is large). Therefore, in this case, the fixed points of neural activity and synaptic strengths become the posteriors; thus, $x_{tj}$ asymptotically becomes the Bayes optimal encoder for a large $t$ limit (provided with $D$ that matches the genuine prior $D^*$).

In other words, we can define perturbation terms $\phi_{j1} \equiv h_{j1} - \ln(\vec{1} - \widehat{W}_{j1}) \cdot \vec{1}$ and $\phi_{j0} \equiv h_{j0} - \ln(\vec{1} - \widehat{W}_{j0}) \cdot \vec{1}$ as functions of $W_{j1}$ and $W_{j0}$, respectively, and can express the cost function as

$$L = \sum_{j=1}^{N_x} \sum_{\tau=1}^{t} \begin{pmatrix} x_{\tau j} \\ 1 - x_{\tau j} \end{pmatrix}^T \left\{ \begin{pmatrix} \ln x_{\tau j} \\ \ln(1 - x_{\tau j}) \end{pmatrix} - \begin{pmatrix} \ln \widehat{W}_{j1} & \ln(\vec{1} - \widehat{W}_{j1}) \\ \ln \widehat{W}_{j0} & \ln(\vec{1} - \widehat{W}_{j0}) \end{pmatrix} \begin{pmatrix} o_\tau \\ \vec{1} - o_\tau \end{pmatrix} - \begin{pmatrix} \phi_{j1} \\ \phi_{j0} \end{pmatrix} \right\} + \mathcal{O}(1). \quad (17)$$

Here, without loss of generality, we can suppose that the constant terms in $\phi_{j1}$ and $\phi_{j0}$ are chosen to ensure that $\exp(\phi_{j1}) + \exp(\phi_{j0}) = 1$ . Under this condition, $(\exp(\phi_{j1}), \exp(\phi_{j0}))$ can be viewed as the prior belief about hidden states

$$\begin{cases} \phi_{j1} = \ln D_1^{(j)} \\ \phi_{j0} = \ln D_0^{(j)} \end{cases} \quad (18)$$

and thus Equation (17) is formally equivalent to the accuracy and state complexity terms of variational free energy.

This means that when the prior belief about states $(D^{(j)})$ is a function of the parameter posteriors $(\mathbf{A}^{(\cdot,j)})$, the generic cost function under consideration can be expressed in the form of variational free energy, up to the $\mathcal{O}(\ln t)$ term. A generic cost function $L$ is sub-optimal from the perspective of Bayesian inference unless $\phi_{j1}$ and $\phi_{j0}$ are tuned appropriately to express the unbiased (i.e., optimal) prior belief. In this BSS setup, $\phi_{j1} = \phi_{j0} = \text{const}$ is optimal; thus, a generic $L$ would asymptotically give an upper bound of variational free energy with the optimal prior belief about states when $t$ is large.

***2.6 Analysis on synaptic update rules.*** To explicitly solve the fixed points of $W_{j1}$ and $W_{j0}$ that provide the global minimum of $L$, we suppose $\phi_{j1}$ and $\phi_{j0}$ as linear functions of $W_{j1}$ and $W_{j0}$, respectively, given by

$$\begin{cases} \phi_{j1} = \alpha_{j1} + W_{j1}\beta_{j1} \\ \phi_{j0} = \alpha_{j0} + W_{j0}\beta_{j0} \end{cases}, \quad (19)$$

where $\alpha_{j1}, \alpha_{j0} \in \mathbb{R}$ and $\beta_{j1}, \beta_{j0} \in \mathbb{R}^{N_o}$ are constants. By solving the variation of $L$ with respect to $W_{j1}$ and $W_{j0}$, we find the fixed point of synaptic strengths as

$$\begin{cases} W_{j1} = \text{sig}^{-1}\left( \dfrac{\overline{x_{tj} o_t}}{\overline{x_{tj}}} + \beta_{j1} \right) \\ W_{j0} = \text{sig}^{-1}\left( \dfrac{\overline{(1 - x_{tj}) o_t}}{\overline{1 - x_{tj}}} + \beta_{j0} \right) \end{cases}. \quad (20)$$

Since the update from $t$ to $t$+1 is expressed as $\text{sig}(W_{j1} + \Delta W_{j1}) - \text{sig}(W_{j1}) = \widehat{W}_{j1} \odot (\vec{1} - \widehat{W}_{j1}) \odot \Delta W_{j1} + \mathcal{O}\left(|\Delta W_{j1}|^2\right)$ and $\text{sig}(W_{j1} + \Delta W_{j1}) - \text{sig}(W_{j1}) \approx x_{(t+1)j} o_{t+1} / \overline{x_{tj}} - x_{(t+1)j} \overline{x_{tj} o_t} / \overline{x_{tj}}^2 = x_{(t+1)j} o_{t+1} / \overline{x_{tj}} - (\widehat{W}_{j1} - \beta_{j1}) x_{(t+1)j} / \overline{x_{tj}}$, we recover the following synaptic plasticity:

$$\begin{cases} \Delta W_{j1} = \underbrace{\dfrac{\widehat{W}_{j1}^{\odot -1} \odot (1-\widehat{W}_{j1})^{\odot -1}}{\overline{x_{tj}}}}_{\text{adaptive learning rate}} \odot \left\{ \underbrace{x_{(t+1)j} o_{t+1}}_{\text{Hebbian plasticity}} - \underbrace{\left(\widehat{W}_{j1} - \beta_{j1}\right) x_{(t+1)j}}_{\text{homeostatic plasticity}} \right\} \\ \Delta W_{j0} = \underbrace{\dfrac{\widehat{W}_{j0}^{\odot -1} \odot (1-\widehat{W}_{j0})^{\odot -1}}{\overline{1-x_{tj}}}}_{\text{adaptive learning rate}} \odot \left\{ \underbrace{\left(1 - x_{(t+1)j}\right) o_{t+1}}_{\text{anti-Hebbian plasticity}} - \underbrace{\left(\widehat{W}_{j0} - \beta_{j0}\right)\left(1 - x_{(t+1)j}\right)}_{\text{homeostatic plasticity}} \right\} \end{cases}, \qquad (21)$$

where $\odot$ denotes the element-wise (Hadamard) product and $\widehat{W}_{j1}^{\odot -1}$ denotes the element-wise inverse of $\widehat{W}_{j1}$. This synaptic plasticity rule is a subclass of the generic synaptic plasticity rule in Equation (11).

In summary, we demonstrated that under a few minimal assumptions and ignoring small contributions to weight updates, the neural network under consideration can be regarded as minimising an approximation to model evidence, because the cost function can be formulated in terms of variational free energy. In what follows, we will rehearse our analytic results and then use numerical analyses to illustrate Bayes optimal inference (and learning) in a neural network when, and only when, it has the right priors.

## 3. Results

***3.1 Analytical form of neural network cost functions.*** The analysis in the preceding section rests on the following assumptions:

*(1) Updates of neural activity and synaptic weights are determined by a gradient descent on a cost function L.*

*(2) Neural activity is updated by the weighted sum of sensory inputs, and its fixed point is expressed as the sigmoid function.*

Under these assumptions, we can express the cost function for a neural network as follows (see Equation (17)):

$$L = \sum_{j=1}^{N_x} \sum_{\tau=1}^{t} \begin{pmatrix} x_{\tau j} \\ 1 - x_{\tau j} \end{pmatrix}^T \left\{ \begin{pmatrix} \ln x_{\tau j} \\ \ln\left(1 - x_{\tau j}\right) \end{pmatrix} - \begin{pmatrix} \ln \widehat{W}_{j1} & \ln\left(\vec{1} - \widehat{W}_{j1}\right) \\ \ln \widehat{W}_{j0} & \ln\left(\vec{1} - \widehat{W}_{j0}\right) \end{pmatrix} \begin{pmatrix} o_\tau \\ \vec{1} - o_\tau \end{pmatrix} - \begin{pmatrix} \phi_{j1} \\ \phi_{j0} \end{pmatrix} \right\} + \mathcal{O}(1),$$

where $\widehat{W}_{j1} = \text{sig}\left(W_{j1}\right)$ and $\widehat{W}_{j0} = \text{sig}\left(W_{j0}\right)$ hold, and $\phi_{j1}$ and $\phi_{j0}$ are functions of $W_{j1}$ and $W_{j0}$, respectively. The log likelihood function (accuracy term) and divergence of hidden states (complexity term) of variational free energy emerge naturally under the assumption of a sigmoid activation function. The cost function above has additional terms denoted by $\phi_{j1}$ and $\phi_{j0}$. In other words, we can say that the cost function *L* is variational free energy under a sub-optimal prior belief about hidden states, depending on $W_{j1}$ and $W_{j0}$: $\ln P\left(s_t^{(j)}\right) = \ln D^{(j)} = \phi_j$, where $\phi_j \equiv \left(\phi_{j1}, \phi_{j0}\right)$. This prior alters the landscape of the cost function in a sub-optimal manner and thus provides a biased solution for neural activities and synaptic strengths, which differ from the Bayes optimal encoders.

For analytical tractability, we further assume the following:

*(3) The perturbation terms ($\phi_{j1}$ and $\phi_{j0}$) that constitute the difference between the cost function and variational free energy with optimal prior beliefs can be expressed as linear equations of $W_{j1}$ and $W_{j0}$.*

From assumption 3, Equation (17) becomes

$$L = \sum_{j=1}^{N_x} \left[ \sum_{\tau=1}^{t} \begin{pmatrix} x_{\tau j} \\ 1 - x_{\tau j} \end{pmatrix}^T \left\{ \begin{pmatrix} \ln x_{\tau j} \\ \ln\left(1 - x_{\tau j}\right) \end{pmatrix} - \begin{pmatrix} \ln \widehat{W}_{j1} & \ln\left(\vec{1} - \widehat{W}_{j1}\right) \\ \ln \widehat{W}_{j0} & \ln\left(\vec{1} - \widehat{W}_{j0}\right) \end{pmatrix} \begin{pmatrix} o_\tau \\ \vec{1} - o_\tau \end{pmatrix} - \begin{pmatrix} \alpha_{j1} + W_{j1}\beta_{j1} \\ \alpha_{j0} + W_{j0}\beta_{j0} \end{pmatrix} \right\} \right] + \mathcal{O}(1), \tag{22}$$

where $\{\alpha_{j1}, \alpha_{j0}, \beta_{j1}, \beta_{j0}\}$ are constants. The cost function has degrees of freedom with respect to the choice of constants $\{\alpha_{j1}, \alpha_{j0}, \beta_{j1}, \beta_{j0}\}$, which correspond to the prior belief about states $D^{(j)}$. The neural activity and synaptic strengths that give the minimum of a generic physiological cost function $L$ are biased by these constants, which may be analogous to physiological constraints (see Discussion for details).

The cost function of the neural networks considered is characterised only by $\phi_j$. Thus, after fixing $\phi_j$ by fixing constrains $\left(\alpha_{j1}, \alpha_{j0}\right)$ and $\left(\beta_{j1}, \beta_{j0}\right)$, the remaining degrees of freedom are the initial synaptic weights. These correspond to the prior distribution of parameters $P(A)$ in the variational Bayesian formulation (please see Supplementary Methods 3).

The fixed point of synaptic strengths that give the minimum of $L$ is given analytically as Equation (20), expressing that $\left(\beta_{j1}, \beta_{j0}\right)$ deviates the centre of the nonlinear mapping—from Hebbian products to synaptic strengths—from the optimal position (shown in Equation (8)). As shown in Equation (14), the derivative of $L$ with respect to $W_{j1}$ and $W_{j0}$ recovers the synaptic update rules that comprise Hebbian and activity-dependent homeostatic terms. Although Equation (14) expresses the dynamics of synaptic strengths that converge to the fixed point, it is consistent with a plasticity rule that gives the synaptic change from $t$ to $t$+1 (Equation (21)).

Hence, based on assumptions 1 and 2, we find that the cost function approximates variational free energy; see also Supplementary Table S1 for their correspondence. Under this condition, neural activity encodes the posterior expectation about hidden states, $x_{\tau j} =$

$\mathbf{s}_{\tau 1}^{(j)} = Q\left(s_\tau^{(j)} = 1\right)$, and synaptic strengths encode the posterior expectation of the parameters, $\widehat{W}_{j1} = \mathrm{sig}(W_{j1}) = \mathbf{A}_{11}^{(\cdot,j)}$ and $\widehat{W}_{j0} = \mathrm{sig}(W_{j0}) = \mathbf{A}_{10}^{(\cdot,j)}$. In addition, based on assumption 3, the accuracy of approximation depends on the deviation of constants $\{\alpha_{j1}, \alpha_{j0}, \beta_{j1}, \beta_{j0}\}$ from their optimal values. From a Bayesian perspective, these constants can be viewed as prior beliefs, $\ln P\left(s_t^{(j)}\right) = \ln D^{(j)} = \left(\alpha_{j1} + W_{j1}\beta_{j1}, \alpha_{j0} + W_{j0}\beta_{j0}\right)$, when we assume that $\left(x_{tj}, 1 - x_{tj}\right)$ represents the state posterior $\mathbf{s}_t^{(j)}$. When and only when $\left(\alpha_{j1}, \alpha_{j0}\right) = (-\ln 2, -\ln 2)$ and $\left(\beta_{j1}, \beta_{j0}\right) = \left(\vec{0}, \vec{0}\right)$, the cost function becomes variational free energy with optimal prior beliefs (for BSS), whose global minimum ensures Bayes optimal encoding.

In short, we identify a class of biologically plausible cost functions from which the update rules for both neural activity and synaptic plasticity can be derived. When the activation function for neural activity is a sigmoid function, a cost function in this class is expressed straightforwardly as variational free energy. With respect to the choice of constants expressing physiological constraints in the neural network, the cost function has degrees of freedom that may be viewed as (potentially sub-optimal) prior beliefs from the Bayesian perspective. Now, we illustrate the implicit inference and learning in neural networks through simulations of BSS.

***3.2 Numerical simulations.*** Here, we simulated the dynamics of neural activity and synaptic strengths when they followed a gradient descent on the cost function in Equation (22). We considered a BSS comprising two hidden sources (or states) and 32 observations (or sensory inputs), formulated as an MDP. The two hidden sources comprised four patterns: $s_t = s_t^{(1)} \otimes s_t^{(2)} = (0{,}0), (1{,}0), (0{,}1), (1{,}1)$. An observation $o_t^{(i)}$ was generated through the likelihood mapping $A^{(i)}$, defined as

$$\begin{cases} P\left(o_t^{(i)} = 1 | s_t, A^{(i)}\right) = A_{1\cdot}^{(i)} = \left(0, \frac{3}{4}, \frac{1}{4}, 1\right) & \text{for } 1 \le i \le 16 \\ P\left(o_t^{(i)} = 1 | s_t, A^{(i)}\right) = A_{1\cdot}^{(i)} = \left(0, \frac{1}{4}, \frac{3}{4}, 1\right) & \text{for } 17 \le i \le 32 \end{cases}. \tag{23}$$

Here, for example, $A_{10}^{(i)} = 3/4$ for $1 \le i \le 16$ is the probability of $o_t^{(i)}$ taking one when $s_t = (1{,}0)$. The simulations continued over $T = 10^4$ time steps. Notably, this simulation setup is exactly the same experimental setup as that we used for *in vitro* neural networks (Isomura et al., 2015; Isomura, Friston, 2018). We leverage this setup to clarify the relationship among our empirical work, a feed-forward neural network model, and variational Bayesian formulations.

First, as in (Isomura & Friston, 2018), we demonstrated that a network with a cost function with optimised constants ($(\alpha_{j1}, \alpha_{j0}) = (-\ln 2, -\ln 2)$ and $(\beta_{j1}, \beta_{j0}) = (\vec{0}, \vec{0})$) can perform BSS successfully (Fig. 2). The responses of neuron 1 came to recognise source 1 after training, indicating that neuron 1 learnt to encode source 1 (Fig. 2A). Meanwhile, neuron 2 learnt to infer source 2 (Fig. 2B). This demonstrates that minimisation of the cost function, with optimal constants, is equivalent to variational free energy minimisation, and hence is sufficient to emulate BSS. Next, we quantified the dependency of BSS performance on the form of the cost function, by varying the above-mentioned constants (Fig. 3).

We varied $(\alpha_{j1}, \alpha_{j0})$ in a range of $0.05 \le \exp(\alpha_{j1}) \le 0.95$, while maintaining $\exp(\alpha_{j1}) + \exp(\alpha_{j0}) = 1$, and found that changing $(\alpha_{j1}, \alpha_{j0})$ from $(-\ln 2, -\ln 2)$ led to a failure of BSS. Because neuron 1 encodes source 1 with optimal $\alpha$, the correlation between source 1 and the response of neuron 1 is close to one, while the correlation between source 2 and the response of neuron 1 is nearly zero. In the case of sub-optimal $\alpha$, these correlations fall to around 0.5, indicating that the response of neuron 1 encodes a mixture of source 1 and source 2 (Fig. 3A). Moreover, a failure of BSS can be induced when the elements of $\beta$ take values far from zero (Fig. 3B). When the elements of $\beta$ are generated from a zero-mean Gaussian distribution, the accuracy of BSS—measured using the correlation between sources and responses—decreases as the standard deviation increases.

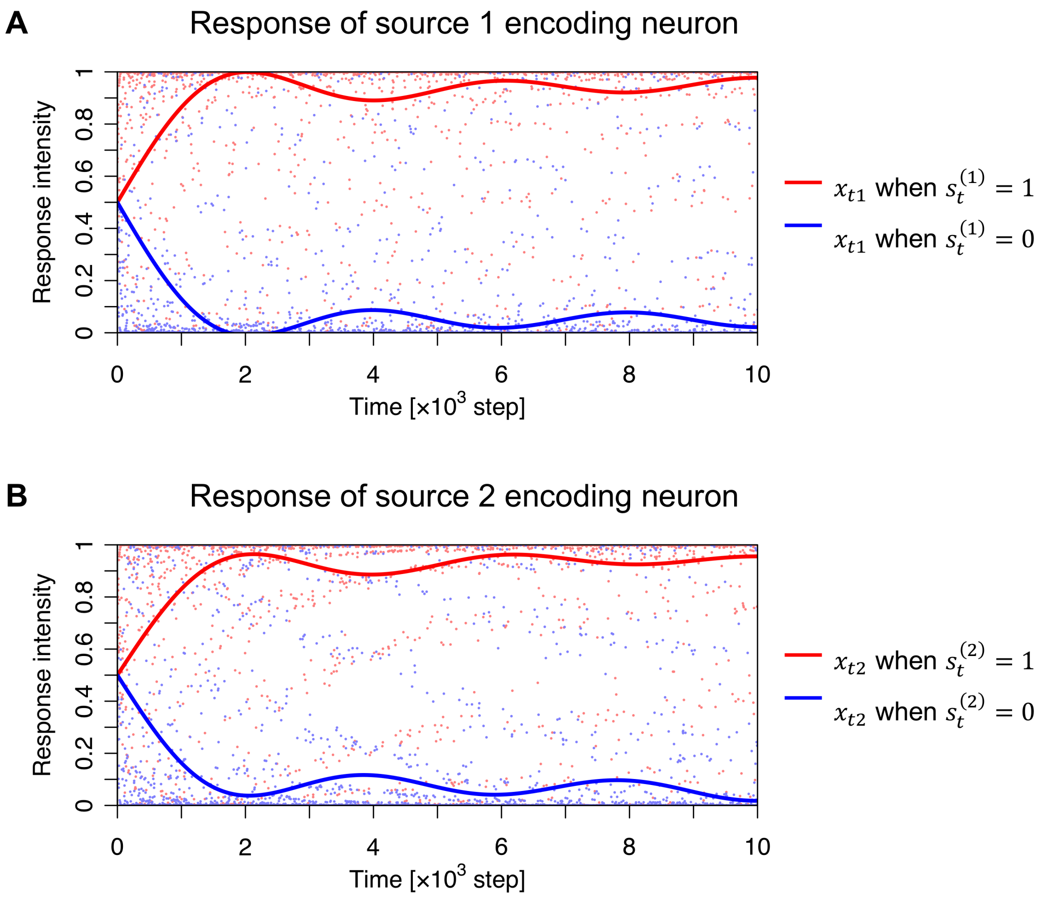


**Figure 2.** Emergence of response selectivity for a source. (**A**) Evolution of neuron 1's responses that learn to encode source 1, in the sense that the response is high when source 1 takes a value of one (red dots), and it is low when source 1 takes a value of zero (blue dots). Lines correspond to smoothed trajectories obtained using a discrete cosine transform. (**B**) Emergence of neuron 2's response that learns to encode source 2. These results indicate that the neural network succeeded in separating two independent sources. The code is provided as Supplementary Source Code.

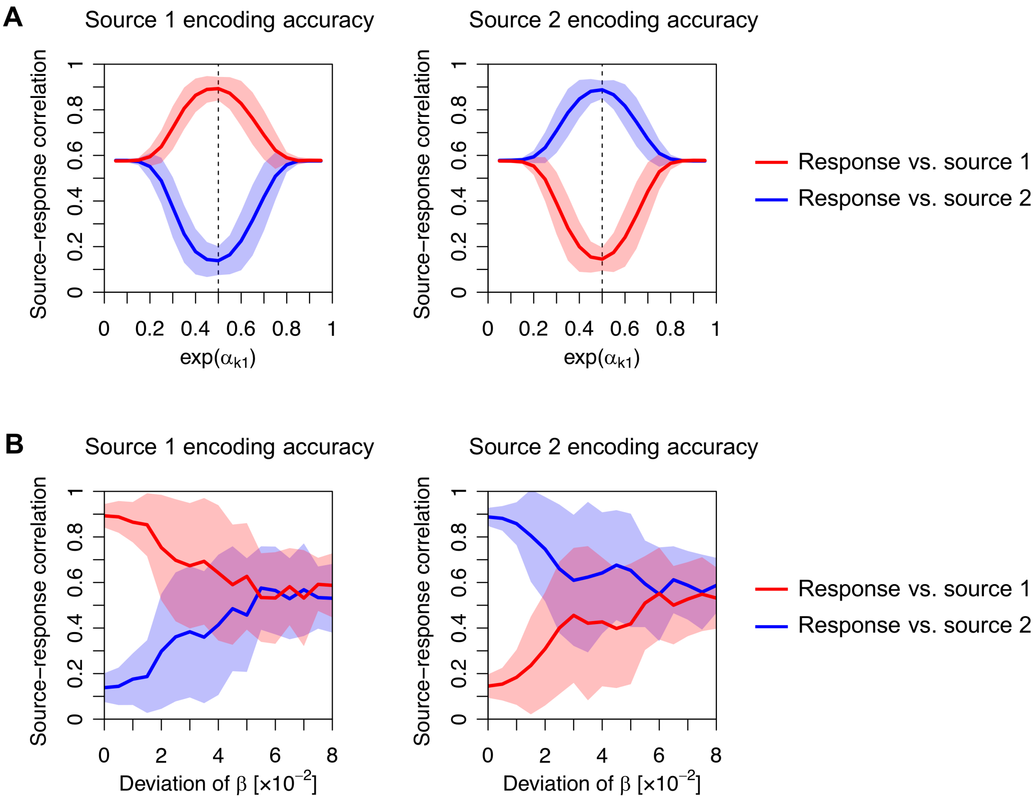


**Figure 3.** Dependence of source encoding accuracy on constants. Left panels show the magnitudes of the correlations between sources and responses of a neuron expected to encode source 1: $\left|\mathrm{corr}\left(s_t^{(1)}, x_{t1}\right)\right|$ and $\left|\mathrm{corr}\left(s_t^{(2)}, x_{t1}\right)\right|$. The right panels show the magnitudes of the correlations between sources and responses of a neuron expected to encode source 2: $\left|\mathrm{corr}\left(s_t^{(1)}, x_{t2}\right)\right|$ and $\left|\mathrm{corr}\left(s_t^{(2)}, x_{t2}\right)\right|$. (**A**) Dependence on the constant $\alpha$ that controls the excitability of a neuron, when $\beta$ is fixed to zero. The dashed line (0.5) indicates the optimal value of $\exp(\alpha_{j1})$. (**B**) Dependence on constant $\beta$, when $\alpha$ is fixed as $(\alpha_{j1}, \alpha_{j0}) = (-\ln 2, -\ln 2)$. Elements of $\beta$ were randomly generated from a Gaussian distribution with zero mean. The standard deviation of $\beta$ was varied (horizontal axis), where zero deviation was optimal. Lines and shaded areas indicate the mean and standard deviation of the source-response correlation, evaluated with 50 different sequences. The code is provided as Supplementary Source Code.

Our numerical analysis, under assumptions 1–3 mentioned above, shows that a network needs to employ a cost function that entails optimal prior beliefs to perform BSS, or equivalently, causal inference. Such a cost function is obtained when its constants, which do not appear in the variational free energy with the optimal generative model for BSS, become negligible. The important message here is that, in this setup, a cost function equivalent to variational free energy is necessary for Bayes optimal inference (Friston et al., 2006; Friston, 2010).

***3.3 Phenotyping networks.*** We have shown that variational free energy (under the MDP scheme) is within the class of biologically plausible cost functions found in neural networks. The neural network's parameters $\phi_j = \ln D^{(j)}$ determine how the synaptic strengths change depending on the history of sensory inputs and neural outputs; thus, the choice of $\phi_j$ provides degrees of freedom in the shape of the generic cost functions under consideration that determine the purpose or function of the neural network. Among various $\phi_j$, only $\phi_j = (-\ln 2, -\ln 2)$ can make the cost function variational free energy with optimal prior beliefs for BSS. Hence, one could regard generic neural networks (of the sort considered in this paper) as performing approximate Bayesian inference under priors that may or may not be optimal. This result is as predicted by the complete class theorem as it implies that any response of a neural network is Bayes optimal under some prior beliefs (and cost function). Therefore, under the theorem, in principle, any neural network of this kind is optimal, when its prior beliefs are consistent with the process that generates outcomes. This perspective indicates the possibility of characterising a neural network model—and indeed a real neuronal network—in terms of its implicit prior beliefs.

These considerations raise the possibility of using empirically observed neuronal responses to infer the prior beliefs implicit in a neuronal network. For example, the synaptic matrix $(W_{j1}, W_{j0})$ can be estimated statistically from response data. By plotting its trajectory over the training period as a function of the history of a Hebbian product, one can estimate the cost function constants. If these constants express a near-optimal $\phi_j$, it can be concluded that the network has, effectively, the right sort of priors for BSS. As we have shown analytically and numerically, a cost function with $(\alpha_{j1}, \alpha_{j0})$ far from $(-\ln 2, -\ln 2)$ or a large deviation of $(\beta_{j1}, \beta_{j0})$ does not provide the Bayes optimal encoder for performing BSS. Since actual neuronal networks can perform BSS (Isomura et al., 2015; Isomura & Friston, 2018), it can be envisaged that the implicit cost function will exhibit a near-optimal $\phi_j$.

One can pursue this analysis further and model the responses or decisions of a neural network using the above-mentioned Bayes optimal MDP scheme under different priors. Thus, the priors in the MDP scheme can be adjusted to maximise the likelihood of empirical responses. This sort of approach has been used in system neuroscience to characterise the choice behaviour in terms of subject specific priors. Please refer to (Schwartenbeck & Friston, 2016) for further details.

Finally, from a practical perspective for optimising neural networks, understanding the formal relationship between cost functions and variational free energy enables us to specify the optimum value of any free parameter to realize some functions. In the present setting, we can effectively optimise the constants by updating the priors themselves, such that they minimise the variational free energy for BSS. Under the Dirichlet form for the priors, the implicit threshold constants of the objective function can then be optimised using the following updates:

$$\phi_j = \ln D^{(j)} = \psi(\mathbf{d}^{(j)}) - \psi\left(\mathbf{d}_1^{(j)} + \mathbf{d}_0^{(j)}\right),$$

$$\mathbf{d}^{(j)} = d^{(j)} + \sum_{\tau=1}^{t} \mathbf{s}_\tau^{(j)} . \quad (24)$$

Please refer to (Schwartenbeck & Friston, 2016) for further details. In effect, this update will simply add the Dirichlet concentration parameters, $\mathbf{d}^{(j)} = \left(\mathbf{d}_1^{(j)}, \mathbf{d}_0^{(j)}\right)$, to the priors in proportion to the temporal summation of the posterior expectations about the hidden states. Therefore, by committing to cost functions that underlie variational inference and learning, any free parameter can be updated in a Bayes optimal fashion when a suitable generative model is available.

## 4. Discussion

In this work, we investigated a class of biologically plausible cost functions for neural networks. A single-layer feed-forward neural network with a sigmoid activation function that receives sensory inputs generated by hidden states (i.e., BSS setup) was considered. We identified a class of cost functions by assuming that neural activity and synaptic plasticity minimise a common function $L$. The derivative of $L$ with respect to synaptic strengths furnishes a synaptic update rule following Hebbian plasticity, equipped with activity-dependent homeostatic term. We have shown that the dynamics of a single-layer feed-forward neural network—that minimises its cost function—is asymptotically equivalent to that of variational Bayesian inference under a particular but generic (latent variable) generative model. Hence, the cost function of the neural network can be viewed as variational free energy, and biological constraints that characterise the neural network—in the form of thresholds and neuronal excitability—become prior beliefs about hidden states. This relationship holds regardless of the true generative process of the external world. We have focused on discrete latent variable models that can be regarded as special (reduced) cases of partially observable Markov decision processes (POMDP). However, because our treatment is predicated on the complete class theorem (Brown, 1981; Wald, 1947), the same conclusions should, in principle, be reached when using continuous state space models. Within the class of discrete state space models, it is fairly straightforward to generate continuous outcomes from discrete latent states; as exemplified by discrete variational autoencoders (Rolfe, 2016) or mixed models, as described in (Friston, Parr et al., 2017).

One can understand the nature of the constants $\{\alpha_{j1}, \alpha_{j0}, \beta_{j1}, \beta_{j0}\}$ from the biological and Bayesian perspectives as follows: $(\alpha_{j1}, \alpha_{j0})$ determines the firing threshold and thus controls the mean firing rates. In other words, these parameters control the amplitude of excitatory and inhibitory inputs, which may be analogous to the roles of GABAergic inputs and neuromodulators in biological neuronal networks (Pawlak et al., 2010; Frémaux & Gerstner, 2016; Kuśmierz et al., 2017). At the same time, $(\alpha_{j1}, \alpha_{j0})$ encodes prior beliefs about states, which exert a large influence on the state posterior. The state posterior is biased if $(\alpha_{j1}, \alpha_{j0})$ is selected in a sub-optimal manner—in relation to the process that generates inputs. Meanwhile, $(\beta_{j1}, \beta_{j0})$ determines the accuracy of synaptic strengths that represent the likelihood mapping of an observation $o_t^{(i)}$ taking 1 (ON state) depending on hidden states (please compare Equation (8) and Equation (20)). Under a usual MDP setup where the state prior does not depend on the parameter posterior, the encoder becomes Bayes optimal when and only when $(\beta_{j1}, \beta_{j0}) = (\vec{0}, \vec{0})$. These constants can represent biological constraints on synaptic strengths, such as the range of spine growth, spinal fluctuations, or the effect of synaptic plasticity induced by spontaneous activity independent of external inputs. Although the fidelity of each synapse is limited due to such internal fluctuations, the accumulation of information over a large number of synapses should allow accurate encoding of hidden states in the current formulation.

In previous reports, we have shown that *in vitro* neural networks—comprising a cortical cell culture—perform BSS when receiving electrical stimulations generated from two hidden sources (Isomura et al., 2015). Furthermore, we showed that minimising variational free energy under an MDP is sufficient to reproduce the learning observed in an *in vitro* network (Isomura & Friston, 2018). Our framework for identifying biologically plausible cost functions could be relevant for identifying the principles that underlie learning or adaptation processes in biological neuronal networks, using empirical response data. Here, we illustrated this potential in terms of the choice of function $\phi_j$ in the cost functions *L*. In particular, if $\phi_j$ is close to a constant $(-\ln 2, -\ln 2)$, the cost function is expressed straightforwardly as a variational free energy with small state prior biases. In the future work, we plan to apply this scheme to empirical data and examine the biological plausibility of variational free energy minimisation.

The correspondence highlighted in this work enables one to identify a generative model (comprising likelihood and priors) that a neural network is using. The formal correspondence between neural network and variational Bayesian formations rests on the asymptotic equivalence between the neural network's cost functions and variational free energy (under some priors). Although variational free energy can take an arbitrary form, the correspondence provides biologically plausible constraints for neural networks that implicitly encode prior distributions. Hence, this formulation is potentially useful for identifying the implicit generative models that underlie the dynamics of real neuronal circuits. In other words, one can quantify the dynamics and plasticity of a neuronal circuit in terms of variational Bayesian inference and learning under an implicit generative model.

The dependence between the likelihood function and the state prior vanishes when the network uses the optimal threshold to perform inference with a generative process that does not involve dependence between the likelihood and the state prior. In other words, the dependence arises from the sub-optimality of the choice of the state prior. This means that the dependence is due to the degrees of freedom in the choice of the threshold that a neural network and its cost function possess. Nevertheless, minimisation of the cost function can render the network Bayes optimal in the variational Bayesian sense, including the choice of the state prior, as described in the previous section. This is because only variational free energy with the optimal priors provides the minimum among a class of neural network cost functions under consideration.

Although we have described the generative process in terms of an MDP, we have ignored state transitions. This means the generative model in this paper reduces to a simple latent variable model, with categorical states and outcomes. As noted above, we refer to MDP models because they predominate in descriptions of variational (Bayesian) belief updating; e.g., (Friston, FitzGerald et al., 2017). Clearly, many generative processes entail state transitions, leading to hidden Markov models (HMM). When state transitions depend upon control variables, we have a POMDP. To deal with such cases, extensions of the current framework are required, which we hope to consider in future work.

In summary, we first identified a class of biologically plausible cost functions for neural networks that underlie changes in both neural activity and synaptic plasticity. We then identified an asymptotic equivalence between these cost functions and the cost functions used in variational Bayesian formations. Given this equivalence, changes in the activity and synaptic strengths of a neuronal network can be viewed as Bayesian belief updating; namely, a process of transforming priors over hidden states and parameters into posteriors, respectively. Hence, a cost function in this class becomes Bayes optimal when activity thresholds correspond to appropriate priors in an implicit generative model. In short, the neural and synaptic dynamics of neural networks can be cast as inference and learning, under a variational Bayesian formation. This is potentially important for two reasons. First, it means that there are some threshold parameters for any neural network (in the class considered) that can be optimised for applications to data, when there are precise prior beliefs about the process generating those data. Second, in virtue of the complete class theorem, one can reverse engineer the priors that any neural network is adopting. This may be interesting when real neuronal networks can be modelled using neural networks of the class that we have considered. In other words, if one can fit neuronal responses—using a neural network model parameterised in terms of threshold constants—it becomes possible to evaluate the implicit priors using the above equivalence. This may find a useful application when applied to *in vitro* (or *in vivo*) neuronal networks (Isomura, Friston, 2018; Levin, 2013) or, indeed, dynamic causal modelling of distributed neuronal responses from non-invasive data (Daunizeau et al., 2011). In this context, the neural network can, in principle, be used as a dynamic causal model to estimate threshold constants and implicit priors. This 'reverse engineering' speaks to estimating the priors used by real neuronal systems, under ideal

Bayesian assumptions; sometimes referred to as meta Bayesian inference (Daunizeau et al., 2010).

## Acknowledgements

T.I. is funded by RIKEN Center for Brain Science. K.J.F. is funded by a Wellcome Principal Research Fellowship (Ref: 088130/Z/09/Z). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## References

Albus, J. S. (1971). A theory of cerebellar function. *Math. Biosci.* **10**, 25-61.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P. & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron* **76**, 695-711.

Belouchrani, A., Abed-Meraim, K., Cardoso, J.F. & Moulines, E. (1997). A blind source separation technique using second-order statistics. *IEEE Trans. Signal Process.* **45**, 434-444.

Brown, G. D., Yamada, S. & Sejnowski, T. J. (2001). Independent component analysis at the neural cocktail party. *Trends Neurosci.* **24**, 54-63.

Brown, L. D. (1981). A complete class theorem for statistical problems with finite-sample spaces. *Ann. Stat.* **9**, 1289-1300.

Cichocki, A., Zdunek, R., Phan, A. H. & Amari, S. I. (2009). *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation.* John Wiley & Sons.

Comon, P. & Jutten, C. (2010). *Handbook of Blind Source Separation: Independent Component Analysis and Applications.* Academic Press.

Daunizeau, J., David, O., & Stephan, K. E. (2011). Dynamic causal modelling: a critical review of the biophysical and statistical foundations. *Neuroimage* **58**, 312-322.

Daunizeau, J., Den Ouden, H. E., Pessiglione, M., Kiebel, S. J., Stephan, K. E., & Friston, K. J. (2010). Observing the observer (I): meta-bayesian models of learning and decision-making. *PLoS One* **5**, e15554.

Dauwels, J. (2007). On variational message passing on factor graphs. *Info. Theory, 2007. ISIT 2007. IEEE Int. Sympo., IEEE*.

Dayan, P., Hinton, G. E., Neal, R. M. & Zemel, R. S. (1995). The Helmholtz machine. *Neural Comput.* **7**, 889-904.

DiCarlo, J. J., Zoccolan, D. & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron* **73**, 415-434.

Forney, G. D. (2001). Codes on graphs: Normal realizations. *IEEE Trans. Info. Theory* **47**, 520-548.

Frémaux, N. & Gerstner, W. (2016). Neuromodulated spike-timing-dependent plasticity, and theory of three-factor learning rules. *Front. Neural Circuits* **9**.

Friston, K. (2005). A theory of cortical responses. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **360**, 815-836.

Friston, K., Kilner, J. & Harrison, L. (2006). A free energy principle for the brain. *J. Physiol. Paris* **100**, 70-87.

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

Friston, K., Mattout, J., & Kilner, J. (2011). Action understanding and active inference. *Biol. Cybern.* **104**, 137-160.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P. & Pezzulo, G. (2016). Active inference and learning. *Neurosci. Biobehav. Rev.* **68**, 862-879.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P. & Pezzulo, G. (2017). Active inference: A process theory. *Neural Comput.* **29**, 1-49.

Friston, K. J., Parr, T. & de Vries, B. D. (2017). The graphical brain: belief propagation and active inference. *Netw. Neurosci.* **1**, 381-414.

George, D. & Hawkins, J. (2009). Towards a mathematical theory of cortical micro-circuits. *PLoS Comput. Biol.* **5**, e1000532.

von Helmholtz, H. (1925). *Treatise on physiological optics (Vol. 3).* The Optical Society of America.

Isomura, T., Kotani, K. & Jimbo, Y. (2015). Cultured cortical neurons can perform blind source separation according to the free-energy principle. *PLoS Comput. Biol.* **11**, e1004643.

Isomura, T. & Friston, K. (2018). In vitro neural networks minimise variational free energy. *Sci. Rep.* **8**, 16926.

Knill, D. C. & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* **27**, 712-719.

Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Stat.* **22**, 79-86.

Kuśmierz, Ł., Isomura, T. & Toyoizumi, T. (2017). Learning with three factors: modulating Hebbian plasticity with errors. *Curr. Opin. Neurobiol.* **46**, 170-177.

Lee, T. W., Girolami, M., Bell, A. J. & Sejnowski, T. J. (2000). A unifying information-theoretic framework for independent component analysis. *Comput. Math. Appl.* **39**, 1-21.

Levin, M. (2013). Reprogramming cells and tissue patterning via bioelectrical pathways: molecular mechanisms and biomedical opportunities. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **5**, 657-676.

Linsker, R. (1988). Self-organization in a perceptual network. *Computer* **21**, 105-117.

Mary, D. (1969). A theory of cerebellar cortex. *J. Physiol.* **202**, 437-470.

Newsome, W. T., Britten, K. H., & Movshon, J. A. (1989). Neuronal correlates of a perceptual decision. *Nature* **341**, 52-54.

Pawlak, V., Wickens, J. R., Kirkwood, A. & Kerr, J. N. (2010). Timing is not everything: neuromodulation opens the STDP gate. *Front. Syn. Neurosci.* **2**, 146.

Rao, R. P. & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79-87.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward.

*Science* **275**, 1593-1599.

Schwartenbeck, P., & K. Friston. (2016). Computational phenotyping in psychiatry: a worked example. *eNeuro* **3**, e0049-16.2016.

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

Tolhurst, D. J., Movshon, J. A., & Dean, A. F. (1983). The statistical reliability of signals in single neurons in cat and monkey visual cortex. *Vision Res.* **23**, 775-785.

Wald, A. (1947). An essentially complete class of admissible decision functions. *Ann. Math. Stat.* **18**, 549-555.

## Supplementary Information

# Reverse engineering neural networks to characterise their cost functions

Takuya Isomura, Karl Friston

## Supplementary Tables

**Table S1.** Correspondence of variables and functions.

| Neural network formation | | | Variational Bayes formation | |
|---|---|---|---|---|
| Neural activity | $x_{tj}$ | $\Leftrightarrow$ | $\mathbf{s}_{t1}^{(j)}$ | State posterior |
| Sensory input | $o_t$ | $\Leftrightarrow$ | $o_t$ | Observation |
| Synaptic strength | $W_{j1}$ | $\Leftrightarrow$ | $\mathrm{sig}^{-1}\left(\mathbf{A}_{11}^{(\cdot,j)}\right)$ | |
| | $\widehat{W}_{j1}$ | $\Leftrightarrow$ | $\mathbf{A}_{11}^{(\cdot,j)}$ | Parameter posterior |
| Perturbation term | $\phi_{j1}$ | $\Leftrightarrow$ | $\ln D_1^{(j)}$ | State prior |
| Threshold | $h_{j1}$ | $\Leftrightarrow$ | $\ln\left(\vec{1}-\mathbf{A}_{11}^{(\cdot,j)}\right)\cdot\vec{1}+\ln D_1^{(j)}$ | |
| Initial synaptic strengths | $\lambda_{j1}\odot\widehat{W}_{j1}^{init}$ | $\Leftrightarrow$ | $a_{11}^{(\cdot,j)}$ | Parameter prior |

## Supplementary Methods

### S1. Order of the parameter complexity

The order of the parameter complexity term

$$\mathcal{D}_A \equiv \sum_{i=1}^{N_o}\sum_{j=1}^{N_s}\sum_{l\in\{1,0\}}\left\{\left(\mathbf{a}_{\cdot l}^{(i,j)}-a_{\cdot l}^{(i,j)}\right)\cdot\ln\mathbf{A}_{\cdot l}^{(i,j)}-\ln\mathcal{B}\left(\mathbf{a}_{\cdot l}^{(i,j)}\right)\right\} \quad (25)$$

is computed. To avoid the divergence of $\ln\mathbf{A}_{\cdot l}^{(i,j)}$, all the elements of $\mathbf{A}_{\cdot l}^{(i,j)}$ are assumed to be larger than a positive constant $\varepsilon$. This means that all the elements of $\mathbf{a}_{\cdot l}^{(i,j)}$ are in the order of $t$. The first term of Equation (25) becomes $\left(\mathbf{a}_{\cdot l}^{(i,j)} - a_{\cdot l}^{(i,j)}\right) \cdot \ln \mathbf{A}_{\cdot l}^{(i,j)} = \mathbf{a}_{\cdot l}^{(i,j)} \cdot \ln \mathbf{A}_{\cdot l}^{(i,j)} + \mathcal{O}(1)$ since $a_{\cdot l}^{(i,j)} \cdot \ln \mathbf{A}_{\cdot l}^{(i,j)}$ is in the order of 1. Moreover, from Equation (3), $\mathbf{a}_{\cdot l}^{(i,j)} \cdot \ln \mathbf{A}_{\cdot l}^{(i,j)} = \mathbf{a}_{\cdot l}^{(i,j)} \cdot \left(\ln \mathbf{a}_{\cdot l}^{(i,j)} - \ln\left(\mathbf{a}_{1l}^{(i,j)} + \mathbf{a}_{0l}^{(i,j)}\right) + \mathcal{O}\left(\left(\mathbf{a}_{\cdot l}^{(i,j)}\right)^{-1}\right)\right) = \mathbf{a}_{\cdot l}^{(i,j)} \cdot \ln\left(\mathbf{A}_{\cdot l}^{(i,j)}\right) + \mathcal{O}(1)$. Meanwhile, the second term of Equation (25) comprises the logarithms of gamma functions as $\ln \mathcal{B}\left(\mathbf{a}_{\cdot l}^{(i,j)}\right) = \ln \Gamma\left(\mathbf{a}_{1l}^{(i,j)}\right) + \ln \Gamma\left(\mathbf{a}_{0l}^{(i,j)}\right) - \ln \Gamma\left(\mathbf{a}_{1l}^{(i,j)} + \mathbf{a}_{0l}^{(i,j)}\right)$. From Stirling's formula,

$$\Gamma\left(\mathbf{a}_{1l}^{(i,j)}\right) = \sqrt{2\pi}\left(\mathbf{a}_{1l}^{(i,j)}\right)^{-\frac{1}{2}}\left(\frac{\mathbf{a}_{1l}^{(i,j)}}{e}\right)^{\mathbf{a}_{1l}^{(i,j)}}\left(1 + \mathcal{O}\left(\left(\mathbf{a}_{\cdot l}^{(i,j)}\right)^{-1}\right)\right) \tag{26}$$

holds. The logarithm of $\Gamma\left(\mathbf{a}_{1l}^{(i,j)}\right)$ is evaluated as

$$\begin{aligned}\ln \Gamma\left(\mathbf{a}_{1l}^{(i,j)}\right) &= \frac{1}{2}\ln 2\pi - \frac{1}{2}\ln \mathbf{a}_{1l}^{(i,j)} + \mathbf{a}_{1l}^{(i,j)}\left(\ln \mathbf{a}_{1l}^{(i,j)} - 1\right) + \ln\left(1 + \mathcal{O}\left(\left(\mathbf{a}_{\cdot l}^{(i,j)}\right)^{-1}\right)\right) \\ &= \mathbf{a}_{1l}^{(i,j)} \ln \mathbf{a}_{1l}^{(i,j)} - \mathbf{a}_{1l}^{(i,j)} + \mathcal{O}(\ln t).\end{aligned} \tag{27}$$

Similarly, $\ln \Gamma\left(\mathbf{a}_{0l}^{(i,j)}\right) = \mathbf{a}_{0l}^{(i,j)} \ln \mathbf{a}_{0l}^{(i,j)} - \mathbf{a}_{0l}^{(i,j)} + \mathcal{O}(\ln t)$ and $\ln \Gamma\left(\mathbf{a}_{1l}^{(i,j)} + \mathbf{a}_{0l}^{(i,j)}\right) = \left(\mathbf{a}_{1l}^{(i,j)} + \mathbf{a}_{0l}^{(i,j)}\right)\ln\left(\mathbf{a}_{1l}^{(i,j)} + \mathbf{a}_{0l}^{(i,j)}\right) - \left(\mathbf{a}_{1l}^{(i,j)} + \mathbf{a}_{0l}^{(i,j)}\right) + \mathcal{O}(\ln t)$ hold. Thus, we obtain

$$\begin{aligned}\ln \mathcal{B}\left(\mathbf{a}_{\cdot l}^{(i,j)}\right) &= \mathbf{a}_{1l}^{(i,j)} \ln \mathbf{a}_{1l}^{(i,j)} + \mathbf{a}_{0l}^{(i,j)} \ln \mathbf{a}_{0l}^{(i,j)} - \left(\mathbf{a}_{1l}^{(i,j)} + \mathbf{a}_{0l}^{(i,j)}\right)\ln\left(\mathbf{a}_{1l}^{(i,j)} + \mathbf{a}_{0l}^{(i,j)}\right) + \mathcal{O}(\ln t) \\ &= \mathbf{a}_{\cdot l}^{(i,j)} \cdot \ln\left(\mathbf{A}_{\cdot l}^{(i,j)}\right) + \mathcal{O}(\ln t).\end{aligned} \tag{28}$$

Hence, Equation (25) becomes

$$\mathcal{D}_A = \sum_{i=1}^{N_o}\sum_{j=1}^{N_s}\sum_{l\in\{1,0\}}\left\{\mathbf{a}_{\cdot l}^{(i,j)} \cdot \ln\left(\mathbf{A}_{\cdot l}^{(i,j)}\right) + \mathcal{O}(1) - \left(\mathbf{a}_{\cdot l}^{(i,j)} \cdot \ln\left(\mathbf{A}_{\cdot l}^{(i,j)}\right) + \mathcal{O}(\ln t)\right)\right\} = \mathcal{O}(\ln t). \tag{29}$$

Therefore, we obtain

$$F\big(\tilde{o}, Q(\tilde{s}), Q(A)\big) = \sum_{j=1}^{N_s}\sum_{\tau=1}^{t} \mathbf{s}_\tau^{(j)} \cdot \left\{\ln \mathbf{s}_\tau^{(j)} - \sum_{i=1}^{N_o} \ln \mathbf{A}^{(i,j)} \cdot o_\tau^{(i)} - \ln D^{(j)}\right\} + \mathcal{O}(\ln t). \tag{30}$$

Under the current generative model comprising binary hidden states and binary observations, the optimal posterior expectation of $\mathbf{A}$ can be obtained up to the order of $\ln t/t$ even when the $\mathcal{O}(\ln t)$ term in Equation (30) is ignored. Solving the variation of $F$ with respect to $\mathbf{A}_{1l}^{(i,j)}$ yields the optimal posterior expectation. From $\mathbf{A}_{0l}^{(i,j)} = 1 - \mathbf{A}_{1l}^{(i,j)}$, we find

$$\delta F = \sum_{i=1}^{N_o}\sum_{j=1}^{N_s}\sum_{\tau=1}^{t} \mathbf{s}_\tau^{(j)} \cdot \left\{-\delta \ln \mathbf{A}_{1\cdot}^{(i,j)}\, o_\tau^{(i)} - \delta \ln\left(\vec{1} - \mathbf{A}_{1\cdot}^{(i,j)}\right)\left(1 - o_\tau^{(i)}\right)\right\}$$

$$= t\sum_{i=1}^{N_o}\sum_{j=1}^{N_s}\left\{-\left(\delta \mathbf{A}_{1\cdot}^{(i,j)} \odot \left(\mathbf{A}_{1\cdot}^{(i,j)}\right)^{\odot -1}\right) \cdot \overline{o_t^{(i)} \otimes \mathbf{s}_t^{(j)}} + \left(\delta \mathbf{A}_{1\cdot}^{(i,j)} \odot \left(\vec{1} - \mathbf{A}_{1\cdot}^{(i,j)}\right)^{\odot -1}\right) \cdot \overline{\left(1 - o_t^{(i)}\right)\mathbf{s}_t^{(j)}}\right\}$$

$$= t\sum_{i=1}^{N_o}\sum_{j=1}^{N_s}\left(\delta \mathbf{A}_{1\cdot}^{(i,j)} \odot \left(\mathbf{A}_{1\cdot}^{(i,j)}\right)^{\odot -1} \odot \left(\vec{1} - \mathbf{A}_{1\cdot}^{(i,j)}\right)^{\odot -1}\right) \cdot \left(\mathbf{A}_{1\cdot}^{(i,j)} \odot \overline{\mathbf{s}_t^{(j)}} - \overline{o_t^{(i)}\mathbf{s}_t^{(j)}}\right) \quad (31)$$

up to the order of $\ln t$. Here, $\left(\mathbf{A}_{1\cdot}^{(i,j)}\right)^{\odot -1}$ denotes the element-wise inverse of $\mathbf{A}_{1\cdot}^{(i,j)}$. From $\delta F = 0$, we find

$$\mathbf{A}_{1\cdot}^{(i,j)} = \overline{o_t^{(i)}\mathbf{s}_t^{(j)}} \odot \left(\overline{\mathbf{s}_t^{(j)}}\right)^{\odot -1} + \mathcal{O}\left(\frac{\ln t}{t}\right). \quad (32)$$

Therefore, we obtain the same result as Equation (8) up to the order of $\ln t/t$.

## S2. Derivation of synaptic plasticity rule

We consider synaptic strengths at time $t$, $W_{j1} = W_{j1}(t)$, and define the change as $\Delta W_{j1} \equiv W_{j1}(t+1) - W_{j1}(t)$. From Equation (15), $h_1'(W_{j1})$ satisfies both

$$h_1'(W_{j1} + \Delta W_{j1}) - h_1'(W_{j1}) = h_1''(W_{j1}) \odot \Delta W_{j1} + \mathcal{O}\left(\left|\Delta W_{j1}\right|^2\right) \quad (33)$$

and

$$h_1'(W_{j1} + \Delta W_{j1}) - h_1'(W_{j1}) = -\frac{x_{(t+1)j}o_{t+1} + t\overline{x_{tj}o_t}}{x_{(t+1)j} + t\overline{x_{tj}}} + \frac{\overline{x_{tj}o_t}}{\overline{x_{tj}}}$$

$$\approx -\frac{x_{(t+1)j}o_{t+1}}{t\overline{x_{tj}}} + \frac{\overline{x_{tj}o_t}}{t\overline{x_{tj}}^2}x_{(t+1)j} = -\frac{1}{t\overline{x_{tj}}}\left(x_{(t+1)j}o_{t+1} - h_1'(W_{j1})x_{(t+1)j}\right). \quad (34)$$

Thus, we find

$$\Delta W_{j1} = \underbrace{-\frac{h_1''(W_{j1})^{\odot -1}}{t\overline{x_{tj}}}}_{\text{adaptive learning rate}} \odot \left( \underbrace{x_{(t+1)j} o_{t+1}}_{\text{Hebbian term}} - \underbrace{h_1'(W_{j1}) x_{(t+1)j}}_{\text{homeostatic term}} \right). \tag{35}$$

Similarly,

$$\Delta W_{j0} = \underbrace{-\frac{h_0''(W_{j0})^{\odot -1}}{t\overline{1 - x_{tj}}}}_{\text{adaptive learning rate}} \odot \left( \underbrace{(1 - x_{(t+1)j}) o_{t+1}}_{\text{anti-Hebbian term}} - \underbrace{h_0'(W_{j0})(1 - x_{(t+1)j})}_{\text{homeostatic term}} \right). \tag{36}$$

These plasticity rules express (anti-) Hebbian plasticity with a homeostatic term.

**S3. Correspondence between parameter prior distribution and initial synaptic strengths**

In general, optimising a model of observable quantities—including a neural network—can be cast inference, if there exists a learning mechanism that updates the hidden states and parameters of that model based on observations. (Exact and variational) Bayesian inference treats the hidden states and parameters as random variables, and thus transforms prior distributions $P(s_t), P(A)$ into posteriors $Q(s_t), Q(A)$. In other words, Bayesian inference is a process of transforming the prior to the posterior based on observations $o_1, \ldots, o_t$ under a generative model. From this perspective, the incorporation of prior knowledge about the hidden states and parameters is an important aspect of Bayesian inference.

The minimisation of a cost function by a neural network updates its activity and synaptic strengths based on observations under the given network properties (e.g., activation function and thresholds). According to the complete class theorem, this process can always be viewed as Bayesian inference. In the main text, we demonstrated that a class of cost functions—for a single-layer feed-forward network with a sigmoid activation function—has a form equivalent to variational free energy under a particular latent variable model. Here, neural activity $x_t$ and synaptic strengths $W$ come to encode the posterior distributions over hidden states $Q'(s_t)$ and parameters $Q'(A)$, respectively, where $Q'(s_t)$ and $Q'(A)$ follow categorical and Dirichlet distributions, respectively. Moreover, we identified that the perturbation factors $\phi_j$—that characterise the threshold function—correspond to the logarithm of the state prior $P(s_t)$ expressed as a categorical distribution.

However, one might ask whether the posteriors obtained using the network $Q'(s_t), Q'(A)$ are formally different from those obtained using variational Bayesian inference $Q(s_t), Q(A)$, since only the latter explicitly considers the prior distribution of parameters $P(A)$. Thus, one may wonder if the network merely influences update rules that are similar to variational Bayes but does not transform the priors $P(s_t), P(A)$ into the posteriors $Q(s_t), Q(A)$, despite the asymptotic equivalence of the cost functions.

Below, we show that the initial values of synaptic strengths $W_{j1}^{init}, W_{j0}^{init}$ correspond to the parameter prior $P(A)$ expressed as a Dirichlet distribution, to show that a neural network indeed transforms the priors into the posteriors. For this purpose, we specify the order 1 term in Equation (12) to make the dependence on the initial synaptic strengths explicit. Specifically, we modify Equation (12) as

$$L_j = \sum_{\tau=1}^{t} \left( f(x_{\tau j}) - \begin{pmatrix} x_{\tau j} \\ 1 - x_{\tau j} \end{pmatrix}^T \left( \begin{pmatrix} W_{j1} \\ W_{j0} \end{pmatrix} o_\tau + \begin{pmatrix} h_{j1} \\ h_{j0} \end{pmatrix} \right) \right)$$

$$+ (W_{j1}, W_{j0}) \left( \lambda_{j1} \odot \widehat{W}_{j1}^{init}, \lambda_{j0} \odot \widehat{W}_{j0}^{init} \right)^T$$

$$+ \left( \ln(\vec{1} - \widehat{W}_{j1}), \ln(\vec{1} - \widehat{W}_{j0}) \right) (\lambda_{j1}, \lambda_{j0})^T, \quad (37)$$

where $\widehat{W}_{j1}^{init} \equiv \mathrm{sig}(W_{j1}^{init})$ and $\widehat{W}_{j0}^{init} \equiv \mathrm{sig}(W_{j0}^{init})$ are the sigmoid functions of the initial synaptic strengths, and $\lambda_{j1}, \lambda_{j0} \in \mathbb{R}^{N_o}$ are row vectors of the inverse learning rate factors that express the insensitivity of the synaptic strengths to the activity-dependent synaptic plasticity. The third term of Equation (37) expresses the integral of $\widehat{W}_{j1}$ and $\widehat{W}_{j0}$ (with respect to $W_{j1}$ and $W_{j0}$, respectively). This ensures that when $t = 0$ (i.e., when the first term on the right-hand side of Equation (37) is zero), the derivative of $L_j$ is given by $\partial L_j / \partial W_{j1} = \lambda_{j1} \odot \widehat{W}_{j1}^{init} - \lambda_{j1} \odot \widehat{W}_{j1}$, and thus $(W_{j1}, W_{j0}) = (W_{j1}^{init}, W_{j0}^{init})$ provides the fixed point of $L_j$.

Similar to the transformation from Equation (12) to Equation (17), we compute Equation (37) as

$$L = \sum_{j=1}^{N_x} \sum_{\tau=1}^{t} \begin{pmatrix} x_{\tau j} \\ 1 - x_{\tau j} \end{pmatrix}^T \left\{ \begin{pmatrix} \ln x_{\tau j} \\ \ln(1 - x_{\tau j}) \end{pmatrix} - \begin{pmatrix} \ln \widehat{W}_{j1} & \ln(\vec{1} - \widehat{W}_{j1}) \\ \ln \widehat{W}_{j0} & \ln(\vec{1} - \widehat{W}_{j0}) \end{pmatrix} \begin{pmatrix} o_\tau \\ \vec{1} - o_\tau \end{pmatrix} - \begin{pmatrix} \phi_{j1} \\ \phi_{j0} \end{pmatrix} \right\}$$

$$+ \sum_{j=1}^{N_x} \left\{ \left( \ln \widehat{W}_{j1}, \ln(\vec{1} - \widehat{W}_{j1}) \right) \left( \lambda_{j1} \odot \widehat{W}_{j1}^{init}, \lambda_{j1} \odot (\vec{1} - \widehat{W}_{j1}^{init}) \right)^T \right.$$

$$\left. + \left( \ln \widehat{W}_{j0}, \ln(\vec{1} - \widehat{W}_{j0}) \right) \left( \lambda_{j0} \odot \widehat{W}_{j0}^{init}, \lambda_{j0} \odot (\vec{1} - \widehat{W}_{j0}^{init}) \right)^T \right\}. \quad (38)$$

Note that we used $W_{j1} = \ln \widehat{W}_{j1} - \ln(\vec{1} - \widehat{W}_{j1})$. Crucially, analogous to the correspondence between $\widehat{W}_{j1}$ and the Dirichlet parameters of the parameter posterior $\mathbf{a}_{11}^{(\cdot,j)}$, $\lambda_{j1} \odot \widehat{W}_{j1}^{init}$

can be formally associated with the Dirichlet parameters of the parameter prior $a_{11}^{(\cdot,j)}$. Hence, one can see the formal correspondence between the second and third terms on the right-hand side of Equation (38) and the expectation of the log parameter prior in Equation (4):

$$\mathrm{E}_{Q(A)}[\ln P(A)] = \sum_{i=1}^{N_o}\sum_{j=1}^{N_s} \ln \mathbf{A}^{(i,j)} \cdot a^{(i,j)}$$

$$= \sum_{i=1}^{N_o}\sum_{j=1}^{N_s} \left\{ \ln \mathbf{A}_{\cdot 1}^{(i,j)} \cdot a_{\cdot 1}^{(i,j)} + \ln \mathbf{A}_{\cdot 0}^{(i,j)} \cdot a_{\cdot 0}^{(i,j)} \right\}. \tag{39}$$

Furthermore, the synaptic update rules are derived from Equation (38) as

$$\begin{cases} \dot{W}_{j1} \propto -\dfrac{1}{t}\dfrac{\partial L}{\partial W_{j1}} = \overline{x_{tj}o_t} - \overline{x_{tj}}\widehat{W}_{j1} + \overline{x_{tj}}\phi'_{j1} + \dfrac{1}{t}\left(\lambda_{j1} \odot \widehat{W}_{j1}^{init} - \lambda_{j1} \odot \widehat{W}_{j1}\right) \\ \dot{W}_{j0} \propto -\dfrac{1}{t}\dfrac{\partial L}{\partial W_{j0}} = \overline{(1-x_{tj})o_t} - \overline{1-x_{tj}}\widehat{W}_{j0} + \overline{1-x_{tj}}\phi'_{j0} + \dfrac{1}{t}\left(\lambda_{j0} \odot \widehat{W}_{j0}^{init} - \lambda_{j0} \odot \widehat{W}_{j0}\right) \end{cases} \tag{40}$$

The fixed point of Equation (40) is provided as

$$\begin{cases} W_{j1} = \mathrm{sig}^{-1}\left( \left( t\overline{x_{tj}\vec{1}} + \lambda_{j1} \right)^{\odot -1} \odot \left( t\overline{x_{tj}o_t} + t\overline{x_{tj}}\phi'_{j1} + \lambda_{j1} \odot \widehat{W}_{j1}^{init} \right) \right) \\ W_{j0} = \mathrm{sig}^{-1}\left( \left( t\overline{1-x_{tj}\vec{1}} + \lambda_{j0} \right)^{\odot -1} \odot \left( t\overline{(1-x_{tj})o_t} + t\overline{1-x_{tj}}\phi'_{j0} + \lambda_{j0} \odot \widehat{W}_{j0}^{init} \right) \right) \end{cases} \tag{41}$$

Note that the synaptic strengths at $t = 0$ are computed as $W_{j1} = \mathrm{sig}^{-1}\left( \left(\lambda_{j1}\right)^{\odot -1} \odot \left(\lambda_{j1} \odot \widehat{W}_{j1}^{init}\right) \right) = W_{j1}^{init}$. Again, one can see the formal correspondence between the final values of the synaptic strengths given by Equation (41) in the neural network formation and the parameter posterior given by Equation (8) in the variational Bayesian formation. As the Dirichlet parameter of the posterior $\mathbf{a}_{11}^{(\cdot,j)}$ is decomposed into the outer product $\overline{o_t \otimes \mathbf{s}_{t1}^{(j)}}$ and the prior $a_{11}^{(\cdot,j)}$, they are associated with $\overline{x_{tj}o_t}$ and $\lambda_{j1} \odot \widehat{W}_{j1}^{init}$, respectively. Thus, Equation (8) corresponds to Equation (41). Hence, for a given constant set $\left\{W_{j1}^{init}, W_{j0}^{init}, \lambda_{j1}, \lambda_{j0}\right\}$, we identify the corresponding parameter prior $P\left(A^{(\cdot,j)}\right) = \mathrm{Dir}\left(a^{(\cdot,j)}\right)$, given by

$$a^{(\cdot,j)} \equiv \begin{pmatrix} a_{11}^{(\cdot,j)} & a_{10}^{(\cdot,j)} \\ a_{01}^{(\cdot,j)} & a_{00}^{(\cdot,j)} \end{pmatrix} = \begin{pmatrix} \lambda_{j1} \odot \widehat{W}_{j1}^{init} & \lambda_{j0} \odot \widehat{W}_{j0}^{init} \\ \lambda_{j1} \odot \left(\vec{1} - \widehat{W}_{j1}^{init}\right) & \lambda_{j0} \odot \left(\vec{1} - \widehat{W}_{j0}^{init}\right) \end{pmatrix}. \tag{42}$$

In summary, one can establish the formal correspondence between neural network and variational Bayesian formations, in terms of the cost functions (Equation (4) vs. Equation (38)), priors (Equation (18) and Equation (42)), and posteriors (Equation (8) vs. Equation (41)). This means that a neural network successively transforms priors $P(s_t), P(A)$ into posteriors $Q(s_t), Q(A)$, as parameterised with neural activity, and initial and final synaptic strengths (and thresholds). Crucially, when increasing number of observations, this process is asymptotically equivalent to that of variational Bayesian inference, under a specific likelihood function.