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. In brief, we show that such cost functions can be cast as a variational bound on model evidence, or marginal likelihood, 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 generating 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 generating 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

Introduction

Cost functions are used to solve problems in various scientific fields—including physics, chemistry, engineering and machine learning. Furthermore, any optimisation problem that can be specified using a cost function can be formulated as a gradient descent; enabling one to treat neuronal dynamics and plasticity as an optimisation process. Neuroscience commonly uses cost functions to express various types of learning: for instance, supervised learning to minimise the differences between outputs and targets, as in a perceptron (Marr, 1969; Albus, 1971); reinforcement learning to maximise future reward (Schultz et al., 1997; Sutton & Barto, 1998), and unsupervised learning to maximise the efficiency of encoding (Linsker, 1988; Brown et al., 2001). These examples highlight the importance of specifying a problem or function in terms of cost functions, from which neural and synaptic dynamics can be derived. In other words, cost functions offer 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, it is no longer necessary to consider the dynamics. We can, instead, characterise the neural network's behaviour in terms of fixed points, transients, attractors and structural stability—based on and only on the cost function. In short, it is important to identify the cost function for a neural network to understand its dynamics, plasticity, and function.

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 input 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 governing 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 the hidden states 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 optimisation of alternative cost functions.

In principle, any neural network that produces an output or a decision can be regarded as performing some form of inference in terms of Bayesian decision theory. On this reading, the complete class theorem suggests that any neural network could be regarded as performing Bayesian inference under some prior beliefs; therefore, it could be regarded as minimising variational free energy. The complete class theorem (Wald, 1947; Brown, 1981) says 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 an identification is possible for a fairly canonical form of neural network and a generic form of 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. For simplicity, we 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. We have previously observed BSS performed by *in vitro* neural networks (Isomura et al., 2015) and have 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 empirical behaviour of *in vitro* networks.

In this work, we ask whether variational free energy minimisation can account for the normative behaviour of any neural network, by considering all possible cost functions (i.e., possible purposes). 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 with sigmoid activation. The gradient descent on the ensuing cost function leads naturally 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.

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 form of this updating 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.

Generative models. Under an MDP scheme (Fig. 1A), a minimal BSS setup—in a discrete-space—can be expressed as the likelihood mapping (A) from *N* hidden sources or states $s_t \equiv s_{t1} \otimes \cdots \otimes s_{tN}$ to *M* observations $o_t \equiv o_{t1} \otimes \cdots \otimes o_{tM}$ (using the outer product operator $\otimes$). Each source and observation take one (ON state) or zero (OFF state) values for each time step ($s_{tk} \in \{0,1\}, o_{ti} \in \{0,1\}$). Throughout this manuscript, *k* indices the *k*-th hidden state, while *i* indices the *i*-th observation. The probability of s_t follows a categorical distribution $P(s_t) = P(s_{t1}) \cdots P(s_{tN}) = \mathrm{Cat}(D_1) \cdots \mathrm{Cat}(D_N)$, where $D_k \equiv (D_{k1}, D_{k0})$ with $D_{k1} + D_{k0} = 1$.

The probability of an outcome is determined by the likelihood mapping, from hidden states to observations, in terms of a categorical distribution *A*: $P(o_{ti}|s_t, A_i) = \mathrm{Cat}(A_i)$, where the elements of A_i are given by $A_{ijl_1 \cdots l_N} = P(o_{ti} = j | s_{t1} = l_1, \dots, s_{tN} = l_N, A)$. This determines the probability of o_{ti} takes $j \in \{0,1\}$ when $s_t = (l_1, \dots, l_N)$. The prior distribution of A_i is defined by Dirichlet distribution $P(A_i) = \mathrm{Dir}(a_i)$ with sufficient statistics (concentration parameter) a_i. We use $\tilde{o} \equiv (o_1, \dots, o_t)$ to denote a sequence of observations and $\tilde{s} \equiv (s_1, \dots, s_t)$ a sequence of hidden states. Formally, the generative model (i.e., the joint distribution over outcomes, hidden states and the parameters of their likelihood mapping) is expressed as

$$P(\tilde{o}, \tilde{s}, A) = P(A) \prod_{\tau=1}^{t} P(o_\tau | s_\tau, A) P(s_\tau) = \prod_{i=1}^{M} P(A_i) \cdot \prod_{\tau=1}^{t} P(s_\tau) \prod_{i=1}^{M} P(o_{\tau i} | s_\tau, A_i) . \qquad (1)$$

[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).

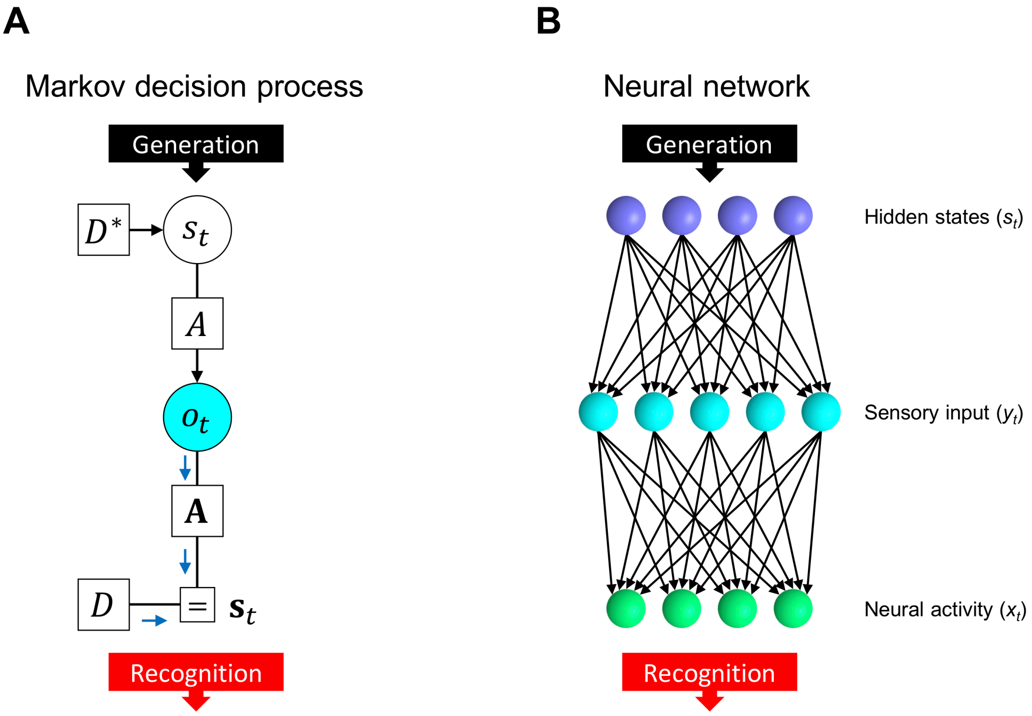

Figure 1. Comparison between an MDP scheme and a neural network. (**A**) MDP scheme expressed as a Forney factor graph (Forney, 2001; Dauwels, 2007) based upon the formulation in (Friston, Parr et al., 2017). In this BSS setup, the prior D determines hidden states s_t and s_t determines observation o_t 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 that the network operates under. If $D = D^*$, inference would be optimal and biased otherwise. (**B**) Neural network comprising a single layer feed-forward network with a sigmoid activation function. The network receives sensory inputs $y_t = (y_{t1}, \dots, y_{tM})$ that are generated from hidden states $s_t = (s_{t1}, \dots, s_{tN})$ and outputs neural activities $x_t = (x_{t1}, \dots, x_{tN})$. Here, y_{ti} corresponds to a binary outcome o_{ti} , and x_{tk} should encode the posterior expectation about a binary state s_{tk}.

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

$$Q(\tilde{s}, A) = Q(A)Q(\tilde{s}) = \prod_{i=1}^{M} Q(A_i) \cdot \prod_{\tau=1}^{t} Q(s_\tau), \tag{2}$$

where the marginal posterior distributions of s_t and A are categorical $Q(s_\tau) = \mathrm{Cat}(\mathbf{s}_\tau)$ and Dirichlet $Q(A_i) = \mathrm{Dir}(\mathbf{a}_i)$ distributions, respectively. Note that $\mathbf{s}_\tau$ and $\mathbf{a}_i$ denote sufficient statistics (i.e., $\mathbf{s}_\tau$ gives expectations between zero and one, and $\mathbf{a}_i$ expresses the concentration parameter). Below, we use the posterior expectation of $\ln A_i$ to encode posterior beliefs about the likelihood, which is given by

$$\ln \mathbf{A}_i \equiv \mathrm{E}_{Q(A_i)}[\ln A_i] = \psi(\mathbf{a}_i) - \psi(\mathbf{a}_{i1} + \mathbf{a}_{i0}) = \ln \mathbf{a}_i - \ln(\mathbf{a}_{i1} + \mathbf{a}_{i0}) + \mathcal{O}(\mathbf{a}_i^{-1}) \tag{3}$$

using the digamma function $\psi(\cdot)$. Here $\mathrm{E}_{Q(A_i)}[\cdot]$ denotes the expectation over $Q(A_i)$. The variational free energy of this generative model is then given by:

$$F\big(\tilde{o}, Q(\tilde{s}), Q(A)\big)$$

$$\equiv \sum_{\tau=1}^{t} \{\mathrm{E}_{Q(s_\tau)Q(A)}[-\ln P(o_\tau|s_\tau, A)] + \mathcal{D}_{\mathrm{KL}}[Q(s_\tau)||P(s_\tau)]\} + \mathcal{D}_{\mathrm{KL}}[Q(A)||P(A)]$$

$$= \underbrace{\sum_{\tau=1}^{t} \mathbf{s}_\tau \cdot \{-\ln \mathbf{A} \cdot o_\tau + \ln \mathbf{s}_\tau - \ln D\}}_{\text{accuracy+state complexity}} + \underbrace{\sum_{i=1}^{M} \{(\mathbf{a}_i - a_i) \cdot \ln \mathbf{A}_i - \ln \mathcal{B}(\mathbf{a}_i)\}}_{\text{parameter complexity}}, \tag{4}$$

where $\mathcal{D}_{\mathrm{KL}}[\cdot||\cdot]$ is complexity as scored by the Kullback-Leibler divergence (Kullback & Leibler, 1951) and $\mathcal{B}(\mathbf{a}_i) \equiv \Gamma(\mathbf{a}_{i1})\Gamma(\mathbf{a}_{i0})/\Gamma(\mathbf{a}_{i1} + \mathbf{a}_{i0})$ is the beta function. The first term in the final equality comprises accuracy and the state complexity, which increases in proportion to time t. Conversely, the second term—the complexity of parameters—increases in the order of $\ln t$ and is thus negligible when t is large (see Supplementary Methods S1 for details). In what follows, we will therefore drop parameter complexity, under the assumption that the scheme has experienced a sufficient number of outcomes.

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:

$$\mathbf{s}_{tk} = \sigma(\ln \mathbf{A}_{\cdot\cdot k} \cdot o_t + \ln D_k), \tag{5}$$

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

$$\mathbf{s}_{tk1} = \frac{\exp(\ln \mathbf{A}_{\cdot\cdot k1} \cdot o_t + \ln D_{k1})}{\exp(\ln \mathbf{A}_{\cdot\cdot k1} \cdot o_t + \ln D_{k1}) + \exp(\ln \mathbf{A}_{\cdot\cdot k0} \cdot o_t + \ln D_{k0})}$$

$$= \mathrm{sig}(\ln \mathbf{A}_{\cdot\cdot k1} \cdot o_t - \ln \mathbf{A}_{\cdot\cdot k0} \cdot o_t + \ln D_{k1} - \ln D_{k0}) \tag{6}$$

using the sigmoid function $\mathrm{sig}(z) \equiv 1/(1+\exp(-z))$. (The posterior expectation of s_{tk} taking a value 0 (OFF state) is thus $\mathbf{s}_{tk0} = 1 - \mathbf{s}_{tk1}$.) Here, D_{k1} and D_{k0} 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., $D_{k1} = D_{k0} = 0.5$ 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 expectations can be associated with nonnegative firing rates, while the arguments of the sigmoid (softmax) function can be associated with neuronal depolarisation; rendering the softmax function a voltage-firing rate activation function

In terms of learning, the optimal posterior expectations about the parameters are given by:

$$\mathbf{a}_i = a_i + \sum_{\tau=1}^{t} o_{\tau i} \otimes \mathbf{s}_\tau = a_i + \overline{o_{ti} \otimes \mathbf{s}_t}, \tag{7}$$

where a_i is the prior, $\otimes$ expresses the operator of outer product, and $\overline{o_{ti} \otimes \mathbf{s}_t} \equiv \sum_{\tau=1}^{t} o_{\tau i} \otimes \mathbf{s}_\tau$. Thus, the optimal posterior expectation of matrix A is $\mathbf{A}_{\cdot\cdot k} = \mathbf{a}_{\cdot\cdot k} / (\mathbf{a}_{\cdot 1k} + \mathbf{a}_{\cdot 0k})$, or equivalently

$$\begin{cases} \mathbf{A}_{\cdot 1k1} = \dfrac{\mathbf{a}_{\cdot 1k1}}{\mathbf{a}_{\cdot 1k1} + \mathbf{a}_{\cdot 0k1}} = \dfrac{\overline{o_{t\cdot 1} \mathbf{s}_{tk1}}}{\overline{\mathbf{s}_{tk1}}} + \mathcal{O}\left(\dfrac{1}{t}\right) \\ \mathbf{A}_{\cdot 1k0} = \dfrac{\mathbf{a}_{\cdot 1k0}}{\mathbf{a}_{\cdot 1k0} + \mathbf{a}_{\cdot 0k0}} = \dfrac{\overline{o_{t\cdot 1}(1 - \mathbf{s}_{tk1})}}{\overline{1 - \mathbf{s}_{tk1}}} + \mathcal{O}\left(\dfrac{1}{t}\right) \end{cases}, \tag{8}$$

where $\overline{o_{t\cdot 1} \mathbf{s}_{tk1}} = \sum_{\tau=1}^{t} o_{\tau\cdot 1} \mathbf{s}_{\tau k1}$, $\overline{\mathbf{s}_{tk1}} = \sum_{\tau=1}^{t} \mathbf{s}_{\tau k1}$, $\overline{o_{t\cdot 1}(1 - \mathbf{s}_{tk1})} = \sum_{\tau=1}^{t} o_{\tau\cdot 1}(1 - \mathbf{s}_{\tau k1})$, and $\overline{1 - \mathbf{s}_{tk1}} = \sum_{\tau=1}^{t} (1 - \mathbf{s}_{\tau k1})$. Whereas, $\mathbf{A}_{\cdot 0k1} = \vec{1} - \mathbf{A}_{\cdot 1k1}$ and $\mathbf{A}_{\cdot 0k0} = \vec{1} - \mathbf{A}_{\cdot 1k0}$. Here $\vec{1} = (1, \dots, 1) \in \mathbb{R}^M$ is a vector of ones. The prior of parameters a_i is in the order of 1 and thus negligible when t is large. The four vectors $(\mathbf{A}_{\cdot 1k1}, \mathbf{A}_{\cdot 1k0}, \mathbf{A}_{\cdot 0k1}, \mathbf{A}_{\cdot 0k0})$ express the optimal posterior expectations of o_t taking ON state when s_{tk} is ON ($\mathbf{A}_{\cdot 1k1}$) or OFF ($\mathbf{A}_{\cdot 1k0}$), or o_t taking OFF state when s_{tk} is ON ($\mathbf{A}_{\cdot 0k1}$) or OFF ($\mathbf{A}_{\cdot 0k0}$). Although this expression may look 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 see the associative or Hebbian aspect of this belief updating; expressed here in terms of the outer products between (presynaptic) expectations about states and (postsynaptic) outcomes in Equation (7). We now turn to the equivalent updating for activities and parameters or weights of a neural network.

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

$$\dot{x}_{tk} \propto -f'(x_{tk}) + W_{k1} y_t - W_{k0} y_t + h_{k1} - h_{k0}, \tag{9}$$

where $y_t \equiv (y_{t1}, \dots, y_{tM})^T = (o_{t11}, \dots, o_{tM1})^T$ is a column vector of inputs that encodes the

ON states of o_t. We suppose $W_{k1} \in \mathbb{R}^M$ and $W_{k0} \in \mathbb{R}^M$ comprise row vectors of synapses, and $h_{k1} \in \mathbb{R}$ and $h_{k0} \in \mathbb{R}$ are adaptive thresholds that depend on the values of W_{k1} and W_{k0}, respectively. One may think W_{k1} and W_{k0} express excitatory and inhibitory synapses, respectively. We will further assume that the nonlinear leakage $f'(\cdot)$ is the inverse of the sigmoid function, such that the fixed point of x_{tk} is given by

$$\begin{aligned} x_{tk} &= \mathrm{sig}(W_{k1} y_t - W_{k0} y_t + h_{k1} - h_{k0}) \\ &= \frac{\exp(W_{k1} y_t + h_{k1})}{\exp(W_{k1} y_t + h_{k1}) + \exp(W_{k0} y_t + h_{k0})}. \end{aligned} \tag{10}$$

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

$$\begin{cases} \Delta W_{k1}(t) \equiv W_{k1}(t+1) - W_{k1}(t) \propto Hebb_1(x_{tk}, y_t, W_{k1}) + Home_1(x_{tk}, W_{k1}) \\ \Delta W_{k0}(t) \equiv W_{k0}(t+1) - W_{k0}(t) \propto Hebb_0(x_{tk}, y_t, W_{k0}) + Home_0(x_{tk}, W_{k0}) \end{cases}, \tag{11}$$

where $Hebb_1$ and $Hebb_0$ mediate Hebbian plasticity as determined by the product of sensory inputs and neural outputs, and $Home_1$ and $Home_0$ are 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 see a formal similarity between the solutions for expectations about states (Equation (6)) and activity in the neural network (Equation (10)). By this analogy, x_{tk} can be regarded as encoding the posterior expectation of the ON state $\mathbf{s}_{\tau k1}$, and W_{k1} and W_{k0} correspond to $\ln \mathbf{A}_{\cdot 1k1} - \ln(\vec{1} - \mathbf{A}_{\cdot 1k1}) = \mathrm{sig}^{-1}(\mathbf{A}_{\cdot 1k1})$ and $\ln \mathbf{A}_{\cdot 1k0} - \ln(\vec{1} - \mathbf{A}_{\cdot 1k0}) = \mathrm{sig}^{-1}(\mathbf{A}_{\cdot 1k0})$, respectively; in the sense that they express the amplitude of $y_t = o_{t\cdot 1}$ influencing x_{tk}.

The optimal posterior expectation 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 is assured when $f'(x_{\tau k})$ is the inverse of the sigmoid function (see Equation (13) below). Under this condition the fixed point or solution for x_{tk} is given by Equation (10), which compares inputs from ON and OFF pathways. This means x_{tk} encodes the posterior expectation of the *k*-th hidden state being ON—that is, $x_{tk} \to \mathbf{s}_{tk1}$. In short, the 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 neural activity and synaptic strengths asymptotically encode Bayesian optimal posterior beliefs about hidden states ($x_{tk} \to \mathbf{s}_{tk1}$) and parameters ($W_{k1} \to \mathrm{sig}^{-1}(\mathbf{A}_{\cdot 1k1})$), respectively? To address this, 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.

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 k-th neuron's activity (Equation (9)) is determined by the gradient of cost function L_k, $\dot{x}_{tk} \propto -\partial L_k / \partial x_{tk}$. By integrating the right-hand side of Equation (9), we obtain a class of cost functions as

$$L_k = \sum_{\tau=1}^{t} \left(f(x_{\tau k}) - x_{\tau k} W_{k1} y_\tau - (1 - x_{\tau k}) W_{k0} y_\tau - x_{\tau k} h_{k1} - (1 - x_{\tau k}) h_{k0} \right) + \mathcal{O}(1)$$

$$= \sum_{\tau=1}^{t} \left(f(x_{\tau k}) - \begin{pmatrix} x_{\tau k} \\ 1 - x_{\tau k} \end{pmatrix}^T \left(\begin{pmatrix} W_{k1} \\ W_{k0} \end{pmatrix} y_\tau + \begin{pmatrix} h_{k1} \\ h_{k0} \end{pmatrix} \right) \right) + \mathcal{O}(1), \tag{12}$$

where $\mathcal{O}(1)$ depends on W_{k1} and W_{k0} that is of smaller order than $\mathcal{O}(t)$ and thus negligible when t is large. The cost function of the entire network is defined by $L \equiv \sum_{k=1}^{N} L_k$. When $f'(x_{\tau k})$ is the inverse of the sigmoid function, we have

$$f(x_{\tau k}) = x_{\tau k} \ln x_{\tau k} + (1 - x_{\tau k}) \ln(1 - x_{\tau k}) \tag{13}$$

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

$$\begin{cases} \dot{W}_{k1} \propto -\dfrac{\partial L_k}{\partial W_{k1}} = \overline{x_{tk} y_t} + \overline{x_{tk}} h'_{k1} \\ \dot{W}_{k0} \propto -\dfrac{\partial L_k}{\partial W_{k0}} = \overline{(1 - x_{tk}) y_t} + \overline{1 - x_{tk}} h'_{k0} \end{cases}, \tag{14}$$

where $\overline{x_{tk} y_t} \equiv \sum_{\tau=1}^{t} x_{\tau k} y_\tau$, $\overline{x_{tk}} \equiv \sum_{\tau=1}^{t} x_{\tau k}$, $\overline{(1 - x_{tk}) y_t} \equiv \sum_{\tau=1}^{t} (1 - x_{\tau k}) y_\tau$, $\overline{1 - x_{tk}} \equiv \sum_{\tau=1}^{t} (1 - x_{\tau k})$, $h'_{k1} \equiv \partial h_{k1} / \partial W_{k1}$, and $h'_{k0} \equiv \partial h_{k0} / \partial W_{k0}$. Note that the update of W_{k1} is not directly influenced by W_{k0}, and *vice versa*; since they encode parameters in physically distinct pathways (i.e., the updates are local learning rules (Lee et al., 2000)). The update rule for W_{k1} can be viewed as Hebbian plasticity mediated by an additional activity-dependent term expressing homeostatic plasticity. Moreover, the update of W_{k0} can be viewed as anti-Hebbian plasticity with a homeostatic term, in the sense that W_{k0} is reduced when input (y_t) and output (x_{tk}) fire together. The fixed points of W_{k1} and W_{k0} are given by

$$\begin{cases} W_{k1} = h_1'^{-1} \left(-\dfrac{\overline{x_{tk} y_t}}{\overline{x_{tk}}} \right) \\ W_{k0} = h_0'^{-1} \left(-\dfrac{\overline{(1 - x_{tk}) y_t}}{\overline{1 - x_{tk}}} \right) \end{cases}. \tag{15}$$

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

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

Comparison with variational free energy. Here, we establish a formal relationship between the cost function L and variational free energy. We define $V_{k1} \equiv \mathrm{sig}(W_{k1})$ and $V_{k0} \equiv \mathrm{sig}(W_{k0})$ as nonlinear functions of synaptic strengths. We consider the case where neural activity is expressed as a sigmoid function and thus Equation (13) holds. From $W_{k1} = \ln V_{k1} - \ln(\vec{1} - V_{k1})$, Equation (12) becomes

$$L = \sum_{k=1}^{N}\sum_{\tau=1}^{t} \begin{pmatrix} x_{\tau k} \\ 1 - x_{\tau k} \end{pmatrix}^{T} \left\{ \begin{pmatrix} \ln x_{\tau k} \\ \ln(1 - x_{\tau k}) \end{pmatrix} - \begin{pmatrix} \ln V_{k1} & \ln(\vec{1} - V_{k1}) \\ \ln V_{k0} & \ln(\vec{1} - V_{k0}) \end{pmatrix} \begin{pmatrix} y_{\tau} \\ \vec{1} - y_{\tau} \end{pmatrix} - \begin{pmatrix} h_{k1} \\ h_{k0} \end{pmatrix} \right.$$

$$\left. + \begin{pmatrix} \ln(\vec{1} - V_{k1}) \\ \ln(\vec{1} - V_{k0}) \end{pmatrix} \vec{1} \right\} + \mathcal{O}(1). \qquad (16)$$

One can immediately see a formal correspondence between this cost function and variational free energy (Equation (4)). That is, when we assume $x_{tk} = \mathbf{s}_{tk1}$, $V_{k1} = \mathbf{A}_{\cdot 1k1}$, and $V_{k0} = \mathbf{A}_{\cdot 1k0}$, Equation (16) has exactly the same form as the sum of the accuracy and state complexity (see the first term in the last equality of Equation (4)), which is the leading order term of variational free energy.

Specifically, when the threshold $h_{k1} = \ln(\vec{1} - V_{k1}) \cdot \vec{1} + \ln D_{k1}$ and $h_{k0} = \ln(\vec{1} - V_{k0}) \cdot \vec{1} + \ln D_{k0}$ hold, Equation (16) becomes equivalent to Equation (4) up to the $\ln t$ order term (that disappears when t is large). Therefore, in this special case, the fixed points of neural activity and synaptic strengths become posterior expectations; thus, x_{tk} asymptotically becomes the Bayes optimal encoder for a large t limit (provided D_k matches the genuine prior D_k^*).

We can define $\phi_{k1} \equiv h_{k1} - \ln(\vec{1} - V_{k1}) \cdot \vec{1}$ and $\phi_{k0} \equiv h_{k0} - \ln(\vec{1} - V_{k0}) \cdot \vec{1}$ as functions of W_{k1} and W_{k0}, respectively, and express the cost function as

$$L = \sum_{k=1}^{N}\sum_{\tau=1}^{t} \begin{pmatrix} x_{\tau k} \\ 1 - x_{\tau k} \end{pmatrix}^{T} \left\{ \begin{pmatrix} \ln x_{\tau k} \\ \ln(1 - x_{\tau k}) \end{pmatrix} - \begin{pmatrix} \ln V_{k1} & \ln(\vec{1} - V_{k1}) \\ \ln V_{k0} & \ln(\vec{1} - V_{k0}) \end{pmatrix} \begin{pmatrix} y_{\tau} \\ \vec{1} - y_{\tau} \end{pmatrix} - \begin{pmatrix} \phi_{k1} \\ \phi_{k0} \end{pmatrix} \right\} + \mathcal{O}(1). \qquad (17)$$

Here, we suppose—without loss of generality—that the constant terms in ϕ_{k1} and ϕ_{k0} are chosen to ensure $\exp(\phi_{k1}) + \exp(\phi_{k0}) = 1$. Under this condition, $(\exp(\phi_{k1}), \exp(\phi_{k0}))$ can be viewed as the prior belief about hidden states

$$\begin{cases} \phi_{k1} = \ln D_{k1} \\ \phi_{k0} = \ln D_{k0} \end{cases} \tag{18}$$

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

In other words, when the prior belief about states (D_{k1}, D_{k0}) is a function of the posterior expectation about parameters $(\mathbf{A}_{\cdot 1k1}, \mathbf{A}_{\cdot 1k0})$, the generic cost function under consideration can be expressed in the form of variational free energy, up to $\mathcal{O}(\ln t)$ term. A generic cost function *L* is sub-optimal from the perspective of Bayesian inference unless ϕ_{k1} and ϕ_{k0} are tuned appropriately to express the unbiased (i.e., optimal) prior belief. In this BSS setup, $\phi_{k1} = \phi_{k0} = \text{const}$ 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.

Analysis on synaptic update rules. For the purpose of explicitly solving the fixed point of W_{k1} and W_{k0} that provide the global minimum of *L*, we suppose ϕ_{k1} and ϕ_{k0} as linear functions of W_{k1} and W_{k0}, respectively, given as

$$\begin{cases} \phi_{k1} = \alpha_{k1} + W_{k0}\beta_{k1} \\ \phi_{k0} = \alpha_{k0} + W_{k0}\beta_{k0} \end{cases}, \tag{19}$$

where $\alpha_{k1}, \alpha_{k0} \in \mathbb{R}$ and $\beta_{k1}, \beta_{k0} \in \mathbb{R}^M$ are constants. By solving the variation of *L* with respect to W_{k1} and W_{k0}, we find the fixed point of synaptic strengths as

$$\begin{cases} W_{k1} = \text{sig}^{-1}\left(\dfrac{\overline{x_{tk}y_t}}{\overline{x_{tk}}} + \beta_{k1}\right) \\ W_{k0} = \text{sig}^{-1}\left(\dfrac{\overline{(1-x_{tk})y_t}}{\overline{1-x_{tk}}} + \beta_{k0}\right) \end{cases}. \tag{20}$$

Since the update from *t* to *t*+1 is expressed as $\text{sig}(W_{k1} + \Delta W_{k1}) - \text{sig}(W_{k1}) = \text{sig}(W_{k1})(1 - \text{sig}(W_{k1}))\Delta W_{k1} + \mathcal{O}(|\Delta W_{k1}|^2)$ and $\text{sig}(W_{k1} + \Delta W_{k1}) - \text{sig}(W_{k1}) \approx x_{(t+1)k}y_{t+1}/\overline{x_{tk}} - x_{(t+1)k}\overline{x_{tk}y_t}/\overline{x_{tk}}^2 = x_{(t+1)k}y_{t+1}/\overline{x_{tk}} - (\text{sig}(W_{k1}) - \beta_{k1})x_{(t+1)k}/\overline{x_{tk}}$, we recover the following synaptic plasticity:

$$\begin{cases} \Delta W_{k1} = \underbrace{\dfrac{\text{sig}(W_{k1})^{\odot -1} \odot (1-\text{sig}(W_{k1}))^{\odot -1}}{\overline{x_{tk}}}}_{\text{adaptive learning rate}} \odot \left\{ \underbrace{x_{(t+1)k}y_{t+1}}_{\text{Hebbian plasticity}} - \underbrace{(\text{sig}(W_{k1}) - \beta_{k1})x_{(t+1)k}}_{\text{homeostatic plasticity}} \right\} \\ \Delta W_{k0} = \underbrace{\dfrac{\text{sig}(W_{k0})^{\odot -1} \odot (1-\text{sig}(W_{k0}))^{\odot -1}}{\overline{1-x_{tk}}}}_{\text{adaptive learning rate}} \odot \left\{ \underbrace{(1 - x_{(t+1)k})y_{t+1}}_{\substack{\text{anti}-\text{Hebbian} \\ \text{plasticity}}} - \underbrace{(\text{sig}(W_{k0}) - \beta_{k0})(1 - x_{(t+1)k})}_{\text{homeostatic plasticity}} \right\} \end{cases}, \tag{21}$$

where $\odot$ denotes the element-wise (Hadamard) product and $\text{sig}(W_{k1})^{\odot -1}$ indicates the element-wise inverse of $\mathrm{sig}(W_{k1})$. This synaptic plasticity rule is a subclass of the generic synaptic plasticity rule in Equation (11).

In summary, 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 or marginal likelihood because the cost function can be formulated in terms of variational free energy. In what follows, we will rehearse these 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.

Results

Analytical form of neural network cost functions. The analysis of 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)):

$$L = \sum_{k=1}^{N} \left[\sum_{\tau=1}^{t} \begin{pmatrix} x_{\tau k} \\ 1 - x_{\tau k} \end{pmatrix}^{T} \left\{ \begin{pmatrix} \ln x_{\tau k} \\ \ln(1 - x_{\tau k}) \end{pmatrix} - \begin{pmatrix} \ln V_{k1} & \ln(\vec{1} - V_{k1}) \\ \ln V_{k0} & \ln(\vec{1} - V_{k0}) \end{pmatrix} \begin{pmatrix} y_{\tau} \\ \vec{1} - y_{\tau} \end{pmatrix} - \begin{pmatrix} \phi_{k1} \\ \phi_{k0} \end{pmatrix} \right\} \right] + \mathcal{O}(1),$$

where $V_{k1} = \mathrm{sig}(W_{k1})$ and $V_{k0} = \mathrm{sig}(W_{k0})$ hold, and ϕ_{k1} and ϕ_{k0} are functions of W_{k1} and W_{k0}, 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 ϕ_{k1} and ϕ_{k0}. In other words, we can say that the cost function L is variational free energy under a sub-optimal prior belief about hidden states—depends on W_{k1} and W_{k0}: $\ln P(s_{tk}) = \ln D_k = \phi_k$, where $\phi_k \equiv (\phi_{k1}, \phi_{k0})$. This prior alters the landscape of 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 (ϕ_{k1} and ϕ_{k0}) that constitute the difference between cost function and variational free energy with optimal prior beliefs can be expressed as linear equations of W_{k1} and W_{k0}.

From assumption 3, Equation (17) becomes:

$$L = \sum_{k=1}^{N} \left[\sum_{\tau=1}^{t} \begin{pmatrix} x_{\tau k} \\ 1 - x_{\tau k} \end{pmatrix}^{T} \left\{ \begin{pmatrix} \ln x_{\tau k} \\ \ln(1 - x_{\tau k}) \end{pmatrix} - \begin{pmatrix} \ln V_{k1} & \ln(\vec{1} - V_{k1}) \\ \ln V_{k0} & \ln(\vec{1} - V_{k0}) \end{pmatrix} \begin{pmatrix} y_{\tau} \\ \vec{1} - y_{\tau} \end{pmatrix} - \begin{pmatrix} \alpha_{k1} + W_{k1}\beta_{k1} \\ \alpha_{k0} + W_{k0}\beta_{k0} \end{pmatrix} \right\} \right] + \mathcal{O}(1), \tag{22}$$

where $\{\alpha_{k1}, \alpha_{k0}, \beta_{k1}, \beta_{k0}\}$ are constants. The cost function has degrees of freedom with respect to the choice of constants $\{\alpha_{k1}, \alpha_{k0}, \beta_{k1}, \beta_{k0}\}$, which correspond to the prior belief about states D_k. 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.

The fixed point of synaptic strengths—that give the minimum of L—is given analytically as Equation (20), expressing that (β_{k1}, β_{k0}) 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_{k1} and W_{k0} 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: $x_{\tau k} = \mathbf{s}_{\tau k1} = Q(s_{\tau k} = 1)$ and synaptic strengths encode the posterior expectation of the parameters: $V_{k1} = \mathrm{sig}(W_{k1}) = \mathbf{A}_{\cdot 1k1}$ and $V_{k0} = \mathrm{sig}(W_{k0}) = \mathbf{A}_{\cdot 1k0}$. In addition, based on assumption 3, the accuracy of approximation depends on the deviation of constants $\{\alpha_{k1}, \alpha_{k0}, \beta_{k1}, \beta_{k0}\}$ from their optimal values. From a Bayesian perspective, these constants can be viewed as prior beliefs, $\ln P(s_{tk}) = \ln D_k = (\alpha_{k1} + W_{k1}\beta_{k1}, \alpha_{k0} + W_{k0}\beta_{k0})$, when we assume $(x_k, 1 - x_k)$ represents the state posterior $\mathbf{s}_{tk}$. When and only when $(\alpha_{k1}, \alpha_{k0}) = (-\ln 2, -\ln 2)$ and $(\beta_{k1}, \beta_{k0}) = (\vec{0}, \vec{0})$, does 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—from the Bayesian perspective—may be viewed as (potentially sub-optimal) prior beliefs. We now illustrate the implicit inference and learning in neural networks, using simulations of BSS.

Numerical simulations. Here, we simulate the dynamics of neural activity and synaptic strengths when they perform a gradient descent on the cost function in Equation (22). We consider a BSS comprising two hidden sources (or states) and 32 observations (or sensory inputs), formulated as an MDP. The two hidden sources comprise four patterns: $s_t = s_{t1} \otimes s_{t2} = (0,0), (1,0), (0,1), (1,1)$. An observation o_{ti} is generated through the likelihood mapping A_i. We defined A_i with:

$$\begin{cases} P(o_{ti} = 1 | s_t, A_i) = A_{i1\cdot\cdot} = \left(0, \frac{3}{4}, \frac{1}{4}, 1\right) & \text{for } 1 \le i \le 16 \\ P(o_{ti} = 1 | s_t, A_i) = A_{i1\cdot\cdot} = \left(0, \frac{1}{4}, \frac{3}{4}, 1\right) & \text{for } 1 \le i \le 32 \\ P(o_{ti} = 0 | s_t, A_i) = A_{i0\cdot\cdot} = (1,1,1,1) - A_{i1\cdot\cdot} & \text{for } 1 \le i \le 32 \end{cases} . \tag{23}$$

Here, for example, $A_{i110} = 3/4$ for $1 \le i \le 16$ is the probability of o_{ti} taking one when $s_t = (1,0)$. The simulations continue over $T = 10^4$ time steps.

First, we show that—as in (Isomura & Friston, 2018)—a network with a cost function with optimised constants $((\alpha_{k1}, \alpha_{k0}) = (-\ln 2, -\ln 2)$ and $(\beta_{k1}, \beta_{k0}) = (\vec{0}, \vec{0}))$ performs 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). Conversely, 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 is sufficient to emulate BSS. Next, we quantified the dependency of BSS performance on the form of the cost function, by varying the above constants (Fig. 3).

We found that changing $(\alpha_{k1}, \alpha_{k0})$ from $(-\ln 2, -\ln 2)$ led to failure of BSS. Because neuron 1 encodes source 1 with optimal α, the correlation between source 1 and the response of neuron 1 is close to one, while the correlation between source 2 and its response is almost zero. In case of sub-optimal α, 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 β take values far from zero (Fig. 3B). When the elements of β 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.

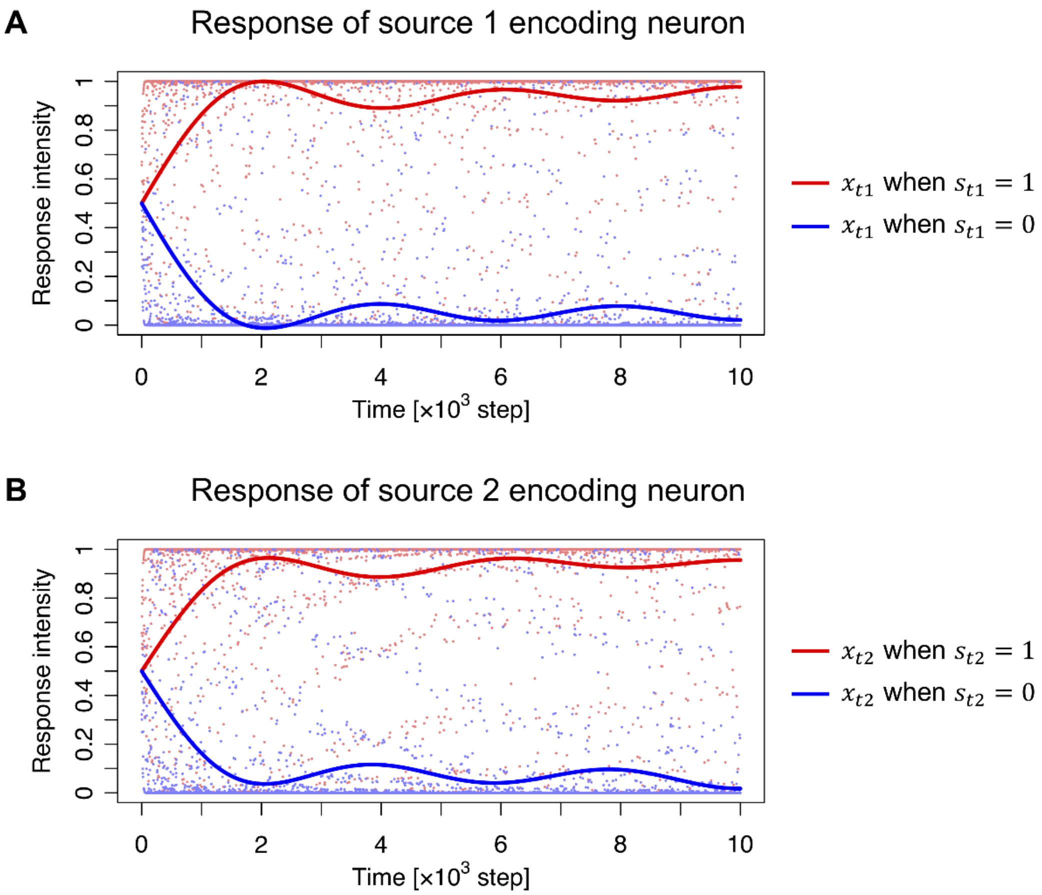

Figure 2. Emergence of response selectivity for a source. (**A**) Evolution of neuron 1's responses that learnt to encode source 1, in the sense that the response is high when source 1 takes a value of one (red dots), and it is low when source 1 takes a value of zero (blue dots). Lines correspond to smoothed trajectories obtained using a discrete cosine transform. (**B**) Emergence of neuron 2's response that learnt to encode source 2. Codes are appended as Supplementary Source Codes.

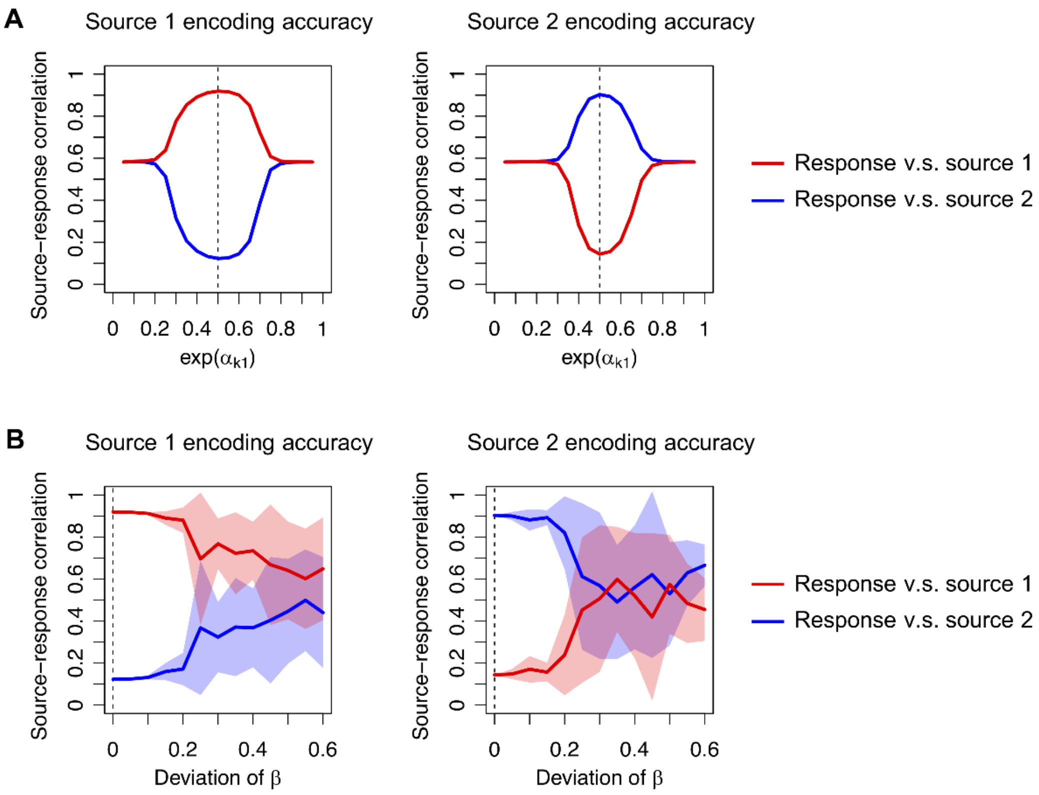

Figure 3. Dependence of source encoding accuracy on constants. Left panels show the magnitudes of correlations between sources and response of a neuron expected to encode source 1: $|\mathrm{corr}(s_{t1}, x_{t1})|$ and $|\mathrm{corr}(s_{t2}, x_{t1})|$. The right panels show the magnitudes of the correlations between sources and response of a neuron expected to encode source 2: $|\mathrm{corr}(s_{t1}, x_{t2})|$ and $|\mathrm{corr}(s_{t2}, x_{t2})|$. (**A**) Dependence on the constant α that controls the excitability of a neuron, when β is fixed to zero. The dashed line (0.5) indicates the optimal value of $\exp(\alpha_{k1})$. (**B**) Dependence on constant β, when α is fixed as $(\alpha_{k1}, \alpha_{k0}) = (-\ln 2, -\ln 2)$. Elements of β were randomly generated from a Gaussian distribution with zero mean. The standard deviation of β was varied (horizontal axis), where zero deviation was optimal (dashed line). Lines and shaded areas indicate the mean and standard deviation of the source-response correlation, evaluated with 10 different values of β. Codes are appended as Supplementary Source Codes.

Our numerical analysis, under assumptions 1–3 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, 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).

Phenotyping networks. The parameters $\phi_k = \ln D_k$ determine how the synaptic strengths change depending on the history of sensory inputs and neural outputs. The generic cost functions under consideration have free parameters regarding the choice of ϕ_k. When ϕ_k is close to $(-\ln 2, -\ln 2)$, the cost function becomes variational free energy with optimal prior beliefs for BSS. We have therefore shown that variational free energy (under the MDP scheme) is within the class of biologically plausible cost functions found in neural networks. Conversely, one could regard neural networks—of the sort considered in this paper—as performing approximate Bayesian inference under priors that may or may not be optimal. Under the complete class theorem, this means that, in principle, any neural network of this kind is optimal, when its prior beliefs are consistent with the process generating outcomes. This perspective speaks of the possibility of characterising a neural network model—and indeed a real neuronal network—in terms of its implicit prior beliefs. It should be noted again that the complete class theorem suggests that any response of a neural network is Bayes optimal under some prior beliefs (and cost function). This means that a neural network can be characterised in terms of its implicit priors.

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_k) 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 ϕ_k, 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 $(\alpha_{k1}, \alpha_{k0})$ far from $(-\ln 2, -\ln 2)$ or a large deviation of (β_{k1}, β_{k0}) 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 ϕ_k.

One can pursue this analysis even 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 systems neuroscience to characterise the choice behaviour in terms of subject specific priors. Please see (Schwartenbeck & Friston, 2016) for greater details.

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

$$\phi_k = \ln D_k = \psi(\mathbf{d}_k) - \psi(\mathbf{d}_{k1} + \mathbf{d}_{k0}),$$

$$\mathbf{d}_k = d_k + \sum_{\tau=1}^{t} \mathbf{s}_{\tau k} \, . \qquad (24)$$

Please see (Schwartenbeck & Friston, 2016) for more details. In effect, this update will simply add the Dirichlet concentration parameters, $\mathbf{d}_k = (\mathbf{d}_{k1}, \mathbf{d}_{k0})$, to the priors in proportion to the temporal summation of posterior expectations about the hidden states. Therefore, by committing to cost functions that underwrite variational inference and learning, any free parameter can be updated in a Bayes optimal fashion when a suitable generative model is available.

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. This cost function can be viewed as variational free energy, where prior beliefs about hidden states are expressed as biological constraints, in the form of thresholds and neuronal excitability.

One can understand the nature of the constants $\{\alpha_{k1}, \alpha_{k0}, \beta_{k1}, \beta_{k0}\}$ from biological and Bayesian perspectives as follows: $(\alpha_{k1}, \alpha_{k0})$ determines the firing threshold and thus controls the mean firing rates. Expressed differently, 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, $(\alpha_{k1}, \alpha_{k0})$ encodes prior beliefs about states, which exert a large influence on the state posterior. The state posterior is biased if $(\alpha_{k1}, \alpha_{k0})$ is picked in a sub-optimal way—in relation to the process generating inputs. In contrast, (β_{k1}, β_{k0}) determines the accuracy of synaptic strengths representing the likelihood mapping of an observation o_{ti} taking 1 (ON state) depending on hidden states (please compare Equation (8) and Equation (20)). Only when $(\beta_{k1}, \beta_{k0}) = (\vec{0}, \vec{0})$, the encoder becomes Bayesian optimal. These constants could 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 enable the accurate encoding of hidden states in the current formulation.

We have shown in previous reports that *in vitro* neural networks—comprising a cortical cell culture—performed 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 was 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 in identifying the principles that underwrite learning or adaptation processes in biological neuronal networks, using empirical response data. Herein, we have illustrated this potential in terms of the choice of function ϕ_k in the cost functions above: if ϕ_k is close to a constant $(-\ln 2, -\ln 2)$, the cost function is expressed straightforwardly as a variational free energy with small prior biases. In the subsequent work, we intend to apply this scheme to empirical data and examine the biological plausibility of variational free energy minimisation.

In summary, we have identified a class of biologically plausible cost functions that explain neural activity and synaptic plasticity. A cost function in the class becomes Bayes optimal when, and only when, several constants expressing biological constraints correspond to appropriate priors in implicit generative model. Estimating these constants from empirical data may be a useful approach to characterise learning and inference in terms of a neural network's priors.

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.

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.

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.

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.

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

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.

Supplementary Information

Supplementary Tables

Table S1. Correspondence of variables and functions.

Neural network formation			Variational Bayes formation	
Neural activity	x_{tk}	$\Leftrightarrow$	$\mathbf{s}_{tk1}$	State posterior
Sensory input	y_t	$\Leftrightarrow$	$o_{t\cdot 1}$	Observation
Synaptic strength	W_{k1}	$\Leftrightarrow$	$\mathrm{sig}^{-1}(\mathbf{A}_{\cdot 1k1})$	
	V_{k1}	$\Leftrightarrow$	$\mathbf{A}_{\cdot 1k1}$	Parameter posterior
Perturbation term	ϕ_{k1}	$\Leftrightarrow$	$\ln D_{k1}$	State prior
Threshold	h_{k1}	$\Leftrightarrow$	$\ln\left(\vec{1} - \mathbf{A}_{\cdot 1k1}\right) \cdot \vec{1} + \ln D_{k1}$	

Supplementary Methods

S1. The order of the parameter complexity

The order of the complexity term: $(\mathbf{a}_i - a_i) \cdot \ln \mathbf{A}_i - \ln \mathcal{B}(\mathbf{a}_i)$ is calculated. The first term becomes $(\mathbf{a}_i - a_i) \cdot \ln \mathbf{A}_i = \mathbf{a}_i \cdot \ln \mathbf{A}_i + \mathcal{O}(1)$ since $a_i \cdot \ln \mathbf{A}_i$ is in the order of 1. Moreover, since $\Gamma(\mathbf{a}_{i1}) = \mathbf{a}_{i1} \cdot \ln \mathbf{a}_{i1} + \mathcal{O}(\ln t)$, we get $\ln \mathcal{B}(\mathbf{a}_i) = \ln \Gamma(\mathbf{a}_{i1}) + \ln \Gamma(\mathbf{a}_{i0}) - \ln \Gamma(\mathbf{a}_{i1} + \mathbf{a}_{i0}) = \mathbf{a}_{i1} \cdot \ln \mathbf{a}_{i1} + \mathbf{a}_{i0} \cdot \ln \mathbf{a}_{i0} - (\mathbf{a}_{i1} + \mathbf{a}_{i0}) \cdot \ln(\mathbf{a}_{i1} + \mathbf{a}_{i0}) + \mathcal{O}(\ln t) = \mathbf{a}_i \cdot \ln \mathbf{A}_i + \mathcal{O}(\ln t)$. Thus, it holds that $(\mathbf{a}_i - a_i) \cdot \ln \mathbf{A}_i - \ln \mathcal{B}(\mathbf{a}_i) = \mathcal{O}(\ln t)$. Hence, we obtain

$$F\left(\tilde{o}, Q(\tilde{s}), Q(A)\right) = \sum_{\tau=1}^{t} (\mathbf{s}_\tau \cdot \ln \mathbf{s}_\tau - \mathbf{s}_\tau \cdot \ln \mathbf{A} \cdot o_\tau - \mathbf{s}_\tau \cdot \ln D) + \mathcal{O}(\ln t). \qquad (25)$$

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 $\mathcal{O}(\ln t)$ term in Equation (25) is neglected. Solving the variation of *F* with respect to $\mathbf{A}_{\cdot 1}$ yields the optimal posterior expectation. From $\mathbf{A}_{\cdot 0} = \vec{1} - \mathbf{A}_{\cdot 1}$, we find

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

$$= -\left(\delta \mathbf{A}_{\cdot 1} \odot \mathbf{A}_{\cdot 1}^{\odot -1}\right) \cdot \overline{o_{t\cdot 1} \otimes \mathbf{s}_t} + \left(\delta \mathbf{A}_{\cdot 1} \odot \left(\vec{1} - \mathbf{A}_{\cdot 1}\right)^{\odot -1}\right) \cdot \overline{\left(\vec{1} - o_{t\cdot 1}\right) \otimes \mathbf{s}_t}$$

$$= \left(\delta\mathbf{A}_{\cdot 1} \odot \mathbf{A}_{\cdot 1}^{\odot -1} \odot \left(\vec{1}-\mathbf{A}_{\cdot 1}\right)^{\odot -1}\right) \cdot \left(-\left(\vec{1}-\mathbf{A}_{\cdot 1}\right) \odot \overline{o_{t\cdot 1} \otimes \mathbf{s}_t} + \mathbf{A}_{\cdot 1} \odot \overline{\left(\vec{1}-o_{t\cdot 1}\right) \otimes \mathbf{s}_t}\right)$$

$$= \left(\delta\mathbf{A}_{\cdot 1} \odot \mathbf{A}_{\cdot 1}^{\odot -1} \odot \left(\vec{1}-\mathbf{A}_{\cdot 1}\right)^{\odot -1}\right) \cdot \left(\mathbf{A}_{\cdot 1} \odot \overline{\vec{1} \otimes \mathbf{s}_t} - \overline{o_{t\cdot 1} \otimes \mathbf{s}_t}\right) \qquad (26)$$

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

$$\mathbf{A}_{\cdot 1} = \overline{o_{t\cdot 1} \otimes \mathbf{s}_t} \odot \left(\overline{\vec{1} \otimes \mathbf{s}_t}\right)^{\odot -1} + \mathcal{O}\left(\frac{\ln t}{t}\right). \qquad (27)$$

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: $W_{k1} = W_{k1}(t)$ and define the change as $\Delta W_{k1} \equiv W_{k1}(t+1) - W_{k1}(t)$. From Equation (15), $h_1'(W_{k1})$ satisfies both

$$h_1'(W_{k1} + \Delta W_{k1}) - h_1'(W_{k1}) = h_1''(W_{k1}) \odot \Delta W_{k1} + \mathcal{O}(|\Delta W_{k1}|^2) \qquad (28)$$

and

$$h_1'(W_{k1} + \Delta W_{k1}) - h_1'(W_{k1}) = -\frac{x_{(t+1)k} y_{t+1} + \overline{x_{tk} y_t}}{x_{(t+1)k} + \overline{x_{tk}}} + \frac{\overline{x_{tk} y_t}}{\overline{x_{tk}}}$$

$$= -\frac{x_{(t+1)k} y_{t+1}}{\overline{x_{tk}}} + \frac{\overline{x_{tk} y_t}}{\overline{x_{tk}}^2} x_{(t+1)k} = -\frac{1}{\overline{x_{tk}}}\left(x_{(t+1)k} y_{t+1} - h_1'(W_{k1}) x_{(t+1)k}\right). \qquad (29)$$

Thus, we find

$$\Delta W_{k1} = \underbrace{-\frac{h_1''(W_{k1})^{\odot -1}}{\overline{x_{tk}}}}_{\text{adaptive learning rate}} \odot \left(\underbrace{x_{(t+1)k} y_{t+1}}_{\text{Hebbian term}} - \underbrace{h_1'(W_{k1}) x_{(t+1)k}}_{\text{homeostatic term}}\right). \qquad (30)$$

Similarly,

$$\Delta W_{k0} = \underbrace{-\frac{h_0''(W_{k0})^{\odot -1}}{\overline{1 - x_{tk}}}}_{\text{adaptive learning rate}} \odot \left(\underbrace{\left(1 - x_{(t+1)k}\right) y_{t+1}}_{\text{anti}-\text{Hebbian term}} - \underbrace{h_0'(W_{k0})\left(1 - x_{(t+1)k}\right)}_{\text{homeostatic term}}\right). \qquad (31)$$

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